materiál pro kurs

Statistická analýza dat

(Jak pracovat s daty a zadávat výpočty v SPSS)

katedra sociologie
FSS MU v Brně

0. lekce (na zopakování)

ZÁKLADNÍ STRATEGIE ANALÝZY: VÝZKUMNÝ PROBLÉM, VÝZKUMNÉ OTÁZKY A PROMĚNNÉ.

MORGAN George A. and Griego V. ORLANDO. 1998. Easy Use and Interpretation of SPSS for Windows: Answering Research Questions With Statistics. Mahwah: Lawrence Erlbaum Associates.

CHAPTER 1

Research Problems, Approaches, and Questions

Research Problems

The research process begins with a problem. *What is a research problem?* Kerlinger (1986) formally describes a problem as "...an interrogative sentence or statement that asks: *What relation exists between two or more variables?*" (p. 16). Note that almost all research studies have more than two variables. Kerlinger suggests that prior to the problem statement "...the scientist will usually experience an obstacle to understanding, a vague unrest about observed and unobserved phenomena, a curiosity as to *why something is as it is*" (p. 11). Appendix A provides templates to help you phrase your research problem, and provides examples from the high school and beyond (HSB) data set.

Variables

A *variable* has one defining quality. *It must be able to vary or have different values*. For example, *gender* is a variable because it has two values, female or male. *Age* is a variable that has a large number of values. *Type of treatment/intervention* (or *type of curriculum*) is a variable if there is more than one treatment or a treatment and a control group. *Number of days to learn something or to recover from an ailment*, common measures of the effect of a treatment, are also variables. Similarly, *amount of mathematics knowledge* is a variable because it can vary from none to a lot. If a concept has one value in a particular study it is not a variable, e.g., ethnic group is not a variable if all participants are Caucasian.

Definition of a variable. We can define the term "variable" as a characteristic of the participants or situation of a given study that has different values in that study. In quantitative research, variables are defined operationally and are commonly divided into independent variables (active or attribute), dependent variables, and extraneous variables. Each of these topics will be dealt with in the following sections.

Operational definitions of variables. An operational definition describes or defines a variable in terms of the operations or techniques used to elicit or measure it. When quantitative researchers describe the variables in their study, they specify what they mean by demonstrating how they measured the variable. Demographic variables like age, gender, or ethnic group are usually measured simply by asking the participant to choose the appropriate category from a list. Types of treatment (or curriculum) are usually described/defined much more extensively so the reader can understand what the researcher meant by, for example, a cognitively enriching curriculum or sheltered work. Likewise, abstract concepts like mathematics knowledge, self-concept, or mathematics anxiety need to be defined operationally by spelling out in some detail how they were measured in a particular study. To do this, the investigator may provide sample questions, append the actual instrument, or provide a reference where more information can be found.

Independent Variables

Active independent variables. This first type of variable is often called a manipulated independent variable. A frequent goal of research is to investigate the effect of a particular intervention. An example might be the effect of a new kind of therapy compared to the traditional treatment. A second example might be the effect of a new teaching method, such as cooperative learning, on student performance. In the two examples provided above, the variable of interest was something that was *given to* the participants. Therefore, an *active independent variable* is a variable, such as a workshop, new curriculum, or other intervention, one level of which *can be given to a group of participants*, usually within a specified period of time *during the study*.

In traditional experimental research, independent variables are those that the *investigator can manipulate*; they presumably cause a change in some resulting behavior, attitude, or physiological measure of interest. An independent variable is considered to be manipulated or active when the investigator has the option to give one value to one group (experimental condition), and another value to another group (control condition).

However, there are many circumstances, especially in applied research, when we have an active independent variable but this variable *is not directly manipulated by the investigator*. Consider the situation where the investigator is interested in a new type of treatment. In order to carry out the study, it turns out that rehabilitation center *A* will be using that treatment. Rehabilitation center *B* will be using the traditional treatment. The investigator will compare the two centers to determine if one treatment works better than the other. Notice that the independent variable is active but has *not* been manipulated *by the investigator*.

Thus, active independent variables are *given* to the participants in the study but are not necessarily manipulated by the experimenter. They may be given by a clinic, school, or someone, other than the investigator. From the participants' point of view the situation was manipulated.

Attribute independent variables. Unlike some authors of research methods books, we do not restrict the term "independent variable" to those variables that are manipulated or active. We define an independent variable more broadly to include any predictors, antecedents, or *presumed* causes or influences under investigation in the study. Attributes of the participants as well as active independent variables fit within this definition. For the social sciences, education, and disciplines dealing with special needs populations, attribute independent variables are especially important. Type of disability or level of disability is often the major focus of a study. Disability certainly qualifies as a variable since it can take on different values even though they are not "given" in the study. For example, cerebral palsy is different from Down syndrome which is different from spina bifida, yet all are disabilities. Also, there are different levels of the same disability. People already have defining characteristics or *attributes* which place them into one of two or more categories. The different disabilities are already present when we begin our study. Thus, we are also interested in studying a class of variables that cannot be given during the study, even by other persons, schools, or clinics.

MORGAN George A. and Griego V. ORLANDO. 1998. *Easy Use and Interpretation of SPSS for Windows: Answering Research Questions With Statistics*. Mahwah: Lawrence Erlbaum Associates.

A variable which cannot be given, yet is a major focus of the study, is called an attribute independent variable (Kerlinger, 1986). In other words, the values of the independent variable are attributes of the persons or the environment that are not manipulated during the study. For example, *gender*, *age*, *ethnic group*, or *disability* are attributes of a person.

Other labels for the independent variable. SPSS uses a variety of terms such as **factor** (chapters 5, 15, 16, 17 and 18), **covariate** (chapter 13), and **grouping variable** (chapters 14, 15). In other cases (chapters 5, 9) SPSS does not make a distinction between the independent and dependent variable, just labeling them variables. Another common label for an attribute independent variable is a measured variable. However, we prefer attribute so it is not easily confused with the dependent variable, which is also measured. Sometimes variables such as gender or ethnic group are called moderator or mediating variables because they serve these functions; however, SPSS does not use these terms so we will not either in this book.

Type of independent variable and inferences about cause and effect. When we analyze data from a research study, the statistical analysis does not differentiate whether the independent variable is an active independent variable or an attribute independent variable. However, even though SPSS and most statistics books use the label independent variable for both active and attribute variables, there is a crucial difference in interpretation. A significant change or difference following manipulation of the active independent variable may reasonably lead the investigator to infer that the independent variable *caused* the change in the dependent variable.

However, a significant change or difference between or among values of an attribute independent variable should *not* lead one to the interpretation that the attribute independent variable caused the dependent variable to change. A major goal of scientific research is to be able to identify a causal relationship between two variables. For those in applied disciplines, the need to demonstrate that a given intervention or treatment causes change in behavior or performance is extremely important. Only the approaches that have an active independent variable (the randomized experimental and to a lesser extent the quasi-experimental) can be successful in providing data that allow one to infer that the independent variable caused the dependent variable.

Although studies with attribute independent variables are limited in what can be said about causation, they can lead to solid conclusions about the differences between groups and about associations between variables. Furthermore, they are the *only* available approach if the focus of your research is on attribute independent variables. The descriptive approach, as we define it, does not attempt to identify relationships. It focuses on describing variables.

As implied above, this distinction between active and attribute independent variables is important because terms such as *main effect* and *effect size* used by SPSS and most statistics books might lead one to believe that if you find a significant difference the independent variable *caused* the difference. These terms are misleading when the independent variable is an attribute.

Values of the independent variable. In defining a variable, we said that it must have more than one value. When describing the different categories of an independent variable, SPSS uses the

word *values*. This does *not* necessarily imply that the values are ordered.[1] Suppose that an investigator is performing a study to investigate the effect of a treatment. One group of participants is assigned to the treatment group. A second group does not receive the treatment. The study could be conceptualized as having one independent variable (*treatment type*), with two values or levels (treatment and no treatment). The independent variable in this example would be classified as an active independent variable. Instead, suppose the investigator was interested primarily in comparing two different treatments but decided to include a third no-treatment group as a control group in the study. The study still would be conceptualized as having one active independent variable (treatment type), but with three values (the two treatment conditions and the control condition). This variable could be diagrammed as follows:

Variable Label	Values	Value Labels
	1	= Treatment 1
Treatment type	2	= Treatment 2
	3	= No treatment (control)

As an additional example, consider gender, which is an attribute independent variable with two values, male and female. It could be diagrammed as follows:

	1	= Male
Gender		
	2	= Female

Note that in SPSS each variable is given a label; the values, which are numbers, may also have labels. It is especially important to know the value labels when the variable is nominal; i.e., when the values of the variable are just names and, thus, are not ordered.

Dependent Variables

The dependent variable is the presumed outcome or criterion. It is assumed to measure or assess the effect of the independent variable. Dependent variables are often test scores, ratings on questionnaires, readings from instruments (electrocardiogram, galvanic skin response, etc.), or measures of physical performance. When we discuss measurement in chapter 3, we are usually referring to the dependent variable. SPSS also uses a number of other terms for the dependent variable. The most common is **dependent list**, used in cases where you can do the same statistic several times, for a list of dependent variables. In discriminant analysis (chapter 13), the dependent variable is called the **grouping variable.** The term **test variable** is used in several of the chapters on *t* tests and analysis of variance.

[1] The terms categories, levels, groups, or samples are sometimes used interchangeably with the term values, especially in statistics books. Likewise the term factor is often used instead of independent variable.

MORGAN George A. and Griego V. ORLANDO. 1998. *Easy Use and Interpretation of SPSS for Windows: Answering Research Questions With Statistics*. Mahwah: Lawrence Erlbaum Associates.

Basic comparative approach. The comparative research approach differs from the experimental and quasi-experimental approaches because the investigator *cannot randomly assign participants* to groups and because there is *not an active independent variable*. Table 1.1 shows that, like experiments and quasi-experiments, comparative designs usually have a few levels or categories for the independent variable and make comparisons between groups. Studies that use the comparative approach examine the presumed effect of an *attribute independent variable*.

An example of the comparative approach is a study that compared two groups of children on a series of motor performance tests. The investigators attempted to determine whether the differences between the two groups were due to perceptual or motor processing problems. One group of children, who had motor handicaps, was compared to a second group of children who did not have motor problems. Notice that the independent variable in this study was an attribute independent variable with two levels, motor handicapped and not handicapped. Thus, it is not possible for the investigator to randomly assign participants to groups, or "give" the independent variable; the independent variable was not active. The independent variable had only two values

or categories so a statistical comparison between the groups would be performed. It is, of course, possible for comparisons to be made between three or more groups.[2]

Basic associational approach. Now, we would like to consider an approach to research where the independent variable is usually continuous or has several ordered categories, usually five or more. Suppose that the investigator is interested in the relationship between giftedness and self-perceived confidence in children. Assume that the dependent variable is a self-confidence scale for children. The independent variable is giftedness. If giftedness had been divided into high, average, and low groups (a few values or levels), we would have called the research approach comparative because the logical thing to do would be to compare the groups. However, in the typical associational approach, the independent variable is continuous or has at least five ordered levels or values.[3] All participants would be in a single group with two continuous variables--giftedness and self-concept. A correlation coefficient could be performed to determine the strength of the relationship between the two variables.

As implied above, it is somewhat arbitrary whether a study is considered to be comparative or associational. For example, a continuous variable such as age can always be divided into a small number of levels such as young and old. However, we make this distinction for two reasons. First, we think it is usually unwise to divide a variable with many ordered levels into a few because information is lost. For example, if the cut point for "old age" was 65, persons 66 and 96 would be lumped together as would persons 21 and 64. Second, different types of statistics are usually used with the two approaches (see Fig. 1.1). We think this distinction and the similar one made in the section on research questions will help you decide on an appropriate statistic, which we have found is one of the hardest parts of the research process for students.

Basic descriptive approach. This approach is different from the other four in that only one variable is considered at a time so that no relationships are made. Table 1.1 shows that this lack of comparisons or associations is what distinguishes this approach from the other four. Of course, the descriptive approach does not meet any of the other criteria such as random assignment of participants to groups.

Most research studies include some descriptive questions (at least to describe the sample), but do not stop there. It is rare these days for published quantitative research to be purely descriptive; we almost always study several variables and their relationships. However, political polls and consumer surveys are sometimes *only* interested in describing how voters *as a whole* react to issues or what products a group of consumers will buy. Exploratory studies of a new topic may just describe what people say or feel about that topic.

Most research books use a considerably broader definition for descriptive research. Some use the phrase "descriptive research" to include all research that is not randomized experimental or

[2] It is also possible to compare relatively large numbers of groups (e.g., 5 or 10) if one has enough participants that the group sizes are adequate, but this is atypical.

[3] It is possible, as we will see in chapters 7 and 8, to use the associational approach and statistics when one has fewer than five ordered values of the variables and even with unordered nominal variables, but this is not typical.

MORGAN George A. and Griego V. ORLANDO. 1998. *Easy Use and Interpretation of SPSS for Windows: Answering Research Questions With Statistics*. Mahwah: Lawrence Erlbaum Associates.

quasi-experimental. Others do not seem to have a clear definition, using descriptive almost as a synonym for exploratory or sometimes "correlational" research. We think it is clearer and less confusing to students to restrict the term descriptive research to questions and studies that use only *descriptive statistics*, such as averages, percentages, histograms, and frequency distributions, and do not test null hypotheses with inferential statistics.

Complex Research Approaches

It is important to note that most studies are more complex than implied by the above examples. In fact, almost all studies have more than one hypothesis or research question and may utilize more than one of the above approaches. It is common to find a study with one active independent variable (e.g., type of treatment) and one or more attribute independent variables (e.g., gender). This type of study combines the randomized experimental approach (if the participants were randomly assigned to groups) and the comparative approach. Most "survey" studies include both the associational and comparative approaches. As mentioned above, most studies also have some descriptive questions so it is common for published studies to use three or even more of the approaches.

Research Questions/Hypotheses

Next, we divide research questions into three broad types: *difference, associational,* and *descriptive*. For the difference type of question, we compare groups or values of the independent variable on their scores on the dependent variable. This type of question typically is used with the randomized experimental, quasi-experimental, and comparative approaches. For an associational question, we associate or relate the independent and dependent variables. Descriptive questions are not answered with inferential statistics; they merely describe or summarize data.

Basic Difference Versus Associational Research Questions or Hypotheses

Hypotheses are defined as *predictive statements about the relationship between variables*. Fig. 1.1 shows that both difference and associational questions/hypotheses have as a *general purpose* the exploration of relationships between variables: This similarity is in agreement with the statement by statisticians that all parametric inferential statistics are relational, and it is consistent with the notion that the distinction between the comparative and associational approach is somewhat arbitrary.[4] However, we believe that the distinction is educationally useful. Note that difference and associational questions differ in specific purpose and the kinds of statistics they use to answer the question.

[4] We use the term associational for this type of research question, approach, and statistics rather than relational or correlational to distinguish them from the *general purpose* of both difference and associational questions/hypotheses described above. Also we wanted to distinguish between correlation, as a specific statistical technique, and the broader types of approach, questions, and group of statistics.

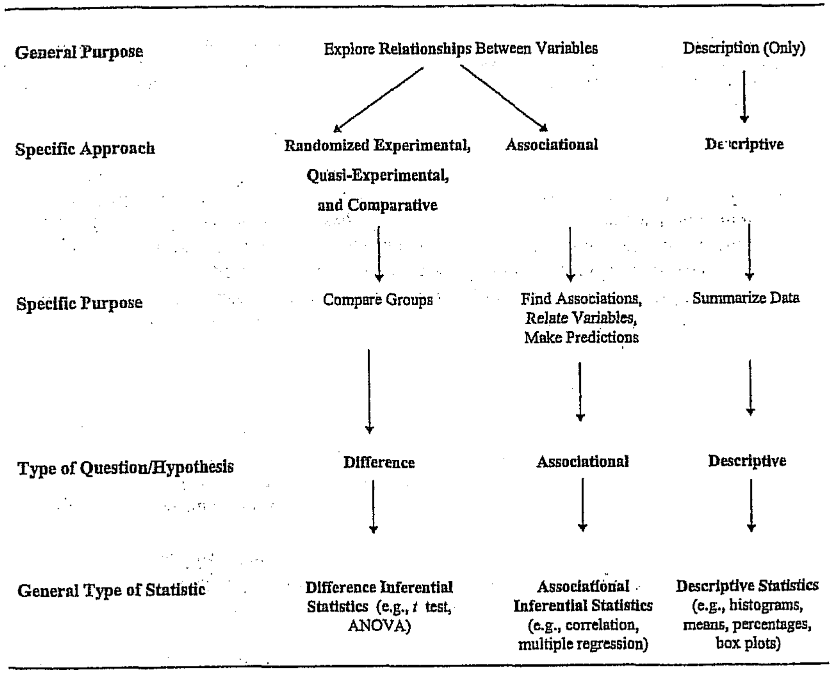

General Purpose
Explore Relationships Between Variables
Description (Only)
Specific Approach
Randomized Experimental, Quasi-Experimental, and Comparative
Associational
Descriptive
Specific Purpose
Compare Groups
Find Associations, Relate Variables, Make Predictions
Summarize Data
Type of Question/Hypothesis
Difference
Associational
Descriptive
General Type of Statistic
Difference Inferential Statistics (e.g., t test, ANOVA)
Associational Inferential Statistics (e.g., correlation, multiple regression)
Descriptive Statistics (e.g., histograms, means, percentages, box plots)

Fig. 1.1. Schematic diagram showing how the purpose, approach and type of research question correspond to the general type of statistic used in a study.

Table 1.2 provides the general format and one example of a basic *difference hypothesis* and of a basic *associational hypothesis*. Research questions are similar to hypotheses, but they are stated in question format. We think it is advisable to use the question format when one does not have a clear directional prediction and for the descriptive approach. More details and examples are given in Appendix A.

Table 1.2. ***Examples of Basic Difference and Associational Hypotheses***

1. ***Difference (group comparison) Hypothesis***

- For this type of hypothesis, the levels or values of the independent variable (e.g., gender) are used to divide the participants into groups (male and female) which are then compared to see if they differ in respect to the average scores on the dependent variable (e.g., empathy).
- An example of a directional research hypothesis is: Women will score higher than men on empathy scores. In other words, the average empathy scores of the women will be significantly higher than the average empathy scores for men.

2. ***Associational (relational) Hypothesis***

- For this type of hypothesis, the scores on the independent variable (e.g., self-esteem) are associated with or related to the dependent variable (e.g., empathy). *It is often arbitrary which variable is considered the independent variable* but most researchers have an idea about what they think is the predictor (independent) and what is the outcome (dependent) variable.
- An example of a directional research hypothesis is: There will be a positive association (relation) between self-esteem scores and empathy scores. In other words, those persons who are high on self-esteem will tend to have high empathy, those with low self-esteem will tend also to have low empathy, and those in the middle on the independent variable will tend to be in the middle on the dependent variable.

Six Types of Research Questions

Table 1.3 expands our overview of research questions to include both basic and complex questions of each of the three types: *descriptive, difference,* and *associational.* The table also includes references to the tables in chapters 3 and 7, designed to help you select an appropriate statistic and examples of the types of statistics that we include under each of the six types of questions. Appendix A and the last section in this chapter provide examples of research questions for each of the six types. We use the terms basic and complex because the more common names, univariate and multivariate, are not used consistently in the literature.

Note that some complex descriptive statistics (e.g., a cross-tabulation table) could be tested for significance with inferential statistics; if they were so tested they would no longer be considered descriptive. We think that most qualitative/constructivist researchers ask complex descriptive questions because they consider more than one variable/concept at a time but do not use inferential/hypothesis testing statistics. Furthermore, complex descriptive statistics are used to check reliability (e.g., Cronbach's alpha) and to reduce the number of variables (e.g., factor analysis).

Table 1.3. ***Summary of Types of Research Questions***

Type of Research Questions (Number of Variables)	Statistics (Example)
1) **Basic Descriptive Questions** – 1 variable	See Table 3.2 (mean, standard deviation, frequency distribution)
2) **Complex Descriptive Questions** – 2 or more variables, but no use of inferential statistics	(box plots, cross-tabulation tables, factor analysis, measures of reliability)
3) **Basic Difference Questions** – 1 independent and 1 dependent variable. Independent variable usually has a few values (ordered or not).	Table 7.1 (*t* test, one-way ANOVA)
4) **Complex Difference Question** – 3 or more variables. Usually 2 or a few independent variables and 1 or more dependent variables considered together.	Table 7.3 (factorial ANOVA, MANOVA)
5) **Basic Associational Questions** – 1 independent variable and 1 dependent variable. Usually at least 5 ordered values for both variables. Often they are continuous.	Table 7.2 (correlation tested for significance)
6) **Complex Associational Questions** – 2 or more independent variables and 1 or more dependent variables. Usually 5+ ordered values for all variables but some or all can be dichotomous variables.	Table 7.4 (multiple regression)

Difference versus associational inferential statistics. We think it is educationally useful, although not common in statistics books, to divide inferential statistics into two types corresponding to difference and associational hypotheses/questions. Difference inferential statistics are used for the experimental, quasi-experimental, and comparative approaches, which test for *differences between groups* (e.g., using analysis of variance). Associational inferential statistics test for *associations or relationships between variables* and use correlation or multiple regression analysis.[5] We will utilize this contrast between difference and associational inferential statistics in chapter 7 and later in this book.

[5] We realize that all parametric inferential statistics are relational so this dichotomy of using one type of data analysis procedure to test for differences (when there are a few values or levels of the independent variables) and another type of data analysis procedure to test for associations (when there are continuous independent variables) is somewhat artificial. Both continuous and categorical independent variables can be used in a general linear model

A Sample Research Problem - The High School and Beyond (HSB) Study

Imagine that you are interested in the general problem of what factors influence mathematics achievement at the end of high school. You might have some hunches or hypotheses about such factors based on your experiences and your reading of the research and popular literature. Some factors that might influence mathematics achievement are commonly called demographics; e.g., gender, ethnic group, and mother's and father's education. A probable influence would be the mathematics courses that the student has taken. We might speculate that grades in math and in other subjects could have an impact on math achievement.[6] However, other "third" variables, such as students' IQ or parent encouragement and assistance, could be the actual causes of high math achievement. Such extraneous variables could influence what courses one took, the grades one received, and might be correlates of the demographic variables. We might wonder how spatial performance scores such as pattern/mosaic score and visualization score might enter into a more complete understanding of the problem and whether these skills seem to be influenced by the same factors as math achievement. Finally, students' attitudes about mathematics might be factors affecting these math achievement scores.

Before we state the research problem and questions in more formal ways, we need to step back and discuss the types of variables and the approaches that might be used to study the above problem. Think about what are the *independent/antecedent* (presumed causes) *variables* and what are the *dependent/outcomes variable(s)* in the above problem. Hopefully, it is obvious that math achievement is the primary dependent variable.

Given the above research problem, which focuses on achievement tests at the end of the senior year, the number of math courses taken is best considered to be an antecedent or independent variable in this study. What about father's and mother's education and gender? How would you classify ethnic group in terms of the type of variable? What about grades? Like IQ and parent encouragement they would be independent variables, but, as with any study, we were not able to measure all the variables that might be of interest. Visualization and mosaic pattern scores could probably be either independent or dependent variables depending upon the specific research question. Finally, the math attitude questions and the resulting composite or scale scores derived from them also could be either independent or dependent variables, but probably independent/antecedent variables in this study. Note that student's class or grade level is not a variable in this study because all the participants are high school seniors (i.e., it does not vary; it is the population of interest).

As we have discussed, independent variables can be *active* (given to the participant or manipulated by the investigator) or *attributes* of the participants or their environments. Are there any *active* independent variables in this study? No! There is no intervention, new curriculum, or something similar. All the independent variables, then, are attribute variables because they are attributes or characteristics of these high school students. Given that all the independent variables are attributes, the research approach *cannot be experimental or quasi-experimental*. The proposed study is basically an individual differences one that will use the *comparative*, *associational*, and *descriptive approaches*. This means that we will *not* be able to draw definite conclusions about cause and effect (i.e., we will find out what is related to math achievement, but we will not know for sure what *causes* math achievement).

(regression) approach to data analysis. However, the practical implications are that most researchers adhere to the above dichotomy in data analysis.

[6] We have decided to use the short version of mathematics (i.e., math) throughout the book to save space, because it is used in common language, and because it is the name of several variables (e.g., *mathach*, *mathgr*) in the sample study.

Research Questions for the Modified HSB Study[7]

We will generate a large number of research questions from the modified HSB data set for Assignments A - L and N. Assignment M uses a different data set that you will enter. In this section, we will list one research question to be answered in each of the assignments to give you an idea of the range of types of questions that one might have in a typical research project like a thesis or dissertation. In addition to the *difference* and *associational questions* that are commonly seen in a research report, we have asked *descriptive questions* and questions about assumptions in the early assignments. Templates for writing the research problem and research questions/ hypotheses are given in Appendix A; it should help you write questions for your own research. The questions below correspond to the lab assignments in Chapters 4-18.

1) Often, we start with basic *descriptive questions* about the demographics of the sample. Thus, we could answer, with the results of Assignment A, the following basic descriptive question: "What is the average educational level of the fathers of the students in this sample?"

2) Additional basic *descriptive questions* about the sample will be answered in Assignment B. For example, "What percentages of the students are male and female?"

3) In Assignment C, we produce a number of new/transformed variables such as three summated scales assessing math attitudes. In this assignment we will examine whether the dependent and continuous independent variables (those that might be used to answer associational questions) are distributed normally, an *assumption* of many statistics. The question is, "Are the frequency distributions of the three math attitude scales markedly different from the normal curve distribution?"

4) We will produce cross-tabulation tables in Assignment D and ask "Is the association between gender and math grades statistically significant?" This is a basic associational question.

[7] The High School and Beyond (HSB) study was conducted by the National Opinion Research Center (1980). The example, discussed here and throughout the book, is based on 13 variables obtained from a random sample of 75 out of 28,240 high school seniors. These variables include achievement scores, grades, and demographics. The raw data for the 13 variables were obtained from an appendix in Hinkle, Wiersma, and Jurs (1994). Note that additional variables (ethnicity and math attitudes) with realistic but fictitious data have been added to the HSB data set in order to provide examples of common additional types of analysis (e.g., summated scales and Cronbach's alpha).

MORGAN George A. and Griego V. ORLANDO. 1998. *Easy Use and Interpretation of SPSS for Windows: Answering Research Questions With Statistics*. Mahwah: Lawrence Erlbaum Associates.

5) In Assignment E, we will answer additional basic *associational* research questions (using Pearson product-moment correlation coefficients) such as, "Is there a positive association/relationship between grades in high school and a math achievement?"

This assignment also will produce a correlation matrix of all the associations among seven key variables including math achievement. Similar matrixes will provide the basis for the answers to the issues raised in Assignments F, G, and H.

6) Assignments F and G are not really intended to provide answers to the research problem posed at the beginning of this section. Assignment F will deal with the issue of whether our conceptualization that there are three aspects of attitudes about mathematics (pleasure, motivation, and competence) is consistent with the ways the students answered the 13 attitude items. The research question might be phrased, "Using the SPSS factor analysis program, will the 13 math attitude items/questions cluster into the same three sets of questions that we proposed conceptually?" This is a complex descriptive question.

7) Whether there is internal consistency reliability of the summated scale scores (determined conceptually or from factor analysis) is another important *assumption* to test before proceeding with the formal research questions. This issue could be phrased, "Are the three scale scores computed from the math attitude questions internally consistent?" There are also other important measures of reliability that will be computed in Assignment G.

8) Assignment H will ask and answer a key research question which is a *complex associational question*: "Is there a combination of math attitudes (motivation, competence, and pleasure), grades, father's and mother's education, and gender that predicts math achievement better than any one of them alone, and, if so, what is the best combination?" Assignment I will answer similar questions.

9) Several basic *difference questions* will be asked in Assignment J. For example, "Do males and females differ on math achievement and grades in high school?"

10) *Basic difference questions* in which the independent variable has three or more values will be asked in Assignment K. For example, "Are there differences among Euro-American, African-American, Hispanic-American, and Asian-American students on math achievement?"

11) *Complex difference questions* will be asked in Assignment L. One *set* of three questions is as follows: (1) "Is there a difference between students who have fathers with no college, some college, and a BS or more with respect to the student's math achievement?" (2) "Is there a difference between students who had an A or B math grade average and those with less than a B average on a math achievement test at the end of high school?" and (3) "Is there an interaction between father's education and math grades with respect to math achievement?"

12) Assignment M will deal with repeated measures and mixed ANOVA questions using a different data set that you will enter into the computer.

13) Finally, Assignment N will answer *complex difference questions* similar to those in Assignments J and K when more than one dependent variable is considered simultaneously.

Another way to group these research questions that we have found useful is as follows:

a) Descriptive statistics about the *demographics of the sample.*
b) *Tests of assumptions* such as that the key variables are distributed normally and the instruments are assessed reliably.
c) Tests of the specific *research questions* posed by the researcher, based on the research problem. These can be *descriptive, associational,* and/or *difference* questions.
d) In addition, we often test other *supplementary questions*, which may be side issues or may arise after we have written the proposal or even after the data have been collected and analyzed.

This introduction to the research problem and questions raised by the HSB data set should help make the assignments meaningful, and it should provide a guide and examples for your own research.

MORGAN George A. and Griego V. ORLANDO. 1998. *Easy Use and Interpretation of SPSS for Windows: Answering Research Questions With Statistics.* Mahwah: Lawrence Erlbaum Associates.

CHAPTER 3

Measurement and Descriptive Statistics

According to S. S. Stevens (1951), "In its broadest sense measurement is the assignment of numerals to objects or events according to rules" (p.1). As we have seen in chapter 1, the process of research begins with a problem that is made up of a question about the relationship between two, or usually more, variables. Measurement is introduced when these variables are operationally defined by certain rules which determine how the participants' responses will be translated into numerals. These numbers can represent nonordered categories in which the numerals do not indicate a greater or lesser degree of the characteristic of the variable. Stevens went on to describe four scales or levels of measurement that he labeled: nominal, ordinal, interval, and ratio. Stevens and most writers since then have argued that the level or scale of measurement used to collect data is one of the most important determinants of the types of statistics that can be done appropriately with that data. As implied by the phrase "levels of measurement," these types of measurements vary from the most basic (nominal) to the highest level (ratio). However, since none of the statistics that are commonly used in social sciences or education require the use of ratio scales we will not discuss them to any extent.

Nominal Scales/Variables

These are the most basic or primitive forms of scales in which the numerals assigned to each category stand for the name of the category, but have no implied order or value. Males may be assigned the numeral 1 and females may be coded as 2. This does not imply that females are higher than males or that two males equal a female or any of the other typical mathematical uses of the numerals. The same reasoning applies to many other true nominal categories such as ethnic groups, type of disability, section number in a class schedule, or marital status (e.g., never married, married, divorced, or widowed). In each of these cases the categories are distinct and nonoverlapping, but not ordered, thus each category in the variable marital status is different from each other but there is no necessary order to the categories. Thus, the four categories could be numbered 1 for never married, 2 for married, 3 for divorced, and 4 for widowed or the reverse, or any combination of assigning a number to each category. What this obviously implies is that you must *not* treat the numbers used for identifying the categories in a nominal scale as if they were numbers that could be used in a formula, added together, subtracted from one another, or used to compute an average. Average marital status makes no sense. However, if one asks a computer to do average marital status, it will blindly do so and give you meaningless information. The important thing about nominal scales is to have clearly defined, nonoverlapping or mutually exclusive categories which can be coded reliably by observers or by self-report.

Qualitative or naturalistic researchers rely heavily, if not exclusively, on nominal scales and on the process of developing appropriate codes or categories for behaviors, words, etc. Although using qualitative/nominal scales does dramatically reduce the types of statistics that can be used with your data, it does not altogether eliminate the use of statistics to summarize your data and make inferences. Therefore, even when the data are nominal or qualitative categories, one's research may benefit from the use of appropriate statistics. We will return shortly to discuss the types of statistics, both descriptive and inferential, that are appropriate for nominal data.

Dichotomous Variables

It is often hard to tell whether a dichotomous variable, one with two values or categories (e.g., Yes or No, Pass or Fail), is nominal or ordered and researchers disagree. We argue that, although some such dichotomous variables are clearly nominal (e.g., gender) and others are clearly ordered (e.g., math grades--high and low), all dichotomous variables form a special case. Statistics such as the mean or variance would be meaningless for a three or more category nominal variable (e.g., ethnic group or marital status, as described above). However, such statistics do have meaning when there are only two categories. For example, in the HSB data the average gender is 1.55 (with males = 1 and females = 2). This means that 55% of the participants were females. Furthermore, we will see in Chapter 12, multiple regression, that dichotomous variables, called dummy variables, can be used as independent variables along with other variables that are interval scale. Thus, it is not necessary to decide whether a dichotomous variable is nominal, and it can be treated as if it were interval scale.

Table 3.1. *Descriptions of Scales of Measurement With Dichotomous Variables Added*

Scale	Description
Nominal	= 3 or more unordered or nominal categories
Dichotomous	= 2 categories either nominal or ordered (special case)
Ordinal	= 3 or more ordered categories, but *clearly unequal intervals* between categories or *ranks*
Interval	= 3 or more ordered categories, and *approximately equal intervals* between categories
Ratio	= 3 or more ordered categories, with equal intervals between categories and a true zero

Ordinal Scales/Variables (i.e., Unequal Interval Scales)

In ordinal scales there are not only mutually exclusive categories as in nominal scales, but the categories are ordered from low to high in much the same way that one would *rank* the order in which horses finished a race (i.e., first, second, third, ...last). Thus, in an ordinal scale one knows which participant is highest or most preferred on a dimension but the intervals between the various ranks are not equal. For example, the second place horse may finish far behind the winner but only a fraction of a second in front of the third place finisher. Thus, in this case there

MORGAN George A. and Griego V. ORLANDO. 1998. *Easy Use and Interpretation of SPSS for Windows: Answering Research Questions With Statistics*. Mahwah: Lawrence Erlbaum Associates.

are unequal intervals between first, second, and third place with a very small interval between second and third and a much larger one between first and second.

Interval and Ratio Scales/Variables (i.e., Equal Interval Scales)

Interval scales have not only mutually exclusive categories that are ordered from low to high, but also the categories are equally spaced (i.e., have equal intervals between them). Most physical measurements (length, weight, money, etc.) are ratio scales because they not only have equal intervals between the values/categories, but also have a true zero, which means in the above examples, no length, no weight, or no money. Few psychological scales have this property of a true zero and thus even if they are very well constructed equal interval scales, it is not possible to say that one has no intelligence or no extroversion or no attitude of a certain type. While there are differences between interval and ratio scales, the differences are not important for us because we can do all of the types of statistics that we have available with interval data. As long as the scale has equal intervals, it is not necessary to have a true zero.

Distinguishing Between Ordinal and Interval Scales

It is usually fairly easy to tell whether three categories are ordered or not, so students and researchers can distinguish between nominal and ordinal data, except perhaps when there are only two categories, and then it does not matter. The distinction between nominal and ordinal makes a lot of difference in what statistics are appropriate. However, it is considerably harder to distinguish between ordinal and interval data. While almost all *physical* measurements provide either ratio or interval data, the situation is less clear with regard to psychological measurements.

When we come to the measurement of psychological characteristics such as attitudes, often we cannot be certain about whether the intervals between the ordered categories are equal, as required for an interval level scale. Suppose we have a five-point scale on which we are to rate our attitude about a certain statement from strongly agree as 5 to strongly disagree as 1. The issue is whether the intervals between a rating of 1 and 2, 2 and 3, 3 and 4, and 4 and 5 are all equal or not. One could argue that because the numbers are equally spaced on the page, and because they are equally spaced in terms of their numerical values, the subjects will view them as equal intervals. However, especially if the in-between points are identified (e.g., strongly agree, agree, neutral, disagree, and strongly disagree), it could be argued that the difference between strongly agree and agree is not the same as between agree and neutral; this contention would be hard to disprove. Some questionnaire or survey items have response categories that are not exactly equal intervals. For example, let's take the case where the subjects are asked to identify their age as one of five categories: 21 to 30, 31 to 40, 41 to 50, 51 to 60, and 61 and above. It should be clear that the last category is larger in terms of number of years covered than the other four categories. Thus, the age intervals are not exactly equal. However, we would consider this scale and the ones above to be at least *approximately interval.*

On the other hand, an example of an ordered scale that is clearly not interval would be one that asked how frequently subjects do something. The answers go something like this: every day, once a week, once a month, once a year, once every 5 years. You can see that the categories

become wider and wider and, therefore, are not equal intervals. There is clearly much more difference between 1 year and 5 years than there is between 1 day and 1 week. Most of the above information is summarized in the top of Table 3.2.

Table 3.2. *Selection of Appropriate Descriptive Statistics for One Dependent Variable*

	Level/Scale of Measurement of Variable		
	Nominal	**Ordinal**	**Interval or Ratio**
Characteristics of the Variable	- Qualitative data - Not ordered - True categories: only names, labels	- Quantitative data - Ordered data - Rank order only	- Quantitative data - Ordered data - Equal intervals between values
Examples	Gender, school, curriculum type, hair color	1st, 2nd, 3rd place, ranked preferences	Age, height, good test scores, good rating scales
Frequency Distribution	Redhead- III Blond - IIII Brunette- II	Best - II Better - III Good - III	5 - I 4 - II 3 - III 2 - III 1 - II
Frequency Polygon/ Histogram	No	Yes	Yes
Bar Graph or Chart	Yes	Yes	Yes
Central Tendency			
Mean	No	Mean Rank	Yes
Median	No	Yes	Yes
Mode	Yes	Yes	Yes
Variability			
Standard Deviation	No	of Ranks	Yes
Range	No	Yes, but[1]	Yes
How many categories	Yes	Yes	Yes
Percent in each	Yes	Yes	Yes
Shape			
Skewness	No	No	Yes
Kurtosis	No	No	Yes

[1] The range of ordinal data may well be misleading

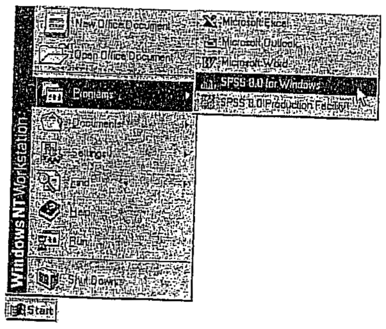
1. lekce

POVAHA HROMADNÝCH DAT A LOGIKA SURVEY. PRÁCE S HROMADNÝMI DATY PŘED JEJICH ANALÝZOU (Modul FILES: procedury), PRÁCE S PROSTŘEDÍM (Moduly Edit, View, Utilities) A VÝSTUPY Z ANALÝZY (Modul : Output).

Starting SPSS for Windows

The easiest way to run SPSS for Windows is by using the Start button. During the installation of SPSS, the Setup procedure adds SPSS to the menu that appears when you click the Start button, as shown in Figure 2.1.

Figure 2.1 SPSS on the Start menu

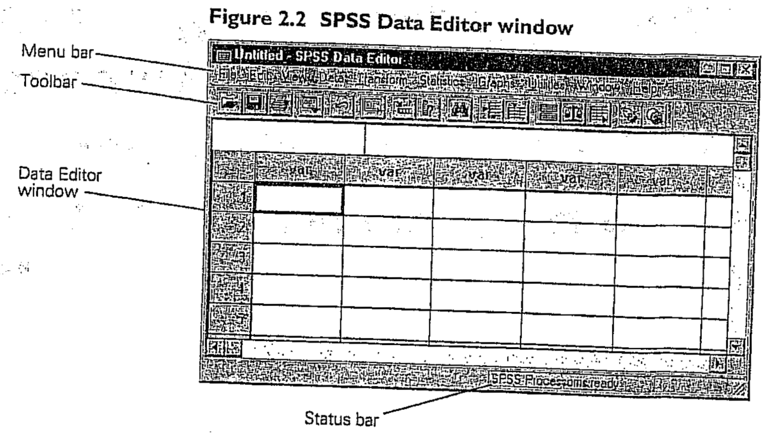


Always use the left mouse button unless the right one is specifically indicated.

▶ To start SPSS, click Start to display the Start menu, then click SPSS 8.0 for Windows.

The SPSS Data Editor window is displayed, as shown in Figure 2.2. You can move it, like any other window, by clicking and dragging its title bar, or resize it by clicking and dragging its sides or corners.

Figure 2.2 SPSS Data Editor window


Menu bar
Toolbar
Data Editor window
Status bar
Untitled - SPSS Data Editor
File Edit View Data Transform Statistics Graphs Utilities Window Help
var
SPSS Processor is ready

Opening a Data File

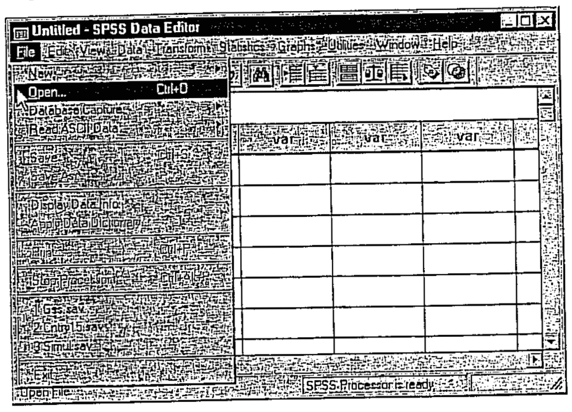
The SPSS Data Editor window displays your working data file. You don't have one yet—that's why the Data Editor is empty. If you have data of your own that are not in the computer yet, you can type the numbers right into the Data Editor. If the data are already in a spreadsheet or database file, you can probably read that file into SPSS. The data used in this book are already in the form of SPSS data files. To use them for the exercises, or just to follow along in the analysis, simply open the appropriate data file. To open a data file:

▶ Click the left mouse button on the word File on the SPSS Data Editor menu bar, as shown in Figure 2.3.

The File menu is displayed.

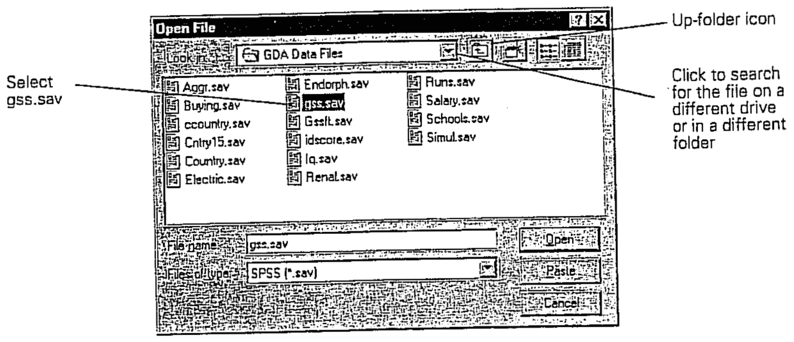
▶ On the File menu, click Open.

Figure 2.3 Opening a data file



When you click Open on the File menu, the Open File dialog box appears, as shown in Figure 2.4.

Figure 2.4 Open File dialog box


Select gss.sav
Up-folder icon
Click to search for the file on a different drive or in a different folder
Open File
Open
Paste
Cancel

Look in: GDA Data Files

Aggr.sav, Buying.sav, ccountry.sav, Cntry15.sav, Country.sav, Electric.sav, Endorph.sav, gss.sav, Gssft.sav, idscore.sav, Iq.sav, Renal.sav, Runs.sav, Salary.sav, Schools.sav, Simul.sav

File name: gss.sav

Files of type: SPSS (*.sav)

▶ Click the *gss.sav* data file where it appears in the list.

▶ Click Open.

What if the gss.sav file doesn't appear? Only files in the current drive and directory are listed. The file you want may either be in another directory or saved on a different drive.

To look in a parent folder (one that contains the current folder), click the up-folder icon, as shown in Figure 2.4.

To look in a subfolder (one contained in the current folder), double-click it in the list.

To look on a different drive, click the up-folder icon repeatedly until you reach My Computer, then double-click the desired drive icon and continue down through the folder hierarchy on that drive. ■■■

When SPSS has finished reading the data file, it displays the data in the Data Editor, as shown in Figure 2.5. This particular data file contains selected information for 1500 people who were interviewed in the 1993 General Social Survey, which annually asks a broad range of questions to a sample of adults in the United States population.

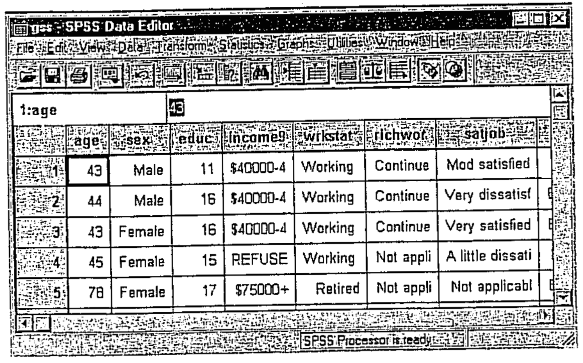
To view the data in the Data Editor, from the menus choose:

Window
gss - SPSS Data Editor

If your screen displays all numbers rather than value labels such as Male and Female in the cells, from the menus choose:

View
Value Labels

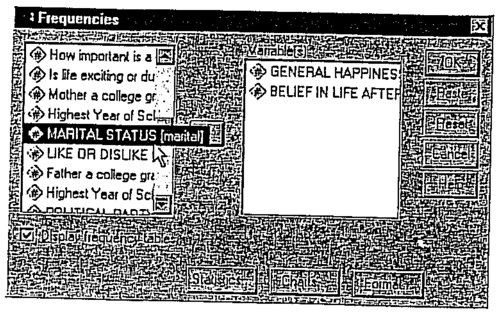
Figure 2.5 Data Editor window with GSS data


gss - SPSS Data Editor
File Edit View Data Transform Statistics Graphs Utilities Window Help
SPSS Processor is ready

1:age 43

	age	sex	educ	income9	wrkstat	richwor	satjob
1	43	Male	11	$40000-4	Working	Continue	Mod satisfied
2	44	Male	16	$40000-4	Working	Continue	Very dissatisf
3	43	Female	16	$40000-4	Working	Continue	Very satisfied
4	45	Female	15	REFUSE	Working	Not appli	A little dissati
5	78	Female	17	$75000+	Retired	Not appli	Not applicabl

Figure 2.7 Frequencies dialog box with default variable labels



To make this book easier to read, we'll use variable names instead of labels in dialog boxes, as shown in Figure 2.7. To display variable names rather than labels in your dialog boxes (so you can follow along with the text), you need to change one of SPSS's default options.

- From the menus select:

  Edit
  Options...

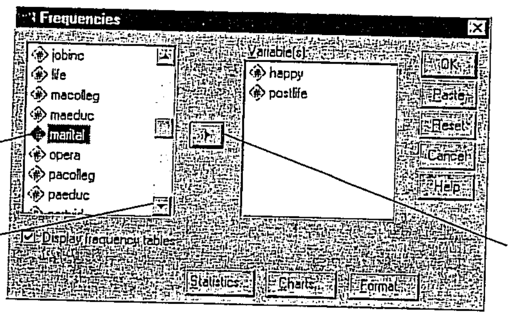
- In the Options dialog box, click the General tab.

- In the Variable Lists group box, click Display names.

- Click OK.

This change doesn't take effect until the next time you open a data file. The effect of the changed option is shown in Figure 2.8.

Figure 2.8 Frequencies dialog box



Variable selected in source list

Click to scroll

Click to move variables between lists

To use this dialog box:

- Click *happy* in the scroll list and then click ▶.

  This moves *happy* into the Variable(s) list.

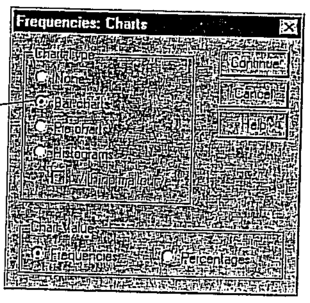
As a shortcut to scroll the source list, click in the list and type the letter p. This scrolls to the first variable beginning with p.

- Scroll down the source list until you see *postlife* and move it into the Variable(s) list as well.

- Click Charts.

This opens the Frequencies Charts dialog box, as shown in Figure 2.9. Here you can request charts along with your frequency tables.

Figure 2.9 Frequencies Charts dialog box



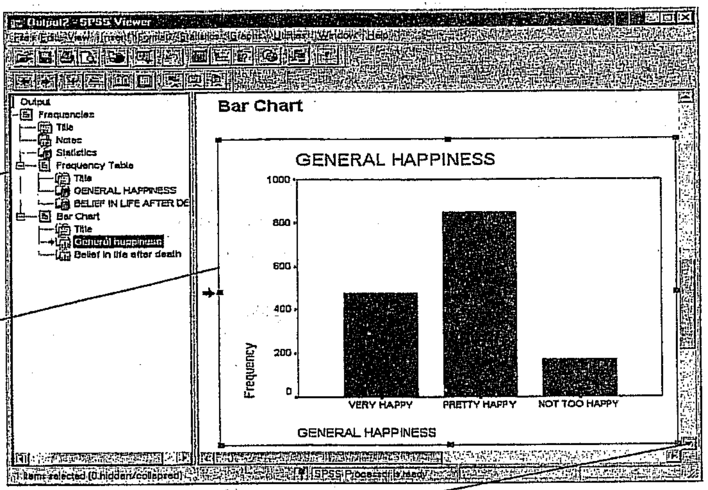
Select Bar charts

- Select Bar charts, as shown in Figure 2.9.

The Viewer Window

The Viewer window is where you see the statistics and graphics—the output—from your work in SPSS. As shown in Figure 2.10, the Viewer window is split into two parts, or **panes**. (A piece of a window is often called a pane in computer software, just as it is at your local hardware store.)

Figure 2.10 Viewer window



Outline pane

Display pane

Click here to scroll through output

The left side (the **outline pane**) contains an outline view of all the different pieces of output in the Viewer, whether they are currently visible or not. The right side (the **display pane**) contains the output itself.

- To change the sizes of the two panes (for example, to make the display pane wider), just point the mouse at the line that divides them, press the left mouse button, and drag the line to the left or right.

It's possible to ignore the outline pane and simply scroll through the output displayed in the display pane on the right side of the Viewer. The outline view offers some handy tricks, however.

The Outline Pane

Individual portions of output are associated with "book" icons in the outline pane. Each icon represents a particular piece of output, such as a table of statistics or a chart.

- If you click one of these icons in the outline pane, the associated piece of output appears instantly in the display pane. (But it may be hidden! See below.)

These icons are the quickest navigational controls in the Viewer.

The book icons are also used to hide or display pieces of output temporarily. Notice that most of them in the outline pane are "open book" icons, while a few look more like closed books. A "closed book" icon represents a hidden piece of output. Hidden output doesn't appear in the display pane but can be recovered any time you want to look at it.

- To hide a single piece of output, double-click the open book icon. This closes the icon and hides the output associated with it.

- To display a hidden piece of output, double-click the closed book icon. This opens the icon and displays the output associated with it.

- To hide *all* of the output from a procedure such as Frequencies, click the little box containing a minus sign to the left of the procedure name. That whole part of the outline collapses, and the minus sign changes to a plus sign to show you that more output is hiding there. Click the plus sign to show it all again.

You will find that you can do lots of things in fairly obvious ways by playing with the outline pane. Try rearranging the output (press the left mouse button on a book icon, drag it to a different place in the outline, and then release the mouse button), or deleting part of the output (click the icon and press the Delete key). The SPSS Help system can tell you all the details.

The Display Pane

The display pane shows as much of the SPSS output as can fit in it. To see more, you can either scroll the pane or use the outline pane to jump around.

The output in the display pane includes several different kinds of objects: tables of numbers (actually a special kind of tables, called pivot tables); charts; and bits of text such as titles. You have complete control over the appearance, and even the content, of most of these objects.

- To change something about an object, double-click it in the display pane.

Double-clicking an object opens an editor that is specially designed to modify it. The appearance of the object changes to show that you are editing it. The menu bar may change. If the object is a chart, a special chart editing window opens to offer you a powerful set of tools for changing the chart's appearance.

Let's look at these objects in the Viewer.

Viewer Objects

In the outline panel, the first line is a container for the entire batch of output. It's simply called Output. There might be a line below it called Log, which isn't going to be discussed in this book. The next line, *Frequencies*, is a heading that contains all the various kinds of output produced by the Frequencies procedure that you just ran. In order, they are:

- Title. The title of the procedure, which is simply text.
- Notes. Notes are usually hidden, so this probably looks like a closed book in the outline pane.
- Statistics. This is a pivot table, which reports the number of cases, or "observations," that were processed by the Frequencies procedure. Most procedures start by producing such a table. The icon is an open book, so if you click it, the display pane will show you what it looks like.
- A frequency table for the first variable processed (*happy*). Frequency tables are discussed in Chapter 3. Note that the icon in the outline pane is labeled *GENERAL HAPPINESS*, which is a descriptive label that was assigned to the variable *happy* when the data file was set up.

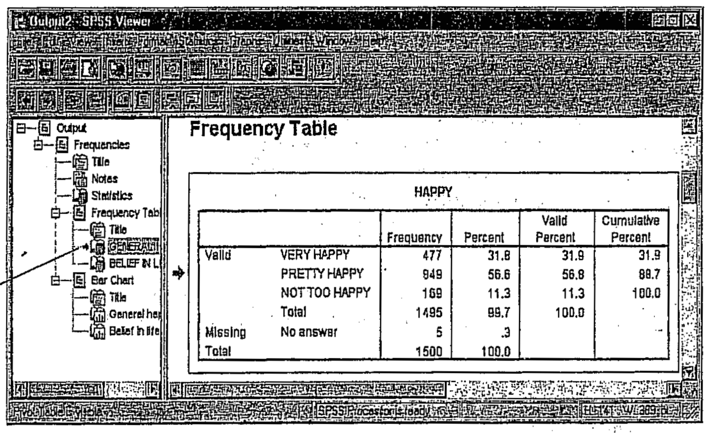
- A frequency table for the next variable, *postlife*, whose icon is labeled *BELIEF IN LIFE AFTER DEATH* in the outline pane.
- A bar chart for *happy*.
- A bar chart for *postlife*.

Let's see what these pivot tables and charts are like.

Pivot Tables

First, a pivot table. Most of SPSS's tabular and statistical output appears in the Viewer in the form of pivot tables.

Figure 2.11 Pivot tables in the Viewer



Frequency Table

HAPPY

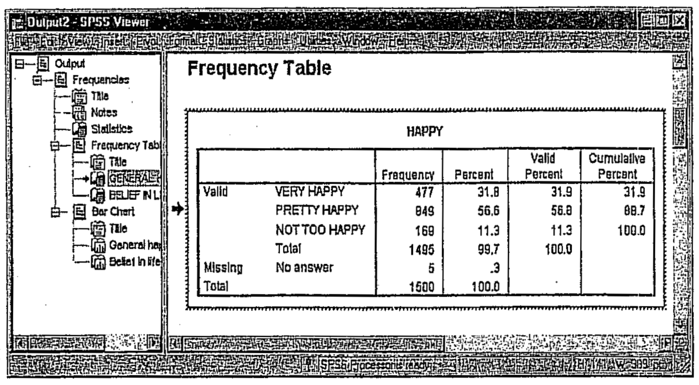
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	VERY HAPPY	477	31.8	31.9	31.9
	PRETTY HAPPY	849	56.6	56.8	88.7
	NOT TOO HAPPY	169	11.3	11.3	100.0
	Total	1495	99.7	100.0	
Missing	No answer	5	.3		
Total		1500	100.0		

Click here to display the frequencies table for GENERAL HAPPINESS

▶ In the outline, click the icon for the pivot table labeled *GENERAL HAPPINESS*. The table instantly appears in the display pane, with an arrow pointing to it, as shown in Figure 2.11.

▶ Move the mouse over to the display pane and double-click on the table itself to indicate that you want to edit it.

Figure 2.12 An activated pivot table



Frequency Table

HAPPY

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	VERY HAPPY	477	31.8	31.9	31.9
	PRETTY HAPPY	849	56.6	56.8	88.7
	NOT TOO HAPPY	169	11.3	11.3	100.0
	Total	1495	99.7	100.0	
Missing	No answer	5	.3		
Total		1500	100.0		

Not a lot seems to happen in Figure 2.12. The pivot table is now surrounded by a cross-hatched line to indicate that it is active in the Pivot Table Editor. The SPSS toolbar vanishes, and if you watch carefully, the menu bar changes—there is now a Pivot menu.

- Double-clicking a pivot table lets you edit it "in place"; that is, right where it sits in the display pane of the Viewer. If you need more room, select the pivot table by clicking once with the left mouse button, and from the menus choose:

Edit
  SPSS Pivot Table Object ▸
    Open...

This command opens the pivot table into a window of its own.

When you are editing a pivot table either way, you can change almost anything about it you want. If you don't like the label, just double-click it. It reappears as highlighted text. Type in the label the way you want it, perhaps *Happiness in General*, and click somewhere else to enter the new label. To change the font of the title or make it bold or italic, click the title and from the menus choose:

Format
  Font...

Then choose a different font or a bold or italic style.

If you don't like the way numbers are displayed in the pivot table, make sure the Pivot Table Editor is active (by double-clicking the table in the display pane), and then either make a selection from the Format menu, or *right-click* one of the numbers in the table to pop up a context menu for it. Most of the things you might want to change can be found in either of these menus under Table Properties or Cell Properties. Check out Table Looks, too, to see how you can apply consistent sets of formatting to whole tables.

Changing fonts and styles and even the text of the labels in a table can make a big difference in the way the table looks. An SPSS pivot table lets you do much more than that, however. You can change the basic organization of the data presented in the table. The Pivot menu (which appears only when you have double-clicked a pivot table in the display frame to activate it) gives you access to powerful tools for reorganizing the table. To get a feel for these tools, activate a pivot table, and from the menus choose:

Pivot
  Transpose Rows and Columns

The same information is displayed. The different codes or responses to the question, which were laid out vertically, are now laid out horizontally; the different types of statistical summaries, which were laid out horizontally, are now laid out vertically.

To see the pivot table as it was before, simply transpose the rows and columns again.

This example is a very simple pivot table. Multidimensional tables offer many more structural possibilities. You can explore those in the SPSS online Help system, or if you like, to see how things work you can build a complex table and start pivoting.

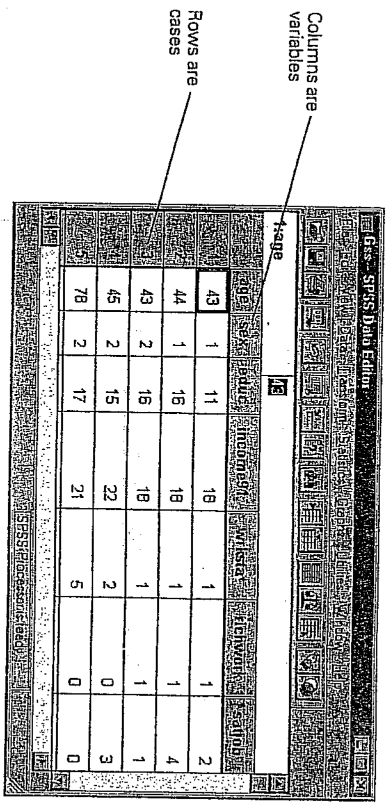
NORUŠIS, Marija J. 1998. *Guide to Data Analysis*. Upper Saddle River, Prentice Hall.

The Data Editor Window

Let's take a closer look at the Data Editor (see Figure 2.14). You can select it from the Window menu or simply click on it if any part of it is visible on your screen.

If you've ever used a spreadsheet, the Data Editor should look familiar. It's just an array of rows and columns. In the Data Editor, each row is a case, and each column is a variable. Cases and variables are fundamental concepts in data analysis. It's time we stopped to define them.

Figure 2.14 Data Editor window



1:age 43

age	sex	educ	income91	wrkstat	rincwork	satjob
43	1	11	18	1	1	2
44	1	16	18	1	1	4
43	2	16	18	1	1	1
45	2	15	22	2	0	3
78	2	17	21	5	0	0

Cases (rows) are the people who participate in a survey or experiment. (Another word often used is **observation.**) Actually, a case need not be a person. It can be anything. If you're doing experiments on rats, the case is the individual rat. If you're studying the beef content of hamburgers, each hamburger is a case. Generally speaking, the case is the unit for which you take measurements.

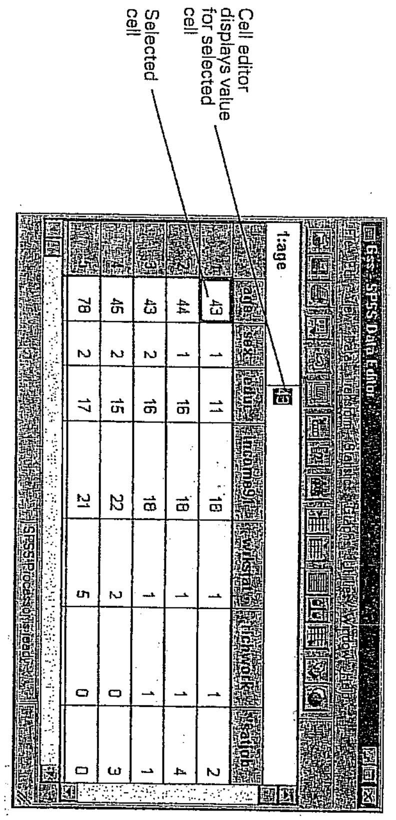
Variables (columns) are the different items of information you collect for your cases. Think about the way you conduct a survey. You ask each person for the same type of information: date of birth, sex, marital status, education, views on whatever subjects your survey is about. Each item for which you record an answer is known as a variable. The answer a particular person gives is known as the **value** for that variable. Year of birth is a variable; responses such as 1952 or 1899 are values for that variable.

The intersection of the row and the column is called a **cell.** Each cell holds the value of a particular case for a particular variable. You can edit values in the Data Editor, as follows:

▶ Click in one of the cells with the mouse.

The cell editor displays the value for the selected cell, as shown in Figure 2.15.

Figure 2.15 Data Editor with cell selected



1:age 43

age	sex	educ	income91	wrkstat	rincwork	satjob
43	1	11	18	1	1	2
44	1	16	18	1	1	4
43	2	16	18	1	1	1
45	2	15	22	2	0	3
78	2	17	21	5	0	0

▶ Type a number to replace the existing value and press ↵Enter.

The new value appears in the cell editor as you type it, but the value in the cell is not updated until you press ↵Enter.

▶ Change another value in the cell editor, but instead of pressing ↵Enter, press Esc.

When you press Esc rather than ↵Enter, the original value in the cell remains unchanged.

MORGAN George A. and Griego V. ORLANDO. 1998. *Easy Use and Interpretation of SPSS for Windows: Answering Research Questions With Statistics.* Mahwah: Lawrence Erlbaum Associates.

CHAPTER 2

Overview of the High School and Beyond (HSB) Data Set and SPSS 7.5

The Modified Hsbdata File

The file name of the data set used with this manual is hsbdata; it stands for high school and beyond data. It is based on a national sample of data from more than 28,000 high school students. The current data set is a sample of 75 students drawn randomly from the larger population. The data that we have from this sample includes school outcomes such as grades and the number of mathematics courses of different types that the students took in high school. Also there are several kinds of standardized test data and demographic data such as gender and mother's and father's education. To provide an example of questionnaire type data, we have included 13 questions about math attitudes. These data were developed for this manual and, thus, are not really the math attitudes of the 75 students in this sample. The questions, however, are based on ones used by the authors to study mastery motivation. Also we made up ethnic group data which, although somewhat realistic overall, do not represent the actual ethnic groups of the 75 students in this sample. This enables us to do some additional analyses.

We have provided you with a disk which contains the data for each of the 75 participants on 28 variables. The hsbdata file, shown in Table 2.1, has already been entered and labeled to enable you to get started on analyses quickly. In Assignments A and M, you will enter some additional data to practice entering it yourself. Also you will, in several assignments, label variables and their values so that your printouts will include the new variable names and the value labels.

The Raw HSB Data and Data Editor

Notice the short variable names at the top of the hsbdata file. (Actually we have transferred the HSB file from the SPSS data editor to Excel and reduced it so that it would fit on two pages, but in SPSS it will look very similar to Table 2.1.) Be aware that the subjects/participants are listed down the page from ID 1 to ID 75 at the bottom of the second page, and the variables are listed across the top. You will always enter data this way. If a variable is measured more than once, such as a pretest and posttest, it will be entered as two variables perhaps called Pre and Post. This method of entering data follows that suggested in chapter 7. Note that most of the values are single digits but that *visual, mosaic,* and *mathach* include some decimals and even minus numbers. Notice also that some cells like variable Q09 for participant ID 1 are blank because a datum is missing. Perhaps participant 1 did not answer question 9 and participant 2 did not answer question 4, etc. Blank is the "system missing" value that can be used for any missing data in an SPSS data file. However, other values also can be used for missing data. Notice that for father's and mother's education level we have used -1 for the missing values, and for ethnic group we have defined 9 as missing. For your purposes, however, we suggest that you leave missing data blank, but you may run across "user defined" missing data codes like -1 or 9 in other researchers' data.

Table 2.1. *Hsbdata Data Set in the SPSS Data Editor*

ID	VISUAL	MOSAIC	GRADES	MATHGR	ALG1	ALG2	GEO	TRIG	CALC	MATHACH	FAED	MAED	GEND	Q01	Q02	Q03	Q04	Q05	Q06	Q07	Q08	Q09	Q10	Q11	Q12	Q13	ETHNIC
1	8.75	31.0	4	0	0	0	0	0	0	9.00	10	10	2	10	4	3	1	4	2	2	1	.	2	1	3	3	2
2	4.75	56.0	5	0	0	0	0	0	0	10.33	2	2	2	3	40	3	.	1	2	3	3	3	1	1	4	2	3
3	4.75	25.0	6	1	0	0	0	0	0	7.67	2	2	2	1	1	1	4	4	4	1	2	4	1	1	3	2	2
4	1.00	22.0	3	0	1	0	0	0	0	5.00	3	3	1	1	1	1	2	4	3	1	3	4	3	4	4	3	1
5	2.25	17.5	3	0	0	0	0	0	0	-1.67	-1	3	2	1	4	1	4	4	4	1	4	1	1	4	2	2	2
6	1.00	23.5	5	1	0	0	0	0	0	1.00	3	2	2	1	1	1	4	4	4	1	4	1	1	4	1	2	3
7	2.50	28.5	6	0	1	1	1	0	0	12.00	9	6	1	.	1	3	1	1	3	3	1	3	1	1	3	3	1
8	3.50	29.5	4	0	1	0	0	0	0	8.00	5	3	2	1	1	3	2	4	2	1	2	3	2	1	3	3	1
9	3.50	28.0	7	1	1	0	0	0	0	13.00	3	3	2	3	1	.	1	1	2	2	2	4	2	1	3	2	1
10	3.75	27.5	5	0	0	0	0	0	0	3.67	8	2	1	1	3	1	4	4	4	1	3	3	2	4	3	4	2
11	11.00	27.0	6	0	1	1	1	1	1	21.00	3	4	1	4	4	4	3	1	4	4	1	4	4	2	4	4	4
12	4.75	26.5	8	1	1	1	1	1	1	23.67	8	9	1	3	4	4	2	1	2	3	2	4	2	1	3	3	1
13	1.00	13.0	5	0	1	0	0	0	0	4.00	2	2	2	1	4	1	3	4	2	1	1	4	1	1	4	3	3
14	1.00	18.0	2	0	1	0	0	0	0	9.00	6	3	1	3	4	3	2	1	1	4	1	4	1	1	4	3	2
15	4.75	25.0	3	0	1	1	0	0	0	5.33	2	3	2	4	4	3	2	1	3	3	1	4	1	1	3	3	1
16	13.50	33.0	7	1	1	1	1	0	0	19.67	-1	3	1	3	4	3	2	1	3	4	2	4	1	2	4	3	2
17	-.25	33.0	7	0	1	0	1	0	0	7.67	3	3	1	4	4	3	2	1	3	3	1	4	1	1	4	3	1
18	6.00	27.0	5	0	1	0	1	0	0	14.33	9	7	1	3	4	3	2	1	1	4	2	3	1	1	3	3	2
19	8.75	26.0	6	1	1	1	1	0	0	14.33	3	3	1	3	4	3	1	2	2	4	1	4	1	1	4	2	3
20	8.75	25.0	8	1	1	1	1	1	0	19.67	8	3	2	3	2	1	2	2	4	3	2	3	3	2	4	3	4
21	-.25	41.0	5	0	1	0	1	0	0	17.00	5	5	1	4	4	4	1	1	3	3	1	3	1	1	4	4	1
22	7.25	27.0	3	0	0	0	0	0	0	7.67	2	3	2	2	4	3	1	2	1	4	2	1	1	1	3	2	1
23	4.75	22.0	5	0	1	1	1	1	0	19.67	9	7	1	4	4	3	1	1	3	3	2	3	2	1	4	3	1
24	1.00	53.0	4	0	0	0	0	0	0	6.67	2	2	2	3	3	2	2	2	1	1	2	2	2	2	3	3	1
25	9.75	27.0	7	1	1	1	1	1	0	14.33	6	6	2	3	4	4	2	1	3	3	2	3	1	1	3	2	2
26	7.25	31.0	5	0	1	0	0	0	0	10.67	5	3	2	3	4	3	3	2	3	3	2	3	1	2	2	3	3
27	2.25	29.0	6	0	1	0	1	0	0	14.33	3	6	2	4	4	3	3	1	1	2	3	3	1	1	3	3	4
28	2.25	26.0	5	1	1	1	0	0	0	14.33	3	2	1	3	4	4	1	1	1	3	1	4	1	1	4	3	1
29	1.00	22.0	6	1	1	0	0	0	0	14.33	5	3	2	4	4	3	2	1	2	3	2	3	1	2	3	3	1
30	13.50	31.0	3	0	0	0	0	0	0	9.00	2	3	1	2	3	2	3	1	3	4	2	3	1	1	3	3	3
31	9.75	27.0	7	1	1	1	1	1	0	17.00	2	3	2	2	4	4	2	1	1	3	2	4	1	1	3	2	2
32	9.50	34.0	6	0	1	1	1	0	0	23.67	9	5	1	3	4	2	2	2	3	2	2	4	1	1	4	3	1
33	9.75	24.5	4	0	1	1	1	1	0	10.33	5	7	1	4	4	4	1	1	2	4	1	4	1	1	3	4	4
34	6.00	23.0	7	1	1	1	1	1	0	15.67	8	8	1	3	4	3	2	2	4	2	2	3	1	1	3	3	3
35	2.25	32.5	8	1	1	0	0	0	0	17.00	4	2	2	3	4	3	3	2	2	2	2	3	2	2	3	2	1
36	9.75	29.5	6	0	1	1	1	1	0	15.67	8	6	1	3	3	3	3	2	3	2	2	3	2	2	1	1	1
37	-.25	33.0	7	0	1	0	1	0	0	7.67	3	3	1	3	4	3	1	1	2	4	1	4	1	1	4	3	1
38	3.75	24.0	7	1	1	1	1	1	1	22.67	2	2	2	3	4	4	1	1	2	4	4	4	1	1	4	4	1
39	6.00	11.0	7	1	1	1	0	0	0	14.33	4	2	2	4	4	3	3	1	3	2	4	4	1	1	4	4	2
40	6.50	23.0	5	0	1	1	1	0	0	21.00	2	2	1	2	4	2	2	3	1	3	3	4	1	1	2	2	4
41	1.00	26.0	6	1	0	0	0	0	0	6.33	2	2	2	2	4	3	2	1	4	3	3	4	1	1	2	2	1
42	6.00	22.5	6	1	1	0	1	0	0	3.67	3	6	2	4	2	4	1	1	3	4	1	2	2	1	3	4	1
43	3.50	20.5	7	0	1	1	0	0	0	10.33	10	10	1	3	4	3	4	1	3	3	3	4	1	2	3	3	2
44	3.50	28.0	8	1	1	1	1	1	1	23.67	8	8	2	3	4	3	2	1	1	4	1	4	1	1	4	3	1

MORGAN George A. and Griego V. ORLANDO. 1998. *Easy Use and Interpretation of SPSS for Windows: Answering Research Questions With Statistics*. Mahwah: Lawrence Erlbaum Associates.

Table 2.1. *Hsbdata Data Set in the SPSS Data Editor* (continued)

ID	VISUAL	MOSAIC	GRADES	MATHGR	ALG1	ALG2	GEO	TRIG	CALC	MATHACH	FAED	MAED	GEND	Q01	Q02	Q03	Q04	Q05	Q06	Q07	Q08	Q09	Q10	Q11	Q12	Q13	ETHNIC
45	1.00	16.0	7	1	1	0	0	0	0	4.00	2	3	2	4	4	3	3	1	2	2	3	3	1	1	3	2	2
46	-.25	30.0	6	0	1	1	0	0	.	17.00	4	6	2	3	4	3	1	1	2	4	1	3	1	1	3	3	1
47	-.25	23.5	6	0	1	1	1	0	0	6.33	10	9	2	2	4	2	4	1	3	1	3	3	1	1	2	2	1
48	-.25	20.5	5	0	1	0	0	0	0	3.67	2	3	2	3	4	2	3	2	1	4	1	4	1	1	2	3	1
49	4.75	22.5	5	0	1	0	0	0	0	9.00	3	5	2	4	3	4	3	1	2	4	2	4	2	1	2	2	1
50	5.00	24.0	6	0	1	1	1	1	0	20.33	8	7	2	3	4	3	3	1	3	3	2	4	1	1	3	3	1
51	4.75	37.0	6	0	1	0	0	0	0	13.00	2	2	2	4	4	4	2	1	3	4	1	3	1	2	3	3	4
52	4.75	31.0	5	1	1	0	0	0	0	12.00	2	3	2	2	4	4	3	1	3	1	3	4	1	1	2	2	9
53	5.00	56.0	6	1	0	0	0	0	0	17.00	6	2	2	3	3	3	2	1	2	2	2	3	2	1	3	2	2
54	11.00	35.5	8	1	1	1	1	1	1	23.67	5	7	2	3	4	2	2	2	2	3	2	3	1	2	2	3	3
55	11.00	28.0	8	1	1	1	1	1	0	21.00	2	2	1	3	3	3	2	1	2	3	2	3	1	1	3	3	1
56	7.25	36.0	7	1	1	1	1	1	1	22.33	10	9	1	3	4	3	1	1	1	4	1	3	1	1	3	2	1
57	3.50	35.0	3	0	0	0	0	0	0	2.33	2	3	2	4	4	3	1	1	2	4	1	4	1	1	2	3	2
58	7.25	44.0	8	1	1	1	1	1	0	22.33	10	6	2	4	4	3	3	2	1	2	3	2	4	1	1	1	3
59	3.50	-4.0	7	1	1	0	0	0	0	5.00	2	2	2	2	3	2	2	2	3	1	3	3	1	1	1	1	1
60	6.00	24.5	5	1	0	0	0	0	0	5.00	2	3	2	3	4	2	2	2	1	3	1	3	1	1	3	3	2
61	9.75	25.0	8	1	1	1	1	0	0	10.33	2	3	2	3	3	3	2	2	3	3	2	3	2	2	3	2	3
62	2.25	51.5	3	0	0	0	0	0	0	5.00	3	3	1	4	4	3	4	1	1	1	4	4	1	1	2	1	1
63	1.00	25.0	7	0	1	0	0	0	0	1.00	10	3	2	4	4	4	1	1	2	4	1	4	1	1	4	4	1
64	3.50	30.5	4	0	1	0	1	0	0	13.00	2	3	1	3	4	2	2	2	3	2	2	3	1	1	3	3	4
65	4.75	29.0	4	0	0	0	0	0	0	9.33	3	2	2	3	4	3	2	1	1	3	2	4	1	1	4	3	1
66	-.25	20.0	7	0	1	1	0	0	0	10.33	2	3	2	4	3	4	2	1	3	3	1	4	3	1	3	3	9
67	4.75	22.0	7	1	1	1	1	1	1	23.67	9	3	1	4	4	3	2	1	3	3	1	3	1	1	3	3	1
68	6.00	7.5	3	0	0	0	0	0	0	5.00	5	5	1	3	4	2	1	1	1	3	1	3	1	1	2	2	1
69	7.25	32.0	6	1	1	1	1	1	1	23.67	9	6	1	4	4	4	3	1	3	3	1	4	1	1	4	2	1
70	1.00	23.0	5	0	1	1	0	0	0	14.67	3	3	1	4	4	3	1	1	2	4	1	4	1	1	4	4	1
71	11.00	24.5	4	0	1	1	1	1	0	18.67	7	3	1	3	4	1	2	3	4	4	1	2	1	2	3	1	1
72	8.50	23.0	7	0	1	1	1	0	0	11.67	5	4	1	4	4	4	3	1	4	1	3	3	3	1	2	1	1
73	8.50	30.5	4	0	1	0	0	0	0	14.33	3	6	1	2	2	2	2	2	3	3	2	3	3	2	3	3	1
74	14.75	13.5	5	0	1	0	1	0	0	11.67	3	3	1	3	3	2	3	2	3	2	3	3	2	2	3	2	1
75	14.75	30.0	7	1	1	1	1	1	0	14.33	6	3	1	3	3	3	1	1	3	3	1	4	2	1	3	3	1

FIELD Andy. 2000. *Discovering Statistics Using SPSS for Windows. Advanced Techniques for the Beginner*. London: Sage.

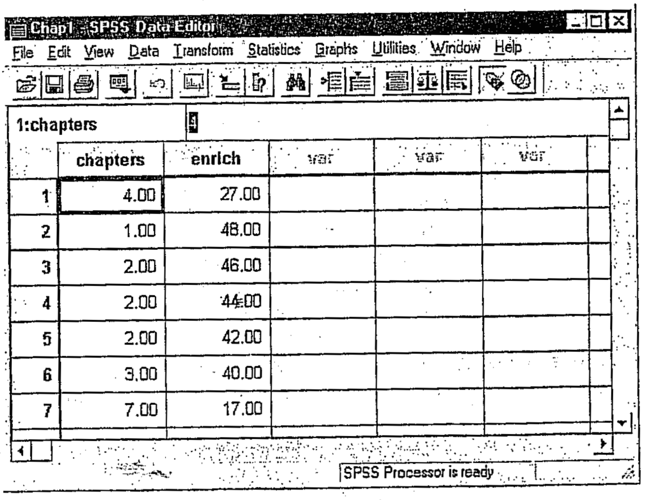
1.2. The SPSS Environment

There are several excellent texts that give introductions to the general environment within which SPSS operates. The best ones include Kinnear and Gray (1997) and Foster (1998). These texts are well worth reading if you are unfamiliar with Windows and SPSS generally because I am assuming at least some knowledge of the system. However, I appreciate the limited funds of most students and so to make this text usable for those inexperienced with SPSS I will provide a brief guide to the SPSS environment—but for a more detailed account see the previously cited texts and the SPSS manuals. This book is based primarily on version 9.0 of SPSS (at least in terms of the diagrams); however, it also caters for versions 7.0, 7.5 and 8.0 (there are few differences between versions 7.0, 8.0 and 9.0 and any obvious differences are highlighted where relevant).

Once SPSS has been activated, the program will automatically load two windows: the data editor (this is where you input your data and carry out statistical functions) and the output window (this is where the results of any analysis will appear). There are a number of additional windows that can be activated. In versions of SPSS earlier than version 7.0, graphs appear in a separate window known as the *chart carousel;* however, versions 7.0 and after include graphs in the output window, which is called the *output navigator* (version 7.0) and the *output viewer* (version 8.0 and after). Another window that is useful is the syntax window, which allows you to enter SPSS commands manually (rather than using the window-based menus). At most levels of expertise, the syntax window is redundant because you can carry out most analyses by clicking merrily with your mouse. However, there are various additional functions that can be accessed using syntax and sick individuals who enjoy statistics can find numerous uses for it! I will pretty much ignore syntax windows because those of you who want to know about them will learn by playing around and the rest of you will be put off by their inclusion (interested readers should refer to Foster, 1998, Chapter 8).

1.2.1. *The Data Editor*

The main SPSS window includes a data editor for entering data. This window is where most of the action happens. At the top of this screen is a menu bar similar to the ones you might have seen in other programs (such as Microsoft Word). Figure 1.6 shows this menu bar and the data editor. There are several menus at the top of the screen (e.g. *File*, *Edit* etc.) that can be activated by using the computer mouse to move the on-screen arrow onto the desired menu and then pressing the left mouse button once (pressing this button is usually known as *clicking*). When you have clicked on a menu, a menu box will appear that displays a list of options that can be activated by moving the on-screen arrow so that it is pointing at the desired option and then clicking with the mouse. Often, selecting an option from a menu makes a window appear; these windows are referred to as *dialog boxes*. When referring to selecting options in a menu I will notate the action using bold type with arrows indicating the path of the mouse (so, each arrow represents placing the on-screen arrow over a word and clicking the mouse's left button). So, for example, if I were to say that you should select the *Save As* ... option in the *File* menu, I would write this as select **File⇒Save As ...**.


Chap1 - SPSS Data Editor
File Edit View Data Transform Statistics Graphs Utilities Window Help

1:chapters

	chapters	enrich	var	var	var
1	4.00	27.00			
2	1.00	48.00			
3	2.00	46.00			
4	2.00	44.00			
5	2.00	42.00			
6	3.00	40.00			
7	7.00	17.00			

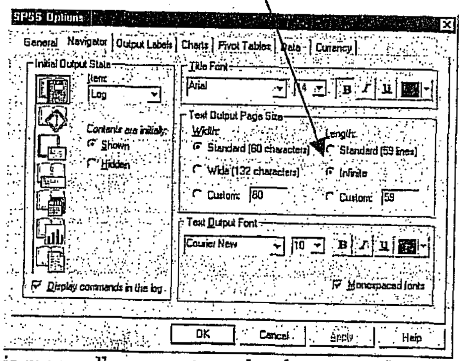
SPSS Processor is ready

Figure 1.6: The SPSS data editor

Within these menus you will notice that some letters are underlined: these underlined letters represent the *keyboard shortcut* for accessing that function. It is possible to select many functions without using the mouse, and the experienced keyboard user may find these shortcuts faster than manoeuvring the mouse arrow to the appropriate place on the screen. The letters underlined in the menus indicate that the option can be obtained by simultaneously pressing ALT on the keyboard and the underlined letter. So, to access the *Save As*... option, using only the keyboard, you should press ALT and F on the keyboard simultaneously (which activates the *File* menu) then, keeping your finger on the ALT key, press A (which is the underlined letter).

Below is a brief reference guide to each of the menus and some of the options that they contain. This is merely a summary and we will discover the wonders of each menu as we progress through the book.

- File: This menu allows you to do general things such as saving data, graphs, or output. Likewise, you can open previously saved files and print graphs, data or output. In essence, it contains all of the options that are customarily found in *File* menus.
- Edit: This menu contains edit functions for the data editor. In SPSS for Windows it is possible to *cut* and *paste* blocks of numbers from one part of the data editor to another (which can be very handy when you realize that you've entered lots of numbers in the wrong place). You can also use the *Options* to select various preferences such as the font that is used for the output. The default preferences are fine for most purposes, the only thing you might want to change (for the sake of the environment) is to set the text output page size length of the viewer to infinite (this saves hundreds of trees when you come to print things).



Item: Log

Contents are initially: ● Shown ○ Hidden

Title Font: Arial 14

Width: ● Standard (80 characters) ○ Wide (132 characters) ○ Custom: 80

Length: ○ Standard (59 lines) ● Infinite ○ Custom: 59

Text Output Font: Courier New 10

☑ Display commands in the log

☑ Monospaced fonts

- Data: This menu allows you to make changes to the data editor. The important features are *insert variable*, which is used to insert a new variable into the data editor (i.e. add a column); *insert case*, which is used to add a new row of data between two existing rows of data; *split file*, which is used to split the file by a grouping variable (see section 2.4.1); and *select cases*, which is used to run analyses on only a selected sample of cases.
- Transform: You should use this menu if you want to manipulate one of your variables in some way. For example, you can use *recode* to change the values of certain variables (e.g. if you wanted to adopt a slightly different coding scheme for some reason). The *compute* function is also useful for transforming data (e.g. you can create a new variable that is the average of two existing variables). This function allows you to carry out any number of calculations on your variables (see section 6.2.2.1).

- Analyze: This menu is called Statistics in version 8.0 and earlier. The fun begins here, because the statistical procedures lurk in this menu. Below is a brief guide to the options in the statistics menu that will be used during the course of this book (this is only a small portion of what is available):
  (a) Descriptive Statistics: This menu is called Summarize in version 8.0 and earlier. This menu is for conducting descriptive statistics (mean, mode, median etc.), frequencies and general data exploration. There is also a command called *crosstabs* that is useful for exploring frequency data and performing tests such as chi-square, Fisher's exact test and Cohen's kappa.
  (b) Compare Means: This is where you can find *t*-tests (related and unrelated—Chapter 6) and one-way independent ANOVA (Chapter 7).
  (c) General Linear Model: This is called *ANOVA Models* in version 6 of SPSS. This menu is for complex ANOVA such as two-way (unrelated, related or mixed), one-way ANOVA with repeated measures and multivariate analysis of variance (MANOVA).
  (d) Correlate: It doesn't take a genius to work out that this is where the correlation techniques are kept! You can do bivariate correlations such as Pearson's *R*, Spearman's rho (ρ) and Kendall's tau (τ) as well as partial correlations (see Chapter 3).
  (e) Regression: There are a variety of regression techniques available in SPSS. You can do simple linear regression, multiple linear regression (Chapter 4) and more advanced techniques such as logistic regression (Chapter 5).
  (f) Data Reduction: You find factor analysis here (Chapter 11).
  (g) Nonparametric: There are a variety of non-parametric statistics available such the chi-square goodness-of-fit statistic, the binomial test, the Mann-Whitney test, the Kruskal-Wallis test, Wilcoxon's test and Friedman's ANOVA (Chapter 2).
- Graphs: SPSS comes with its own, fairly versatile, graphing package. The types of graphs you can do include: bar charts, histograms, scatterplots, box-whisker plots, pie charts and error bar graphs to name but a few. There is also the facility to edit any graphs to make them look snazzy—which is pretty smart if you ask me.
- View: This menu deals with system specifications such as whether you have grid lines on the data editor, or whether you display value labels (exactly what value labels are will become clear later).
- Window: This allows you to switch from window to window. So, if you're looking at the output and you wish to switch back to your data sheet, you can do so using this menu. There are icons to shortcut most of the options in this menu so it isn't particularly useful.

- Help: This is an invaluable menu because it offers you on-line help on both the system itself and the statistical tests. Although the statistics help files are fairly useless at times (after all, the program is not supposed to teach you statistics) and certainly no substitute for acquiring a good knowledge of your own, they can sometimes get you out of a sticky situation.

As well as the menus there are also a set of *icons* at the top of the data editor window (see Figure 1.6) that are shortcuts to specific, frequently used, facilities. All of these facilities can be accessed via the menu system but using the icons will save you time. Below is a brief list of these icons and their function:

This icon gives you the option to open a previously saved file (if you are in the data editor SPSS assumes you want to open a data file, if you are in the output viewer, it will offer to open a viewer file).

This icon allows you to save files. It will save the file you are currently working on (be it data or output). If the file hasn't already been saved it will produce the *save data as* dialog box.

This icon activates a dialog box for printing whatever you are currently working on (either the data editor or the output). The exact print options will depend on the printer you use. One useful tip when printing from the output window is to highlight the text that you want to print (by holding the mouse button down and dragging the arrow over the text of interest). In version 7.0 onwards, you can also select parts of the output by clicking on branches in the viewer window (see section 1.2.4). When the *print* dialog box appears remember to click on the option to print only the selected text. Selecting parts of the output will save a lot of trees because by default SPSS will print everything in the output window.

Clicking this icon will activate a list of the last 12 dialog boxes that were used. From this list you can select any box from the list and it will appear on the screen. This icon makes it easy for you to repeat parts of an analysis.

This icon allows you to go directly to a case (i.e. a subject). This is useful if you are working on large data files. For example, if you were analysing a survey with 3000 respondents it would get pretty tedious scrolling down the data sheet to find a particular subject's responses. This icon can be used to skip directly to a case (e.g. case 2407). Clicking on this icon activates a dialog box that requires you to type in the case number required.

Clicking on this icon will give you information about a specified variable in the data editor (a dialog box allows you to choose which variable you want summary information about).

This icon allows you to search for words or numbers in your data file and output window.

Clicking on this icon inserts a new case in the data editor (so, it creates a blank row at the point that is currently highlighted in the data editor). This function is very useful if you need to add new data or if you forget to put a particular subject's data in the data editor.

Clicking this icon creates a new variable to the left of the variable that is currently active (to activate a variable simply click once on the name at the top of the column).

Clicking on this icon is a shortcut to the **Data⇒Split File ...** function (see section 2.4.1). Social scientists often conduct experiments on different groups of people. In SPSS we differentiate groups of people by using a coding variable (see section 1.2.3.1), and this function lets us divide our output by such a variable. For example, we might test males and females on their statistical ability. We can code each subject with a number that represents their gender (e.g. 1 = female, 0 = male). If we then want to know the mean statistical ability of each gender we simply ask the computer to split the file by the variable **gender**. Any subsequent analyses will be performed on the men and women separately.

This icon shortcuts to the **Data⇒Weight Cases ...** function. This function is necessary when we come to input frequency data (see section 2.8.2) and is useful for some advanced issues in survey sampling.

This icon is a shortcut to the **Data⇒Select Cases ...** function. If you want to analyze only a portion of your data, this is the option for you! This function allows you to specify what cases you want to include in the analysis.

Clicking this icon will either display, or hide, the value labels of any coding variables. We often group people together and use a coding variable to let the computer know that a certain subject belongs to a certain group. For example, if we coded gender as 1 = female, 0 = male then the computer knows that every time it comes across the value 1 in the **gender** column, that subject is a female. If you press this icon, the coding will appear on the data editor rather than the numerical values; so, you will see the words *male* and *female* in the **gender** column rather than a series of numbers. This idea will become clear in section 1.2.3.1.

1.2.2. Inputting Data

When you first load SPSS it will provide a blank data editor with the title *New Data*. When inputting a new set of data, you must input your data in a logical way. The SPSS data editor is arranged such that *each row represents data from one subject while each column represents a variable*. There is no discrimination between independent and dependent variables: both types should be placed in a separate column. The key point is that each row represents one participant's data. Therefore, any information about that case should be entered across the data editor. For example, imagine you were interested in sex differences in perceptions of pain created by hot and cold stimuli. You could place some people's hands in a bucket of very cold water for a minute and ask them to rate how painful they thought the experience was on a scale of 1 to 10. You could then ask them to hold a hot potato and again measure their perception of pain. Imagine I was a subject. You would have a single row representing my data, so there would be a different column for my name, my age, my gender, my pain perception for cold water, and my pain perception for a hot potato: Andy, 25, male, 7, 10. The column with the information about my gender is a grouping variable: I can belong to either the group of males or the group of females, but not both. As such, this variable is a between-group variable (different people belong to different groups). Therefore, between-group variables are represented by a single column in which the group to which the person belonged is defined using a number (see section 1.2.3.1). Variables that specify to which of several groups a person belongs can be used to split up data files (so, in the pain example you could run an analysis on the male and female subjects separately—see section 2.4.1). The two measures of pain are a repeated measure (all subjects were subjected to hot and cold stimuli). Therefore, levels of this variable can be entered in separate columns (one for pain to a hot stimulus and one for pain to a cold stimulus).

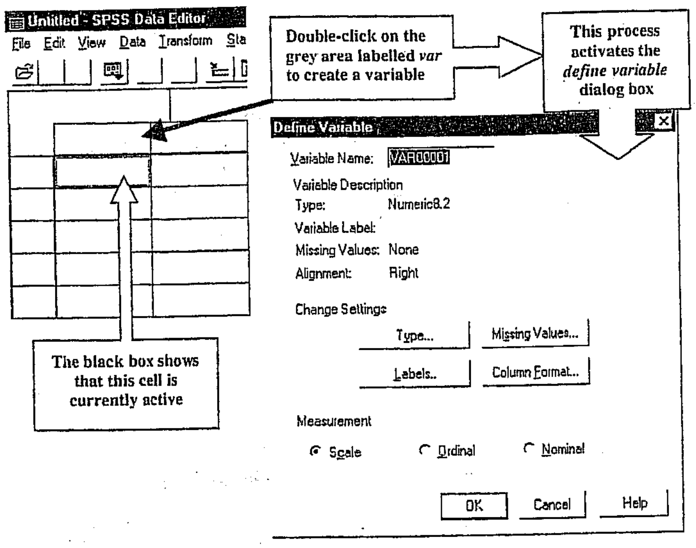
In summary, any variable measured with the same subjects (a repeated measure) should be represented by several columns (each column representing one level of the repeated measures variable). However, when a between-group design was used (e.g. different subjects were assigned to each level of the independent variable) the data will be represented by two columns: one that has the values of the dependent variable and one that is a coding variable indicating to which group the subject belonged. This idea will become clearer as you learn about how to carry out specific procedures.

The data editor is made up of lots of *cells*, which are just boxes in which data values can be placed. When a cell is active it becomes highlighted with a black surrounding box (as in Figure 1.7). You can move around the data editor, from cell to cell, using the arrow keys ← ↑ ↓ → (found on the right of the keyboard) or by clicking the mouse on the cell that you wish to activate. To enter a number into the data editor simply move to the cell in which you want to place the data value, type the value, then press the appropriate arrow button for the direction in which you wish to move. So, to enter a row of data, move to the far left of the row, type the value and then press → (this process inputs the value and then moves you into the next cell on the left).

1.2.3. Creating a Variable

There are several steps to creating a variable in the SPSS data editor (see Figure 1.7):

- Move the on-screen arrow (using the mouse) to the grey area at the top of the first column (the area labelled *var*.
- Double-click (i.e. click two times in quick succession) with the left button of the mouse.
- A dialog box should appear that is labelled *define variable* (see Figure 1.7).
- In this dialog box there will be a default variable name (something like **var00001**) that you should delete. You can then give the variable a more descriptive name. There are some general rules about variable names, such as that they must be 8 characters or less and you cannot use a blank space. If you violate any of these rules the computer will tell you that the variable name is invalid when you click on OK. Finally, the SPSS data editor is not case sensitive, so if you use capital letters in this dialog box it ignores them. However, SPSS is case sensitive to labels typed into the *Variable Label* part of the *define labels* dialog box (see section 1.2.3.1); these labels are used in the output.
- If you click on OK at this stage then a variable will be created in the data editor for you. However, there are some additional options that you might find useful.


Untitled - SPSS Data Editor
File Edit View Data Transform Sta
Double-click on the grey area labelled *var* to create a variable
This process activates the *define variable* dialog box
The black box shows that this cell is currently active
Define Variable
Type...
Missing Values...
Labels..
Column Format...
OK
Cancel
Help

Variable Name: VAR00001

Variable Description
Type: Numeric8.2
Variable Label:
Missing Values: None
Alignment: Right

Change Settings

Measurement
● Scale ○ Ordinal ○ Nominal

Figure 1.7: Creating a variable

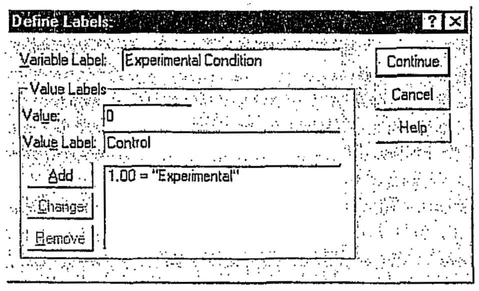
In versions 8 and 9 of SPSS, the *define variable* dialog box contains three options for selecting the level of measurement at which the variable was measured (earlier versions do not have these options). If you are using the variable as a coding variable (next section) then the data are categorical (also called *nominal*) and so you should click on the *Nominal* option. For example, if we asked people whether reading this chapter bores them they will answer *yes* or *no*. Therefore, people fall into two categories: bored and not bored. There is no indication as to exactly how bored the bored people are and therefore the data are merely labels, or categories into which people can be placed. Interval data are scores that are measured on a scale along the whole of which intervals are equal. For example, rather than asking people if they are bored we could measure boredom along a 10-point scale (0 being very interested and 10 being very bored). For data to be interval it should be true that the increase in boredom represented by a change from 3 to 4 along the scale should be the same as the change in boredom represented by a change from 9 to 10. Ratio data have this property, but in addition we should be able to say that someone who had a score of 8 was twice as bored as someone who scored only 4. These two types of data are represented by the *Scale* option. It should be obvious that in some social sciences (notably psychology) it is extremely difficult to establish whether data are interval (can we really tell whether a change on the boredom scale represents a genuine change in the experience of boredom?). A lower level of measurement is ordinal data, which does not quite have the property of interval data, but we can be confident that higher scores represent higher levels of a construct. We might not be sure that an increase in boredom of 1 on the scale represents the same change in experience between 1 and 2 as it does between 9 and 10. However, we can be confident that someone who scores 9 was, in reality, more bored than someone who scored only 8. These data would be ordinal and so you should select *Ordinal*. The *define variable* dialog box also has four buttons that you can click on to access other dialog boxes and these functions will be described in turn.

1.2.3.1. Creating Coding Variables

In the previous sections I have mentioned coding variables and this section is dedicated to a fuller description of this kind of variable (it is a type of variable that you will use a lot). A coding variable (also known as a grouping variable) is a variable consisting of a series of numbers that represent levels of a treatment variable. In experiments, coding variables are used to represent independent variables that have been measured between groups (i.e. different subjects were assigned to different groups). So, if you were to run an experiment with one group of subjects in an experimental condition and a different group of subjects in a control group, you might assign the experimental group a code of 1, and the control group a code of 0. When you come to put the data into the data editor, then you would create a variable (which you might call **group**) and type in the value 1 for any subjects in the experimental group, and 0 for any subject in the control group. These codes tell the computer that all of the cases that have been assigned the value 1 should be treated as belonging to the same group, and likewise for the subjects assigned the value 0.

There is a simple rule for how variables should be placed in the SPSS data editor: levels of the between-group variables go down the data editor whereas levels of within-subject (repeated measures) variables go across the data editor. We shall see exactly how we put this rule into operation in chapter 6.

To create a coding variable we create a variable in the usual way, but we have to tell the computer which numeric codes we are assigning to which groups. This can be done by using the Labels.. button in the *define variable* dialog box (see Figure 1.7) to open the *define labels* dialog box (see Figure 1.8). In the *define labels* dialog box there is room to give your variable a more descriptive title. For the purposes of the data editor itself, I have already mentioned that variable labels have to be 8 characters or less and that they have to be lower case. However, for the purposes of the output, it is possible to give our variable a more meaningful title (and this label can also have capital letters and space characters too—great!). If you want to give a variable a more descriptive title then simply click with the mouse in the white space next to where it says *Variable Label* in the dialog box. This will place the cursor in that space, and you can type a title: in Figure 1.8 I have chosen the title *Experimental Condition*. The more important use of this dialog box is to specify group codings. This can be done in three easy steps. First, click with the mouse in the white space next to where it says *Value* (or press ALT and U at the same time) and type in a code (e.g. 1). These codes are completely arbitrary: for the sake of convention people usually use 1, 2 and 3 etc., but in practice you could have a code of 495 if you were feeling particularly arbitrary. The second step is to click the mouse in the white space below, next to where it says *Value Label* (or press ALT and E at the same time) and type in an appropriate label for that group. In Figure 1.8 I have typed in 0 as my code and given this a label of *Control*. The third step is to add this coding to the list by clicking on Add. In Figure 1.8 I have already defined my code for the experimental group, to add the coding for the control group I must click on Add. When you have defined all of your coding values simply click on OK; if you click on OK and have forgotten to add your final coding to the list, SPSS will display a message warning you that any pending changes will be lost. In plain English this simply tells you to go back and click on Add.


Define Labels

Variable Label: Experimental Condition

Value Labels

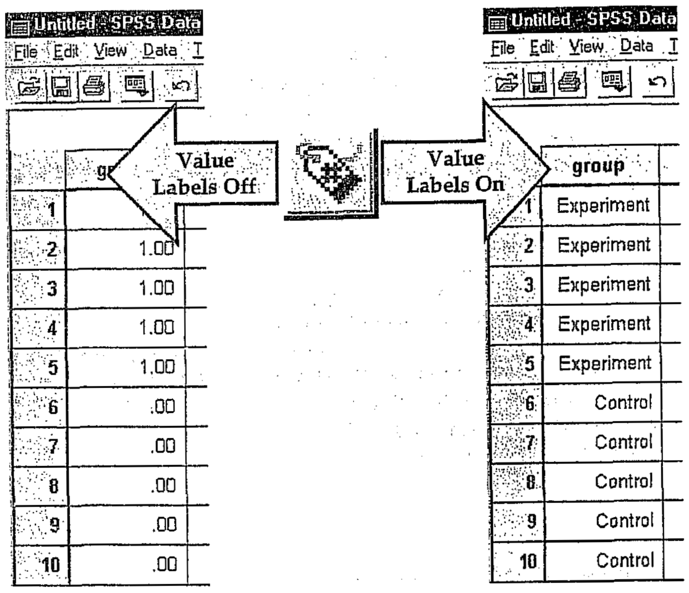
Value: 0

Value Label: Control

1.00 = "Experimental"

Figure 1.8: Defining coding values in SPSS

Having defined your codings, you can then go to the data editor and type these numerical values into the appropriate column. What is really groovy is that you can get the computer to display the codings themselves, or the value labels that you gave them by clicking on (see Figure 1.9). Figure 1.9 shows how the data should be arranged for a coding variable. Now remember that each row of the data editor represents one subject's data and so in this example it is clear that the first five subjects were in the experimental condition whereas subjects 6–10 were in the control group. This example also demonstrates why grouping variables are used for variables that have been measured between subjects: because by using a coding variable it is impossible for a subject to belong to more than one group. This situation should occur in a between-group design (i.e. a subject should not be tested in both the experimental and the control group). However, in repeated measures designs (within subjects) each subject is tested in every condition and so we would not use this sort of coding variable (because each subject does take part in every experimental condition).



	gr
1	
2	1.00
3	1.00
4	1.00
5	1.00
6	.00
7	.00
8	.00
9	.00
10	.00

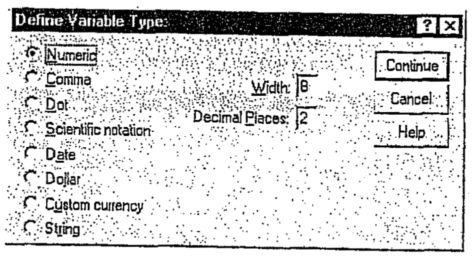
	group
1	Experiment
2	Experiment
3	Experiment
4	Experiment
5	Experiment
6	Control
7	Control
8	Control
9	Control
10	Control

Figure 1.9: Coding values in the data editor with the value labels switched off and on

1.2.3.2. Types of Variables

There are different types of variables that can be used in SPSS. In the majority of cases you will find yourself using numeric variables. These variables are ones that contain numbers and include the type of coding variables that have just been described. However, one of the other options when you create a variable is to specify the type of variable and this is done by clicking on Type... in the *define variable* dialog box. Clicking this button will activate the dialog box in Figure 1.10, which shows the default settings. By default, a variable is set up to store 8 digits, but you can change this value by typing a new number in the space labelled *Width* in the dialog box. Under normal circumstances you wouldn't require SPSS to retain any more than 8 characters unless you were doing calculations that need to be particularly precise. Another default setting is to have 2 decimal places displayed (in fact, you'll notice by default that when you type in whole numbers SPSS will add a decimal place with two zeros after it—this can be disconcerting initially!). It is easy enough to change the number of decimal places for a given variable by simply replacing the 2 with a new value depending on the level of precision you require.

The *define variable type* dialog box also allows you to specify a different type of variable. For the most part you will use numeric values. However, the other variable type of use is a string variable. A string variable is simply a line of text and could represent comments about a certain subject, or other information that you don't wish to analyze as a grouping variable (such as the subject's name). If you select the string variable option, SPSS lets you specify the width of the string variable (which by default is 8 characters) so that you can insert longer strings of text if necessary.


Define Variable Type

● Numeric
○ Comma
○ Dot
○ Scientific notation
○ Date
○ Dollar
○ Custom currency
○ String

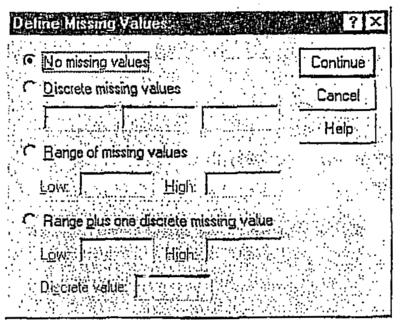
Width: 8

Decimal Places: 2

Figure 1.10: Defining the type of variable being used

1.2.3.3. Missing Values

Although as researchers we strive to collect complete sets of data, it is often the case that we have missing data. Missing data can occur for a variety of reasons: in long questionnaires participants accidentally miss out questions; in experimental procedures mechanical faults can lead to a datum not being recorded; and in research on delicate topics (e.g. sexual behaviour) subjects may exert their right not to answer a question. However, just because we have missed out on some data for a subject doesn't mean that we have to ignore the data we do have (although it sometimes creates statistical difficulties). However, we do need to tell the computer that a value is missing for a particular subject. The principle behind missing values is quite similar to that of coding variables in that we choose a numeric value to represent the missing data point. This value simply tells the computer that there is no recorded value for a participant for a certain variable. The computer then ignores that cell of the data editor (it does not use the value you select in the analysis). You need to be careful that the chosen code doesn't correspond with any naturally occurring data value. For example, if we tell the computer to regard the value 9 as a missing value and several subjects genuinely scored 9, then the computer will treat their data as missing when, in reality, it is not.


Define Missing Values

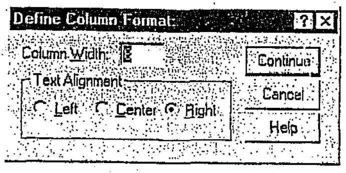
● No missing values
○ Discrete missing values
○ Range of missing values
Low: High:
○ Range plus one discrete missing value
Low: High:
Discrete value:

Figure 1.11: Defining missing values

To specify missing values you simply click on Missing Values... in the *define variable* dialog box to activate the *define missing values* dialog box (see Figure 1.11). By default SPSS assumes that no missing values exist but if you do have data with missing values you can choose to define them in one of three ways. The first is to select discrete values (by clicking on the circle next to where it says *Discrete missing values*) which are single values that represent missing data. SPSS allows you to specify up to three discrete values to represent missing data. The reason why you might choose to have several numbers to represent missing values is that you can assign a different meaning to each discrete value. For example, you could have the number 8 representing a response of 'not applicable', a code of 9 representing a 'don't know' response, and a code of 99 meaning that the subject failed to give any response. As far as the computer is concerned it will ignore any data cell containing these values; however, using different codes may be a useful way to remind you of why a particular score is missing. Usually, one discrete value is enough and in an experiment in which attitudes are measured on a 100-point scale (so scores vary from 1 to 100) you might choose 999 to represent missing values because this value cannot occur in the data that have been collected. The second option is to select a range of values to represent missing data and this is useful in situations in which it is necessary to exclude data falling between two points. So, we could exclude all scores between 5 and 10. The final option is to have a range of values and one discrete value.

1.2.3.4. Changing the Column Format

The final option available to us when we define a variable is to adjust the formatting of the column within the data editor. Click on Column Format... in the *define variable* dialog box and the dialog box in Figure 1.12 will appear. The default option is to have a column that is 8 characters wide with all numbers and text aligned to the right-hand side of the column. Both of these defaults can be changed: the column width by simply deleting the value of 8 and replacing it with a value suited to your needs, and the alignment by clicking on one of the deactivated circles (next to either *Left* or *Center*). It is very useful to adjust the column width when you have a coding variable with value labels that exceed 8 characters in length.


Define Column Format

Column Width: 8

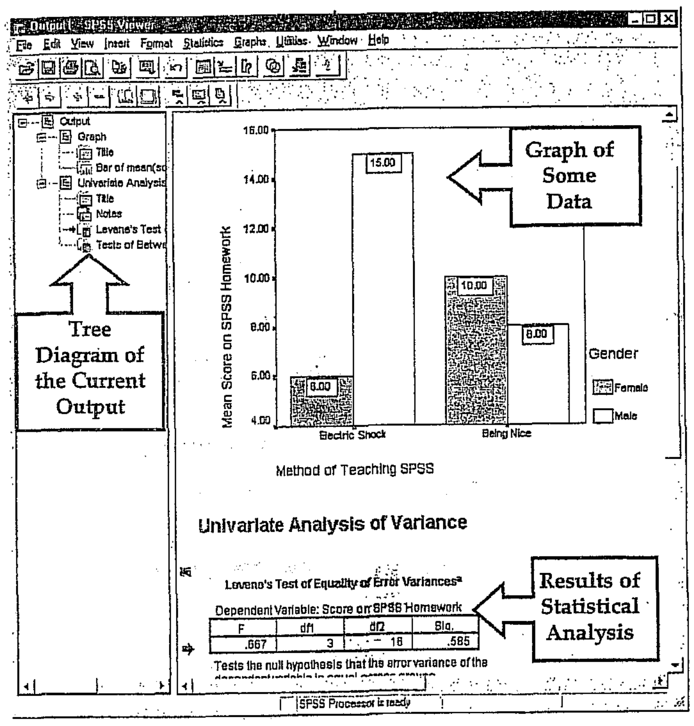
Text Alignment: ○ Left ○ Center ● Right

Figure 1.12: Defining the format of the column

1.2.4. The Output Viewer

Alongside the main SPSS window, there is a second window known as the output viewer (or *output navigator* in versions 7.0 and 7.5). In earlier versions of SPSS this is simply called the output window and its function is, in essence, the same. However, whereas the output window of old displayed only statistical results (in a very bland font I might add), the new, improved and generally amazing output viewer will happily display graphs, tables and statistical results and all in a much nicer font. Rumour has it that future versions of SPSS will even include a tea-making facility in the output viewer (I live in hope!).

Figure 1.13 shows the basic layout of the output viewer. On the right-hand side there is a large space in which the output is displayed. SPSS displays both graphs and the results of statistical analyses in this part of the viewer. It is also possible to edit graphs and to do this you simply double-click on the graph you wish to edit (this creates a new window in which the graph can be edited). On the left-hand side of the output viewer there is a tree diagram illustrating the structure of the output. This tree diagram is useful when you have conducted several analyses because it provides an easy way of accessing specific parts of the output. The tree structure is fairly self-explanatory in that every time you conduct a procedure (such as drawing a graph or running a statistical procedure), SPSS lists this procedure as a main heading.



Univariate Analysis of Variance

Levene's Test of Equality of Error Variances[a]

Dependent Variable: Score on SPSS Homework

F	df1	df2	Sig.
.667	3	16	.585

Tests the null hypothesis that the error variance of the

Figure 1.13 The output viewer

In Figure 1.13 I conducted a graphing procedure and then conducted a univariate analysis of variance (ANOVA) and so these names appear as main headings. For each procedure there are a series of sub-procedures, and these are listed as branches under the main headings. For example, in the ANOVA procedure there are a number of sections to the output such as a Levene's test (which tests the assumption of homogeneity of variance) and the between-group effects (i.e. the *F*-test of whether the means are significantly different). You can skip to any one of these sub-components of the ANOVA output by clicking on the appropriate branch of the tree diagram. So, if you wanted to skip straight to the between-group effects you should move the on-screen arrow to the left-hand portion of the window and click where it says *Tests of Between-Subjects Effects*. This action will highlight this part of the output in the main part of the viewer. You can also use this tree diagram to select parts of the output (which is useful for printing). For example, if you decided that you wanted to print out a graph but you didn't want to print the whole output, you can click on the word *Graph* in the tree structure and that graph will become highlighted in the output. It is then possible through the print menu to select to print only the selected part of the output. In this context it is worth noting that if you click on a main heading (such as *Univariate Analysis of Variance*) then SPSS will highlight not only that main heading but all of the sub-components as well. This is extremely useful when you want to print the results of a single statistical procedure.

There are a number of icons in the output viewer window that help you to do things quickly without using the drop-down menus. Some of these icons are the same as those described for the data editor window so I will concentrate mainly on the icons that are unique to the viewer window.

As with the data editor window, this icon activates the print menu. However, when this icon is pressed in the viewer window it activates a menu for printing the output. When the print menu is activated you are given the default option of printing the whole output, or you can choose to select an option for printing the output currently visible on the screen, or most useful is an option to print a selection of the output. To choose this last option you must have already selected part of the output (see above).

This icon returns you to the data editor in a flash!

This icon takes you to the last output in the viewer (so, it returns you to the last procedure you conducted).

This icon *promotes* the currently active part of the tree structure to a higher branch of the tree. For example, in Figure 1.13 the *Tests of Between-Subjects Effects* are a sub-component under the heading of *Univariate Analysis of Variance*. If we wanted to promote this part of the output to a higher level (i.e. to make it a main heading) then this is done using this icon.

This icon is the opposite of the above in that it *demotes* parts of the tree structure. For example, in Figure 1.13 if we didn't want the *Univariate Analysis of Variance* to be a unique section we could select this heading and demote it so that it becomes part of the previous heading (the *Graph* heading). This button is useful for combining parts of the output relating to a specific research question.

This icon collapses parts of the tree structure, which simply means that it hides the sub-components under a particular heading. For example, in Figure 1.13 if we selected the heading *Univariate Analysis of Variance* and pressed this icon, all of the sub-headings would disappear. The sections that disappear from the tree structure don't disappear from the output itself; the tree structure is merely condensed. This can be useful when you have been conducting lots of analyses and the tree diagram is becoming very complex.

This icon expands any collapsed sections. By default all of the main headings are displayed in the tree diagram in their expanded form. If, however, you have opted to collapse part of the tree diagram (using the icon above) then you can use this icon to undo your dirty work.

This icon and the following one allow you to show and hide parts of the output itself. So, you can select part of the output in the tree diagram and click on this icon and that part of the output will disappear. It isn't erased, but it is hidden from view. So, this icon is similar to the collapse icon listed above except that it affects the output rather than the tree structure. This is useful for hiding less relevant parts of the output.

This icon undoes the previous one, so if you have hidden a selected part of the output from view and you click on this icon, that part of the output will reappear. By default, all parts of the output are shown and so this icon is not active: it will become active only once you have hidden part of the output.

Although this icon looks rather like a paint roller, it unfortunately does not paint the house for you. What it does do is to insert a new heading into the tree diagram. For example, if you had several statistical tests that related to one of many research questions you could insert a main heading and then demote the headings of the relevant analyses so that they all fall under this new heading.

Assuming you had done the above, you can use this icon to provide your new heading with a title. The title you type in will actually appear in your output. So, you might have a heading like 'Research Question number 1' which tells you that the analyses under this heading relate to your first research question.

This final icon is used to place a text box in the output window. You can type anything into this box. In the context of the previous two icons, you might use a text box to explain what your first research question is (e.g. 'My first research question is whether or not boredom has set in by the end of the first chapter of my book. The following analyses test the hypothesis that boredom levels will be significantly higher at the end of the first chapter than at the beginning').

1.2.5. Saving Files

Although most of you should be familiar with how to save files in Windows it is a vital thing to know and so I will briefly describe what to do. To save files simply use the 🖫 icon (or use the menus: **File⇒Save** or **File⇒Save As...**). If the file is a new file, then clicking this icon will activate the *Save As* ... dialog box (see Figure 1.14). If you are in the data editor when you select *Save As* ... then SPSS will save the data file you are currently working on, but if you are in the viewer window then it will save the current output.

There are a number of features of the dialog box in Figure 1.14. First, you need to select a location at which to store the file. Typically, there are two types of locations where you can save data: the hard drive (or drives) and the floppy drive (and with the advent of rewritable CD-ROM drives, zip drives, jaz drives and the like you may have many other choices of location on your particular computer). The first thing to do is select either the floppy drive, by double clicking on 🖫, or the hard drive, by double clicking on 🖴. Once you have chosen a main location the dialog box will display all of the available folders on that particular device (you may not have any folders on your floppy disk in which case you can create a folder by clicking on 📁). Once you have selected a folder in which to save your file, you need to give your file a name. If you click in the space next to where it says *File name*, a cursor will appear and you can type a name of up to ten letters. By default, the file will be saved in an SPSS format, so if it is a data file it will have the file extension *.sav*, and if it is a viewer document it will have the file extension *.spo*. However, you can save data in different formats such as

Microsoft Excel files and tab-delimited text. To do this just click on ▼ where it says *Save as type* and a list of possible file formats will be displayed. Click on the file type you require. Once a file has previously been saved, it can be saved again (updated) by clicking 🖫. This icon appears in both the data editor and the viewer, and the file saved depends on the window that is currently active. The file will be saved in the location at which it is currently stored.

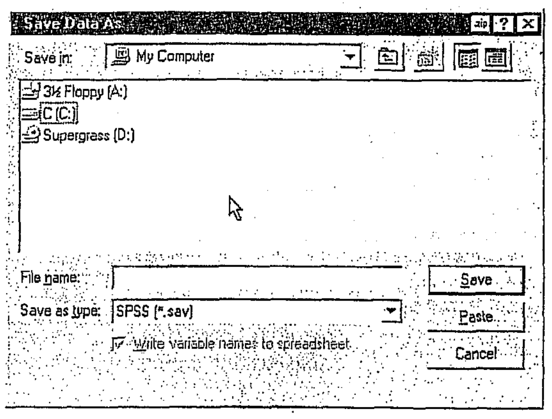


Figure 1.14: The *save data as* dialog box

1.2.6. Retrieving a File

Throughout this book you will work with data files that have been provided on a floppy disk. It is, therefore, important that you know how to load these data files into SPSS. The procedure is very simple. To open a file, simply use the 📂 icon (or use the menus: **File⇒Open**) to activate the dialog box in Figure 1.15. First, you need to find the location at which the file is stored. If you are loading a file from the floppy disk then access the floppy drive by clicking on ▼ where it says *Look in* and a list of possible location drives will be displayed. Once the floppy drive has been accessed you should see a list of files and folders that can be opened. As with saving a file, if you are currently in the data editor then SPSS will display only SPSS data files to be opened (if you are in the viewer window then only output files will be displayed). You can open a folder by double-clicking on the folder icon. Once you have tracked down the required file you can open it either by selecting it with the mouse and then clicking on Open, or by double-clicking on the icon next to the file you want (e.g. double-clicking on ▦). The data/output

will then appear in the appropriate window. If you are in the data editor and you want to open a viewer file, then click on ▼ where it says *Files of type* and a list of alternative file formats will be displayed. Click on the appropriate file type (viewer document (**.spo*), Excel file (**.xls*), text file (**.dat*, **.txt*)) and any files of that type will be displayed for you to open.

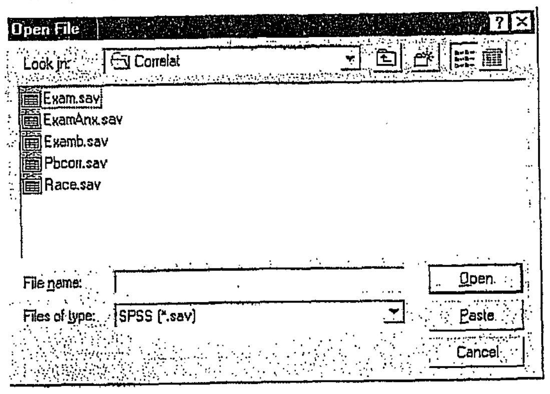


Figure 1.15: Dialog box to open a file

2. lekce

ROZLOŽENÍ KATEGORIZOVANÝCH DAT: ZÁKLADY UNIVARIAČNÍ ANALÝZY (TŘÍDĚNÍ I. STUPNĚ - Modul ANALYZE: procedura Frequencies).

ŘEHÁK, J., ŘEHÁKOVÁ, B.1986. *Analýza kategorizovaných dat v sociologii*. Academia: Praha.

Kapitola 4 Rozložení četností

4.1 Výskyt jevu, četnost, procento

Statistická analýza kategorizovaných dat je založena na studiu výskytů jednotlivých jevů a vztahů mezi nimi. *Jev* ve statistice chápeme obecně jako souhrn určitých projevů, vlastností, vztahů, podmínek, který je empiricky identifikovatelný, o němž můžeme vždy jednoznačně prohlásit, že buď nastal, nebo nenastal. Takový jednoznačný výsledek bývá však v praxi zatížen chybou a my se můžeme dopouštět dvou omylů s různými pravděpodobnostmi: jev nastal. ale prohlásíme, že nenastal, nebo jev nenastal, ale prohlásíme, že nastal. Proto je empirická identifikovatelnost jevu odstupňována podle pravděpodobnosti správného určení. Statistický jev je vždy vztažen k určitému komplexu podmínek, za nichž má odlišení „nastal-nenastal" smysl. Statistická analýza dat vychází z obou uvedených aspektů a jejich rozbor je jedním ze základních východisek postupů i konečné interpretace výsledků.

Jako příklad může sloužit často sledovaný jev **a** ≡ „spokojenost v práci". Místo něho však zjišťujeme jev **b** ≡ „respondent prohlásil, že je v práci spokojen". Vztah mezi oběma je sociologicky i metodologicky velmi složitý. Zatímco výskyty druhého jevu, **b**, zjišťujeme u respondentů vybraného souboru přesně (až na technické chyby v záznamu a přenosu dat), údaje o prvním jevu (**a**) jsou jím reprezentovány jen do jistého (většinou neurčeného) stupně spolehlivosti a platnosti. Respondenti nemusí vyjádřit skutečnou spokojenost, ať už záměrně či neúmyslně, pod vlivem nevhodně voleného sběru dat nebo v důsledku nějaké okamžité situace na pracovišti apod. Negací druhého jevu je jev opačný, „respondent neprohlásil, že je spokojen", který však může znamenat: respondent není spokojen, nechce svou spokojenost vyjádřit, nebo to dokonce vůbec prohlásit nemohl, protože nepracuje. Základní podmínkou sledování výskytu uvedeného jevu je tedy pracovní aktivita respondenta, která je např. u výzkumů pracovních kolektivů automaticky splněna, jindy však musí být zjišťována.

V praxi sociologicko-statistické analýzy je určení kontextu, v němž má smysl o jevu mluvit, totožné s určením souboru, k němuž má význam vztáhnout výskytovost jevu. Takový soubor lze vymezit pomocí vhodně zvoleného doplňkového jevu ($\bar{a}$) k jevu zkoumanému tak, že sjednocení obou charakterizuje kontext.

V uvedeném příkladě je možno doplněk k jevu „respondent prohlásil, že je v práci spokojen" volit jako jev „respondent prohlásil, že je v práci nespokojen". Opačné oběma jevům je „spokojen — nespokojen", společné (a tím i definicí souboru, k němuž vztahujeme výskytovost) je „prohlásil", „v práci" a všechna obecně platná omezení v daném šetření. Logicky tedy z analýzy vynecháme ekonomicky neaktivní respondenty a ty, kteří o spokojenosti nevypověděli.

Kontext, ke kterému vztahujeme výskytovost jevu, může být pro různé cíle definován různými způsoby, v rozmanité šíři a podmíněnosti. Každá statistická analýza je podmíněná: zvolené omezující podmínky jsou základním kvalitativním východiskem pro interpretaci statistických výsledků. Prakticky je podmíněná analýza prováděna vhodnou redukcí souboru dat. Obzvlášť silně se nutnost určení významových souvislostí projeví u komparace souborů, která je možná jen při srovnatelném základu. V běžné praxi se setkáváme s celou řadou rušivých vlivů, s nimiž není vždy lehké se vyrovnat: mateřská dovolená a vojenská služba jako důvody absence v podniku pro ženy a muže a různé věkové skupiny; otázka po důvodech změny zaměstnání pro osoby, které nejsou zaměstnány, nebo u nichž ke změně nedošlo; zjišťování postojů a názorů u osob, které si je nevytvořily, či si je dokonce ani vytvořit nemohly; typy školního vzdělání pro různé věkové kategorie apod.

Statistické jevy, jejich identifikovatelnost i způsob identifikace a komplex podmínek, za kterých má smysl o nich mluvit, se určují nejen u každé analýzy, ale už při přípravě sběru dat, při jejich záznamu a přenosu, při tvorbě dotazníků a záznamových listů, instrukcí pro pozorování, tazatele apod. Interpretace výsledků se opírá nejen o číselné závěry, ale i o rozbor empirické situace, metodologie sběru i teorii vztahů mezi podstatovými jevy, jež nás zajímají, a zjišťovanými empirickými jevy, které jsou vlastním předmětem statistické analýzy.

Statistický jev se váže ke statistické jednotce, u níž nastává, k jejímu místu, času, kontextu. Při sběru dat zjišťujeme u každé statistické jednotky, zda u ní jev nastal, nenastal, či nastat nemůže. U n^* jednotek souboru tak máme empirický údaj: m = počet jednotek, u nichž jev nastal, $\bar{m}$ = počet jednotek, u nichž jev nenastal, m^* = počet jednotek, u nichž jev nemá smysl, m' = počet jednotek, u nichž chybí informace, nebo je zjevně chybná. Z analýzy vynecháme $M = m^* + m'$ jednotek (tzv. vynechávaná data) a pro daný jev pracujeme se souborem o velikosti $n = m + \bar{m} = n^* - m^* - m'$.

Rozdíl mezi absolutním a poměrovým ukazatelem výskytu je dán otázkami: „kolik?" a „jaká část? (jaký podíl?)". V sociologické analýze ve většině případů pracujeme s poměrovými údaji, které jsou charakterizovány *relativními četnostmi* jevů $f = \frac{m}{n}$, tj. podílem souboru, u něhož jev nastal. Doplňková relativní četnost opačného jevu je $g = 1 - f = \frac{\bar{m}}{n}$. Někdy určujeme také podíl vynechávaných dat pro daný jev: $v = \frac{m' + m^*}{n^*}$. V praxi většinou uvádíme stonásobky relativních četností, kterým říkáme *procenta*. Vlastnosti relativních četností jsou velmi jednoduché:

a) Relativní četnost můžeme určovat vždy, existuje-li neprázdný ($n > 0$) soubor jednotek, pro něž má jev smysl.

b) $f = 0$ právě když jev vůbec nenastal.

c) $f = 1$ právě když jev nastal u všech jednotek.

d) Čím vyšší je f, tím častější je jev.

Při analýze více jevů **a**, **b**, **c**, ... značíme obvykle četnosti n_a, n_b, n_c, ... resp. f_a, f_b, f_c, ...

Absolutní četnost *m má někdy sama o sobě praktický význam* (např. počet osob, které odešly z pracovního kolektivu, musí být nahrazen bez ohledu na velikost skupiny), většinou však nás zajímá výskytovost jako podíl počtu výskytů v souboru (onemocní-li pět osob v třicetičlenném kolektivu, je to méně závažné, než onemocní-li stejný počet osob v patnáctičlenném kolektivu).

Každý jev **a** určuje jednoznačně dichotomickou proměnnou $\boldsymbol{A} = (a, \bar{a})$ = („jev **a** nastal", „jev **a** nenastal"); proto analýzu výskytovosti jevu provádíme také pomocí metod dalších paragrafů. Rozložení proměnné $\boldsymbol{A}$ je $(f, 1-f)$ resp. $(f_a, f_{\bar{a}})$.

4.2 Rozložení četností

Kategorizovanou proměnnou můžeme statisticky chápat jako soubor jevů, pro který platí:

— každé dva jevy jsou neslučitelné (žádné dva nemohou nastat současně);
— soubor jevů je úplný (alespoň jeden z jevů musí nastat);
— každý z jevů má smysl (každý z jevů může nastat);
— každý z jevů je identifikovatelný (v určitém stupni spolehlivosti);
— při identifikaci určujeme jednoznačně, který z jevů nastal.

Každá kategorie pak odpovídá jednomu z jevů; určení toho z jevů, který u statistické jednotky nastal, je totožné s určením kategorie, do které ji zařadíme. Proto kategorie znaku $\boldsymbol{A} = \{a_1, a_2, ..., a_K\}$ považujeme za soubor možných jevů, které lze zjistit.

Statistická analýza vychází ze vztahů všech K četností $\{n_1, n_2, ..., n_K\}$ resp. $\{f_1, f_2, ..., f_K\}$, a navíc z typu znaku, tj. z relací, které platí mezi $\{a_k\}$ tak, jak byly určeny vnějším sociologickometodologickým kritériem. Jde-li o prostý seznam jevů, hovoříme o nominálním znaku, jsou-li jevy uspořádány, jde o ordinální znak, přiřazujeme-li jevům čísla, dostáváme kardinální kategorizovaný znak. Zvláštní roli hraje znak *dichotomický, jehož dvě hodnoty se vzájemně vylučují a k jehož* statistickému popisu postačuje údaj o jedné kategorii, tj. f_1 nebo f_2 (druhý údaj plyne automaticky, $f_2 = 1 - f_1$).

Tabulka četností v třídění 1. stupně zahrnuje K nezávislých parametrů. Buď je to rozložení $\{n_k\}_K$, z něhož plyne výběrový rozsah $n = \Sigma n_k$, nebo $\{n, f_k\}_K$, kde jedna z relativních četností je odvoditelná z ostatních ($\Sigma f_k = 1$). V praxi analýzy je

vhodné využít *grafická zobrazení rozložení četností, která mají celou řadu tvarů.* Nejvhodnější je *histogram* (*sloupkový graf*) a pro nominální znaky také *kruhový graf*. Existuje celá řada dalších vhodných i méně vhodných ilustrativních metod, které lze vidět v publikacích statistické služby, v odborných článcích a knihách.

Příklad 4.1. *Důvody změny zaměstnání.* Ve výzkumu „Životní dráhy mládeže" byla položena otázka: „Změnil jste zaměstnání? Jestliže ano, jaký jste k tomu měl důvod?" Při záznamu odpovědí byly kódovány statistické jevy, které odpovídaly předem určeným kategoriím znaku, „*důvody změn*", a doplňkové jevy: „absence změny", „chybějící informace". Výsledky třídění 1. stupně jsou uvedeny v tab. 4.1a.

Tabulka 4.1. *Změna zaměstnání*
a) *Rozložení četností pro změnu zaměstnání a její důvody* (soubor mládeže ČSSR, 18—29 let)

Kód	Kategorie	Absolutní četnost	Relativní četnost	Procento
1	neměnil zaměstnání	1 628	0.8555	86
2	rodinné důvody	74	0.0389	4
3	finanční důvody	57	0.0300	3
4	zlepšení podmínek resp. výhodnější dojíždění	48	0.0252	3
5	nová práce lépe odpovídá zájmům a schopnostem	22	0.0116	1
6	lepší možnost růstu a postupu	8	0.0042	0
7	zdravotní důvody	17	0.0089	1
8	reorganizace	6	0.0032	0
9	ostatní	12	0.0063	1
0	chybí informace	31	0.0163	2
	Celkem	1 903	1.0001	101

(Zdroj: V. Dubský, Životní dráhy mládeže, výzkumný soubor, ÚFS ČSAV, Praha 1978).

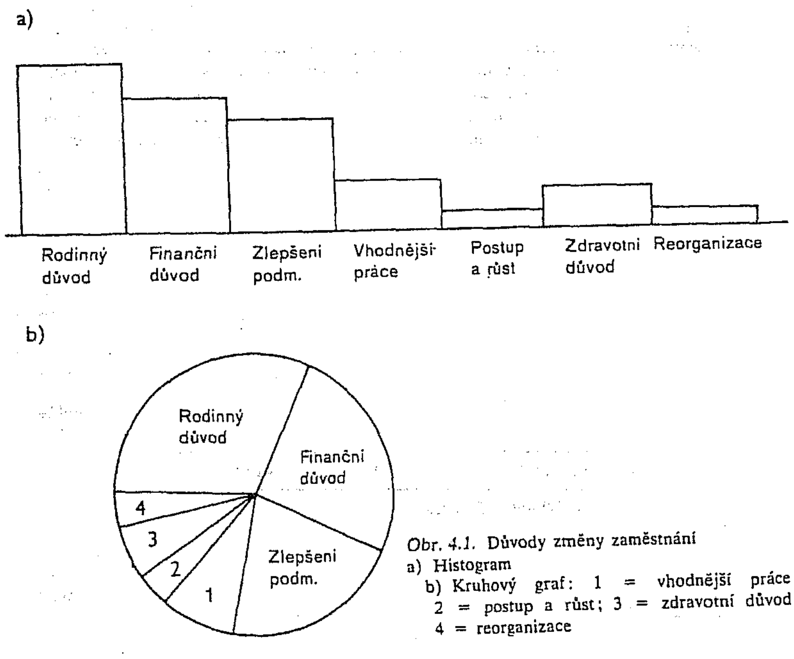
b) *Důvody změny zaměstnání* (Výzkum „Životní dráhy mládeže", soubor mládeže 15—29 let, $n = 232$).

Důvod	Rodinný důvod	Finanční důvod	Zlepšení podmínek	Vhodnější práce	Postup a růst	Zdravotní důvody	Reorganizace	Celkem
Procentní zastoupení	32%	25%	21%	9%	3%	7%	3%	100%

Proměnná „důvody změny zaměstnání" obsahuje však jen kategorie, které mají význam za podmínky, že respondent změnil zaměstnání. Proto z analýzy vynecháme kategorii 1 a 0 (= kód pro chybějící informaci). Nakonec vynecháme i málo obsazenou kategorii „ostatní důvody", která nemá interpretační význam (z důvodů obsahové heterogenity i nízkého procenta zastoupení). Po redukci

dostaneme tab. 4.1b, která charakterizuje výskytovost důvodů změny na redukovaném souboru, a která je vhodným východiskem pro analýzu dat.

Rozložení z tab. 4.1b můžeme zobrazit histogramem nebo kruhovým grafem (viz. obr. 4.1).



Obr. 4.1. Důvody změny zaměstnání
a) Histogram
b) Kruhový graf: 1 = vhodnější práce; 2 = postup a růst; 3 = zdravotní důvod; 4 = reorganizace

Pro publikaci tabulek rozložení četností platí obvyklé zásady:

1. Každá tabulka je plně informativní a vypovídá sama o sobě. Obsahuje název nebo přesnou charakteristiku proměnné, charakteristiku souboru, místa, času, kontextu, případně i metodu, která je využita, a výsledky statistické analýzy.

2. Řádky a sloupce jsou jasně označeny slovním popisem (především jde o význam kategorií proměnné, význam charakteristik a čísel v tabulce), pouze obecně přijaté a dobře definované statistické symboly mohou být výjimkou.

3. Hlavní informace se umisťuje do záhlaví tabulky, doplňková informace do poznámek k tabulce (nikoliv pod čáru).

4. V poznámkách pod tabulkou (případně v záhlaví) je uveden zdroj dat, pokud nejde o data, která patří autorům, o data určená jinde (např. v rejstříku použitých dat) nebo společná pro celou publikaci.

5. Absolutní četnosti se uvádějí pouze tehdy, mají-li vlastní informativní hodnotu. Relativní čísla jsou většinou vyjádřena v procentech, a to zaokrouhleně

na celá čísla (méně často na jedno desetinné místo), vždy k nim uvádíme velikost souboru n.

6. Zaokrouhlování při dělení $\frac{n_i}{n}$ vede k tomu, že součet procent nemusí být přesně 100, ale může dávat 99, 101, či 100,1, 99,9 apod. Dříve se procenta v jednotlivých kategoriích upravovala tak, aby součet dával 100%, v současné době se od takových úprav upouští.

7. V poznámkách u tabulky (nebo i přímo v tabulce) zpravidla uvádíme procento vynechávaných hodnot.

Obdobná pravidla platí pro přípravu grafů: plná informativnost, vhodné měřítko, které zajišťuje přehlednost, slovní popis, případná slovní informace přímo v grafu nesmí rušit vjem, uvedení zdroje.

V tabulce rozložení četností pro *nominální* znak můžeme kategorie řadit sestupně podle četností jejich obsazení. Tím získáváme větší přehled a rychlejší informaci. Někdy uvádíme jen ty kategorie, které hrají v rozložení výraznou a interpretovatelnou roli. Takovou formu volíme, především jde-li o tzv. dlouhé znaky (velké K), a u znaků s předem neomezeným počtem hodnot. Typickými příklady proměnných, které většinou tabelujeme tímto způsobem, jsou: respondentův *nejoblíbenější* zpěvák (sportovec, kniha, film, opera), příčina pracovní neschopnosti, záměr trávení dovolené. Uvedená forma se však nehodí pro ordinální a kardinální znaky, neboť by porušila vztahy mezi kategoriemi.

4.3 Kumulativní četnosti (distribuční funkce)

U ordinálních a kardinálních znaků jsou kategorie seřazeny podle vnějšího kritéria určeného obsahem. Z tohoto jednoznačného řazení vychází řada analytických metod založených na kumulativních četnostech, vyjadřujících postupné přibývání výskytů podél stupnice uvažované proměnné. Používáme *absolutní* i *relativní kumulativní četnosti*

$$M_k = \sum_{j=1}^{k} n_j = \text{počet jednotek v kategoriích } 1, 2, ..., k,$$

$$F_k = \frac{M_k}{n} = \sum_{j=1}^{k} f_j = \text{podíl jednotek v kateg. } 1, 2, ..., k, \tag{4.1}$$

$$P_k = \sum_{j=1}^{k} p_j = \text{podíl jednotek v kateg. } 1, 2, ..., k \text{ v základním souboru.}$$

Pro popis vzorců v dalších částech zavedeme úmluvu

$$M_0 = F_0 = P_0 = 0. \tag{4.2}$$

Poznamenejme, že $M_K = n$, $F_K = P_K = 1$. Souboru relativních čísel $\{F_k\}_K$ resp. $\{P_k\}_K$ říkáme *distribuční funkce*. Smysl a využití kumulativních četností ilustruje příklad 4.2.

Příklad 4.2. Příchody do zoo.

Při sociologickém šetření struktury návštěvníků Pražské zoo a délky jejich pobytu bylo zjišťováno rozložení příchodů (metodou načítání příchozích u vchodu). Četnosti získané během jednoho výzkumného dne uvádí tab. 4.2.

Tabulka 4.2. *Příchody do ZOO Praha v sobotu 12. 8. 1978*
(charakteristika dne: skoro zataženo, chladno)

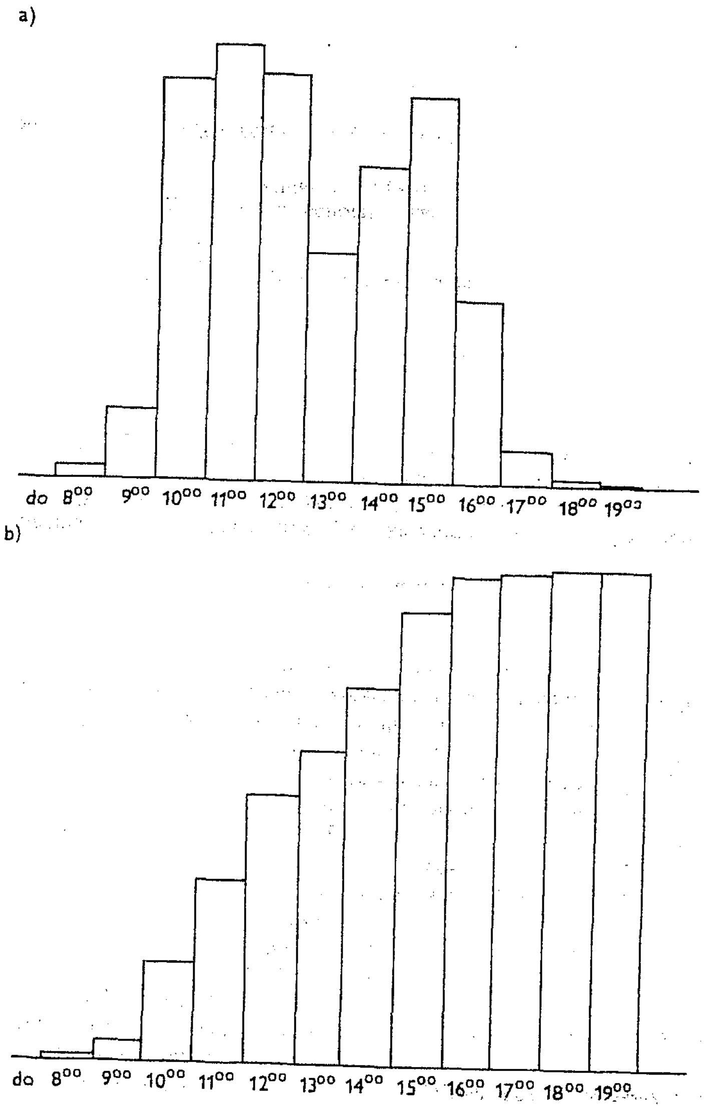
	Příchody v hodinách						
	do 08.00	do 09.00	do 10.00	do 11.00	do 12.00	do 13.00	do 14.00
Počet příchozích	20	219	956	1 034	971	547	759
Procento	0.3	3.7	16.1	17.4	16.3	9.2	12.7
Kumulativní četnost	20	239	1 195	2 229	3 200	3 747	4 506
Kumulativní procento	0.3	4.0	20.1	37.5	53.8	63.0	75.7

	Příchody v hodinách					
	do 15.00	do 16.00	do 17.00	do 18.00	do 19.00	Celkem
Počet příchozích	924	438	75	11	0	5 954
Procento	15.5	7.3	1.3	0.2	0.0	100.0
Kumulativní četnost	5 430	5 868	5 943	5 954	5 954	—
Kumulativní procento	91.2	98.5	99.8	100.0	100.0	—

Absolutní četnosti jsou důležité pro zhodnocení náporu na pokladnu, pro služby uvnitř zahrady, pro požadavky na městskou dopravu. Relativní četnosti dobře ukazují rozložení náporu během dne a umožňují porovnání podobných údajů z jiných dnů. Kumulativní četnosti skýtají okamžitou informaci o tom, kolik lidí již do zoo přišlo, ale také podíl, kolik jich do určité hodiny přišlo, a tudíž jaká část jich ještě přijde. Znak „hodina příchodu" může být chápán nominálně (charakteristika určité části dne), ordinálně (průběžný posun během dne) i kardinálně (*kvantifikovat můžeme např. časovým odstupem* od otevírací nebo zavírací hodiny, od vrcholného zatížení restaurace apod.). Kumulativní četnosti umožňují také rychlý výpočet četností výskytů v určitém intervalu složeném ze sousedních kategorií: relativní četnost kategorií

$$(i, i+1, \ldots, j) = f_i + f_{i+1} + \ldots + f_j = F_j - F_{i-1}. \qquad (4.3)$$

Tak např. mezi 10. a 14. hod. přišlo 75.7% − 20.1% = 55.6% návštěvníků. Četnosti lze graficky zobrazit pomocí *obr. 4.2.*



Obr. 4.2. Přehled příchodů do Zoo Praha, 12. 8. 1978
a) Četnosti příchodů v hodinových intervalech
b) Kumulativní četnosti příchodů

4.5 Zvláštní případ tabelací: vícenásobné výběrové otázky

Tabelace vícenásobných výběrových otázek není ve striktním slova smyslu tříděním 1. stupně. Vzhledem k častému výskytu v sociologických šetřeních se o ní však zmíníme. Vícenásobné výběrové otázky jsou instrukce typu: „Z přiloženého seznamu vyberte dvě položky, které považujete za nejdůležitější", „Jmenujte

Tabulka 4.3. *Názor na důležitost cílů v zaměstnání*
(Pokyn: „Vyberte dva z předložených cílů, které považujete za nejdůležitější",
$n = 1903$)

Cíl	Počet voleb	Procento z počtu voleb	Procento z počtu respondentů
Řídit lidi, být vedoucím	164	5	9
Materiální zajištění	949	26	50
Společenská úcta, vážnost, prestiž	198	5	10
Možnost přinést maximální užitek lidem, společnosti	552	15	29
Tvůrčí činnost, možnost vytvářet nové	313	9	16
Každodenní svědomité plnění svých povinností	427	12	22
Radost z vykonané práce	828	23	44
Možnost rozšiřovat obzor	200	6	11
Celkem	3 631	101%	191%

(Zdroj: V. Dubský, Životní dráhy mládeže, výzkumný soubor, ÚFS ČSAV, Praha 1978)

nejvýše tři oblíbené zpěváky", „Uveďte tři nejpodstatnější příčiny jevu". Přitom instrukce neobsahuje pokyn k seřazení položek. Analýza těchto dat je složitá, neboť jevy jsou specifickým způsobem závislé. Třídění se provádí tak, že zjišťujeme

$$m_j = \text{počet voleb, které dostala položka „}j\text{“}, \qquad (4.8)$$

a odhad četnosti pro každou z J položek (J je počet buď předložených, nebo jmenovaných možností):

$$r_j = \frac{m_j}{n}. \qquad (4.9)$$

Jiným způsobem je tabelace jednorozměrné tabulky relativních četností vzhledem k počtu realizovaných voleb

$$g_j = \frac{m_j}{M}, \quad M = \sum_{j=1}^{J} m_j, \qquad (4.10)$$

tj. podílu voleb kategorie „j" na všech realizovaných volbách.

Můžeme též uvést index R:

$$R = \frac{M}{Ln} = \frac{\sum_{j=1}^{J} r_j}{L}, \qquad (4.11)$$

který vyjadřuje, do jaké míry respondenti využili povolených L voleb.

Častou chybou při zpracování odpovědí na vícenásobné výběrové otázky je to, že děláme třídění 1. stupně pomocných a jen formálně zavedených znaků, které vzniknou tak, že např. 3 možné volby kódujeme: 1. znak = první zatržená hodnota v seznamu, 2. znak = druhá zatržená hodnota v seznamu, 3. znak = třetí zatržená hodnota v seznamu. Rozložení těchto pomocných znaků nemá smysl a žádný interpretační význam. Např. lze snadno ověřit, že kód první položky se nemůže vůbec vyskytnout u 2. a 3. znaku, kód druhé položky se nemůže vyskytnout u 3. znaku, kód třetí položky se může vyskytnout u 1. znaku jen tehdy, využil-li respondent pouze jednu volbu. Hodnoty $\{m_j\}_J$ vznikají součtem rozložení uvedených pomocných znaků.

Příklad 4.3. Cíle v povolání.

Mladým lidem ve věku *18—29 let byl dán tazatelem* pokyn: „Ve svém povolání se lidé snaží dosáhnout nejrůznějších cílů. Vyberte dva z nich, které jsou podle Vašeho názoru nejdůležitější." Osm důvodů bylo předloženo na kartě, uvedené volby byly zakroužkovány v záznamovém listě. Kódování bylo provedeno pomocí dvou pomocných znaků ***A*** = kód první zakroužkované kategorie, ***B*** = kód druhé zakroužkované kategorie. Tabulka 4.3 vznikla jako součet absolutních *četností* znaku ***A*** a ***B*** (dílčí tabulky nemají smysl).

(Může překvapit nízké procento kategorie, „materiální zajištění", neboť při samostatném dotazu bychom očekávali téměř stoprocentní odpověď „ano, je to důležité".) Součet 191% ukazuje, že využití

Counting Responses 3

How can you summarize the various responses people give to a question?

- What is a frequency table, and what can you learn from it?
- How can you tell from a frequency table if there have been errors in coding or entering data?
- What are percentages and cumulative percentages?
- What are pie charts and bar charts, and when do you use them?
- When do you use a histogram?
- What are the mode and the median?
- What do percentiles tell you?

Whenever you ask a number of people to answer the same questions, or when you measure the same characteristics for several people or objects, you want to know how frequently the possible responses occur. This can be as simple as just counting up the number of yes or no responses to a question. Or it can be considerably more complicated if, for example, you've asked people to report their annual income to the nearest penny. In this case, simply counting the number of times each unique income occurs may not be a useful summary of the data. In this chapter, you'll use the Frequencies procedure to summarize and display values for a single variable. You'll also learn to select appropriate statistics and charts for different types of data.

▶ The data analyzed in this chapter are in the *gss.sav* data file. For instructions on how to obtain the Frequencies output shown in the chapter, see "How to Obtain a Frequency Table" on p. 48.

Describing Variables

To see what's actually involved in examining and summarizing data, you'll use the nine variables from the General Social Survey described in Table 3.1. (You will use data from only 1500 respondents, since the SPSS student system is restricted in the number of cases in a data file.)

? *What's the General Social Survey?* The General Social Survey is administered yearly by the National Opinion Research Center to a sample of about 1500 persons 18 years of age and older. The sample represents the population of non-institutionalized adults living in the United States. (College dormitories are excluded from the survey!) Questions on many different topics—from how often you pray to where you were living at age 16—are included. Data from the General Social Survey are distributed at a nominal cost and are widely used by researchers and students (Davis & Smith, 1993). ■■■

All of these variables are defined as numeric in SPSS, but in most cases the numbers are just codes for non-numeric information. Value labels for each variable specify what the codes really mean.

In the SPSS Data Editor, to display (or hide) value labels, from the menus choose:

View
Value Labels

Table 3.1 Variables from the General Social Survey

Variable Name	Description
age	Age of respondent in years
sex	1=Male, 2=Female
educ	Years of education
income91	Total family income in 1993 (classified into one of 21 income categories)
wrkstat	Work status (1=Full-time work, 2=Part-time work, 3=Temporarily not working, 4=Unemployed (laid off), 5=Retired, 6=In school, 7=Keeping house, 8=Other)
richwork	"Would you continue or stop working if you became rich?" (1=Continue, 2=Stop)
satjob	Job satisfaction (1=Very satisfied, 2=Moderately satisfied, 3=A little dissatisfied, 4=Very dissatisfied)
life	"Do you find life exciting, pretty routine, or dull?" (1=Dull, 2=Routine, 3=Exciting)
impjob	"How important to your life is having a fulfilling job?" (1=One of the most important, 2=Very important, 3=Somewhat important, 4=Not too important, 5=Not at all important)

Start by looking at the variable *impjob*, which tells you how important a fulfilling job is to the respondent. Since there are only five possible responses, you can easily count how many people gave each of them.

A Simple Frequency Table

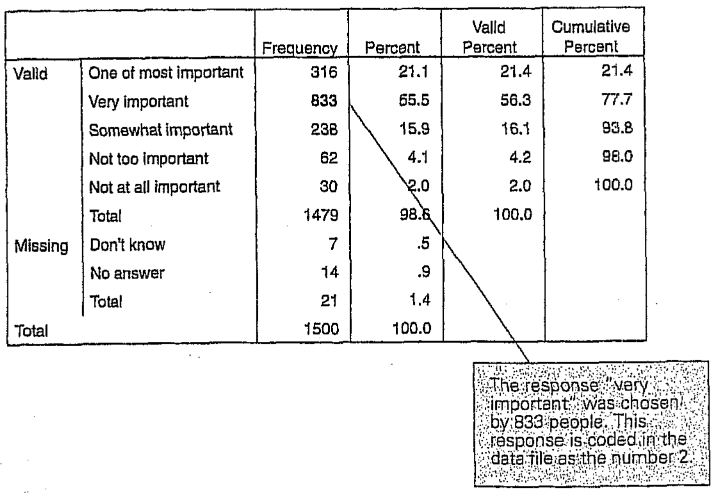
In Figure 3.1, you see the frequency table for the job importance variable.

Figure 3.1 Frequency table of job importance

To obtain this frequency table, from the menus choose:

Statistics
Summarize ▸
Frequencies...

In the Frequencies dialog box, select the variables impjob, as shown in Figure 3.11.



		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	One of most important	316	21.1	21.4	21.4
	Very important	833	55.5	56.3	77.7
	Somewhat important	238	15.9	16.1	93.8
	Not too important	62	4.1	4.2	98.0
	Not at all important	30	2.0	2.0	100.0
	Total	1479	98.6	100.0	
Missing	Don't know	7	.5		
	No answer	14	.9		
	Total	21	1.4		
Total		1500	100.0		

The response "very important" was chosen by 833 people. This response is coded in the data file as the number 2.

From a frequency table, you can tell how frequently people gave each response. The first row is for the response *one of the most important* (coded in the data with the value 1). The second row is for the response *very important* (coded in the data with the number 2). To determine how many people gave each response, look at the column labeled *Frequency*. For example, you find that 316 people find a fulfilling job to be *one of the most important* things to them, and 238 find it to be *somewhat important.* Only 30 people find having a fulfilling job *not at all important*. In the row labeled *Total*, you see that 1479 people selected one of the five possible valid responses.

The second part of the table tells you how many people did not select one of the five choices. There are two rows in the frequency table for the responses *don't know* and *no answer*. *Don't know* is used for people unwilling to commit themselves to a response. *No answer* is used when the response is illegible, lost, or not recorded by the interviewer. When the data file was defined, both *don't know* and *no answer* were identified as missing-value codes. That is, you don't have a valid answer for people whose responses are coded as *don't know* or *no answer*. In the *Frequency* column, you see that the response *don't know* was selected by 7 people and that the response was not available for 14 people. A total of 21 failed to select a valid response; that is, their response was identified as missing.

In the last row of the frequency table, you see that a total of 1500 people participated in the survey. Of these, 21 failed to select one of the five available responses; that is, their response was identified as *missing*. The other 1479 provided a valid response.

? *Why do you use different codes for* don't know *and* no answer? It's important to pinpoint why data values are missing. A response of *don't know* tells you that a person probably doesn't have strong feelings about the topic. It's unlikely that they find a job to be very important. A response of *no answer* doesn't tell you anything about a person's opinion of the importance of a job. The number of *no answer* responses tells you whether the survey was carefully conducted. You'll see later that if there are many cases with missing values, you may have serious problems in drawing conclusions from your data. ■■■

In the frequency table, value labels, which are descriptions of the codes assigned when you define a variable, are used to identify rows. If you don't assign these descriptions, the actual codes are shown. If your codes are not inherently meaningful, you should assign value labels to them so that the output is easier to understand. Assigning a value label once is much easier than repeatedly having to look up the meanings of codes.

Only responses actually selected by the participants are included in the frequency table. If no one selected the response *not at all important*, it would not be included in the table. Similarly, if you accidentally enter a code that does not correspond to a valid response—say a code of 0, 6, or 7 for the job importance variable—you will find it as a row in the frequency table. That's why frequency tables are useful for detecting mistakes in the data file. If you find wrong codes in your data values, you must correct the data file before proceeding.

Percentages

A frequency count alone is not a very good summary of the data. For example, if you want to compare your results to those of another survey, it won't do you much good to know simply that 762 people in that survey chose the response *very important*. From the count alone, you can't tell if the other survey's results are similar to yours. To compare the two surveys, you must convert the observed counts to percentages.

From a percentage, you can tell what proportion of people in the survey gave each of the responses. Unlike counts, you can compare percentages across surveys with different numbers of cases. You compute a percentage by dividing the number of cases that gave a particular response by the total number of cases. Then you multiply the result by 100.

In Figure 3.1, you find percentages in the column labeled *Percent*. Note that the 316 people who gave the response *one of the most important* are 21.1% of the 1500 people in your survey. Similarly, the 238 people who gave the response *somewhat important* are 15.9% of your sample. The 7 people who *don't know* are 0.5% of the total sample. (The actual percentage is 0.47%, but by default only one decimal place is shown.) The sum of the percentages over all the possible responses, including *don't know* and *no answer*, is 100%.

To change the number of decimal places shown in the output, double-click the pivot table to activate it, select the cell or column of interest, then choose:

Format
Cell Properties...

and change the Decimals specification.

Percentages Based on Valid Responses

To get the numbers in the column labeled *Percent*, you divide the observed frequency by the total number of cases in the sample and multiply by 100. Cases with codes identified as *missing* are included in the denominator. That can be a problem. For example, the General Social Survey does not ask all questions of all people. The question "Would you continue or stop working if you became rich?" was asked of only two-thirds

of people who were working or temporarily unemployed. Figure 3.2 shows the responses of people to this question.

Figure 3.2 Frequency table of continue working

To obtain this table, in the Frequencies dialog box select the variable richwork, as shown in Figure 3.11.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Continue working	448	29.9	69.8	69.8
	Stop working	194	12.9	30.2	100.0
	Total	642	42.8	100.0	
Missing	Not applicable	842	56.1		
	Don't know	11	.7		
	No answer	5	.3		
	Total	858	57.2		
Total		1500	100.0		

The percentage of people giving the response *continue working* is 29.9. What does that mean? Does it mean that about 30% of people in the survey would continue working if they became rich? No. It means that about 30% of the people in the sample, regardless of whether they were asked the question or volunteered an answer, gave the response *continue working*. Of the 1500 people in the survey, 56.1% weren't even asked the question (recorded in the table as *Not applicable)*. An additional 1% were asked and either gave the response *don't know* or their response was lost (*no answer*). All of these missing people are included in the denominator of the *Percent* calculation.

If you want to know what percentage of people who gave an acceptable answer selected *continue working*, look at the *Valid Percent* column. Almost 70% of people who answered the question claim that they would continue working if they struck it rich. (It's up to you whether you believe that percentage!) That's quite different from 30%. To calculate the entries in the *Valid Percent* column, you must exclude all people who gave an answer identified as *missing*. Valid percentages sum to 100 over all possible answers that are not missing. In this example, there are only two valid answers: *continue working* and *stop working*. Of the people who gave one of these answers, 69.8% selected the first and 30.2% selected the second. These two percentages sum to 100.

Problems with Missing Data

Removing people who aren't asked a question from the calculation of percentages is not troublesome. They don't make interpretation of the results difficult. However, if a lot of people who are asked the question refuse to answer, that can be a problem. In Figure 3.2, you see that only 11 people gave an answer of *don't know*. They represent fewer than 2% of the 653 people who were actually asked the question. So, you don't have to worry much about their impact on any conclusions you draw.

In contrast, however, consider the following situation. You conduct an employee satisfaction survey among 100 employees and find that 55 of them rate themselves as satisfied, 4 rate themselves as unsatisfied, and the remaining 41 decline to answer your question. That means that 55% of the polled employees consider themselves satisfied. However, if you exclude those who refused to answer from the denominator, 93% of the employees who answered the question consider themselves satisfied.

Which is the correct conclusion? Unfortunately, you don't know. It's possible that you have a company full of satisfied employees, many of whom don't like to answer questions. It's also possible that almost half of your employees are unhappy but are wary of voicing their dissatisfaction. When your data have many missing values because of people refusing to answer questions, it may be difficult, if not impossible, to draw correct conclusions. When you report percentages based on cases with nonmissing values, you should also report the percentage of cases that refused to give an answer.

Cumulative Percentages

There's one more percentage of interest in the frequency table. It's called the cumulative percentage. For each row of the frequency table, the **cumulative percentage** tells you the percentage of people who gave that response and any response that precedes it in the frequency table. It is the sum of the valid percentages for that row and all rows before it. Since there are only two possible valid answers for the continue working variable, the cumulative percentages in Figure 3.2 are of little interest. Instead, consider Figure 3.1 again. The cumulative percentage for *somewhat important* is 93.8. This means that over 93% of the people who answered the question said that a fulfilling job was at least somewhat important to their lives. Only 6.2% of the people rated the importance of a fulfilling job as less than *somewhat important*. Cumulative percentages are most useful when there is an underlying order to the codes assigned to a variable.

Sorting Frequency Tables

Unless you specify otherwise, SPSS produces a frequency table in which the order of the rows corresponds to the values of the codes you assign to the responses. The first row is for the smallest number found in the data values, and the last is for the largest. Codes that have been declared missing are at the end of the table. For example, if you had assigned the code 1 to *stop working*, it would have appeared first in the frequency table in Figure 3.2.

When you have several possible responses and the codes are not arranged in a meaningful order, you may want to rearrange the frequency table so that it's easier to use. You can determine the order of the rows in the table based on the frequency of values in the data. For example, Figure 3.3 shows a frequency table for the work status variable when the table is sorted in descending order of frequencies. Look at the column labeled *Frequency*. The frequencies go from largest to smallest.

Figure 3.3 Frequency table sorted by counts

To obtain this output, select Format in the Frequencies dialog box. Then select Descending counts, as shown in Figure 3.12.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Working fulltime	747	49.8	49.8	49.8
	Retired	231	15.4	15.4	65.2
	Keeping house	200	13.3	13.3	78.5
	Working parttime	161	10.7	10.7	89.3
	Unempl, laid off	51	3.4	3.4	92.7
	School	42	2.8	2.8	95.5
	Other	36	2.4	2.4	97.9
	Temp not working	32	2.1	2.1	100.0
	Total	1500	100.0	100.0	

Table is sorted by the counts in the Frequency column

Sorting a frequency table will usually change the values in the *Cumulative Percent* column, since the cumulative percentages depend on the order of the rows in the table. When the work status table is sorted by decreasing frequency, the cumulative percentage for *retired* is the percentage of people retired or working full time. In the default frequency table, however, in which the rows are sorted by the values of the codes,

NORUŠIS, Marija J. 1998. *Guide to Data Analysis*. Upper Saddle River, Prentice Hall.

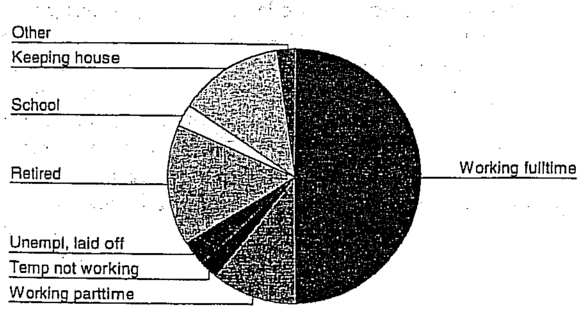
the cumulative percentage for *retired* is the sum of the valid percentages for codes 1 through 5.

Pie Charts

The information in a frequency table is easier to see if you turn it into a visual display, such as a bar chart or a pie chart. In Figure 3.4, you see a pie chart of the frequency table in Figure 3.3. There is a "slice" for each row of the frequency table. From the pie chart, you can easily see that almost half of your sample is *working full time*. It's also easy to see that the number of people who are *retired, keeping house,* and *working part time* are roughly equal. If you have many small slices in a pie chart, you can combine them into an *other* category. For example, Figure 3.5 is the pie chart for the same frequency table, except that all slices that have fewer than 5% of the cases (*in school, temporarily not working, unemployed,* and *other*) are combined into a single slice.

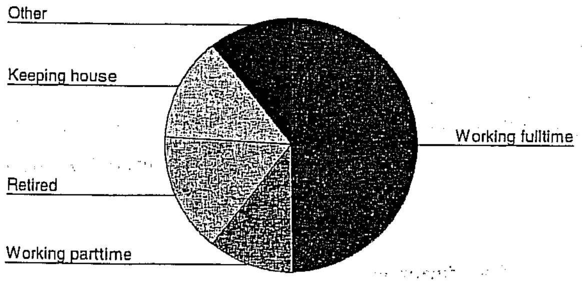
To obtain a pie chart, select Pie charts in the Frequencies Charts dialog box, as shown in Figure 3.14.

Figure 3.4 Pie chart of work status



You can collapse categories in a pie chart after it has been created. See "Modifying Chart Options" on p. 518 in Appendix A.

Figure 3.5 Work status with categories collapsed

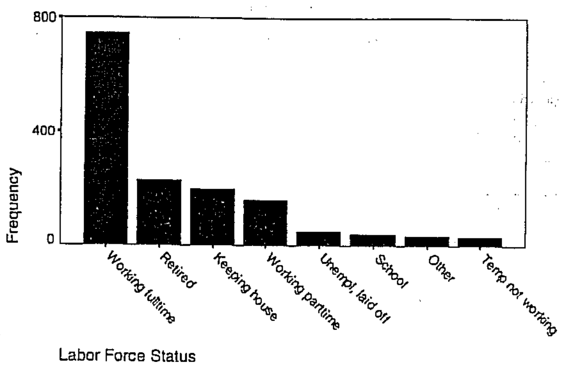


would expect, the tallest bar is for the *working full time* category. It's about three times as tall as the next largest bar, which represents *retired.*

To obtain this output, select Bar charts in the Frequencies Charts dialog box, as shown in Figure 3.14.

You can also obtain bar charts using the Graphs menu, as discussed in Appendix A.

Figure 3.6 Bar chart of work status



3. lekce

ROZLOŽENÍ SPOJITÝCH DAT: ZÁKLADY UNIVARIAČNÍ ANALÝZY (TŘÍDĚNÍ I. STUPNĚ - Modul ANALYZE: procedury Frequencies, Descriptives, Explore).

NORUŠIS, Marija J. 1998. *Guide to Data Analysis*. Upper Saddle River, Prentice Hall.

Summarizing the Age Variable

Although you can produce frequency tables for any kind of data, a frequency table becomes less useful as the number of possible responses increases. For example, you can construct a frequency table for the variable *age*, which tells you the ages of the people in your survey, but you will have as many rows in the frequency table as there are different ages in the data file. In Figure 3.7 you see that there is a row for every age from 18 to 89.

What's this? Nobody in the General Social Survey sample was 90 years or older? Actually, this is just a quirk in the way ages are coded in the General Social Survey. For obscure historical reasons, the General Social Survey assigns an age of 89 to everyone with an age of 89 or older. The code 99 indicates that the age is not known. Because so few people are that old, this quirk has very little effect on analyses that use the age variable. When you design a study, record the actual age—or better yet, the birth date, since it's harder to fudge. (To remain 30 forever, you have to remember to change your birth year annually!)

The Pivot Table Editor is used to hide the Percent column.

Figure 3.7 Frequency table of age

		Frequency	Valid Percent	Cumulative Percent
Valid	18	5	.3	.3
	19	17	1.1	1.5
	20	18	1.2	2.7
	21	22	1.5	4.1
	22	15	1.0	5.2
	23	28	1.9	7.0
	24	23	1.5	8.6
	25	30	2.0	10.6
	26	27	1.8	12.4
	27	22	1.5	13.8
	28	42	2.8	16.7
	29	30	2.0	18.7
	30	36	2.4	21.1
	31	31	2.1	23.1
	32	28	1.9	25.0
	33	33	2.2	27.2
	34	25	1.7	28.9
	35	41	2.7	31.6
	36	42	2.8	34.4
	37	37	2.5	36.9
	38	41	2.7	39.7
	39	38	2.5	42.2
	40	36	2.4	44.6
	41	36	2.4	47.0
	42	30	2.0	49.0
	43	39	2.6	51.6
	44	28	1.9	53.5
	45	30	2.0	55.5
	46	29	1.9	57.5
	47	32	2.1	59.6
	48	20	1.3	60.9
	49	27	1.8	62.7
	50	21	1.4	64.1
	51	26	1.7	65.9
	52	21	1.4	67.3
	53	18	1.2	68.5
	54	19	1.3	69.8
	55	22	1.5	71.2
Valid	56	12	.8	72.0
	57	18	1.2	73.2
	58	25	1.7	74.9
	59	14	.9	75.9
	60	16	1.1	76.9
	61	11	.7	77.7
	62	17	1.1	78.8
	63	19	1.3	80.1
	64	13	.9	80.9
	65	17	1.1	82.1
	66	19	1.3	83.3
	67	11	.7	84.1
	68	16	1.1	85.2
	69	19	1.3	86.4
	70	9	.6	87.0
	71	15	1.0	88.0
	72	19	1.3	89.3
	73	20	1.3	90.6
	74	18	1.2	91.8
	75	17	1.1	93.0
	76	13	.9	93.8
	77	15	1.0	94.8
	78	14	.9	95.8
	79	7	.5	96.3
	80	6	.4	96.7
	81	9	.6	97.3
	82	10	.7	97.9
	83	3	.2	98.1
	84	3	.2	98.3
	85	4	.3	98.6
	86	5	.3	98.9
	87	6	.4	99.3
	88	3	.2	99.5
	89	7	.5	100.0
	Total	1495	100.0	
Missing	NA	5		
	Total	5		
Total		1500		

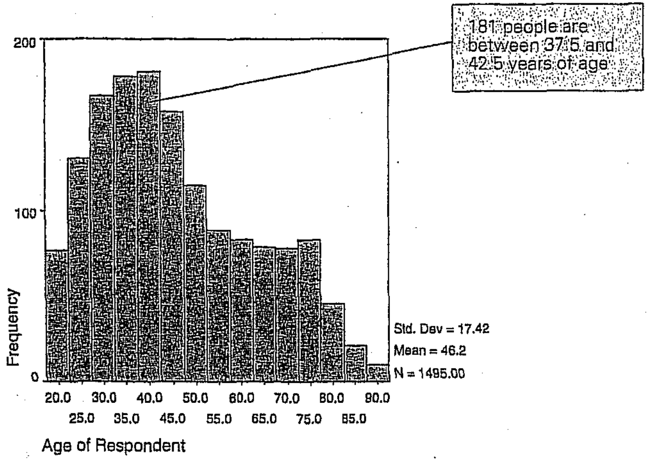
NORUŠIS, Marija J. 1998. *Guide to Data Analysis*. Upper Saddle River, Prentice Hall.

Histograms

You won't find pie charts and bar charts of the age variable to be useful either. There will be as many slices and bars as there are distinct ages. The arrangement of the values in the charts can be troublesome as well. Both bar charts and pie charts arrange bars and slices in ascending order of the values. However, if a particular age doesn't occur, an empty space is not left for it. That means that in a bar chart, the bar for 46 years may be right next to the bar for 50 years. You won't see a gap to remind you that ages 47 through 49 don't occur in your data.

A better display for a variable like *age*, for which it makes sense to group adjacent values, is a histogram. A **histogram** looks like a bar chart, except that each bar represents a range of values. For example, a single bar may represent all people in their twenties. In a histogram, the bars are plotted on a numerical scale that is determined by the observed range of your data.

Figure 3.8 Histogram of age



To obtain this output, select Charts in the Frequencies dialog box. Then select Histograms, as shown in Figure 3.14.

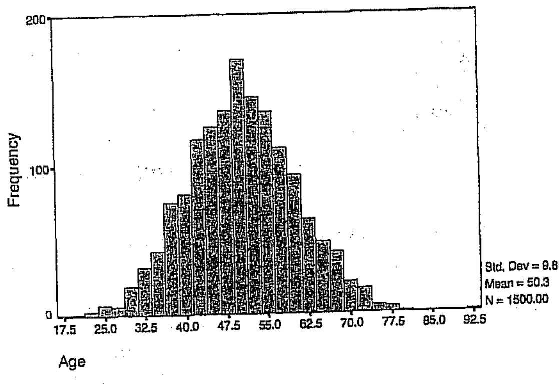
You can also obtain histograms using the Graphs menu, as discussed in Appendix A.

You see a histogram for the age variable in Figure 3.8. Age values are on the horizontal axis, and frequencies are on the vertical axis. The first bar represents cases with ages between 17.5 and 22.5. The middle value in this interval, the **midpoint**, is 20, which becomes the label used for the bar. From the histogram, you see that about 80 cases fall into this interval. Similarly, the second bar represents cases with ages between 22.5 and 27.5. This bar represents 130 cases.

A histogram tells you about the distribution of the data values. That is, it tells you how likely various values are. From it, you can see whether the cases cluster around a central value. You can also see whether large and small values are equally likely and whether there are values far removed from the rest. This is important not only to understand the data you've collected, but also for choosing appropriate statistical techniques for analyzing them. In Figure 3.8, you see that the age distribution has a peak corresponding to the interval 37.5 to 42.5. Additionally, you can see that the distribution of ages is not symmetric but has a "tail" extending to the older ages. That's because the General Social Survey interviews only respondents who are 18 or older.

What's a symmetric distribution? A distribution is symmetric if a vertical line going through its center divides it into two halves that are mirror images of each other. Figure 3.9 shows what a symmetric distribution of a hypothetical age variable might look like. Note that small and large values of age are equally likely.

Figure 3.9 Symmetric distribution



Mode and Median

You can use a variety of statistics to further summarize the information in a frequency table. In Chapter 4, you'll learn about a large number of such summary statistics. In the remainder of this chapter, you'll focus on summary measures that are easily obtained from the frequency table.

The mode is defined as the most frequently occurring value in your data. From the frequency table in Figure 3.7, you see that two ages (28 and 36) are tied for the mode. There are 42 people with each of these ages. Although these ages occur most frequently, they represent a small percentage of the total cases. Less than 3% of the total has an age of 28. Knowing the mode tells you very little about the data.

? *What are the modes for the job importance and What if you were rich variables?* In Figure 3.1, you see that *very important* is the most frequently occurring response to the job importance question. That makes it the mode. Similarly, for the if rich variable shown in Figure 3.2, *Continue working* is the most common response, so it's the mode. ■ ■ ■

Scales on which variables are measured are discussed in Chapter 4.

If you can meaningfully order your data values from smallest to largest, you can compute additional summary measures. These measures are better than the mode, since they make use of the additional information about the order of the data values. For example, the **median** is the value that is greater than half the data values and less than the other half.

To display the median and mode along with your frequency table, in the Frequencies dialog box, select Statistics. Then click Median and Mode (see Figure 3.13).

You calculate the median by finding the middle value when values for all cases are ordered from smallest to largest. If you have an odd number of cases, the median is just the middle value. If you have an even number of cases, the median is the value midway between the two middle ones. For example, the median of the five values 12, 34, 57, 92, and 100 is 57. For the six numbers 13, 20, 40, 60, 89, and 123, the median is 50, because 50 is the value midway between 40 and 60, the two middle numbers. (Add the two middle values and divide by two.)

You can calculate the median very easily from a frequency table. Find the first value for which the cumulative percentage exceeds or is equal to 50%. For the age variable in Figure 3.7, the median is 43. That means that half of the people in your sample are less than 43 years of age, and half are older. That's much more useful information than knowing that 28 and 36 years are tied for the mode.

? *What's the median for the work status variable?* Since there is no meaningful order of the codes assigned to the work status variable, it doesn't make sense to talk about median work status. You should use the median only when the data values can be ranked from smallest to largest. ■ ■ ■

Percentiles

When you calculate the median, you find the number that splits the sample into two equal parts. Half of the cases have values smaller than the median, and the other half have values larger than the median. You can compute values that split the sample in other ways. For example, you can find the value below which 25% of the data values fall. Such values are called percentiles, since they tell you the percentage of cases with values below and above them. Twenty-five percent of the cases have values smaller than the 25th percentile, and 75% of the cases have values larger than the 25th percentile. The median is the 50th percentile, since 50% of the cases have values less than the median, and 50% have values greater than the median.

To obtain this output, select Statistics in the Frequencies dialog box. Then select Quartiles, as shown in Figure 3.13.

Figure 3.10 Quartiles for age

	N		Percentiles		
	Valid	Missing	25	50	75
AGE	1495	5	32.00	43.00	59.00

See "Percentiles" on p. 95 in Chapter 6 for further discussion of percentiles.

You can compute percentiles from a frequency table by finding the first value with a cumulative percentage larger than or equal to the percentile you're interested in. You can see the 25th, 50th, and 75th percentiles for the age variable in Figure 3.10. From these, you know that 25% of the cases are 32 or younger, 50% are 43 or younger, and 75% are 59 or younger. Together, the 25th, 50th, and 75th percentiles are known as **quartiles**, since they split the sample into four groups with roughly equal numbers of cases. That is, 25% of the cases are 32 years old or younger, 25% are between 32 and 43, 25% are between 43 and 59, and 25% are 59 or older.

Summary

How can you summarize the various responses people give to a question?

- A frequency table tells you how many people (cases) selected each of the responses to a question. It contains the number and percentage of the people who gave each response, as well as the number of people for whom responses are not available.
- If you find codes in the frequency table that weren't used in your coding scheme, you know that an error in data coding or data entry has occurred.
- A count can be transformed into a percentage by dividing it by the total number of responses and multiplying by 100.
- A cumulative percentage is the percentage of cases with values less than or equal to a particular value.
- Pie charts and bar charts are graphical displays of counts.
- A histogram is a graphical display of counts for ranges of data values.
- The mode is the data value that occurs most frequently.
- The median is the middle value when data values are arranged from smallest to largest.
- Percentiles are values below which and above which a certain percentage of case values fall.

Computing Descriptive Statistics 4

How can you summarize the values of a variable?

- What are scales of measurement, and why are they important?
- How does the arithmetic mean differ from the mode and the median?
- When is the median a better measure of central tendency than the mean?
- What does the variance tell you? The coefficient of variation?
- What are standardized scores, and why are they useful?

In the previous chapter, you used frequency tables, bar charts, pie charts, histograms, and percentiles to examine the distribution of values for a variable. These are essential techniques for getting acquainted with the data. Often, however, you want to summarize the information even further by computing summary statistics that describe the "typical" values, or the **central tendency**, as well as how the data spread out around this value, or the **variability**. In this chapter, you'll learn how to use the Frequencies and Descriptives procedures to compute the most commonly used summary statistics for central tendency and variability.

▶ This chapter continues to use the *gss.sav* data file. For instructions on how to obtain the Descriptives output discussed in the chapter, see "How to Obtain Univariate Descriptive Statistics" on p. 69.

What's a statistic? Often when you collect data, you want to draw conclusions about a broader base of people or objects than are actually included in your study. For example, based on the responses of people included in the General Social Survey, you want to draw conclusions about the population of adults in the United States. The people you observe are called the **sample**. The people you want to draw conclusions about are called the **population**. A **statistic** is some characteristic of the sample. For example, the median age of people in the General Social Survey is a statistic. The term **parameter** is used to describe characteristics of the population. If you had the ages of all adults in the United States, the median age would be called a parameter value. Most of the time, population values, or parameters, are not known. You must estimate them based on statistics calculated from samples.

Summarizing Data

Consider again the data described in the previous chapter. Suppose you want to summarize the data values further. You want to know the typical age for participants in the survey, or their typical status in the workplace, or typical satisfaction with their job. A unique answer to these questions doesn't exist, since there are many different ways to define "typical." For example, you might define it as the value that occurs most often in the data (the mode), or as the middle value when the data are sorted from smallest to largest (the median), or as the sum of the data values divided by the number of cases (the arithmetic mean). To choose among the various measures of central tendency and variability you must consider the characteristics of your data as well as the properties of the measures. Although the mode may be a plausible statistic to report for status in the labor force, it may be a poor selection for a variable like age.

Scales of Measurement

One of the characteristics of your data that you must always consider is the scale on which they are measured. Scales are often classified as nominal, ordinal, interval, and ratio, based on a typology proposed by Stevens (1946). A nominal scale is used only for identification. Data measured on a **nominal scale** cannot be meaningfully ranked from smallest to largest. For example, status in the work force is measured on a nominal scale, since the codes assigned to the categories, although numeric, don't really mean anything. There is no order to *retired, in school, keeping house,* and *other.* Place of birth, hair color, and favorite statistician are all examples of variables measured on a nominal scale.

Variables whose values indicate only order or ranking are said to be measured on an **ordinal scale**. Job satisfaction and job importance are examples of variables measured on an ordinal scale. There are limitations on what you can say about data values measured on an ordinal scale. You can't say that someone who has a job satisfaction rating of 1 (*very satisfied*) is twice as satisfied as someone with a rating of 2 (*moderately satisfied*). All you can conclude is that one person claims to be more satisfied than the other. You can't tell how much more. The variable *income91* described in Chapter 3 is also measured on an ordinal scale. That's because the income is grouped into 21 unequal categories. You can't tell exactly how much more one person earned than another.

If you record people's actual annual incomes, you are measuring income on what is called a **ratio scale**. You can tell how much larger or smaller one value is compared with another. The distances between values are meaningful. For example, the distance between incomes of $20,000 and $30,000 is the same as the distance between incomes of $70,000 and $80,000. You can also legitimately compute ratios of two values. An income of $50,000 is twice as much as an income of $25,000. Age and years of education are both examples of variables measured on a ratio scale.

In the Define Variable dialog box, select Scale as the Measurement alternative for variables measured on a ratio or an interval scale.

An **interval scale** is just like a ratio scale except that it doesn't have an absolute zero. You can't compute ratios between two values measured on an interval scale. The standard example of a variable measured on an interval scale is temperature. You can't say that a 40°F day is twice as warm as a 20°F day. Few variables are measured on an interval scale, and the distinction between interval and ratio scales is seldom, if ever, important in statistical analyses.

Although it is important to consider the scale on which a variable is measured, statisticians argue that Stevens' typology is too strict to apply to real world data (Velleman & Wilkinson, 1993). For example, an identification number assigned to subjects as they enter a study might appear to be measured on a nominal scale. However, if the numbers are assigned sequentially from the first subject to enter the study to the last, the identification number is useful for seeing whether there is a relationship between some outcome of the study and the order of entry of the subjects. If the outcome is a variable like how long it takes a subject to master a particular task, it's certainly possible that instructions have improved during the course of a study and later participants fare better than earlier ones.

It's an oversimplification to conclude that the measurement scale dictates the statistical analyses you can perform. The questions that you want to be answered should direct the analyses. However, you should always make sure that your analysis is sensible. Using the computer, it's easy to calculate meaningless numbers, such as percentiles for place of birth or the median car color. In subsequent discussion, we'll occasionally refer to the scale of measurement of your data when describing various statistical techniques. These are not meant to be absolute rules but useful guidelines for performing analyses.

Mode, Median, and Arithmetic Average

The mode, median, and arithmetic average are the most commonly reported measures of central tendency. In Chapter 3, you saw how to compute the mode and median. You calculate the **mode** by finding the most frequently occurring value. The mode, since it does not require that the values of a variable have any meaning, is usually used for variables measured on a nominal scale. The mode is seldom reported alone. It's a useful statistic to report together with a frequency table or bar chart. You can easily find fault with the mode as a measure of what is typical. Even accompanied by the percentage of cases in the modal category or categories, it tells you very little.

If you are summarizing a variable whose values can be ranked from smallest to largest, the median is a more useful measure of central tendency. You calculate the **median** by sorting the values for all cases and then selecting the middle value. A problem with the median as a summary measure is that it ignores much of the available information. For example, the median for the five values 28, 29, 30, 31, and 32 is 30. For the five values 28, 29, 30, 98, and 190, it is also 30. The actual amounts by which the values fall above and below the median are ignored. The high values in the second example have no effect on the median.

The most commonly used measure of central tendency is the **arithmetic mean**, also known as the average. (For a sample, it's denoted as $\bar{X}$.) The mean uses the actual values of all of the cases. To compute the mean, add up the values of all the cases and then divide by the number of cases. For example, the arithmetic mean of the five values 28, 29, 30, 98, and 190 is

$$\text{Mean} = \frac{28 + 29 + 30 + 98 + 190}{5} = 75 \qquad \textbf{Equation 4.1}$$

Don't calculate the mean if the codes assigned to the values of a variable are arbitrary. For example, average car manufacturer and average religion don't make sense, since the codes are not meaningful.

Can I use the mean for variables that have only two values? Many variables, such as responses to yes/no or agree/disagree questions, have two values. If a variable has only two values, coded as 0 or 1, the arithmetic mean tells you the proportion of cases coded 1. For example, if 5 out of 10 people answered yes to a question and the coding scheme used is 0=no, 1=yes, the arithmetic mean is 0.50. You know that 50% of the sample answered yes.

Comparing Mean and Median

Figure 4.1 contains descriptive statistics from the Frequencies procedure for the age and education variables.

You can obtain these statistics using the Frequencies procedure, as discussed in Chapter 3. In the Frequencies Statistics dialog box (see Figure 3.13), select Mean, Median, and Mode.

Figure 4.1 Mean, median, and mode for age and education

Statistics

	N				
	Valid	Missing	Mean	Median	Mode
AGE Age of Respondent	1495	5	46.23	43.00	28[1]
EDUC Highest Year of School Completed	1496	4	13.04	12.00	12

[1] Multiple modes exist. The smallest value is shown

You see that the average age of the participants of the General Social Survey is 46.23 years. The median is somewhat lower, 43 years. The average number of years of school completed is 13.04, and the median is 12. For both of these variables, the arithmetic mean is somewhat greater than the median. The reason is that both of these variables have a "tail" toward larger values. Remember the histogram for age from Chapter 3. Since the General Social Survey is restricted to adults at least 18 years of age, young ages do not occur in the data. There is no such restriction for older ages. The older ages drive up the mean, which is based on all data values. They have no effect on the median, since it depends only on the values of the middle cases. If the distribution of data values is exactly symmetric, the mean and median are equal. If the distribution has a long tail (that is, the distribution is skewed), the mean is larger than the median if the tail extends toward larger values, and smaller than the median if the tail extends toward smaller values. In this example, the differences between the mean and the median are not very large. This is not always true.

Consider the following example. You ask five employees of a company how much money they earned in the past year. You get the following replies: $45,000, $50,000, $60,000, $70,000, and $1,000,000. The average salary received by these five people is $245,000. The median is $60,000. The arithmetic mean doesn't really represent the data well. The CEO salary makes the employees appear much better compensated than they really are. The median better represents the employees' salaries.

Measures of central tendency that are less affected by extreme values are discussed in Chapter 6.

Whenever you have data values that are much smaller or larger than the others, the mean may not be a good measure of central tendency. It is unduly influenced by extreme values (called **outliers**). In such a situation, you should report the median and mention that some of the cases had extremely small or large values.

Measures of Variability

Measures of central tendency don't tell you anything about how much the data values differ from each other. For example, the mean and median are both 50 for these two sets of ages: 50, 50, 50, 50, 50 and 10, 20, 50, 80, 90. However, the distribution of ages differs markedly between the two sets. **Measures of variability** attempt to quantify the spread of observations. We'll discuss the most common measures of variability in this chapter. Chapter 6 contains discussion of additional measures.

To obtain this output, from the menus choose:

Statistics
Summarize ▸
Descriptives...

Select the variables age and educ, as shown in Figure 4.5. In the Descriptives Options dialog box, select the variance, as shown in Figure 4.6.

Figure 4.2 Descriptive statistics for age and education

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
AGE Age of Respondent	1495	18	89	46.23	17.42	303.386
EDUC Highest Year of School Completed	1496	0	20	13.04	3.07	9.450
Valid N (listwise)	1491					

Range

The range is the simplest measure of variability. It's the difference between the largest and the smallest data values. Since the values for a nominal variable can't be meaningfully ordered from largest to smallest, it doesn't make sense to compute the range for a nominal variable such as status in the work force. In Figure 4.2, you see that for the variable *age*, the smallest value (labeled *Minimum*) is 18. The largest value (labeled *Maximum*) is 89. The range is 71 years. A large value for the range tells you that the largest and smallest values differ substantially. It doesn't tell you anything about the variability of the values between the smallest and the largest.

NORUŠIS, Marija J. 1998. *Guide to Data Analysis.* Upper Saddle River, Prentice Hall.

You can use the Explore procedure, described in Chapter 6, to calculate the range and the interquartile range.

A better measure of variability is the **interquartile range**. It is the distance between the 75th and 25th percentile values. The interquartile range, unlike the ordinary range, is not easily affected by extreme values. In Chapter 3, you calculated the 25th percentile for the age variable as 32 years, the 75th percentile as 59. The interquartile range is therefore 27, the difference between the two.

Variance and Standard Deviation

The most commonly used measure of variability is the variance. It is based on the squared distances between the values of the individual cases and the mean. To calculate the squared distance between a value and the mean, just subtract the mean from the value and then square the difference. (One reason you must use the squared distance instead of the distance is that the sum of distances around the mean is always 0.) To get the variance, sum up the squared distances from the mean for all cases and divide the sum by the number of cases minus 1.

The formula for computing the variance of a sample (denoted s^2)is

You can obtain the variance by selecting Options in the Descriptives dialog box. See Figure 4.6.

$$\text{Variance} = \frac{\text{sum of squared distances from the mean for all cases}}{(\text{number of cases} - 1)}$$

Equation 4.2

For example, to calculate the variance of the numbers 28, 29, 30, 98, and 190, first find the mean. It is 75. The sample variance is then

$$s^2 = \frac{(28-75)^2+(29-75)^2+(30-75)^2+(98-75)^2+(190-75)^2}{4}$$
$$= 5{,}026$$

Equation 4.3

If the variance is 0, all of the cases have the same value. The larger the variance, the more the values are spread out. In Figure 4.2, the variance for the age variable is 303.39 square years; for the education variable it is 9.45 square years. To obtain a measure in the same units as the original data, you can take the square root of the variance and obtain what's known as the **standard deviation**. Again in Figure 4.2, the standard deviation (labeled *Std Dev*) for the age variable is 17.42 years; for the education variable, it is 3.07 years.

Why divide by the number of cases minus 1 when calculating the sample variance, rather than by the number of cases? You want to know how much the data values vary around the population mean, but you don't know the value of the population mean. You have to use the sample mean in its place. This makes the sample values have less variability than they would if you used the population mean. Dividing by the number of cases minus 1 compensates for this.

The Coefficient of Variation

The magnitude of the standard deviation depends on the units used to measure a particular variable. For example, the standard deviation for age measured in days is larger than the standard deviation of the same ages measured in years. (In fact, the standard deviation for age in days is 365.25 times the standard deviation for age in years.) Similarly, a variable like salary will usually have a larger standard deviation than a variable like height.

The **coefficient of variation** expresses the standard deviation as a percentage of the mean value. This allows you to compare the variability of different variables. To compute the coefficient of variation, just divide the standard deviation by the mean and multiply by 100. (Take the absolute value of the mean if it is negative.)

$$\text{coefficient of variation} = \frac{\text{standard deviation}}{|\text{mean}|} \times 100$$

Equation 4.4

The coefficient of variation equals 100% if the standard deviation equals the mean. The coefficient of variation for the age variable is 37.68%. For the education variable, the coefficient of variation is 23.54%. Compared to their means, age varies more than education.

13

Standard Scores

The mean often serves as a convenient reference point to which individual observations are compared. Whenever you receive an examination back, the first question you ask is, How does my performance compare with the rest of the class? An initially dismal-looking score of 65% may turn stellar if that's the highest grade. Similarly, a usually respectable score of 80 loses its appeal if it places you in the bottom quarter of the class. If the instructor just tells you the mean score for the class, you can only tell if your score is less than, equal to, or greater than the mean. You can't say how far it is from the average unless you also know the standard deviation.

For example, if the average score is 70 and the standard deviation is 5, a score of 80 is quite a bit better than the rest. It is two standard deviations above the mean. If the standard deviation is 15, the same score is not very remarkable. It is less than one standard deviation above the mean. You can determine the position of a case in the distribution of observed values by calculating what's known as a standard score, or *z* score.

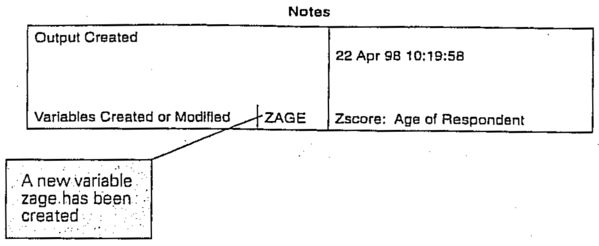
To calculate the standard score, first find the difference between the case's value and the mean and then divide this difference by the standard deviation.

$$\text{standard score} = \frac{\text{value} - \text{mean}}{\text{standard deviation}} \qquad \textbf{Equation 4.5}$$

A standard score tells you how many standard deviation units a case is above or below the mean. If a case's standard score is 0, the value for that case is equal to the mean. If the standard score is 1, the value for the case is one standard deviation above the mean. If the standard score is –1, the value for the case is one standard deviation below the mean. (For many types of distributions, including the normal distribution discussed in Chapter 10, most of the observed values fall within plus or minus two standard deviations of the mean.) The mean of the standard scores for a variable is always 0, and their standard deviation is 1.

You can use the Descriptives procedure in SPSS to obtain standard scores for your cases and to save them as a new variable. Figure 4.3 shows the notes from the Descriptives procedure that indicate that a new variable, the standard score for age, has been created. In addition, a new variable, *zage*, has been saved in the Data Editor, containing the standard scores for age (see Figure 4.4).

Figure 4.3 Descriptive statistics in the Viewer



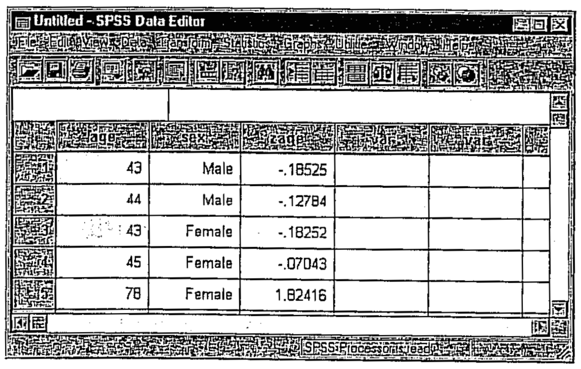
Notes

Output Created		22 Apr 98 10:19:58
Variables Created or Modified	ZAGE	Zscore: Age of Respondent

A new variable zage has been created

To save standardized scores, select Save standardized values as variables in the Descriptives dialog box, as shown in Figure 4.5.

Figure 4.4 Data Editor with standard scores saved as a new variable



Untitled - SPSS Data Editor
File Edit View Data Transform Statistics Graphs Utilities Window Help
SPSS Processor is ready

	age	sex	zage	var	var
1	43	Male	-.18525		
2	44	Male	-.12784		
3	43	Female	-.18252		
4	45	Female	-.07043		
5	78	Female	1.82416		

You see that the first case has an age of 43. From the standard score, you know that the case has an age less than average, but not very much. The age for the case is less than a quarter of a standard deviation below the mean. The fifth case has an observed age of 78, which is almost two standard deviations above the mean.

Standard scores allow you to compare relative values of several different variables for a case. For example, if a person has a standard score of 2 for income, and a standard score of –1 for education, you know that the person has a larger income than most and somewhat fewer years of education. You couldn't meaningfully compare the original values, since the variables all have different units of measurement, different means, and different standard deviations.

Summary

How can you summarize the values of a variable?

- Scales of measurement tell you about the properties of the values of a variable.
- The arithmetic mean is calculated by summing the values of a variable and dividing by the number of cases. Unlike the median and mode, the arithmetic mean uses all of the values of a variable.
- The median is a better measure of central tendency than the mean when there are data values that are far removed from the rest.
- The variance is a measure of the spread of data values around the mean. The coefficient of variation tells you the percentage the standard deviation is of the mean.
- A standardized score tells you how many standard deviation units above or below the mean an observation is.

Looking at Distributions 6

What additional displays are useful for summarizing the distribution of a variable for several groups?

- What is a stem-and-leaf plot?
- How does a stem-and-leaf plot differ from a histogram?
- What is a boxplot?
- What can you tell from the length of a box?
- How is the median represented in a boxplot?

Since most statistical analyses of data involve comparisons of groups, SPSS contains many procedures that help you to examine the distribution of values for individual groups of cases. In Chapter 5, you used the Means procedure to calculate descriptive statistics for education when the cases were subdivided on the basis of job satisfaction and gender. To examine each of the groups in more detail, you need to use the Explore procedure, which contains additional descriptive statistics as well as plots. That's what this chapter is about. (The statistics and displays described in this chapter are also useful for looking at the distribution of values for the entire sample.)

▶ This chapter continues to use the *gssft.sav* data file. For instructions on how to obtain the Explore output shown in the chapter, see "How to Explore Distributions" on p. 102.

Age and Job Satisfaction

In Chapter 5, you looked at the relationship between education and job satisfaction for full-time workers. You found that the average years of education did not differ much among people in the different categories of job satisfaction. Now you'll consider the relationship between age and job satisfaction for the same group of cases. You'll be able to examine the groups in considerably more detail. Consider first the descriptive statistics for age among workers in the different satisfaction categories.

To obtain these descriptive statistics, from the menus choose:

Statistics
Summarize ▸
Explore...

Select the variables age and satjob, as shown in Figure 6.9.

Figure 6.1 Case Processing Summary

		Cases					
		Valid		Missing		Total	
	Job Satisfaction	N	Percent	N	Percent	N	Percent
Age of Respondent	Very satisfied	325	99.4%	2	.6%	327	100.0%
	Mod satisfied	319	99.7%	1	.3%	320	100.0%
	A little dissatisfied	74	100.0%	0	.0%	74	100.0%
	Very dissatisfied	26	100.0%	0	.0%	26	100.0%

From Figure 6.1, you see that there are 325 cases in the *very satisfied* group for whom age is available. The number of missing cases is 2. That means that 2 *very satisfied* cases do not have a valid value for the age variable. Since these cases represent only 0.6% of all the cases in the group, you don't have to worry about the effect of missing age values on your analysis. Age is available for almost all of the cases in the other groups as well. Note that the number of cases varies considerably among the four groups. Over 300 people classify themselves as *very satisfied* and a similar number as *moderately satisfied*. However, 74 people are *a little dissatisfied* and only 26 *very dissatisfied*. You'll have to be careful about what you say about the last two groups since they are based on small numbers of cases.

The Pivot Table Editor was used to hide some statistics and to rearrange the default table

Figure 6.2 Average age and job satisfaction

	Age of Respondent			
	Job Satisfaction			
	Very satisfied	Mod satisfied	A little dissatisfied	Very dissatisfied
Mean	41.50	39.49	40.26	38.58
5% Trimmed Mean	41.05	39.11	39.83	38.19
Median	40.00	39.00	38.00	36.50
Std. Deviation	11.54	10.89	10.72	9.91
Minimum	19	20	23	22
Maximum	82	75	72	63
Range	63	55	49	41
Interquartile Range	15.50	16.00	14.25	17.00

The means of the ages, shown in Figure 6.2, range from a high of 41.5 years in the *very satisfied* group to 38.58 in the *very dissatisfied* group. The median ages are slightly less in all of the groups because the age distributions have tails toward larger values. As you've learned in Chapter 4, one of the shortcomings of the arithmetic mean is that very large or very small values in the data can change its value substantially. The trimmed mean avoids this problem. A **trimmed mean** is calculated just like the usual arithmetic mean, except that a designated percentage of the cases with the largest and smallest values are excluded. This makes the trimmed mean less sensitive to outlying values. The 5% trimmed mean excludes the 5% largest and the 5% smallest values. It's based on the 90% of cases in the middle. The trimmed mean provides an alternative to the median when you have some data values that are far removed from the rest.

In Figure 6.2 you see that the 5% trimmed mean doesn't differ much from the usual mean. That's not surprising since the largest age in all groups (*Maximum*) is 82 and the smallest age (*Minimum*) is 19. The person with an age of 82 is in a group with 325 cases. You'd need a surviving Roman warrior to have a real effect on the mean.

Again from Figure 6.2, you see that the standard deviation ranges from 11.54 years in the *very satisfied* group to 9.91 in the *very dissatisfied* group. The range is largest in the very satisfied group since it contains the 82 year old. The range is based on the largest and smallest values so a single outlying value can have a large effect on the range. Unlike the ordinary range, the interquartile range is not easily affected by extreme values, since the bottom 25% and the top 25% of the data values are excluded from its computation. It's the difference between the 75th and the 25th percentile values. In Figure 6.2, you see that the interquartile ranges, are fairly similar in all of the groups.

Identifying Extreme Values

Since many statistics are affected by data values that are much smaller or larger than most, it's always important to examine your data to see if extreme values are present. You can obtain from the Explore procedure a list of the cases with the five largest and the five smallest values in each group. Always check any suspicious values to make sure they are not the result of an error in data recording or entry. If you find a mistake, it's easy to change the values for a case using the Data Editor. If the extreme values are correct, make sure to select summary measures that are not unduly affected by these outliers.

To obtain extreme values, select Statistics in the Explore dialog box. Then select Outliers, as shown in Figure 6.10.

Figure 6.3 Outliers from the very satisfied group

	Job Satisfaction			Case Number	Value
Age of Respondent	Very satisfied	Highest	1	344	82
			2	223	78
			3	263	77
			4	401	77
			5	208	73
		Lowest	1	173	19
			2	364	20
			3	714	20
			4	320	21
			5	665	21

At 82 years, case 344 is the oldest respondent in the group

Figure 6.3 shows for the *very satisfied* group the cases with the five largest and smallest ages. The column labeled *Case Number* contains the sequence number in the file for each case. That makes it easier for you to track the suspicious values. (SPSS can also show a name or any other identifier to label the cases.) You see that case 344 is the oldest, at age 82. Case 173 is the youngest, at age 19. Neither of these values is usual. Since only five cases with the smallest and largest values are listed, it's possible that not all cases with those values are listed. For example, there may be more than one 73-year-old or more than two 21-year-olds. If that's true, a note is printed beneath the table. Just because a value is included in the extreme value table doesn't mean it really is an outlier. It's only one of the largest or smallest values; you must decide if it is really unusual.

Percentiles

Using Explore, you can also obtain percentiles for each of the groups. Figure 6.4 shows the percentiles for age for the satisfaction subgroups.

To obtain percentiles, select Statistics in the Explore dialog box. Then select Percentiles. (See Figure 6.10.)

Figure 6.4 Age percentiles

			Percentiles						
		Job Satisfaction	5	10	25	50	75	90	95
Weighted Average (Definition 1)	Age of Respondent	Very satisfied	24.00	27.00	33.50	40.00	49.00	57.40	61.00
		Mod satisfied	24.00	26.00	31.00	39.00	47.00	55.00	60.00
		A little dissatisfied	25.00	27.00	32.75	38.00	47.00	55.50	60.25
		Very dissatisfied	24.10	28.00	30.00	36.50	47.00	51.60	59.50
Tukey's Hinges	Age of Respondent	Very satisfied			34.00	40.00	49.00		
		Mod satisfied			31.00	39.00	47.00		
		A little dissatisfied			33.00	38.00	47.00		
		Very dissatisfied			30.00	36.50	47.00		

10% of cases are 27 or younger in the very satisfied group

10% are 57.4 or older in the very satisfied group

Two sets of percentiles are shown: The first set is obtained using a method called *Weighted Average*. Other ways can be used, but most of the time they give pretty much the same results. You see that for the *very satisfied* group 10% of the cases are 27 years of age or younger (the 10th percentile) and 10% are 57.4 or older (the 90th percentile). The percentiles are comparable for the four satisfaction groups.

The second part of the percentile table shows *Tukey's Hinges*, which are quartiles (values that divide the sorted cases into four equal groups) calculated using a slightly different method than the weighted average percentiles.

How can you get different numbers for the same percentiles? Percentiles don't have a single, unique definition. For example, consider the eight numbers 25 26 27 27 27 27 30 31. What's the 25th percentile? Any number between 26 and 27 is a plausible value. One definition of percentiles gives the answer 26.5, since that's the average of 26 and 27, the interval within which the percentile falls. Another definition results in the answer 26, since that's the first value for which the cumulative percentage is equal to or greater than 25%.

For small data sets, especially when several cases have the same values, different percentiles may have the same value. For the previous example, it's possible for percentiles greater than the 25th and less than the 75th to have the value 27. For small data sets, percentile values can vary a lot for samples from the same population, so you shouldn't place too much confidence in their exact values. (You also shouldn't worry about where the "equal" goes. That is, whether 25% of the cases have values less than the 25th percentile, or whether 25% of the cases have values less than or *equal* to the 25th percentile. Statistical software packages implement arbitrary rules about where the "equal" goes.)

Plots

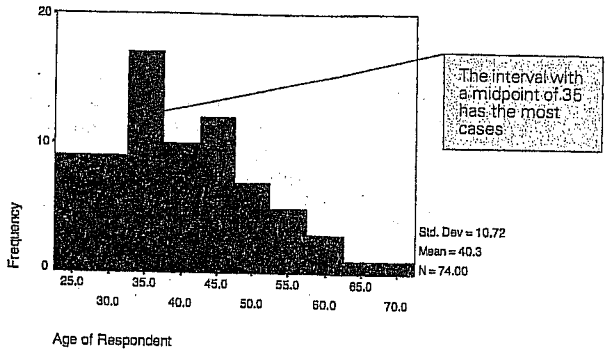
One of the easiest ways to see the distributions of your variables is literally with a picture. The Explore procedure provides several plots that let you evaluate the shape of a distribution. From these plots, you see how often different values of a variable occur in your data. As you will see in Part 3, your choice of the statistical analysis for a particular problem depends on the assumptions you are willing to make about the distributions of the variables of interest. That's why it's important to examine them.

Histograms and Stem-and-Leaf Plots

The Explore procedure can produce separate histograms for groups of cases. The histograms are identical to those produced by the Frequencies procedure, as described in Chapter 3. Figure 6.5 shows a histogram of age for the people *a little dissatisfied* with their jobs.

Figure 6.5 Age histogram for a little dissatisfied

To obtain a histogram, select Plots in the Explore dialog box. Then select Histogram, as shown in Figure 6.11.



The interval with a midpoint of 35 has the most cases

Note the main peak, centered at 35 years of age, with a smaller peak at 45. From the histogram you can only tell the number of cases in each of the intervals: you don't know the actual values of the cases. For example, for the interval centered around 50, all of the cases could be 50 years old, or they could be any combination of 48-, 49-, 50-, 51- and 52-year-olds. From the histogram, you see that the distribution of age values in the groups is not symmetric. There is a tail toward larger values. You know that's because only adults are included in the General Social Survey.

What kinds of things should I look for in a histogram? You already know that you should look for cases with values very different from the rest. In fact, if there are such cases, they can cause most of your data values to bunch in one or two bars of the histogram, since the horizontal axis of the histogram is selected so that all data values can be shown. You should see also whether the distribution is symmetric, since many of the statistical procedures described in Part 3 require that the distribution be more or less symmetric.

You should also look for separate clumps of data values. For example, if young men and mature women made up most of the *a little dissatisfied* group, you would see a bunch of cases with values in the 20's, perhaps, and another bunch of cases with values in the 60's. There wouldn't be many cases in between. That's an important finding, since then you know that a mean age of 40-something for the *a little dissatisfied group* is meaningless. It doesn't represent the data well. In this situation, you'd want to analyze the data for men and for women separately.

A stem-and-leaf plot is a display very much like a histogram. However, more information about the actual data values is preserved. Consider Figure 6.6, which is a stem-and-leaf plot for age in the group *a little dissatisfied* with their jobs. It looks like a histogram, because the length of each line corresponds to the number of cases in the interval. However, the cases are represented with different symbols. Each observed value is divided into two components—the leading digit or digits, called the stem, and a trailing digit, called the leaf. For example, the value 23 has a stem of 2 and a leaf of 3.

In a stem-and-leaf plot, each row corresponds to a stem and each case is represented by its leaf. More than one row can have the same stem. For example, in Figure 6.6 each stem is subdivided into two rows.

Figure 6.6 Stem-and-leaf plot of age for a little dissatisfied

To obtain a stem-and-leaf plot, select Plots in the Explore dialog box. Then select Stem-and-leaf, as shown in Figure 6.11.

```
Age of Respondent Stem-and-Leaf Plot for
SATJOB= A little dissatisfied

 Frequency    Stem &  Leaf
     2.00        2 .  33
    13.00        2 .  5556777899999
     7.00        3 .  0123334
    18.00        3 .  555566666777788899
     7.00        4 .  0012234
    13.00        4 .  5556666677888
     5.00        5 .  02223
     5.00        5 .  55679
     3.00        6 .  013
     1.00 Extremes    (>=72)

 Stem width:   10
 Each leaf:        1 case(s)
```

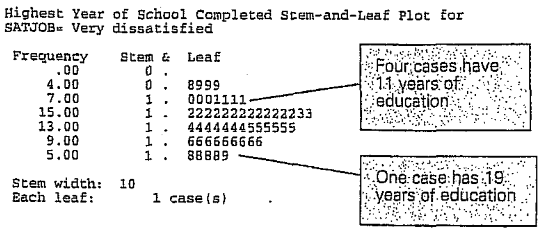
Multiply stem by stem width and add leaf values to get actual data values (60, 61, and 63)

Look at the row with the stem of 6 in Figure 6.6. The three leaves are 0, 1, and 3. What does this mean? In order to translate the stem-and-leaf values into actual numbers, you must look at the stem width given below the plot. In this case it's 10. You multiply each stem value by 10 and then add it to the leaf to get the actual value. The resulting age values are 60, 61, and 63. If the stem width were 100, you would multiply each stem by 100 and each leaf by 10 before adding them together. The values for the indicated row would be 600, 610, and 630.

If there are few values of the stem (for example, if most cases are in one or two decades of age), each stem can be subdivided into more than two rows. Consider, for example, Figure 6.7, which is a stem-and-leaf plot for years of education for all *very dissatisfied* people, regardless of their status in the work force.

Why all of a sudden are we looking at all people instead of full-time workers? The data values determine the type of stem-and-leaf plot that Explore makes. To illustrate this particular version of the plot, we had to look for a set of data that would generate it. Including all cases in the *very dissatisfied* group worked.

Figure 6.7 Stem-and-leaf plot of education for very dissatisfied



```
Highest Year of School Completed Stem-and-Leaf Plot for
SATJOB= Very dissatisfied

 Frequency    Stem &  Leaf
      .00        0 .
     4.00        0 .  8999
     7.00        1 .  0001111
    15.00        1 .  222222222222233
    13.00        1 .  4444444555555
     9.00        1 .  666666666
     5.00        1 .  88889

 Stem width:   10
 Each leaf:        1 case(s)
```

Four cases have 11 years of education

One case has 19 years of education

In Figure 6.7, the stem value 1 is subdivided into five rows—each representing two leaf values. The first row is for leaves of 0 and 1, the second row is for leaves of 2 and 3, the third for leaves of 4 and 5, the fourth for 6 and 7, and the last for 8 and 9. You see that the *very dissatisfied* group is made up of people of various educational levels. Having a college degree is no guarantee of job satisfaction.

How would you make a stem-and-leaf plot of a variable like income? For a variable like income, which has many digits, it's unwieldy and unnecessary to represent each case by the last digit. (Think of how many stems you would have!) Instead, you can look at income to the nearest thousand. For example, you can take a number like 25,323 and divide it into a stem of 2 and a leaf of 5. In this case, the stem is the ten thousands, and the leaf is the thousands. You no longer retain the entire value for the case, but that's not of concern, since income differences in the hundreds seldom matter very much. The Explore procedure always displays the stem width under the plot.

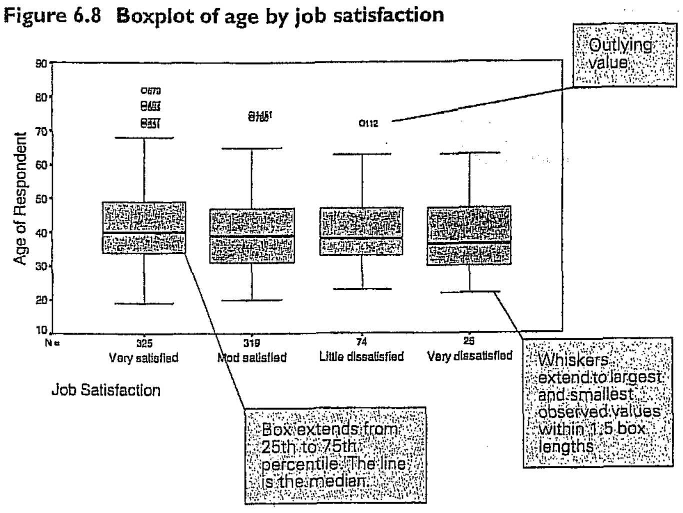
Boxplots

Another display that helps you visualize the distribution of a variable is the boxplot. It simultaneously displays the median, the interquartile range, and the smallest and largest values for a group. A boxplot is more compact than a histogram but doesn't show as much detail. For example, you can't tell if your distribution has a single peak or if there are intervals that have no cases.

You can use the Explore procedure to produce a display that contains boxplots for all the groups of interest. Consider Figure 6.8, which is an annotated boxplot of the age of the respondent for the four categories of job satisfaction.

To obtain this boxplot, select Plots in the Explore dialog box. Then select Factor levels together. (See Figure 6.11.)

Figure 6.8 Boxplot of age by job satisfaction



The lower boundary of the box represents the 25th percentile. The upper boundary represents the 75th percentile. (The percentile values known as Tukey's hinges are used to construct the box.) The vertical length of the box represents the interquartile range. Fifty percent of all cases have values within the box. The line inside the box represents the median. Note that the only meaningful scale in the boxplot is the vertical scale. All values are plotted on this scale. The width of a box doesn't represent anything.

In a boxplot, there are two categories of cases with outlying values. Cases with values between 1.5 and 3 box lengths from the upper or lower edge of the box are called **outliers** and are designated with (O). Cases with values more than 3 box-lengths from the upper or lower edge of the box are called **extreme values**. There aren't any such cases here, but if there were, they would be designated with asterisks (*). Lines are drawn from the edges of the box to the largest and smallest values that are outside the box but within 1.5 box lengths. (These lines are sometimes called **whiskers**, and the plot is sometimes called a **box-and-whiskers plot**.)

What can you tell about your data from a boxplot? From the median, you can get an idea of the typical value (the central tendency). From the length of the box, you can see how much the values vary (the spread or variability). If the line representing the median is not in the center of the box, you can tell that the distribution of your data values is not symmetric. If the median is closer to the bottom of the box than to the top, there is a tail toward larger values (this is also called **positive skewness**). If the line is closer to the top of the box, there is a tail toward small values (negative skewness). The length of the tail is shown by the length of the whiskers and the outlying and extreme points.

The Explore procedure also has more specialized charts and statistics for examining groups. These are discussed in Chapter 14 and Chapter 21.

In Figure 6.8, you see that the *very satisfied* group has the highest median age, though the differences among the groups are small. (Chapter 13 tests the hypothesis that the average age of people who are *very satisfied* is the same as the average age of those who are less than *very satisfied*.) The *very satisfied* and *moderately satisfied* groups have some large outliers. They are identified by case number in the plot. These are the satisfied old-timers who continue to work. If you specify a case label, the extreme and outlying points will be identified with this label.

What should I do if I find outliers and extremes on my boxplots? Use the case numbers to track down the data points and make sure the values are correct. If these points are the results of data entry or coding errors, correct them.

Interpretation Questions

Examine the printouts for the descriptive statistics.

1. Using Output 4.1, look for any obvious errors or problems in the data. What will you look for?

2. Name the nominal variables in Output 4.2. Can you interpret the mean and standard deviation? Explain.

3. Using Output 4.2: a) How many participants are there all together? b) How many have complete data (nothing missing)? c)What percentage took algebra 1 in high school? d) What is the range of father's education scores? Does this agree with the codebook?

Outputs and Interpretations

```
GET
  FILE='A:\hsbdata.sav'.
EXECUTE .
```

These are the instructions or syntax that you produced to retrieve the hsbdata file from the disk.

Output 4.1: Descriptives With Errors

Syntax from Problem 1. If you saved it, you can use it later to modify and rerun the analysis. See Appendix C.

Syntax for the mean, standard deviation, minimum, and maximum for all variables

```
DESCRIPTIVES
  VARIABLES=alg1 alg2 calc ethnic faed gend geo grades id maed mathach mathgr
  mosaic q01 q02 q03 q04 q05 q06 q07 q08 q09 q10 q11 q12 q13 trig visual
  /STATISTICS=MEAN STDDEV MIN MAX .
```

Interpretation of Output 4.1

The Output provides the number of subjects (*N*), the lowest and highest score, mean or average, and standard deviation for each variable. At the beginning of your data analysis, check to make sure that all means seem reasonable (given the information in your codebook), and check to see that the minimum and maximum are within the appropriate range for each variable. For example, note in the codebook that *alg1* has to be **0 = not taken** or **1 = taken** so the minimum should be 0 and maximum 1. If not, you have an error to correct before proceeding. Did you find the two errors in the data? You will correct them in **Problem 2**. Note from the bottom of Output 4.1 that the valid number (*N*) of observations/subjects (listwise) is 67, rather than 75, the number of participants in the data file. This is because the listwise *N* only includes the persons with *no* missing data on any variable. Notice that several variables (e.g., ethnicity) each have a few participants missing.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
algebra 1 in h.s.	75	0	1	.79	.41
algebra 2 in h.s.	75	0	1	.47	.50
calculus in h.s.	75	0	1	.11	.31
ethnicity	73	1	4	1.77	1.02
father's education	73	2	10	4.73	2.83
gender	75	1	2	1.55	.50
geometry in h.s.	75	0	1	.48	.50
grades in h.s.	75	2	8	5.68	1.57
identification	75	1	75	38.00	21.79
mother's education	75	2	10	4.11	2.24
math achievement	75	-1.67	23.67	12.5645	6.6703
math grades	75	0	1	.41	.50
mosaic, pattern test	75	-4.0	56.0	27.413	9.574
question 1	74	1	10	3.08	1.21
question 2	75	1	40	4.00	4.31
question 3	74	1	4	2.82	.90
question 4	74	1	4	2.16	.92
question 5	75	1	4	1.61	.97
question 6	75	1	4	2.43	.98
question 7	75	1	4	2.76	1.05
question 8	75	1	4	1.95	.91
question 9	74	1	4	3.32	.76
question 10	75	1	4	1.41	.74
question 11	75	1	4	1.36	.75
question 12	75	1	4	3.00	.82
question 13	75	1	4	2.67	.79
trigonometry in h.s.	75	0	1	.27	.45
visualization score	75	-.25	14.75	5.2433	3.9120
Valid N (listwise)	67				

Remember, 0=not taken 1=taken

Sometimes you get misleading information. There is no "average" ethnicity.

This N is different because some of the data is missing.

These two are errors and need correcting.

N only includes the persons with no missing data on any variable.

It's good to check this against the codebook. Why do you think the first number is negative?

When variables are 1 and 0, the mean indicates the percent who had 1, i.e., 27% took trig.

algebra 1 in h.s.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	not taken	16	21.3	21.3	21.3
	taken	59	78.7	78.7	100.0
	Total	75	100.0	100.0	
Total		75	100.0		

algebra 2 in h.s.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	not taken	40	53.3	53.3	53.3
	taken	35	46.7	46.7	100.0
	Total	75	100.0	100.0	
Total		75	100.0		

geometry in h.s.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	not taken	39	52.0	52.0	52.0
	taken	36	48.0	48.0	100.0
	Total	75	100.0	100.0	
Total		75	100.0		

trigonometry in h.s.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	not taken	55	73.3	73.3	73.3
	taken	20	26.7	26.7	100.0
	Total	75	100.0	100.0	
Total		75	100.0		

calculus in h.s.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	not taken	67	89.3	89.3	89.3
	taken	8	10.7	10.7	100.0
	Total	75	100.0	100.0	
Total		75	100.0		

Output 5.3: Frequencies, Statistics, and Histograms

Syntax for the frequency distribution, descriptive statistics, and histograms

```
FREQUENCIES
  VARIABLES=mosaic visual grades mathach
  /PERCENTILES= 33 67
  /STATISTICS=STDDEV VARIANCE RANGE MEAN MEDIAN MODE SKEWNESS SESKW KURTOSIS
  SEKURT
  /HISTOGRAM  NORMAL.
```

Interpretation of Output 5.3

The output file provides all the requested statistics for the four variables as a group. Then the four frequency distributions and four histograms with the normal curve superimposed over them are given individually so you can visualize whether the frequency distribution (histogram) looks normal. However, visual inspection can be deceiving because distributions need only be approximately normal. In the statistics tables, note columns for the skewness and kurtosis of the four variables. Divide each of the statistics by its standard error. If the result is not more than 5.5 (which is approximately the .01 level) that skewness or kurtosis is *not* significantly different from normal. Note that, using this measure, none of the four variables is markedly skewed, but the distribution of the *mosaic* scores is too peaked; i.e., it has a positive kurtosis almost six times its standard error. You can see this visually in the histogram.

Notice also the 33rd and 67th percentile columns in the statistic tables. You could use these percentiles if you wanted to divide your participants into three approximately equal size groups such as low, medium, and high. You can see from Output 5.3 that the 33rd and 67th percentiles for *mosaic* are 24.04 and 29.50. Thus, the low mosaic group would have scores from lowest to 24.04, the medium group from 24.04 to 29.50, and the high achievement group from 29.50 to highest. This could be done using the **Recode** command described in Assignment C.

The average, middle, and most frequent score.

Common measure of the variability of the scores.

The standard deviation squared.

Statistics

	N		Mean	Median	Mode	Std. Deviation	Variance	Range
	Valid	Missing						
visualization score	75	0	5.2433	4.7500	1.00ᵃ	3.9120	15.3040	15.00
mosaic, pattern test	75	0	27.413	27.000	25.0ᵃ	9.574	91.658	60.0
grades in h.s.	75	0	5.68	6.00	7	1.57	2.46	6
math achievement	75	0	12.5645	13.0000	14.33	6.6703	44.4930	25.33

a. Multiple modes exist. The smallest value is shown.

The highest minus the lowest score.

Statistics

	Skewness		Kurtosis		Percentiles	
					33.0000	67.0000
	Statistic	Std. Error	Statistic	Std. Error	Statistic	Statistic
visualization score	.536	.277	-.398	.548	3.5000	6.4600
mosaic, pattern test	.529	.277	3.106	.548	24.040	29.500
grades in h.s.	-.332	.277	-.763	.548	5.00	7.00
math achievement	.044	.277	-.940	.548	9.0000	15.5870

Determines if the curve is nonsymmetrical.

Tells the shape of a curve (flat or peaked).

Is the statistic 2.5 times greater than the standard error?

Divided students into three groups: Low, middle, and high thirds.

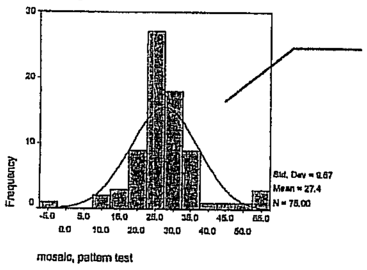
MORGAN George A. and Griego V. ORLANDO. 1998. *Easy Use and Interpretation of SPSS for Windows: Answering Research Questions With Statistics*. Mahwah: Lawrence Erlbaum Associates.

mosaic, pattern test

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	-4.0	1	1.3	1.3	1.3
	-4.0	1	1.3	1.3	2.7
	11.0	1	1.3	1.3	4.0
	13.0	1	1.3	1.3	5.3
	13.5	1	1.3	1.3	6.7
	16.0	1	1.3	1.3	8.0
	17.5	1	1.3	1.3	9.3
	18.0	1	1.3	1.3	10.7
	20.0	1	1.3	1.3	12.0
	20.5	2	2.7	2.7	14.7
	22.0	4	5.3	5.3	20.0
	22.5	2	2.7	2.7	22.7
	23.0	4	5.3	5.3	28.0
	23.5	2	2.7	2.7	30.7
	24.0	2	2.7	2.7	33.3
	24.5	3	4.0	4.0	37.3
	25.0	5	6.7	6.7	44.0
	26.0	3	4.0	4.0	48.0
	26.5	1	1.3	1.3	49.3
	27.0	5	6.7	6.7	56.0
	27.5	1	1.3	1.3	57.3
	28.0	3	4.0	4.0	61.3
	28.5	1	1.3	1.3	62.7
	29.0	2	2.7	2.7	65.3
	29.5	2	2.7	2.7	68.0
	30.0	2	2.7	2.7	70.7
	30.5	2	2.7	2.7	73.3
	31.0	4	5.3	5.3	78.7
	32.0	1	1.3	1.3	80.0
	32.5	1	1.3	1.3	81.3
	33.0	3	4.0	4.0	85.3
	34.0	1	1.3	1.3	86.7
	35.0	1	1.3	1.3	88.0
	35.5	1	1.3	1.3	89.3
	36.0	1	1.3	1.3	90.7
	37.0	1	1.3	1.3	92.0
	41.0	1	1.3	1.3	93.3
	44.0	1	1.3	1.3	94.7
	51.5	1	1.3	1.3	96.0
	53.0	1	1.3	1.3	97.3
	56.0	2	2.7	2.7	100.0
	Total	75	100.0	100.0	
Total		75	100.0		

Do you know what this means?

Histogram

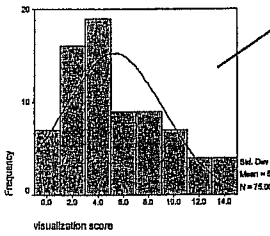


This is not normally distributed. The kurtosis is high.

visualization score

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	-.25	7	9.3	9.3	9.3
	1.00	10	13.3	13.3	22.7
	2.25	5	6.7	6.7	29.3
	2.50	1	1.3	1.3	30.7
	3.50	7	9.3	9.3	40.0
	3.75	2	2.7	2.7	42.7
	4.75	10	13.3	13.3	56.0
	5.00	2	2.7	2.7	58.7
	6.00	6	8.0	8.0	66.7
	6.50	1	1.3	1.3	68.0
	7.25	5	6.7	6.7	74.7
	8.50	2	2.7	2.7	77.3
	8.75	2	2.7	2.7	80.0
	9.50	1	1.3	1.3	81.3
	9.75	6	8.0	8.0	89.3
	11.00	4	5.3	5.3	94.7
	13.50	2	2.7	2.7	97.3
	14.75	2	2.7	2.7	100.0
	Total	75	100.0	100.0	
Total		75	100.0		

Histogram

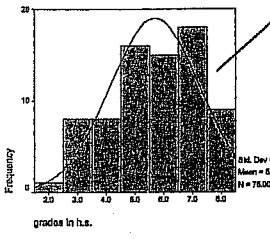


This is approximately normally distributed.

grades in h.s.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	mostly D	1	1.3	1.3	1.3
	half CD	8	10.7	10.7	12.0
	mostly C	8	10.7	10.7	22.7
	half BC	16	21.3	21.3	44.0
	mostly B	15	20.0	20.0	64.0
	half AB	18	24.0	24.0	88.0
	mostly A	9	12.0	12.0	100.0
	Total	75	100.0	100.0	
Total		75	100.0		

Histogram



This is approximately normally distributed.

BLACK, Thomas, R. 1999. *Doing Quantitative Research in the Social Sciences. An Integrated Approach to Research Design, Measurement and Statistics*. London: Sage.

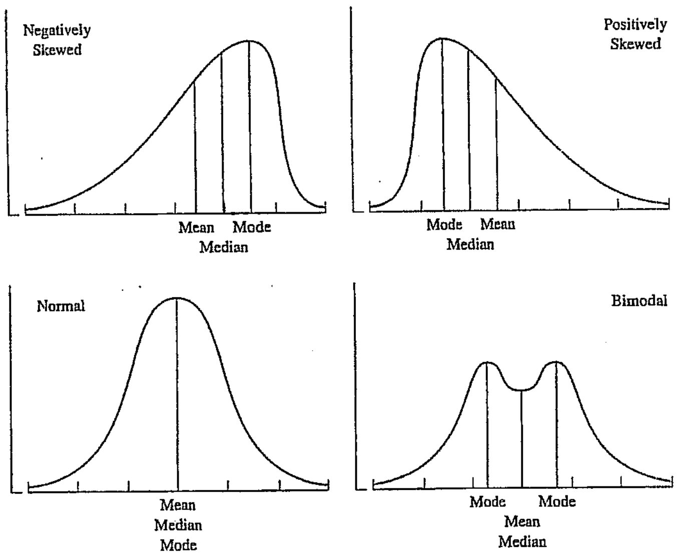


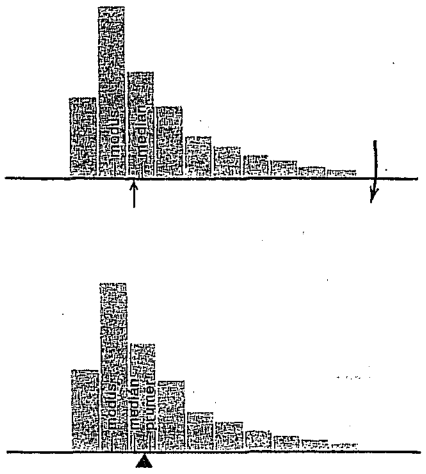
FIGURE 12.20
Mean, median and mode for a variety of distributions

WONNACOT Thomas H a Ronald J. WONNACOT, 1993. *Statistika pro obchod a hospodářství*. Praha: Victoria Publishing.

Co když je ale rozdělení nesymetrické? Na obr. 2-4 a) vidíme, že rozdělení dlouze klesá vpravo. Bude v tomto případě např. medián totožný s modem? Vzhledem k velkému množství pozorování v pravé části je zřejmé, že pro rozdělení výběru na dvě poloviny je třeba, aby medián ležel více vpravo od vrcholu rozdělení, tedy i vpravo od modu.

A kde se bude nacházet průměr? Někde blízko mediánu? V obr. 2-2 a) vidíme, co se stane, když se pokusíme vyvážit rozdělení v mediánu. Na každé straně je stejný počet pozorování, ale pozorování napravo dosahují dále, a celé rozdělení se tedy naklání doprava. K nalezení skutečného těžiště je třeba jít dále doprava, jak vidíme na obr. 2-4 b). Průměr se tedy nachází napravo od mediánu.

Jaké jsou tedy závěry z nesymetrického rozdělení funkce? **Vzhledem k modu leží medián ve směru delší části rozdělení a průměr ještě dále v tomto směru.**



OBRÁZEK 2-4
Modus, medián a průměr v rozdělení klesajícím vpravo.
a) Medián je vpravo od modu.
b) Těžiště (průměr) je vpravo od mediánu. (Rozdělení nebude v rovnováze, umístíme-li těžiště do mediánu, neboť pozorování vpravo převáží.

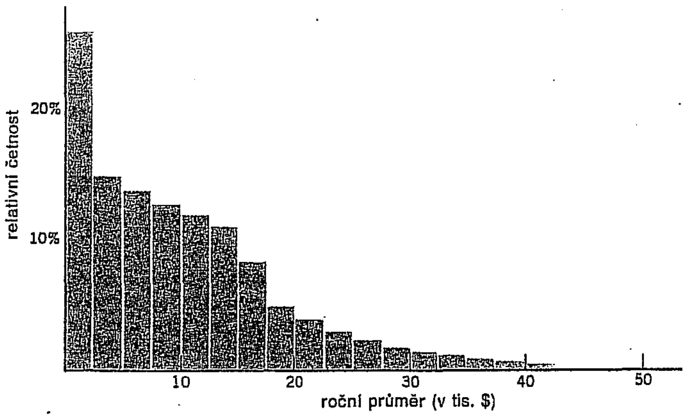
F—KTERÁ CHARAKTERISTIKA POLOHY JE NEJVHODNĚJŠÍ — MODUS, MEDIÁN NEBO PRŮMĚR?

Pro některé účely je vhodná jedna charakteristika, pro jiné jiná. Pro ilustraci se podívejme na rozdělení příjmů 78 miliónů Američanů v roce 1975 znázorněné na obr. 2-5.

Modus se nachází v blízkosti 0, a ukazuje nám pouze, že nejvíce lidí je prakticky bez příjmů – nezaměstnaní a důchodci. Největší příjmy jsou zastoupeny v mnohem menším počtu v rozmezí 2 až 40 tisíc – tento jev však není modem vůbec postižen. Budeme-li používat modus, nebudeme moci srovnat příjmy v roce 1975 s mnohem nižšími příjmy v roce 1875! V tomto případě nám je modus k ničemu.

Medián se nachází v asi 8 tisících dolarech a tato hodnota je mnohem reprezentativnější – 50 % nad a 50 % pod. Možná je to nejlepší hodnota „typického" amerického příjmu. Navíc je **resistentní**, tj. nereaguje na extrémní hodnoty jediného pozorování. Například, zvýšíme-li nejvyšší příjem desetkrát, medián se nezmění.

Konečně průměr je okolo 10 tisíc dolarů. Tato hodnota byla získána rovnocenným zahrnutím všech dolarů — dolarů žebráka i milionáře. To má své výhody i nevýhody — je to nejužitečnější měřítko pro berní úřad, neboť vyjadřuje celkový příjem (78 miliónů lidí × 10 tisíc dolarů = 780 miliard dolarů); přesto není tak dobrým měřítkem pro typický příjem jako medián, neboť se může značně změnit vychýlením jen jediného pozorování (jediným velmi vysokým nebo velmi nízkým příjmem), tj. není rezistentní jako medián.



OBRÁZEK 2-5
Příjmy amerických mužů, 1975. (Stat. Abst. of U.S., 1980, str. 462)

4. lekce
UMĚLÉ PROMĚNNÉ (modul TRANSFORM: procedury Recode, Compute, Count, Rank Cases) .

Transforming and Selecting Data

B

SPSS includes a powerful set of facilities for transforming data values and selecting which cases should be analyzed. This appendix covers two general types of data manipulation: data transformation and case selection.

Data transformation procedures change the actual values of your variables or create new variables. For example, you can create a new variable that contains the natural log of an existing variable. **Case selection** procedures do not change data values but restrict the number of cases used in the analysis. For example, you can restrict your analysis to people who are married or who are holding full-time jobs.

SPSS also provides a number of advanced data manipulation utilities that are not used in this book and are not discussed here. These utilities are described, however, in the online Help system.

Data Transformations

Often you need to make modifications to your data before you can perform your analysis. For small changes, such as the urbanization of Bhutan in Chapter 8, it is easy to enter the corrected value into the Data Editor. But suppose you want to take the natural log of several variables, each with 1500 cases, as you do for the analysis in Chapter 22? SPSS provides data transformation facilities to handle such tasks easily and accurately.

Data transformations affect the values of existing variables or create new variables. Transformations affect only the working data file; the changes do not become permanent unless you save the working data file to your disk.

Transformations at a Glance

This appendix describes the following transformations, available using the SPSS Data Editor's Transform menu:

Compute. Compute calculates data values according to a precise expression. With this option, you can do anything from set a variable to 0 for all cases to calculate an elaborate expression involving the values of other variables. You can assign the computed values to a new variable, or you can assign them to an existing variable (replacing the current values). You can also request that the computation be carried out selectively based on a conditional expression.

Recode. Recode assigns discrete values to a variable, based solely on the present values of the variable being recoded. You can assign the recoded values to the variable being recoded, or you can assign them to a new variable. You can also request that the computation be carried out selectively based on a conditional expression.

Automatic Recode. Automatic recode assigns successive integer codes—1, 2, 3, and so on—to a new variable, based on the existing codes of another variable. This saves you the effort of specifying how the recoding should be carried out.

The following options are also available on the Transform menu but are not discussed in this book. These transformations are described in the online Help system.

Random Number Seed. Lets you reproduce the pseudo-random numbers generated by SPSS for sampling and certain functions in the transformation language.

Count. Creates a new variable that counts for each case the number of times certain specified values occur in other variables. You can count, for example, the number of times that values of 1 or 2 occur in a group of existing variables.

Rank Cases. Creates rank scores, which show each case's rank among all the cases in the file according to the values of a particular variable.

Create Time Series. Creates new time series, containing functions such as the differences between successive cases, in a time series data file.

Replace Missing Values. Supplies nonmissing values to replace missing values, according to any of several functions that might provide plausible values.

Run Pending Transformations. Forces SPSS to execute transformations that are pending as a result of the Transform & Merge Options setting. (See "Delaying Processing of Transformations" below.)

Saving Changes

Bear in mind when transforming your data that you are only changing the working data file.

▶ To make the changes permanent, save the working data file to your hard disk.

▶ To discard the changes, exit SPSS (or open a new data file) without saving the working data file.

Delaying Processing of Transformations

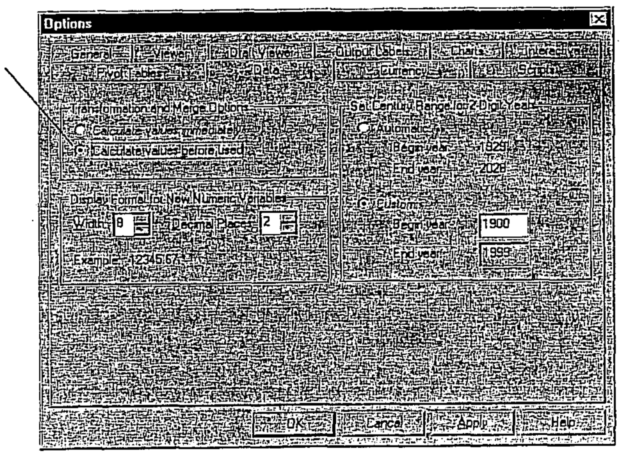
SPSS normally executes transformation commands as soon as you request them. However, since transformations can take several minutes to execute for a very large data file, there are times when you want to enter a dozen or more transformation commands one after another and then let the computer process them all at once.

▶ To prevent SPSS from processing transformations immediately, from the menus choose:

Edit
  Options...

▶ In the tabbed SPSS Options dialog box, click the Data tab. This displays the SPSS Options Data tab, as shown in Figure B.1

Figure B.1 SPSS Options dialog box



Transformation and Merge Options
○ Calculate values immediately
● Calculate values before used

Display Format for New Numeric Variables
Width: 8 Decimal Places: 2
Example: 12345.67

Set Century Range for 2-Digit Years
○ Automatic
Begin year: 1929
End year: 2028
● Custom
Begin year: 1900
End year: 1999

Select Calculate values before used

▶ Set Transformation & Merge Options to Calculate values before used and click OK.

With this setting, SPSS does not execute Compute and Recode transformations until it needs the data. In the meantime, the status bar displays the message Transformations pending and the results of transformations are not yet visible.

▶ To execute pending transformations, run a procedure that requires SPSS to use the data or choose Run Pending Transformations from the Transform menu.

When transformations are pending, the Data Editor will not allow you to make certain changes to your working data file.

Recoding Values

Recoding is done with a series of specifications, of the form, "If the old value is this, assign a new value of that." A case's existing value is checked against each of these specifications until one of them matches. Then the new value is assigned, and SPSS moves on to process the next case.

NORUŠIS, Marija J. 1998. *Guide to Data Analysis*. Upper Saddle River, Prentice Hall.

There are two Recode commands: Recode into Same Variables and Recode into Different Variables. The former changes the values of variables based solely on their existing values, while the latter creates new variables with values that depend only on the existing values of single variables.

- A case is never changed by more than one of a group of recode specifications.
- If a case doesn't match any of the recode specifications, its value remains unchanged (if recoding into same variable) or becomes system-missing (if recoding into a new variable).

Example: Recoding Age into Age Categories

This example recodes the variable *age* (age in integer years) into a new variable that contains age in one of three categories: 14 through 29, 30 through 49, and 50 or older. (If *age* is not an integer you must modify the recode statement so that ages between 29 and 30, and between 49 and 50, are assigned to the proper groups.)

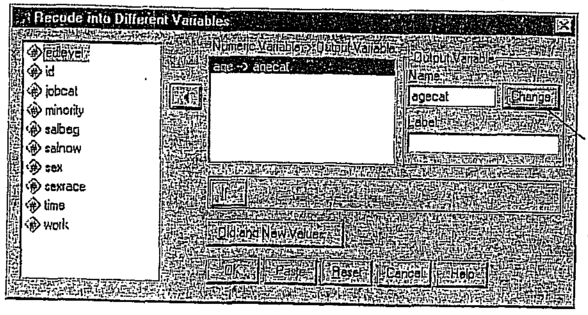
▶ Open the *salary.sav* data file.

▶ From the menus choose:

Transform
 Recode ▶
  Into Different Variables...

This opens the Recode into Different Variables dialog box, as shown in Figure B.2.

Figure B.2 Recode into Different Variables dialog box



age --> agecat

Name: agecat

Type agecat and click Change

▶ Move *age* into the Input Variable -> Output Variable list. The name of the list changes to reflect that a numeric variable has been selected, as shown in Figure B.2.

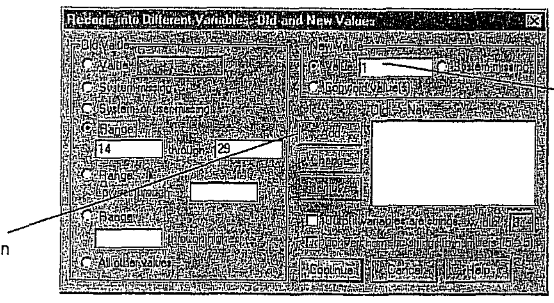
▶ In the Output Variable box, type *agecat* for the output variable and click Change.

This adds agecat to the Numeric Variable -> Output list. A new variable *agecat* will be created, which contains the recoded values of *age*.

▶ Click Old and New Values.

This opens the Old and New Values dialog box, as shown in Figure B.3.

Figure B.3 Old and New Values dialog box



Enter new value

Click Add to add specification to list

▶ In the Old Value group, select the first Range alternative.

▶ Type 14 in the first range box and 29 in the second range box.

▶ Type 1 in the New Value box.

▶ Click Add.

The specification 14 thru 29 -> 1 is added to the Old -> New list. All ages between 14 and 29 will be coded 1 in the new *agecat* variable.

▶ Click again on Range. Type 30 in the first range box and 49 in the second range box.

▶ Type 2 in the New Value box and click Add.

▶ Click Range: through highest and type 50 in the box.

NORUŠIS, Marija J. 1998. *Guide to Data Analysis*. Upper Saddle River, Prentice Hall.

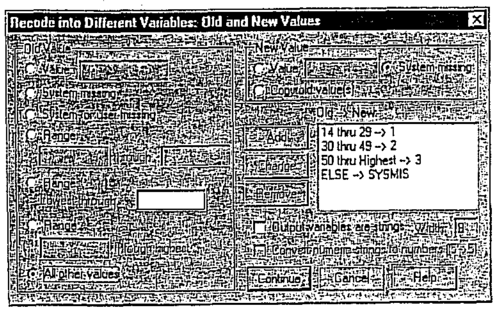
▶ Type 3 in the New Value box and click Add.

That should take care of all age groups in this file. But what if someone is coded with an age less than 14? Since the file contains data about adults who work for a bank, that would surely be a coding mistake, but it could happen. It's best to be safe.

▶ In the Old Values box, click All other values.

▶ In the New Value box, click System-missing and click Add one more time.

Figure B.4 Completed Old and New Values dialog box



14 thru 29 --> 1
30 thru 49 --> 2
50 thru Highest --> 3
ELSE --> SYSMIS

The Old and New Values dialog box should now look like Figure B.4. If it doesn't—if one of your specifications is incorrect—click the incorrect specification in the Old -> New list, make the needed correction, and click Change.

▶ Click Continue to return to the Recode into Different Variables dialog box. Then click OK to execute the transformation.

You have now changed the working data file; however, you don't want to make these changes a permanent part of the *salary.sav* data file.

▶ To avoid saving changes to the *salary.sav* data file, exit SPSS *without* saving changes or clear the Data Editor by selecting New from its File menu.

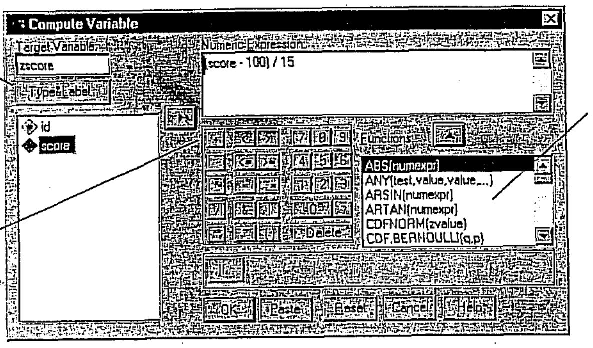
Computing Variables

The Compute Variable dialog box assigns the result of a single expression to a "target variable" for each case. The target variable can be a new variable or an existing variable (in which case the existing values will be overwritten). For example, you can compute standard scores for a variable, as described in the first example below. A great number of functions are available, so expressions can be quite complex.

▶ To open the Compute Variable dialog box, as shown in Figure B.5, from the Data Editor menus choose:

Transform
 Compute...

Figure B.5 Compute Variable dialog box



Target Variable: zscore

Numeric Expression: (score - 100) / 15

Click to specify variable type if other than numeric

Function list

Calculator pad

Unlike a spreadsheet, SPSS does not remember the formula used to compute data values or automatically update them. (In the example mentioned above, if you go back and change the values for the variable *score*, the *zscore* values will not be automatically recalculated to reflect the change.)

The Calculator Pad

The calculator pad allows you to paste operators and functions into your formula. You don't have to use the calculator pad: you can click anywhere in the Numeric Expression box and start typing. Often that's the simplest and quickest way to build an expression. The visual controls in the calculator pad are there to remind you of the possibilities and to reduce the likelihood that you won't remember how to spell one of the many functions available in SPSS.

To use the calculator pad, just click on buttons to paste symbols and operators at the insertion point. Use the mouse to move the insertion point.

See "Example: Cumulative Distribution Function" on p. 535 for an example of using functions in an expression.

To paste a function, select it in the scrolling list and click the ▲ button. You must then fill in the arguments, which are the values that the function operates on.

A few basic calculator pad operators are described in Table B.1. The Help system contains a more detailed description of the calculator pad, with definitions of all the functions.

Table B.1 Calculator pad operators

*	Multiply
/	Divide
**	Raise to power
+	Add
–	Subtract

Example: Computing Z Scores

Suppose you have a sample of IQ scores, and you wish to calculate standard scores (*z* scores) for the sample. Assuming that in the population IQ scores have a mean of 100 and a standard deviation of 15 (as was long assumed to be true), the formula is

zscore = (*score* – 100)/15

To compute standard scores for a variable according to this formula:

▶ Open the *iq.sav* data file.

This file contains IQ scores for a hypothetical group of students.

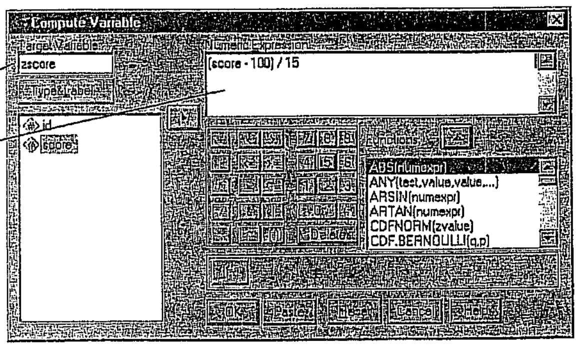
▶ Activate the Data Editor window.

▶ From the menus choose:

Transform
  Compute...

This opens the Compute Variable dialog box, as shown in Figure B.6.

Figure B.6 Compute Variable dialog box


Compute Variable
Target Variable: zscore
Numeric Expression: (score - 100) / 15
Functions: ABS(numexpr), ANY(test,value,value,...), ARSIN(numexpr), ARTAN(numexpr), CDFNORM(zvalue), CDF.BERNOULLI(q,p)

▶ Click in the Target Variable box and type zscore.

Since zscore is a numeric variable, you don't have to click the Type & Label button.

▶ Click in the Numeric Expression box.

You can simply type the expression (score–100)/15 directly in the Numeric Expression box or build it using the calculator pad, as follows:

▶ Select score in the variable list and click ▶.

The variable name *score* is pasted into the expression at the insertion point.

▶ Enter –100.

▶ Select the entire expression score –100 and click the ( ) button.

The expression now reads (score –100).

▶ Enter /15.

The expression now reads (score –100)/15.

▶ Click OK.

SPSS computes *z* scores for all cases in the working data file.

Example: Cumulative Distribution Function

You can calculate the proportion of the population with *z* scores greater in absolute value than each of the *z* scores in your sample, as discussed in question 10 in the exercises for Chapter 10. Assuming the variable *zscore* contains the *z* scores for your sample, the formula is

twotailp = 2 *(1 – cdfnorm(abs(zscore)))

▶ If you want to attempt this example, you can substitute any variable that contains *z* scores for the *zscore* variable named in the formula above. (You can use the Descriptives procedure to save *z* scores for any variable, as described in Chapter 4.)

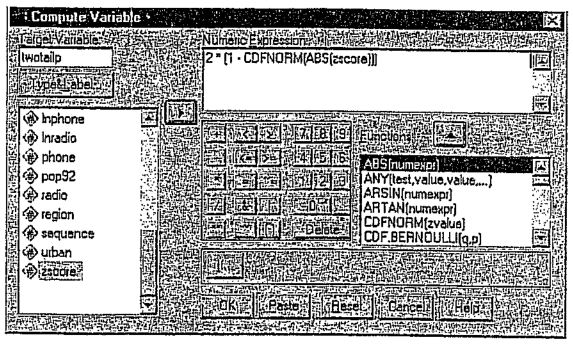
▶ From the menus choose:

Transform
  Compute...

This opens the Compute Variable dialog box.

▶ Type twotailp in the Target Variable box.

Figure B.7 Compute Variable dialog box


Compute Variable
Target Variable: twotailp
Numeric Expression: 2 * (1 - CDFNORM(ABS(zscore)))
Variables: hphone, inradio, phone, pop92, radio, region, sequence, urban, zscore
Functions: ABS(numexpr), ANY(test,value,value,...), ARSIN(numexpr), ARTAN(numexpr), CDFNORM(zvalue), CDF.BERNOULLI(q,p)

You can simply type the expression 2*(1-CDFNORM(ABS(zscore))), as shown in Figure B.7, or build the expression as follows:

▶ Enter 2*(1–).

▶ With the cursor inside the right parenthesis, select CDFNORM(zvalue) in the Functions list and click ▲.

The CDFNORM function is pasted into the formula at the insertion point. The expression now reads 2 * (1– CDFNORM(?)), with the question mark selected. You must replace the question mark with an argument for the CDFNORM function.

▶ Select ABS(numexpr) in the Functions list and click ▲.

The expression now reads 2 * (1– CDFNORM (ABS(?))). Once again, the question mark is selected; you must now supply an argument for the ABS function.

▶ Select *zscore* in the variable list and click ▶.

The variable *zscore* is now pasted in as the argument for the ABS function. The expression is now complete.

▶ Click OK.

SPSS computes the proportions for all cases in the working data file.

Automatic Recoding

Automatic Recode is particularly useful as a way of converting a string variable into a numeric variable.

SPSS's Recode facility is quite useful but requires you to enter detailed specifications. The Automatic Recode facility needs no specifications. It simply converts all the codes of a current variable into new codes—1, 2, 3, and so on—for a new variable.

Example: Creating Numeric Country Codes

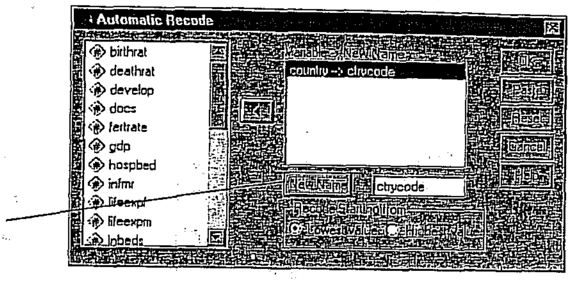
In the *country.sav* data file, the string variable *country* contains the name of each country. Suppose you want to create numeric country codes. You can do this as follows:

▶ From the Data Editor menus choose:

Transform
  Automatic Recode...

This opens the Automatic Recode dialog box, as shown in Figure B.8.

Figure B.8 Automatic Recode dialog box



Type name for new variable and click New Name

▶ Select *country* in the variable list and move it into the Variable -> New Name list.

▶ Type ctrycode in the New Name box and click New Name.

▶ Click OK.

SPSS creates the new variable *ctrycode*, which contains a unique numeric code for each country. The codes are assigned in sequence; the first country will have a code of 1, the second 2, and so on. If there were several cases for the same country, they would all be assigned the same code value.

Since the original variable *country* does not have value labels, the actual values of *country* (Afghanistan, Albania, Algeria, and so on) are used as value labels for the new variable *ctrycode*.

Conditional Transformations

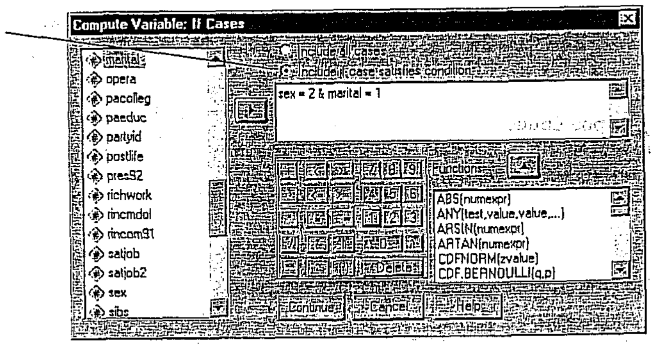
If you want to transform the values of only some cases, depending on their data values, you need a **conditional transformation**, one that is carried out only if a logical condition is true. For example, you might want to transform *only* cases for people who are full-time workers.

The Compute Variable dialog box, both Recode dialog boxes, and the Count dialog box (not shown) allow you to specify such a logical condition.

▶ To specify a logical condition for a transformation, click If in the Compute Variable, Recode, or Count dialog box.

This opens a dialog box where you can specify a logical condition. For example, the Compute Variable If Cases dialog box is shown in Figure B.9.

Figure B.9 Compute Variable If Cases dialog box



sex = 2 & marital = 1

Select to specify a logical condition

See "The Calculator Pad" on p. 532.

This dialog box contains the familiar Calculator Pad. Here you use it to build a logical condition, one that is either true or false for a case, depending on the case's data values. Table B.2 describes some operators that are particularly useful in building logical conditions.

Table B.2 Operators useful in logical expressions

<	Less than
>	Greater than
<=	Less than or equal to
>=	Greater than or equal to
=	Equal to
~=	Not equal to
&	And
|	Or
~	Not

The logical expression sex = 2 & marital = 1, for example, is true only for those cases in which *both* conditions are met: the variable *sex* must equal 2 *and* the variable *marital* must equal 1. The logical expression sex = 2 | marital = 1, by contrast, is true if either of the conditions is met.

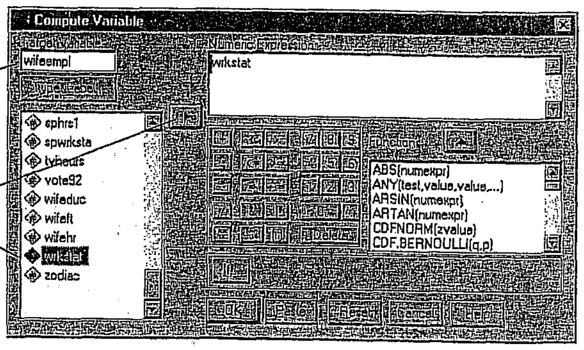
Example: Wife's Employment Status

The General Social Survey contains employment status questions for the respondent and for the respondent's spouse. The respondent could be either husband or wife, depending on who was interviewed. This means that for each household, the wife's work status could be coded in either the variable *wrkstat* (if the wife was interviewed) *or* in the spouse's work status variable *spwrksta* (if the husband was interviewed). To create a variable containing, for all married couples, the wife's employment status, you might proceed as follows:

▶ To open the Compute Variable dialog box (see Figure B.10), from the menus choose:

Transform
  Compute...

Figure B.10 Compute Variable dialog box



wifeempl

wrkstat

Type wifeempl

Paste wrkstat into the expression

▶ In the Compute Variable dialog box, type wifeempl into the Target Variable box.

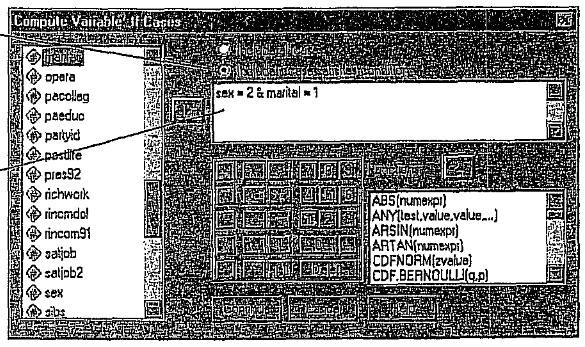
▶ Select *wrkstat* in the variable list and press ▶ to move it into the Numeric Expression box.

The new variable *wifeempl* will have the same value as the variable *wrkstat*. However, you must specify that this expression will only be evaluated for cases where the respondent is a married woman.

▶ Click If.

This opens the Compute Variable If Cases dialog box, as shown in Figure B.11.

Figure B.11 Compute Variable If Cases dialog box



sex = 2 & marital = 1

Select Include if case satisfies condition

Enter expression

▶ Select Include if case satisfies condition.

▶ Using either the calculator pad or the keyboard, enter the condition sex = 2 & marital = 1.

This condition specifies that the new value should be computed only for cases for whom the value of the variable *sex* equals 2 (the code for female) *and* for whom the value of *marital* equals 1 (married). For cases that do not meet this condition, the new variable *wifeempl* will be equal to the system-missing value.

▶ Click Continue to return to the Compute Variables dialog box. Then click OK.

This creates a new variable *wifeempl*, which is equal to *wrkstat* for married women. For cases where the respondent is not married, or is a man, the value of *wifeempl* is not defined (system-missing).

At this point, you're halfway there. But what about respondents who are married men? In that case the wife's employment status would be coded in the variable *spwrksta*, which contains the work status of the respondent's spouse.

▶ Open the Compute Variable dialog box again.

▶ Delete *wrkstat* from the Numeric Expression box.

NORUŠIS, Marija J. 1998. *Guide to Data Analysis*. Upper Saddle River, Prentice Hall.

▶ Select the variable *spwrksta* and paste it into the Numeric Expression box.

▶ Click If.

The logical expression still reads sex = 2 & marital = 1.

▶ Delete the 2 and type 1 in its place.

The expression now reads sex = 1 & marital = 1.

▶ Click Continue and then click OK.

This sets *wifeempl* equal to *spwrksta* for married men. To summarize, the first transformation creates a new variable *wifeempl*, which is equal to *wrkstat* for married women and not defined for others. The second conditional transformation sets *wifeempl* equal to *spwrksta* for married men. The end result is a variable equal to wife's employment status for all married couples. For unmarried respondents, neither transformation is executed and *wifeempl* is never changed. Since it's a new variable, it is assigned the system-missing value for the unmarried respondents.

Bryman Alan and Duncan Cramer. 1997. *Quantitative Data Analysis with SPSS for Windows*. London, Routledge.

RELATIONAL OPERATORS

A relational operator like = compares the value on its left (for example, **gender**) with that on its right (for example, **1**). There are six such operators, which are represented by the following symbols:

=	equal to
~=	not equal to
<	less than
<=	less than or equal to
>	greater than
>=	greater than or equal to

The question of which is the most appropriate operator to use in selecting cases will depend on the selection criteria. To select cases under 40 years of age, we could use less than (<):

age < 40

It would also, of course, have been possible to use less than or equal to (<=) 39 in this instance since we are dealing with whole numbers:

age <= 39

To select non-whites, we could use not equal to (~=) 1 since whites are coded 1:

ethnicgp ~= 1

COMBINING LOGICAL RELATIONS

We can combine logical expressions with the logical operators & (**and**) and | (**or**). For example, we can select white men under 40 with the following conditional expression:

ethnicgp = 1 & gender = 1 & age < 40

To select people of only West Indian and African origin, we would have to use the | (**or**) logical operator:

ethnicgp = 3 | ethnicgp = 4

Note that it is necessary to repeat the full logical relation. It is *not* permissible to abbreviate this command as:

ethnicgp = 3 | 4

An alternative way of doing the same thing is to use the **any** logical function where any case with a value of either **3** or **4** for the variable **ethnicgp** is selected:

any (ethnicgp,3,4)

The variable and the values to be selected are placed in parentheses.

To select people between the ages of 30 and 40 inclusively, we can use the expression:

age >= 30 & age <= 40

Here, we have to use the **&** (**and**) logical operator. If we used | (**or**), we would in effect be selecting the whole sample since everybody is either above 30 or below 40 years of age.

Another way of selecting people aged 30 to 40 inclusively is to use the **range** logical function where any case with a value in the range of 30 to 40 for the variable **age** is selected:

range (age,30,40)

MORGAN George A. and Griego V. ORLANDO. 1998. *Easy Use and Interpretation of SPSS for Windows: Answering Research Questions With Statistics*. Mahwah: Lawrence Erlbaum Associates.

APPENDIX C

Working with SPSS Syntax (Log) Files

Mei-Huei Tsay

It is possible to modify syntax files to run slightly different statistics and/or complex, customized statistics. Sometimes output files are too large to save on a disk or take up too much space on your hard drive. Therefore, it is a good idea to understand how to use the syntax or log files that contain SPSS commands. You can use the SPSS logs from your output file to run these commands.

It is possible to open a syntax window and type in commands, but sometimes it is easier to build your syntax file by using one of the following methods:

- Paste syntax commands from dialog boxes.
- Copy syntax from the output log.
- Copy syntax from the journal file.

Creating Syntax Commands From Dialog Boxes: Using Paste Instead of OK

The easiest way to generate a syntax command file is to make selections in dialogue boxes and paste the syntax of the selections into a syntax window. By pasting the syntax in the syntax window, you can generate a job file which allows you to repeat the analysis, edit it, save the syntax in a syntax file, and copy/cut it into an output log.

To paste syntax commands from a dialog box:

- Retrieve the data file.
- Open the dialog box and make desired selections. For example: **Statistics => Summarize => Frequencies.**
- After making all the desired selections, click **Paste** instead of **OK** (see Fig. C.1). The syntax command is pasted to the syntax window. If you don't have an open syntax window, SPSS will open a new syntax window and paste the syntax there (see Fig. C.2).

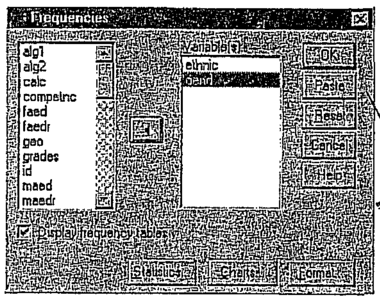
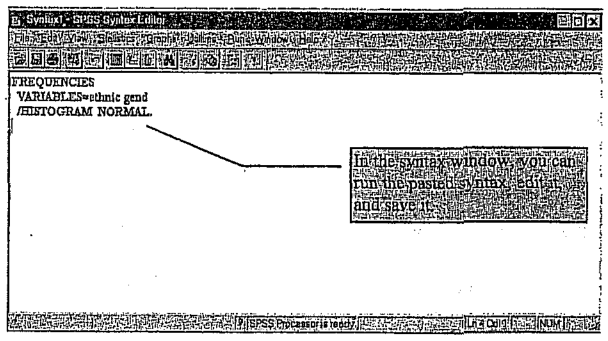


Fig. C.1. Frequencies dialog box.

After making the desired selections, click Paste instead of **OK** which allows you to paste your command selections from the dialog box to a syntax window.


Syntax1 - SPSS Syntax Editor
File Edit View Statistics Graphs Utilities Run Window Help
SPSS Processor is ready
Ln 4 Col 1
NUM

```
FREQUENCIES
 VARIABLES=ethnic gend
 /HISTOGRAM NORMAL.
```

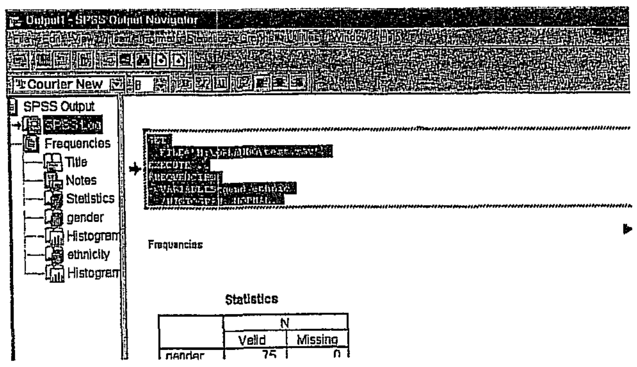
In the syntax window, you can run the pasted syntax, edit it, and save it.

Fig. C.2. The syntax window.

- To run/use a syntax file, when there is only one syntax file in the window, just simply click on **Run** in the menu bar.
- If there are many syntax files in the window, *highlight* the desired syntax first, then click **Run** then **Selection.** The output will show on the output window.
- If you need to repeat all the analyses in the syntax window, you can click on **Run** and then **All.**

MORGAN George A. and Griego V. ORLANDO. 1998. *Easy Use and Interpretation of SPSS for Windows: Answering Research Questions With Statistics*. Mahwah: Lawrence Erlbaum Associates.

- To do this, from the menus click on **Edit => Options => Navigator => Display commands in the log.** Each command you did will then be recorded in the SPSS log as in Fig. C.3.



```
GET
  FILE='M:\ORLANDO\test.sav'.
EXECUTE .
FREQUENCIES
  VARIABLES=gend ethnic
  /HISTOGRAM NORMAL.
```

	N	
	Valid	Missing
gender	75	0

Fig. C.3. Syntax commands in the SPSS log.

- To copy the syntax from the Output Navigator, first *double click* on the syntax file table to activate it (which allows you to edit the table), then **highlight** the desired syntax file (see Fig. C.3); from **Edit,** click on **Copy.**
- Second, open a previously saved syntax file or create a new one.
- To create a new syntax file, from the menus choose **File => New => Syntax;** then in the syntax window, choose **Edit** then **Paste.** You can run or change the pasted syntax as we did above.

Using Syntax From the Journal File

This is a more complicated way of doing things, but it is worth mentioning here.

By default, SPSS records all commands executed during a session in a journal file named *spss.jnl* (set with **Options** on the **Edit** menu). You can edit the journal file and save it as a syntax file that you can use to repeat the previous analysis.

To open the journal file,

- from the menus choose **File => Open**; under the **Files of Type**,
- choose **All files** (*.*);
- then choose *spss.jnl* from the file name box or enter **.jnl* in the **File Name box,**
- then click **Open** (see Fig. C.4).

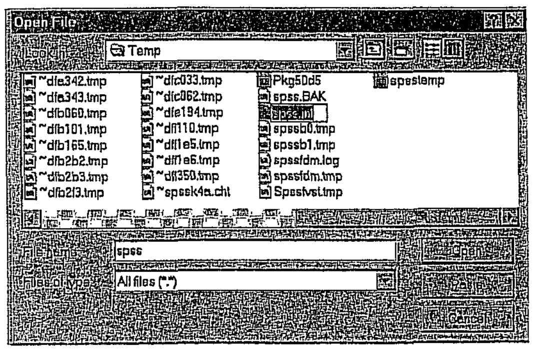
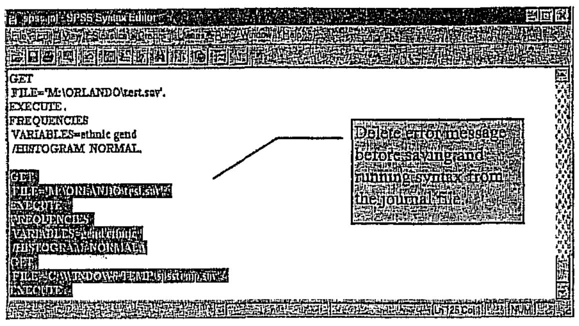


Fig. C.4. Open spss.jnl.

The journal file is a text file that can be edited like any other text file. But notice, because error and warning messages are also recorded in the journal file along with your commands, you must delete any of these messages that appear before saving or running the syntax file (see Fig. C.5).


spss.jnl - SPSS Syntax Editor
File Edit View Statistics Graphs Utilities Run Window Help
Ln 25 Col 1
NUM

```
GET
 FILE='M:\ORLANDO\test.sav'.
EXECUTE .
FREQUENCIES
 VARIABLES=ethnic gend
 /HISTOGRAM NORMAL.

GET
 FILE='M:\ORLANDO\test.sav'.
EXECUTE .
FREQUENCIES
 VARIABLES=gend ethnic
 /HISTOGRAM NORMAL.
GET
 FILE='C:\WINDOWS\TEMP\spsstmp.sav'.
EXECUTE .
```

Delete error message before saving and running syntax from the journal file.

Fig. C.5. Editing the journal file.

To run or edit the journal file, see above.

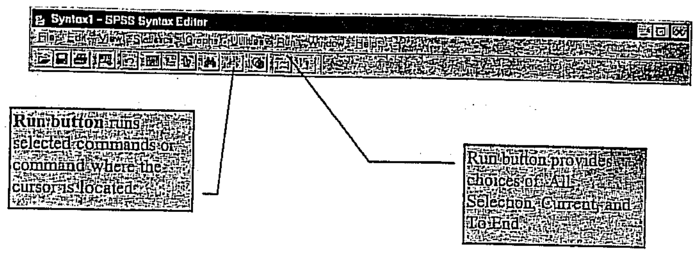
Running Syntax Commands

You can run single commands, selected groups of commands, or all commands in a syntax window. The following options are available on the **Run** menu (see Fig. C.6):

- **All.** Runs all commands in the syntax window.
- **Selection.** Runs the currently selected commands. This includes any commands partially highlighted.
- **Current.** Runs the command where the cursor is currently located.

MORGAN George A. and Griego V. ORLANDO. 1998. *Easy Use and Interpretation of SPSS for Windows: Answering Research Questions With Statistics*. Mahwah: Lawrence Erlbaum Associates.

- **To End**. Runs all commands from the current cursor location to the end of the command syntax file.


Syntax1 - SPSS Syntax Editor
File Edit View Statistics Graphs Utilities Run Window Help

Run button runs selected commands or command where the cursor is located.

Run button provides choices of: All, Selection, Current and To End.

Fig. C.6. Syntax editor toolbar.

5. lekce

NORMÁLNÍ ROZLOŽENÍ A ZÁKLADY TESTOVÁNÍ HYPOTÉZ. STATISTICKÁ INFERENCE ANEB ZOBECŇOVÁNÍ VÝBĚROVÝCH VÝSLEDKŮ NA ZÁKLADNÍ SOUBOR.

3.3 Normální rozdělení

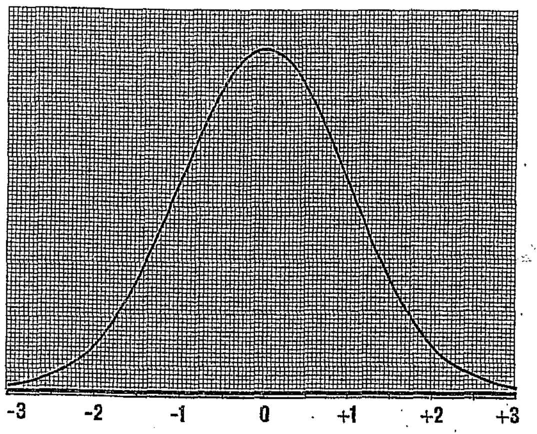
Jsou-li všichni havrani černí, není třeba vytvářet výběrové soubory a uvažovat o tom, kolik by mohlo být havranů šedých nebo světle modrých. Je-li pravděpodobnost, že havrani jsou černí, $p = 1$, jde o určitost.

Něco jiného je, zkoumám-li např. váhu sta, tisíce nebo ještě většího množství havranů. I když se přitom vyloučí mláďata, objeví se uvnitř zkoumaného souboru výkyvy. Podobně se navzájem odlišuje tělesná výška dospělých mužů, kvocient inteligence školních dětí, počet slov na plně potištěných stránkách

knihy, přesná délka sériově vyráběných věšáků na zeď, životnost elektrických žárovek, ba dokonce navzájem nezávislá měření jedné a téže vzdálenosti. Jistě jste si vzpomněli i na rozdělení příjmů našich 25 abiturientů a obyvatel Zbohatlíkova, avšak právě tato rozdělení musíme z našeho pojednání prozatím vyloučit. Naproti tomu se sem velmi dobře hodí úvahy, které jsme rozvíjeli v souvislosti s binomickým rozdělením a zákonem velkých čísel. Výsledky dvaceti hodů mincí stejně jako téměř každého druhu výběrových šetření vykazují totiž také charakteristické znaky tzv. normálního rozdělení, jímž se nyní budeme zabývat.

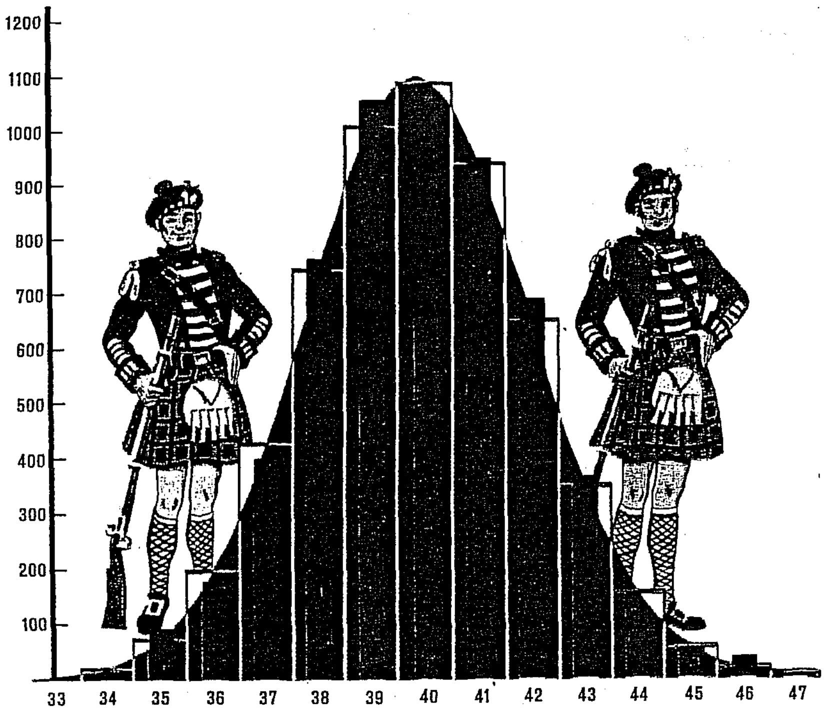
Může být sporné, zda označení *„normální rozdělení"* je zvoleno šťastně (kritika někdy uvádí, že slovo „normální" naznačuje jakési souhlasné hodnocení), nicméně ve všech světových jazycích se přesto mluví o *„normálním rozdělení"*, *„Normalverteilung"*, *„normal distribution"*, *„distribution normale"* — tzn. že statistická terminologie je bez tohoto slova nemyslitelná —, aniž by statistika proto napadlo, že jiná rozdělení než normální jsou „abnormální". Nebude také ani požadovat, aby se daná čísla a měřené hodnoty daly ve všech případech úzkostlivě uvést do souladu s ideálním obrazem normálních rozdělení. *Normální rozdělení* je jako každé jiné statistické rozdělení *především myšlenkovým modelem a početní pomůckou, tedy nikoliv exaktním přírodním zákonem, který by musel být naplněn s malichernou přesností.*

Jestliže se statistika vždy, kdykoli vystupuje jako induktivní statistika, snaží usuzovat o celku na základě vzorků a dílčích pozorování, a proto může poskytnout jen více nebo méně pravděpodobné odhady, bylo by dvojnásob absurdní, kdyby se předstírala přesnost, která se především ve skutečnosti v takovém stupni nikdy nevyskytuje a která je dále vyloučena samým nahodilým charakterem každého výběrového souboru.



Normální křivka nebo lépe jedna z normálních křivek vzhledem k tomu, že by ji bylo možno nakreslit strmější nebo mnohem plošší. Podstatný je jen vztah výše křivky v bodech, které vymezují směrodatnou odchylku. Body ležící na křivce ve vzdálenosti + nebo — 2 σ směrodatné odchylky se nacházejí v $^1/_8$ výšky, kterou má křivka normálního rozdělení ve vrcholu (nad průměrem 0).

Teorie a praxe však již prokázaly správ-

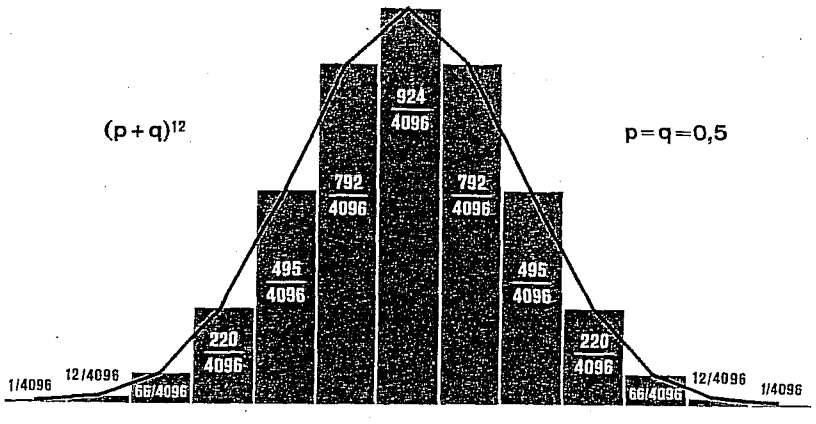


Obvod hrudi skotských vojáků podle Quételetovy statistiky. Shoda pozorovaných hodnot s normálním rozdělením je až zarážející (μ = 39,8 palce).

nost domněnky, že normální rozdělení platí pro téměř všechny výběry a pro velmi mnohá rozdělení podchytitelných souborů.

Normální rozdělení má především velmi příjemnou vlastnost, která lehce vysvětluje jeho oblibu: ať už jde o jakékoliv objekty, úkazy, měření nebo sčítání, *jsou vždy jednoznačně určeny střední hodnotou a rozptylem.* Pokud jde o „střední hodnotu", má se zpravidla, a nikoliv neprávem, na mysli aritmetický průměr $\bar{x}$ nebo μ, a to platí i v případě normálního rozdělení. Avšak stejnou hodnotu jako aritmetický průměr mají i modus a medián (nejčetnější hodnota a prostřední hodnota). Krátká úvaha hned prokáže, že normální rozdělení s největší četností „uprostřed" musí být symetrické. Jak taková normální křivka, která je grafickým znázorněním normálního rozdělení, ve skutečnosti vypadá, ukazuje obrázek na str. 74.

Základem normálního rozdělení je zkušenost, že bezpočetné znaky a hodnoty jsou rozloženy tak, že jeden vý-



Binomické rozdělení $(p + q)^{12}$ dovoluje již jasně poznat podobu normální křivky.

sledek měření nebo sčítání je „nejčetnější" a „na obě strany od něho" jsou výsledky ponenáhlu stále méně četné, až konečně vykazují jen ojedinělou extrémní hodnotu. Jeden z prvních příkladů takového normálního rozdělení podal Quételet na základě měření obvodu prsou 5738 skotských vojáků. Nejčetnější hodnota činila (zaokrouhleno na celé coule) 40 coulů, skoro stejnou četnost vykazovalo 39 coulů, 41 a 38 coulů se vyskytovalo již vzácněji, 42 a 37 byly ještě vzácnější a konečně zjištěných 33, resp. 48 coulů představovalo jen ojedinělé extrémní hodnoty.

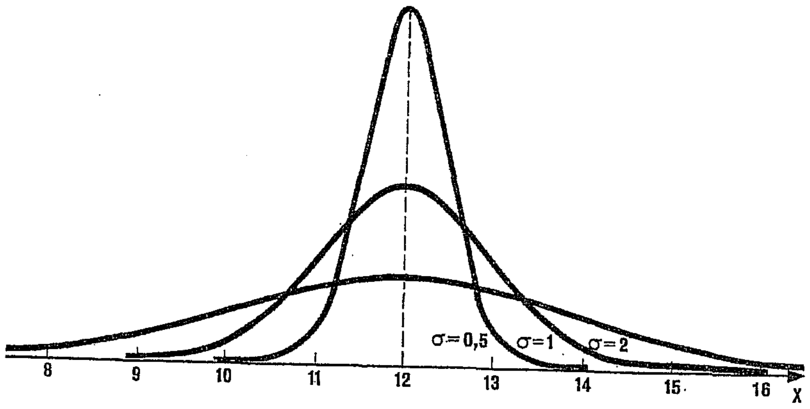
Podobné uspořádání vykazují i výsledky dalších četných měření, např. váha cigaret vyráběných cigaretovým automatem: v nejčetnějších případech vážily 1,18 až 1,20 g nebo 1,20 až 1,22 g, jen málokteré byly lehčí než 1,08 g nebo těžší než 1,32 g.

Podle rozsahu rozptylu i podle měřítka zvoleného pro grafické znázornění vzniká v idealizované formě plošší nebo strmější křivka.

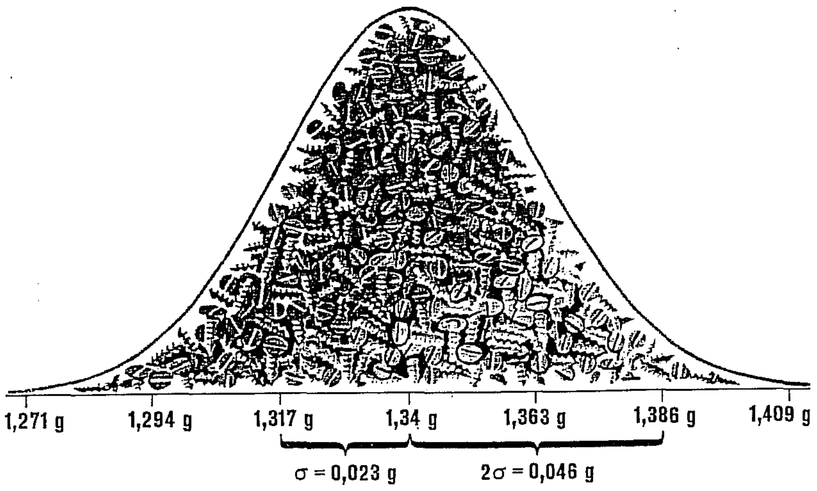
Mluví se o *„normální křivce"*, v němčině o „Glockenkurve" *(zvonovité křivce)* se zřetelem na středně vysokou křivku, která se také uvádí jako model pro normovanou normální křivku, ve Francii o *„courbe en chapeau de gendarme"* (křivku policejního klobouku) se zřetelem na plošší normální křivku. Jako vědecká označení se také používají názvy *„Gaussova křivka rozdělení chyb"* a *„de Moivrova stochastika"* s odvoláním na oba praotce zvonovité křivky.

Ať už se zvolí jakékoliv měřítko a ať je rozptyl k průměru v jakémkoliv poměru, normální křivka má vždy některé charakteristické znaky, z nichž uvedeme ty, které jsou pro praxi nejdůležitější. Jestliže pozorujeme plochu ležící pod „zvonem" jako soubor, *leží na obě strany od maxima* (střední hodnota, nejčetnější hodnota) *vždy přesně stejné části této plochy*, a to v úseku mezi $+\sigma$ a $-\sigma$ leží 68,26 %, tj. nepatrně více než $^2/_3$ celkové plochy, v úseku mezi $+2\sigma$ a -2σ skoro přesně 95 % a mezi $+3\sigma$ a -3σ již 99,7 % plochy.

Hodnoty za třemi směrodatnými odchylkami se proto berou v úvahu již jen velmi zřídka, ačkoliv normální křivka se teoreticky rozkládá od —∞ po +∞, tzn. v jisté míře od nekonečna do nekonečna. Tímto *postupem do nekonečna a plynulým průběhem* se normální křivka liší od binomického rozdělení. Přesto však mezi těmito oběma rozděleními je tak těsná souvislost, že *normální rozdělení* je možno v téměř všech prakticky důležitých případech pokládat za dostatečně přesné *vyjádření binomického rozdělení*, čímž si lze ušetřit svízelné početní operace, které jsme alespoň v náznaku poznali při našich úvahách o binomickém rozdělení. V historii normální křivky nelze zamlčet její původ z teorie hazardní hry. U její kolébky stál stařešina počtu pravděpodobnosti Abraham de Moivre.



Tři normální rozdělení kolem střední hodnoty $\mu = 12$. Má-li rozdělení silný rozptyl ($\sigma = 2$), je křivka plochá a rozložená; je-li rozptyl malý ($\sigma = 0{,}5$), je strmá a vysoká. Střední křivka vykazuje proporce „normovaného normálního rozdělení".



Průměrná váha určitého druhu šroubů je 1,34 g, směrodatná odchylka 0,023 g; protože jde o normální rozdělení, budou se velmi vzácně vyskytovat šrouby, které váží méně než 1,294 nebo více než 1,386 g.

3.32 Normované normální rozdělení

I když jsou normální křivky pravidelné, symetrické a stejnorodé, získávají velký praktický význam teprve dalším procesem *standardizace (normování)*. K tomu, abychom porozuměli procesu standardizace, musíme uvést ještě několik příkladů nestandardizovaného normálního rozdělení. Charakteristické chování křivky je vždy stejné: bod obratu křivky leží vždy ve vzdálenosti $+\sigma$ a $—\sigma$, tečna v bodu obratu vždy protíná souřadnici ve vzdálenosti $+2\sigma$ a $—2\sigma$, teoreticky je vždy křivka rozložena na obě strany donekonečna, avšak již při $+3\sigma$ a $—3\sigma$ se prakticky dotýká osy úseček (souřadnice x).

Nanášené měrné hodnoty jsou však nestejné. Váha šroubů se může někdy odchylovat o směrodatnou odchylku 0,023 g od průměru 1,34 g; jindy mohou psychometrická šetření skupiny studentů vykazovat rozptyl 36 ($\sigma = 6$) kolem kvocientu inteligence 112; tělesná váha, výsledky sklizně, stejně tak jako libovolný počet sčítaných a měřených výsledků mohou být rozděleny přibližně normálně kolem střední hodnoty se směrodatnou odchylkou, která může jednou činit půldruhého dne, jindy 0,85 kg, 3,2 cm, 0,8 ohmů, 35 marek nebo 3,5 mm.

Mají-li být všechna tato rozdílná měření a sčítání opravdu účelně zobrazena normálním rozdělením, je žádoucí, aby byl k dispozici *standardizovaný soubor nástrojů*, které umožňují odpověď na velmi rozličné otázky: „Kolik % šroubů je mimo toleranční meze ± 0,06 g?" — „Odchyluje se výrazně rozptyl inteligenčního kvocientu určité skupiny od rozptylu přibližně dvojnásobně velké skupiny, s níž byl proveden stejný test?" — „Vytvoříme-li výběrový soubor 50 žárovek, bude v něm s pravděpodobností větší než 68 % nejméně jeden zmetek?" — „Jak velká je pravděpodobnost, že v ruletě padne v průběhu nejbližšího tisíce sázek číslo 13 přesně třináctkrát?"

Převedení všech těchto rozmanitých měření, otázek a odpovědí na jediné schéma je možné jen tehdy, když je normální rozdělení *jednoznačně určeno* směrodatnou odchylkou a průměrem a když struktura normálního rozdělení *se nemění*, ať už jde o centimetry, hektolitry, ohmy anebo čísla v ruletě. Použijme pro objasnění daného problému např. šroubů: nejdříve máme průměrnou hodnotu 1,34 g a směrodatnou odchylku 0,023 g. První krok: vynecháme gramy a zůstává $\mu = 1{,}34$ a $\sigma = 0{,}023$. Během druhého kroku provedeme další abstrakci: průměr $\mu = 1{,}34$ je pro standardizované normální rozdělení právě průměrem a průměr standardizované normální křivky

SWOBODA Helmut. 1977. *Moderní statistika*. Praha, Svoboda.

se zásadně rovná nule. To není pouhá libovůle, nýbrž účelná konvence, protože od kterého jiného čísla než od nuly je možno tak lehce zjistit zrcadlově stejné odchylky nahoru a dolů? Odchylce „nahoru" (třeba +2) odpovídá stejná odchylka „dolů" (—2). Z těchto důvodů také tabulky standardizovaného normálního rozdělení, které jsou nepostradatelnou pomůckou při statistické práci, nerozlišují mezi kladnými a zápornými odchylkami. Přesto však je možno z nich vyčíst pravděpodobnost kteréhokoliv jevu a četnosti.

Jak to vypadá v praxi? Tak např. chceme zjistit, kolik šroubů se odchyluje v tom nebo onom směru o více než 0,06 g od průměrné váhy. Známe: $\sigma = 0{,}023$. Víme také, že rozdělení ve standardizovaném normálním rozdělení závisí již jen na směrodatné odchylce, protože jsme průměr posunuli na nulu. Zcela jednoduše proto porovnáme hledané 0,06 g se směrodatnou odchylkou a bez velké námahy vypočítáme, že 0,06 je rovno 2,6 směrodatné odchylky standardizovaného normálního rozdělení.

Danou skutečnost lze označit také jinak. Třeba takto: šrouby nemají být těžší než 1,40 g a lehčí než 1,28 g. Jaká je pravděpodobnost, že odchylky budou přesahovat uvedené váhy? Pak bychom měli podle vzorce pro standardizaci normálního rozdělení uvést hledané hodnoty do spojitosti se střední hodnotou:

$$\text{normovaná hodnota} = \frac{\text{hodnota minus aritmetický průměr}}{\text{směrodatná odchylka}}$$

Jestliže tuto „normovanou hodnotu" označíme z, může být vyjádření ještě kratší:

$$\boxed{z = \frac{x - \mu}{\sigma}}$$

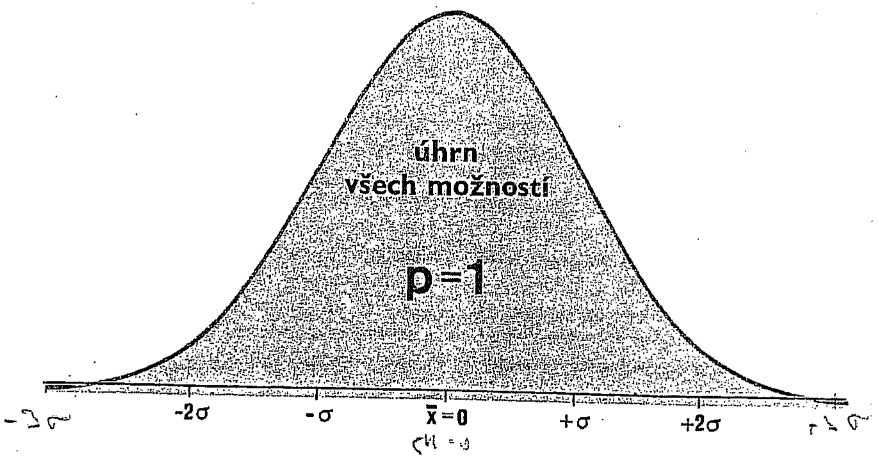
Dosadíme-li naše čísla, dostaneme:

$$\frac{1{,}40 - 1{,}34}{0{,}023} = 2{,}6, \text{ a stejně}$$

$$\frac{1{,}28 - 1{,}34}{0{,}023} = -2{,}6.$$

Záporné znaménko označuje odchylku pod střední hodnotu.

Nyní tedy víme, že $z = 2{,}6$. Ale co vlastně je z (nebo ať už tuto hodnotu nazveme jakkoliv, protože se vyskytuje pod různými názvy)? Chceme-li to tak vyjádřit, je to standardizovaná směrodatná odchylka. Stejně jako jsme zprvu přeměnili reálnou střední hodnotu $\mu = 1{,}34$ g na $\mu = 0$, tak i nyní jsme přeměnili reálnou směrodatnou odchylku $\sigma = 0{,}023$ na standardizovanou (normalizovanou) odchylkovou veličinu $z = 2{,}6$.



Standardizace normálního rozdělení: průměrná hodnota je vždy 0, odchylky se už neudávají v gramech, litrech nebo absolutních četnostech, nýbrž pouze v blíže nespecifikovaných směrodatných odchylkách.

NORUŠIS, Marija J. 1998. *Guide to Data Analysis*. Upper Saddle River, Prentice Hall.

10 The Normal Distribution

What is the normal distribution, and why is it important for data analysis?

- What does a normal distribution look like?
- What is a standard normal distribution?
- What is the Central Limit Theorem, and why is it important?

In Chapter 9, you learned how to evaluate a claim about the mean of a variable that has two possible values. Using the binomial test, you calculated the probabilities of getting various sample results when the probability of a success was assumed to be known. In this chapter, you'll learn how to test claims about the mean of a variable that has more than two values. You'll also learn about the normal distribution and the important role it plays in statistics.

▶ This chapter examines data on serum cholesterol levels from the *electric.sav* data file. In addition, some figures use simulated data sets included in the file *simul.sav*. The histograms and output shown can be obtained using the SPSS Graphs menu (see Appendix A) and the Descriptives procedure (see Chapter 4).

The Normal Distribution

You may have noticed that the shapes of the two stem-and-leaf plots in Chapter 9 are similar. They look like bells (on their sides). The same data are displayed as histograms in Figure 10.1 and Figure 10.2, where a bell-shaped distribution with the same mean and variance as the data is superimposed. You can see that most of the values are bunched in the center. The farther you move from the center, in either direction, the fewer the number of observations. The distributions are also more or less symmetric. That is, if you divide the distribution into two pieces at the peak, the two halves of the distribution are very similar in shape, but mirror images of each other. (The theoretical bell distribution is perfectly symmetric.)

You can obtain histograms using the Graphs menu, as described in Appendix A.

In the Histograms dialog box, select the variables cured10 and cured40.

Figure 10.1 Simulated experiments: sample size 10

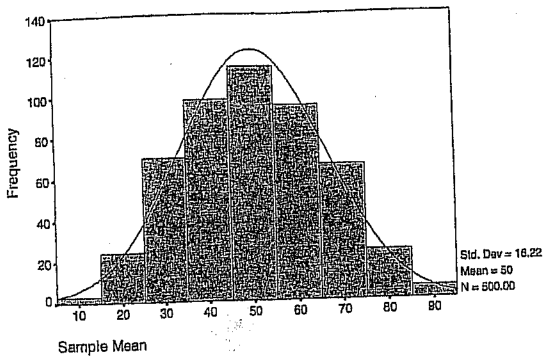
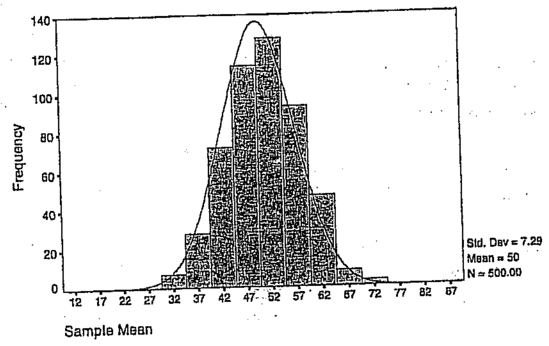


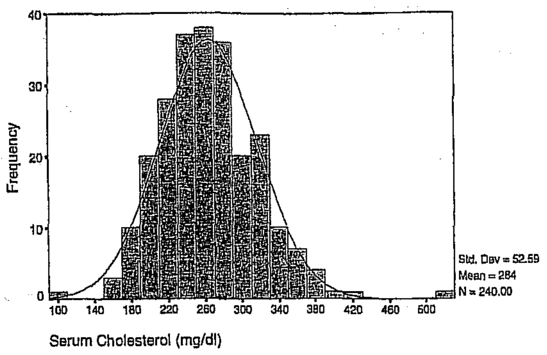
Figure 10.2 Simulated experiments: sample size 40



Many variables—such as blood pressure, weight, and scores on standardized tests—turn out to have distributions that are bell-shaped. For example, look at Figure 10.3, which is a histogram of cholesterol levels for a sample of 239 men enrolled in the Western Electric study (Paul et al., 1963). Note that the shape of the distribution is very similar to that in Figure 10.2. That's a pretty remarkable coincidence, since Figure 10.2 is a plot of many sample means from a distribution that has only two values (1=cured, 0=not cured), while Figure 10.3 is a plot of actual cholesterol values.

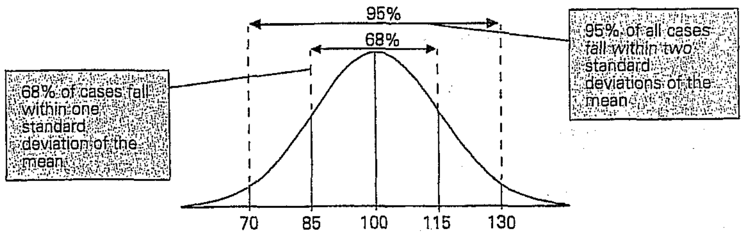
Figure 10.3 Histogram of cholesterol values

To obtain this histogram, open the electric.savdata file and select chol58 in the Histograms dialog box.



The bell distribution that is superimposed on Figure 10.1, Figure 10.2, and Figure 10.3 is called the **normal distribution**. A mathematical equation specifies exactly the distribution of values for a variable that has a normal distribution. Consider Figure 10.4, which is a picture of a normal distribution that has a mean of 100 and a standard deviation of 15. The center of the distribution is at the mean. The mean of a normal distribution has the same value as the most frequently occurring value (the mode), and as the median, the value that splits the distribution into two equal parts.

Figure 10.4 A normal distribution



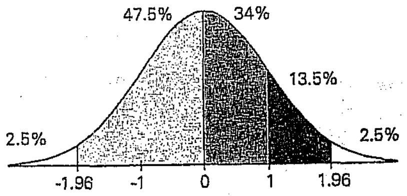
68% of cases fall within one standard deviation of the mean

95% of all cases fall within two standard deviations of the mean

If a variable has exactly a normal distribution, you can calculate the percentage of cases falling within any interval. All you have to know are the mean and the standard deviation. Suppose that scores on IQ tests are normally distributed, with a mean of 100 and a standard deviation of 15, as was once thought to be true. In a normal distribution, 68% of all values fall within one standard deviation of the mean, so you would expect 68% of the population to have IQ scores between 85 (one standard deviation below the mean) and 115 (one standard deviation above the mean). Similarly, 95% of the values in a normal distribution fall within two standard deviations of the mean, so you would expect 95% of the population to have IQ scores between 70 and 130.

Since a normal distribution can have any mean and standard deviation, the location of a case within the distribution is usually given by the number of standard deviations it is above or below the mean. (Recall from Chapter 4 that this is called a **standard score**, or **z score**.) A normal distribution in which all values are given as standard scores is called a **standard normal distribution**. A standard normal distribution has a mean of 0, and a standard deviation of 1. For example, a person with an IQ of 100 would have a standard score of 0, since 100 is the mean of the distribution. Similarly a person with an IQ of 115 would have a standard score of +1, since the score is one standard deviation (15 points) above the mean, while a person with an IQ of 70 would have a standard score of −2, since the score is two standard deviation units (30 points) below the mean.

Figure 10.5 The standard normal distribution



Some of the areas in a standard normal distribution are shown in Figure 10.5. Since the distribution is symmetric, half of the values are greater than 0, and half are less. Also, the area to the *right* of any given positive score is the same as the area to the *left* of the same negative score. For example, 16% of cases have standardized scores greater than +1, and 16% of cases have standardized scores less than −1. Appendix D gives areas of

the normal distribution for various standard scores. The exercises show you how to use SPSS to calculate areas in a normal distribution.

If you're more than two standard deviations from the mean on some characteristic, does that mean you're abnormal? Not necessarily. For example, pediatricians often evaluate a child's size by finding percentile values. They may tell the parents that their child is at the 2.5th percentile, or 97.5th percentile for height. (For a normal distribution, these percentiles correspond to standardized scores of −2 and +2.) The small or large percentile values don't necessarily indicate that something is wrong. Even if you took a group of healthy children and looked at their height distribution, some of them would be more than two standard deviations from the mean. Somebody has to fall into the tails of the normal distribution. This also leads to a convincing argument against grading on the curve. Even in a brilliant, hard-working class, some students will receive scores more than 2 standard deviations below the mean. Does that make their performance unacceptable? Not necessarily. ■ ■ ■

Samples from a Normal Distribution

If you look again at Figure 10.3, you'll see that the normal distribution that is superimposed on the cholesterol data doesn't fit the data values exactly. The observed data are not perfectly normal. Instead, the distribution of the data values can be described as approximately normal. That's not surprising. Even if you assume that cholesterol values have a perfect normal distribution in the population, you wouldn't expect a sample from this distribution to be exactly normal. You know that a sample is not a perfect picture of the population. You expect that samples from a normal population would appear to be more or less bell shaped, but it would be unrealistic to expect that every sample is exactly normal. In fact, even the population distribution of most variables is not exactly normal. Instead, it's usually the case that the normal distribution is a good approximation. Slight departures from the normal distribution have little effect on statistical analyses that assume that the distribution of data values is normal.

Means from a Normal Population

Since we've established that the normal distribution is a reasonable representation of the distribution of data values for many variables, we can use this information in testing statistical hypotheses about such variables. For example, suppose you want to test whether highly paid CEO's have average cholesterol levels which are different from the population as a whole. In 1991, *Forbes* sent out a survey to the 200 most highly compensated CEO's requesting their cholesterol levels. The 21 CEO's who responded had an average cholesterol of 193 mg/dL. Assume that, in the population, cholesterol levels are normally distributed with a mean of 205 and a standard deviation of 35. Based on this information, how would you determine if the CEO's differ from the rest of us not only in their net worth but in average cholesterol as well?

To answer this question, you need to know whether 193 is an unlikely sample value for the mean, when the true population value is 205. To arrive at this information, you'll follow the same procedure as you did in Chapter 9. However, instead of taking samples from a population in which only two values can occur, you'll take repeated samples from a normal population.

Figure 10.6 Distribution of 500 sample means

To obtain this histogram, open the simul.sav file and select the variable normal21 in the Histograms dialog box.

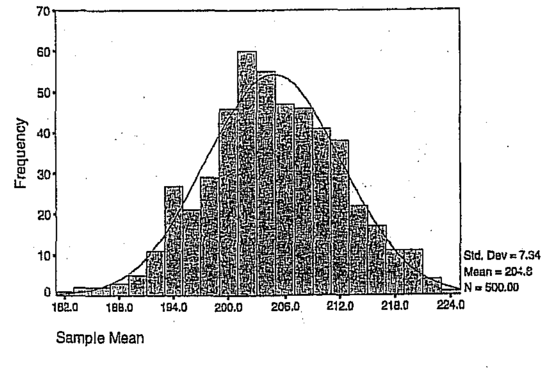


Figure 10.6 shows the distribution of 500 sample means from a normal distribution with a mean of 205 and a standard deviation of 35. Each mean is based on 21 cases. As you can see, the distribution of sample means is also approximately normal. That's always the case when you

NORUŠIS, Marija J. 1998. *Guide to Data Analysis*. Upper Saddle River, Prentice Hall.

calculate sample means for data from a normal population. The mean of the sample means is very close to 205, the population value. In fact, for the theoretical sampling distribution of the means, the value is exactly 205. (Remember, the theoretical distribution of sample means is mathematically derived and tells you precisely what the distribution of the sample means is for all possible samples of a particular size.) In Figure 10.6, the standard deviation of the means, also known as the standard error of the mean, is 7.34.

Standard Error of the Mean

You saw in Chapter 9 that the standard error of the mean tells you how much sample means from the same population vary. It depends on two things: how large a sample you take (that is, the number of cases used to compute the mean) and how much variability there is in the population. Means based on large numbers of cases vary less than means based on small numbers of cases. Means calculated from populations with little variability vary less than means calculated from populations with large variability.

If you know the population standard deviation (or variance) and the number of cases in the sample, you can calculate the standard error of the mean by dividing the standard deviation by the square root of the number of cases. In this example, the population standard deviation is 35 and the number of cases is 21, so the standard error of the mean is:

$$\frac{35}{\sqrt{21}} = 7.64 \qquad \textbf{Equation 10.1}$$

Note that the value we calculated based on the 500 samples with 21 hypothetical CEO's in each sample was not exactly 7.64, but very close. What we obtained was an *estimate* of the true value. That's because we did not take all possible samples from the population, but restricted our attention to 500.

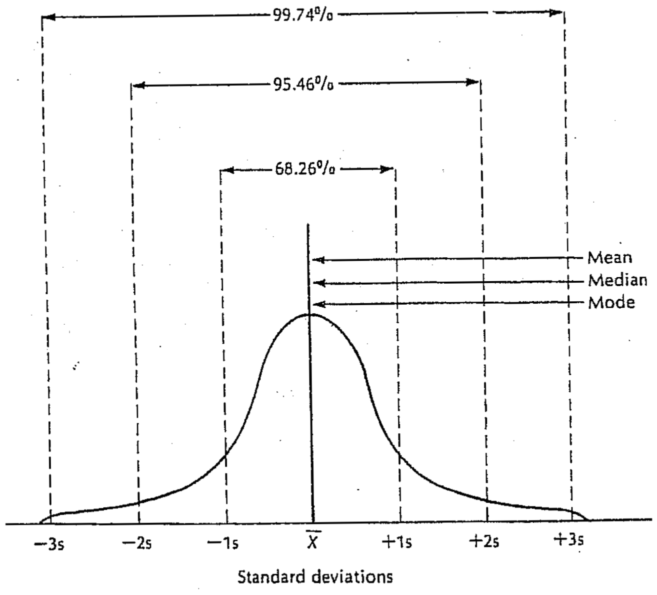
Summary

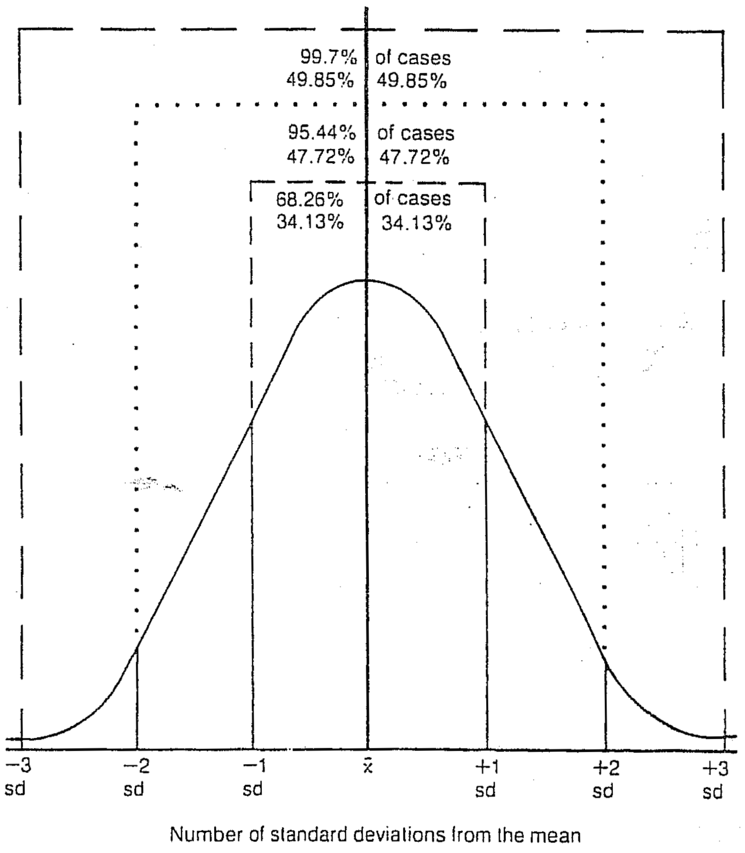
What is the normal distribution, and why is it important for data analysis?

- A normal distribution is bell shaped. It is a symmetric distribution in which the mean, median, and mode all coincide. In the population, many variables, such as height and weight, have distributions that are approximately normal.
- Although normal distributions can have different means and variances, the proportional distribution of the cases about the mean is always the same.
- A standard normal distribution has a mean of 0 and a standard deviation of 1.
- The Central Limit Theorem states that for samples of a sufficiently large size, the distribution of sample means is approximately normal. (That's why the normal distribution is so important for data analysis.)

WILLIAMSONN, J. B., David A. KARP, John R. DALPHIN, Paul S. GRAY. 1982. *The Research Craft. An Introduction to Social Research Methods*. Boston: Little, Brown and Company.

FIGURE 16.3 Percentages of Observations within Various Standard Deviation Units of the Mean for the Normal Curve





2.5. Testing whether a Distribution is Normal

2.5.1. Running the Analysis

It is all very well to look at histograms, but they tell us little about whether a distribution is close enough to normality to be useful. Looking at histograms is subjective and open to abuse (I can imagine researchers sitting looking at a completely distorted distribution and saying 'yep, well Bob, that looks normal to me', and Bob replying 'yep, sure does'). What is needed is an objective test to decide whether or not a distribution is normal. Fortunately, such tests exist: the Kolmogorov-Smirnov and Shapiro-Wilk tests. These tests compare the set of scores in the sample to a normally distributed set of scores with the same mean and standard deviation. If the test is non-significant ($p > 0.05$) it tells us that the distribution of the sample is not significantly different from a normal distribution (i.e. it is probably normal). If, however, the test is significant ($p < 0.05$) then the distribution in question is significantly *different from a normal distribution* (i.e. *it is non-normal*). These tests are great: in one easy procedure they tell us whether our scores are normally distributed (nice!).

The Kolmogorov-Smirnov (K-S from now on) test can be accessed through the *explore* command (**Analyze⇒Descriptive Statistics⇒Explore...**).[2] Figure 2.6 shows the dialog boxes for the *explore* command. First, enter any variables of interest in the box labelled *Dependent List* by highlighting them on the left-hand side and transferring them by clicking on ▸. For this example, just select the exam scores and numeracy scores. It is also possible to select a factor (or grouping variable) by which to split the output (so, if you select **uni** and transfer it to the box labelled *Factor List*, SPSS will produce exploratory analysis for each group—a bit like the *split file* command). If you click on Statistics... a dialog box appears, but the default option is fine (it will produce means, standard deviations and so on). The more interesting option for our purposes is accessed by clicking on Plots... . In this dialog box select the option ☑ Normality plots with tests, and this will produce both the K-S test and normal Q-Q plots for all of the variables selected. By default, SPSS will produce boxplots (split according to group if a factor has been specified) and stem and leaf diagrams as well. Click on Continue to return to the main dialog box and then click OK to run the analysis.

[2] This menu path would be **Statistics⇒Summarize⇒Explore...** in version 8.0 and earlier.

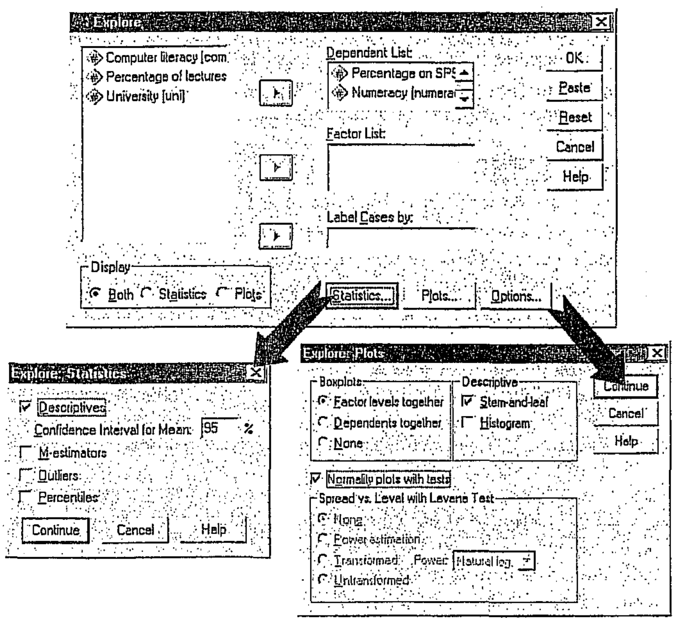
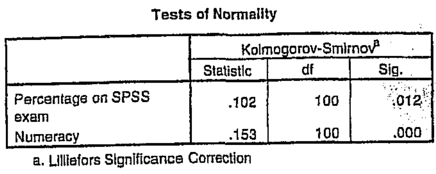


Figure 2.6: Dialog boxes for the *explore* command

2.5.2. Output

The first table produced by SPSS contains descriptive statistics (mean etc.) and should have the same values as the tables obtained using the frequencies procedure. The important table is that of the Kolmogorov-Smirnov test. This table includes the test statistic itself, the degrees of freedom (which should equal the sample size) and the significance value of this test. Remember that a significant value (*Sig.* less than 0.05) indicates a deviation from normality. For both numeracy and SPSS exam, the K-S test is highly significant, indicating that both distributions are not normal. This result is likely to reflect the bimodal distribution found for exam scores, and the positively skewed distribution observed in the numeracy scores. However, these tests confirm that these deviations were *significant*. This finding is important because the histograms tell us only that our sample distributions deviate from normal; they do not tell us whether this deviation is large enough to be important.



Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
Percentage on SPSS exam	.102	100	.012
Numeracy	.153	100	.000

a. Lilliefors Significance Correction

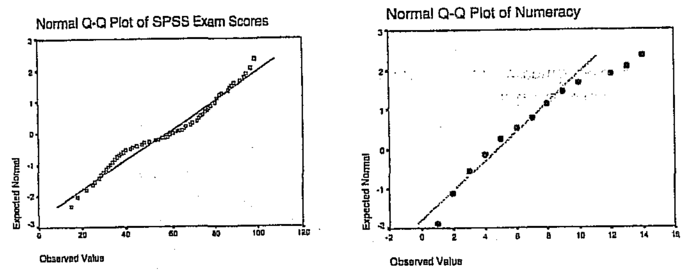


Figure 2.7: Normal Q-Q plots of numeracy and SPSS exam scores

SPSS also produces a normal Q-Q plot for any variables specified (see Figure 2.7). The normal Q-Q chart plots the values you would expect to get if the distribution were normal (expected values) against the values actually seen in the data set (observed values). The expected values are a straight diagonal line, whereas the observed values are plotted as individual points. If the data are normally distributed, then the observed values (the dots on the chart) should fall exactly along the straight line (meaning that the observed values are the same as you would expect to get from a normally distributed data set). Any deviation of the dots from the line represents a deviation from normality. So, if the Q-Q plot looks like a straight line with a wiggly snake wrapped around it then you have some deviation from normality! In both of the variables analysed we already know that the data are not normal, and these plots confirm this observation because the dots deviate substantially from the line. It is noteworthy that the deviation is greater for the numeracy scores, and this is consistent with the higher significance value of this variable on the Kolmogorov-Smirnov test. A deviation from normality such as this tells us that we cannot use a parametric test, because the assumption of normality is not tenable. In these circumstances we can sometimes turn to non-parametric tests as a means of testing the hypothesis of interest. In the next section we shall look at some of the non-parametric procedures available on SPSS.

BLACK, Thomas, R. 1999. *Doing Quantitative Research in the Social Sciences. An Integrated Approach to Research Design, Measurement and Statistics*, London: Sage.

THE NORMAL DISTRIBUTION OF PROBABILITIES AND Z-SCORES

We will now progress to the situation where discrete or continuous data have been collected and we are confident that we are dealing with a normally distributed trait. Even though our sample data may not provide a perfectly normal curve, they are close enough to assure us we can proceed, and we have some independent evidence that the trait is intrinsically normally distributed. The next question is: what can the data reveal?

It is possible to glean a certain amount of information when provided with the mean and standard deviation for a distribution. Such information will assume a normal distribution for the population and therefore will use its intrinsic shape as the basis for discussing probabilities of occurrence of events. For example, IQ tests are actually designed to have a mean of 100 and a standard deviation of 15. Referring to Figure 12.21, one would expect that about 68% of all persons taking an IQ test will have an IQ of between 85 and 115. One way of indicating an individual's performance is to state his or her position on the horizontal axis in terms of the percentage of examinees performing below this position, the *percentile group*. In other words, if John did better than 67% of the other people taking an exam, then John was in the 67th percentile group. If you have an IQ score of 115, one standard deviation above the mean, then your score is better than 84% of all persons taking that examination (50% below the mean plus 34% up to the first standard deviation). This also means that visually, 84% of the area under the curve is to the left, as shown in Figure 13.5.

It is possible to identify where in a distribution an individual score lies when the mean and standard deviation are known. It is relatively easy to convert a raw score into a number of standard deviations, called a *z-score*, which can be found in a table to see exactly in what percentile group that score falls:

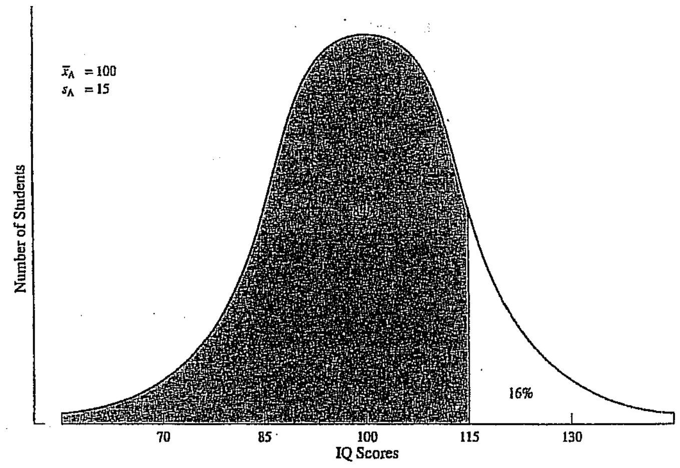


FIGURE 13.5
The 84th percentile group for IQ scores

$$z\text{-score} \equiv \frac{\text{raw score} - \text{mean}}{\text{standard deviation}}$$

Expressed in mathematical symbols,

$$z_i = \frac{x_i - \bar{x}}{S} \qquad (13.1)$$

where x_i is the individual score for person i, $\bar{x}$ is the mean of the distribution of scores and S is the standard deviation of the distribution from equation (12.4). For example, an IQ score of 92 would be:

$$z = \frac{92 - 100}{15} = \frac{-8}{15} = -0.53$$

or 0.53 standard deviations *below* the mean. Looking this up in Table B.1 in the Appendix B, reveals that the score corresponds to a percentage score of 20.19% below the mean. Subtracting this from the 50% total below the mean results in this score being in the 29.81 percentile. In other words, this person scored higher than 29.81% of the persons taking this test and 70.19% did better than this person. This simply tells how an individual with this score performed with respect to all the others. What decisions are made based upon such results is the domain of the researchers or other persons using these data. Now try Activity 13.4, where you are asked to find equivalent z-scores for raw scores.

> Activity 13.4
>
> Find the percentile group for IQ scores of 110, 98 and 120, using Table B.1 in Appendix B.
>
> *Answers (rounded values)*: $z = 0.67$ *thus* 75%, $z = -0.13$ *thus* 45%, $z = 1.33$ *thus* 91%.

The IQ score distribution is based upon population data, whereas in many situations one would be finding z-scores based upon an estimate of the population mean and standard deviation provided by sample data. The assumption is that the distribution will not be greatly different if the sample is truly representative. As noted earlier, in most situations, population data will simply not be available anyway.

LOETHER Herman J. and Donald G. McTAVISH. 1988. *Descriptive and Inferencial Statistics. An Introduction.* Boston, Allyn and Bacon.

15

The Logic of Hypothesis Testing

Suicide is obviously an individual act with psychological overtones. Emile Durkheim, the famous French sociologist, maintained, however, that the regularity and predictability of suicide rates over time could not be explained by psychological variables. He was convinced that suicide rates were explainable, rather, in terms of "social facts." He said,

> If, instead of seeing in them only separate occurrences, unrelated and to be separately studied, the suicides committed in a given society during a given period of time are taken as a whole, it appears that this total is not simply a sum of independent units, a collective total, but is itself a new fact *sui generis*, with its own unity, individuality, and consequently its own nature—a nature, furthermore, dominantly social. (Durkheim, tr. 1951:46.)

With the collective nature of the suicidal act in mind, Durkheim attributed the regularity of its rates for various populations to the social fact of group solidarity, or the lack thereof. He theorized that people lacking support from group solidarity are most vulnerable to suicide. His theory did not seek to explain the dynamics of individual suicides; it sought, rather, to explain suicide *rates* in terms of the differential vulnerability of various cohorts of people.

Although Durkheim's work was completed more than 90 years ago, it is considered a sociological classic because it illustrates the proper relationship between theory and data. In spite of soft spots in his methods, his general theoretical notions about suicide as a sociological phenomenon are still relatively sound.*

* Douglas (1967) critiques Durkheim's work and reviews further studies of suicide.

From his theory Durkheim derived some specific hypotheses about the relationship between social solidarity and suicide rates. Even though he never defined social solidarity in a rigorous manner, he did indicate various indices of it that could be observed and measured. For instance, he felt that social solidarity varied from one religious persuasion to another. He reasoned that social solidarity was highest among Jews, next among Catholics, and lowest among Protestants. He attributed these differences to differences in the degree to which the lives of individual members were dominated by their religions. Because Protestantism allowed for more free inquiry and placed more responsibility on the shoulders of the individual, it fostered less social solidarity than the other two. Accordingly, Durkheim hypothesized that suicide rates would be highest for Protestants and lower for Catholics and Jews. To test his hypothesis, he examined suicide statistics for various European countries. In general, he found that his hypothesis was supported by the data. For example, in the states of Germany suicide rates varied in direct proportion to the number of Protestants and in inverse proportion to the number of Catholics (Durkheim, tr. 1951:153). Moreover, those European countries that were predominantly Catholic (*e.g.*, Portugal, Spain, and Italy) had low suicide rates as compared with high rates for predominantly Protestant countries, and the rates for mixed Catholic–Protestant countries were intermediate (Durkheim, tr. 1951:152). When the rates of Protestant, Catholic, and Jewish groups were compared, Protestant rates were consistently higher than those of the other two, and Jewish rates were generally lower than those for Catholics (Durkheim, tr. 1951:155).*

Another index of social solidarity examined was marital status. Durkheim reasoned that the married enjoyed more social solidarity than the unmarried or widowed; thus, suicide rates would be lower for them. Again, the data tended to support his hypothesis. Rather consistently, for each age category, married persons had lower rates than unmarried or widowed persons (Durkheim, tr. 1951: 176–177). Furthermore, suicide rates for married persons with children were consistently lower than those for married persons without children (Durkheim, tr. 1951:186 ff.).

Durkheim went on to examine several other variables that he took to be indices of social solidarity, in each case comparing suicide rates from several different sources. The data generally supported his theory.†

Durkheim's basic approach was to develop theory to account for the regularity and predictability of suicide rates in Europe in the latter part of the 19th century. His major explanatory concept was social solidarity. He *selected* a number of measurable variables to be indices of social solidarity, *deduced* a number of specific hypotheses relating these indices to suicide rates, *gathered* all available data bearing on his hypotheses, and *examined* them to see whether they lent

* It should be noted that Durkheim's statistics relating suicide rates and religion were for countries or areas of countries and not for individuals. If predominantly Protestant countries have high suicide rates, it does not necessarily follow that those who are committing suicide are Protestants (*see also* Robinson, 1950.)

† Although the statistical techniques known to us were not available to Durkheim, on a raw level he duplicated the reasoning underlying modern statistics. Sir Francis Galton invented correlation earlier (Galton, 1886), but the technique was not generally known nor understood at the time that Durkheim conducted his study.

support to his predictions. He *concluded* that the data, on the whole, supported his hypotheses and theory. Therefore, he offered his theory as a fruitful explanation of suicide rates.

The point of this exercise in social science was to develop a theoretical explanation for a social phenomenon useful for predicting the same class of phenomena in the future. Durkheim did not *prove* his theory any more than any scientific theory is ever proven. He did demonstrate, however, that his findings were useful for predicting suicide rates. This is the way science uses theory. A theory is never proven; it is demonstrated to be useful or not useful. and is used or revised. If a competing theory is developed that predicts better, it is substituted for the previous theory and is used until it is replaced, in turn, by another theory that is an even better predictor.

The process of developing theoretical explanations of social behavior is what sociology is all about. Science is, after all, a continual interplay between theory and data. The examination of data in a systematic manner gives rise to theory; theory gives direction to the collection of new data; and new data either give additional support to the theory or contribute to its refutation. The game of science is concerned with the selection of the most useful theory from a number of competing ones. The *application* of scientific knowledge by practitioners and policymakers involves the practical use of the most fruitful theory currently in vogue.

When Durkheim developed his theory he had in mind an explanation that would transcend particular populations. He did not, however, use sampling techniques nor did he worry whether his data were representative of either general or special populations. The data he used were descriptive of specific geographic areas at specific points in time. He did examine sets of data descriptive of a number of *different* special populations. In effect, he replicated his hypothesis tests. A very important principle of science is that hypothesis tests should be repeated independently a number of times to gauge their soundness. The impressive aspect of Durkheim's work was the consistency with which these separate sets of data upheld his hypotheses.

15.1 STATISTICS AND HYPOTHESIS TESTING

When concepts that appear in sociological theories are not defined in measurable form, the hypotheses linking the concepts are not directly testable. The usual procedure in such cases is to specify measurable indices of the concepts, frame hypotheses relating the indices, and test these **"working hypotheses"** against empirical data.

Figure 15.1 illustrates the relationship between a general theoretical hypothesis and a working hypothesis using Durkheim's study of suicide as an example. Robert E. Clark conducted a study designed to test the general hypothesis that incidence of mental disorders varies with occupational status (Clark, 1949). As indices of mental disorders Clark used diagnostic categories assigned to patients in mental hospitals in the Chicago area. He looked separately at rates for alcoholic psychoses, senile psychoses, paresis, manic-depressive psychoses, and schizophrenia among patients whose occupations were known. He used two indices of occupational status: a measure of the prestige of the occupation on the North-Hatt scale,* and the median income for the occupation in the Chicago area at the time of the study.

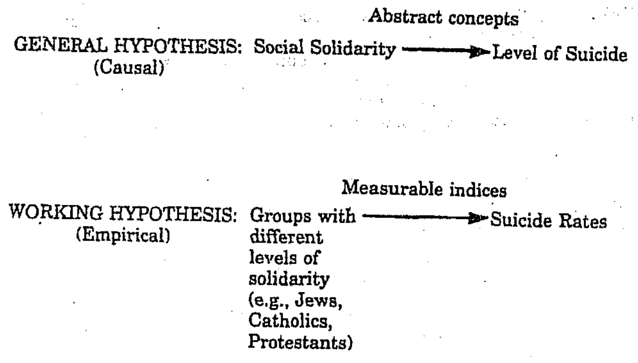

Abstract concepts
GENERAL HYPOTHESIS: Social Solidarity → Level of Suicide
(Causal)
Measurable indices
WORKING HYPOTHESIS: Groups with different levels of solidarity (e.g., Jews, Catholics, Protestants) → Suicide Rates
(Empirical)

FIGURE 15.1 Relationship between a General Theoretical Hypothesis and a Working Hypothesis: Durkheim's Theory of Suicide

From the one general hypothesis Clark framed several working hypotheses, relating indices of mental disorders to indices of occupational status. When he tested his working hypotheses against the data, he found that his general hypothesis had to be qualified. Alcoholic psychoses, senile psychoses, paresis, and schizophrenia were inversely related to occupational status (the higher the occupational status, the lower the incidence of the disorder), but manic-depressive psychoses were unrelated (Clark, 1949:440). Although Clark's substantive findings are interesting in themselves, we are concerned primarily with the procedures used to test the general hypothesis. Clark selected indices of his theoretical concepts and recast his general hypothesis in terms of these indices, thus deriving working hypotheses; then he subjected his working hypotheses to empirical tests and, on the basis of these tests, drew conclusions about the general hypothesis.

When population data are available, the indices serve as parameters; and decisions about hypotheses can be made simply by examining the parameters (no test of significance is needed). When Durkheim compared suicide rates in Catholic Bavaria with those in Protestant Prussia he merely had to take note of the fact that the Prussian rates were higher. Since these rates were parameters for his populations, the differences between them were actual differences (assuming that the data were error free). Therefore, descriptive statistics allow for direct tests of hypotheses without further complications.

Unfortunately, population data are not usually available. We are faced with the necessity of examining sample data and making generalizations about the population. The procedure for testing hypotheses with sample data is as follows:

1. *Specify indices of the concepts included in the general hypothesis.*
2. *Derive working hypotheses that link the indices.*
3. *Draw a sample of data from the population to which the hypotheses apply.*
4. *Use statistical techniques to analyze the sample data.*
5. On the basis of that analysis, *decide whether the data support the working hypotheses.*
6. *Decide whether the general hypothesis fruitfully describes the population.*

When hypotheses are tested using sample data, the complication introduced is that parameters must be estimated, since they cannot be examined directly because they are not available to the researcher. For example, if Durkheim's comparisons of suicide rates in Catholic Bavaria and Protestant Prussia had been based on sample data rather than population data, he would have been faced with the necessity of estimating suicide rates from his sample data, and then deciding whether the estimated parameters actually differed.

We found in Chapter 14 that estimating parameters involves the use of probability theory applied to sampling distributions. As you will see later. there is a close relationship between interval estimates of parameters and statistical tests of hypotheses. The difference between them is a difference in orientation rather than kind.

* For a description of the North-Hatt Scale see Reiss (1961).

BOX 15.1 Sampling Distribution

If the concept of the sampling distribution is not yet quite clear to you, perhaps you should go back and review Chapter 13, particularly the units under Section 13.2. An understanding of the concept of sampling distribution is essential to the discussion that follows.

15.1a The Statistical Hypothesis

When we use sample data to test hypotheses, it is necessary to introduce a third type of hypothesis—the **statistical hypothesis**. *We start with a general hypothesis, which we translate into working hypotheses, and from our working hypotheses we derive statistical hypotheses.* Statistical hypotheses make statements about population parameters, but are tested by examination of statistics computed from sample data. As a result of the outcomes of tests of statistical hypotheses, we decide what conclusions about the working hypotheses are warranted and, in turn, these decisions help us to make decisions about the general hypothesis.

Again, we will rely heavily on sampling distributions to help us make decisions. The questions we will ask, however, will be somewhat different from those raised in estimating parameters. We will examine the following questions:

Is it reasonable to conclude that the statistic we have computed from the sample is an estimate of a specific, given parameter?

Are the statistics we have computed reasonably seen as separate sample estimates of a common parameter or do they estimate different, distinct parameters?

In each case primary concern will be with making decisions about hypotheses that refer to parameters or to relationships between parameters. In classical hypothesis testing we must choose between two competing hypotheses.

15.2 TESTING STATISTICAL HYPOTHESES

Thus far our discussion of hypothesis testing has been fairly abstract. It might be helpful at this point to take a concrete example, run through the process involved in testing a hypothesis, then analyze the procedure involved. In the process of presenting the example we will introduce the concepts that play an integral part in the testing of statistical hypotheses.

One characteristic that particularly distinguishes the developing countries of the world from the others is the rate at which infants and young children die. Infant and childhood diseases that have long since ceased to be serious killers in the industrialized countries still take a terrible toll of babies and small children in developing countries. According to World Health Organization estimates, 3,450,000 children die every year of diseases that are preventable through vaccination (United Nations Children's Fund, 1985). Measles alone is estimated to kill 2 million annually. Is the distinction between the developing and the developed countries merely that vaccinations are more common in the latter than in the former? According to McKeown (1976), malnutrition is an important factor in the whole equation. People who are malnourished are more vulnerable to infection, and thus fall victim to diseases that, in other circumstances, are much less virulent (1976:35).

One might assume that malnutrition would not be a common characteristic of countries that are largely rural because the inhabitants could subsist on foodstuffs they grew themselves. According to Bogue (1969:46), however, predominantly rural settlement patterns are characteristic of developing countries and it is in the developing countries that death rates are high.

To understand better this whole process whereby underdevelopment, malnutrition, infectious disease, and high infant death rates are linked, an informative preliminary step would be to establish the nature of the relationship between rural-urban settlement and the level of nutrition existing nationally. The Food and Agriculture Organization of the United Nations (U.N.) has collected international data on daily per capita calorie supply as a percentage of the requirement necessary for satisfactory nutrition (United Nations Children's Fund, 1985:134–35). These data may be used as a measure of the level of nutrition available for each country covered.

Consistent with what has been said thus far about underdevelopment, nutrition, disease, and death rates, **our working hypothesis will be that countries that are predominantly rural will be less likely to provide the required per capita daily calorie intake than will the countries of the world in general.** In order to test this hypothesis we used the U.N. data to divide the countries of the world into two categories: those that meet or exceed the percentage requirement of daily calorie intake and those that do not. Using this distinction, we found that 65% of the countries of the world met or exceeded the daily percentage requirement.

Furthermore, we singled out the countries of the world in which more than 50% of their populations were rural and drew a simple random sample of 30 of those countries. When these countries were examined to determine whether their per capita daily calorie intake met or exceeded the percentage requirement it was found that 9 of the countries did and 21 did not. The research question that we wish to answer is whether this distribution for the 30 predominantly rural countries provides evidence in support of our working hypothesis.

15.2.1 The Null Hypothesis

Framing a statistical hypothesis in a positive manner, we would come up with some such hypothesis as the following: **Countries that are more than 50% rural are significantly less likely to meet or exceed the daily per capita calorie intake requirement than are the countries of the world in general.** The kind of sample data we would consider as evidence of significance would have to be stated explicitly so that we could test the hypothesis.

Since statistical inference is based upon probability theory, tests of statistical hypotheses are probabilistic rather than absolute. It is not possible to prove or disprove statistical hypotheses in an absolute sense. The best that can be achieved is an estimate of their truth or falsity.

It so happens that the rejection of a statistical hypothesis is much more clear-cut than is its acceptance. Therefore, the usual procedure is to frame a statistical hypothesis contrary to that which we are hoping to prove. Such a hypothesis is known as a **null hypothesis.** If sample data warrant rejection of the null hypothesis, that is regarded as evidence for its alternatives—those hypotheses our theory predicted and those we proposed as explanations in the first place. The null hypothesis gets its name from the fact that it is the hypothesis to be nullified by statistical test.

The advantage of using the null hypothesis is that it serves as a basis for selecting a specific sampling distribution, that is, the sampling distribution that would be found if the null hypothesis were, in fact, true. This sampling distribution is then used to determine whether sample data warrant rejection of the null hypothesis in favor of some set of alternatives to it.

Since we wish to seek evidence in support of the contention that the rural countries of the world are less likely to meet daily calorie intake requirements than are countries in general, we wish to show that significantly fewer than 0.65 of the rural countries meet or exceed those requirements. We use the 0.65 because that is the proportion of *all* of the countries of the world that meet or exceed the requirements. It is, therefore, the parameter of interest to us in generating the sampling distribution necessary to test our null hypothesis. The null hypothesis would be formulated as follows: **The proportion of rural countries that meet or exceed the daily per capita calorie intake requirement will not differ significantly from 0.65.** If this null hypothesis were true, we would have a sampling distribution of proportions with an expected value of 0.65, which is the value for countries in which the daily per capita calorie intake met or exceeded the requirement.

The procedure we follow in testing the null hypothesis is: (a) assume that it is true, (b) generate a sampling distribution from the null hypothesis, (c) draw a random sample, (d) collect the relevant sample data, (e) compute the relevant statistic, and (f) decide whether it is reasonable to assume that the statistic came from the given sampling distribution. If the probability that the statistic came from the given sampling distribution is as small as or smaller than some predetermined level, we reject the null hypothesis in favor of its alternatives. If the probability is not as small as or smaller than the predetermined level, we fail to reject the null hypothesis.

Notice that we *fail to reject* the null hypothesis rather than accept it. If we accepted the null hypothesis, we would be saying, in effect, that it is true. However, if we fail to find reason to reject the null hypothesis, it does not necessarily follow that it is true. We are saying, rather, that the data we collected did not provide us with sufficient basis for concluding that the null hypothesis is false. Perhaps, our data collection was merely inadequate.

By the same reasoning, if we *do* reject the null hypothesis, it does not follow that we accept its alternatives. All we imply by rejecting the null hypothesis is that some set of alternatives to it is more probable than the null hypothesis itself.

Note, also, that both the null hypothesis and its alternatives apply not to sample data but to population data—not to statistics but to parameters. We test the null hypothesis with sample data and generalize from statistics to parameters.

NULL HYPOTHESES

The rather convoluted thinking of null hypotheses is necessary if we are going to set the scene for testing hypotheses using statistical tools. As noted in Chapter 1, theories survive and gain support as a result of not being disproved, rather than being proven conclusively. For sound theories, this does not imply a ticking bomb waiting to explode in the form of some researcher in the future proving it wrong. What it does suggest is that researchers are usually trying out components of a theory in different situations or with different groups; they are looking for the *limits* of applicability or refinements in detail. Hypotheses, as described above, express anticipated outcomes as predicted by a given theory or the expected consequence of an application of principles to a situation, stated in more specific terms than those of a general research question.

When it comes to testing hypotheses, all that statistics can tell us is whether the outcomes we ultimately see could have happened due to some causal relationship *or* simply by chance alone. In other words, the effect has to be big enough, whether it is the difference in average scores on some performance task for two groups, or the size of a correlation coefficient. The null hypothesis simply states that 'no significant difference' is expected between what we obtain and what would happen by chance alone. If the difference observed is greater than some minimum, then it is considered significant and whatever has happened (probably) did not occur by chance alone. It is still up to the researcher to prove through sound design and data collection that nothing could have caused the observed effect other than what is described in the hypothesis.

So the next stage in refining our statement of hypotheses would be to try to express them as null hypotheses related to the data that will be collected. As a consequence of a given study, several types of null hypothesis could be generated – for example, describing differences in scores or frequencies of events between the sample and the population (normative), *or* between two groups or among three or more groups – i.e., they actually belong to the same population, not to separate populations (experimental, quasi-experimental or ex post facto). The statements simply anticipate that any difference(s) will be too small to be attributable to anything but chance.

Alternatively, if one were carrying out a correlational study, the null hypothesis of 'no significant correlation' anticipates correlations that will be so small that they could have happened by chance alone. To illustrate this, the hypotheses of Table 2.2 above are provided in Table 2.3 with corresponding possible null hypotheses.

TABLE 2.3
Hypotheses from Table 2.2 and potential corresponding null hypotheses

Hypotheses	Null hypotheses
A random sample of assembly-line workers in factories in Birmingham will be found to suffer a greater frequency of sleep interruptions, and a longer amount of time awake after going to bed, than the population as a whole.	(Both of the hypotheses assume that population data exist.) There will be no significant difference between the mean number of times per night that assembly-line workers in Birmingham awaken and the mean for the population of employed adults as a whole, or between the mean number of minutes that these workers are awake per night and that for the *population of employed* adults.
One of three counselling approaches, A, B or C, will produce a greater reduction in frequency of return to drinking among alcoholics.	There will be no significant difference. frequencies of 'dry' and return drinkers across three equivalent sets of alcoholics participating in the three counselling approaches, A, B, C.
It is expected that there will be a negative correlation between social class and drug use, and a negative correlation between educational achievement and drug use for a representative selection of 18–24-year-olds.	There will be no significant correlation between social class and frequency of drug use, or between educational achievement and frequency of drug use for a random selection of 18–24-year-olds (i.e., any correlation will not differ from that which could be expected by chance alone).
For a sample of identical twin boys who are the sons of alcoholic fathers *and* fostered or adopted from infancy separately from each other, one to a family with at least one alcoholic parent, one group will show a greater tendency towards alcoholism than the other.	There will be no significant difference in frequency of alcoholism between groups of separated twins, all sons of alcoholics, when one twin goes to a family with at least one alcoholic parent and the other goes to a family with no alcoholic parents.
In a given hospital, patients on 24-hour prescriptions will be expected to feel more rested if they are awakened for medicines at times that follow REM rather than just at equal time intervals.	There will be no significant difference in the perception of feeling rested, as measured by the Bloggs Restedness Scale completed by patients, between two groups: those whose medication was administered at regular time intervals and those whose medication was administered at times close to times prescribed but following a period of REM.

The process of specifying a null hypothesis is one that focuses the attention on what will happen next, stating the implications of the proposed relationship among variables in terms that can be resolved by statistical instruments (see Figure 2.14). At this stage, it is sometimes possible to identify potential difficulties in carrying out the research. For example, where are we going to find the sample of twins implied by the fourth proposal in Table 2.3? Some of the more interesting questions generate very difficult scenarios for resolving them, compelling researchers to rethink the hypotheses resulting from a question. Obviously, it is better to consider such issues early in the research process before too much is invested in an impossible task.

Testing the null hypothesis

For normally distributed traits, those that produce sample means out in either of the tails of a distribution of sampling means are highly unlikely. Social science researchers commonly accept that events which occur less frequently than 5% of the time are unlikely to have occurred by chance alone and consequently are considered statistically significant. To apply this to a normal distribution would mean that the 5% must be divided between the top and the bottom tails of the distribution, with 2.5% for each (there are occasions when all 5% would occur in one tail, but that is the exception, to be discussed later). Consulting Table B.1 in Appendix B, the top 2.5% is from 47.5% onward, or (interpolating) 1.96 standard deviations (SEMs) or more from the mean. The two ranges of sample means that would be considered *statistically significant*, and result in the rejection of the null hypothesis since they probably did not occur as part of the natural chance variation in the means, are shown shaded in Figure 13.8.

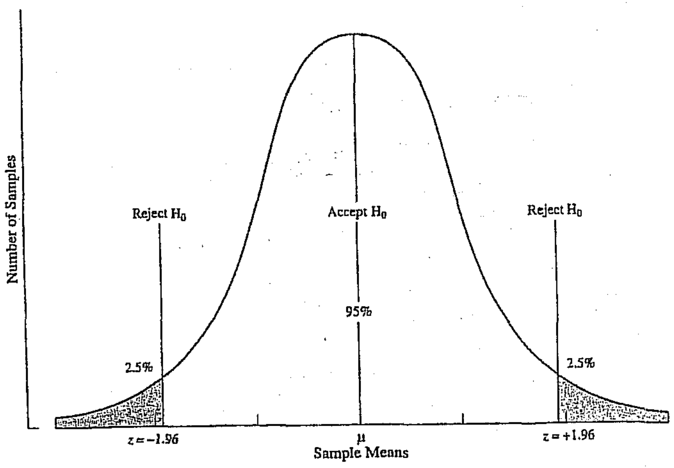


FIGURE 13.8
Normal distribution of sample means with 5% significance levels, where μ is the population mean.

Thus for the situation above involving the mean IQ of the sample of 11-year-olds, the null hypothesis and the statement of expected outcomes need an addition:

> . . . and, is the probability that the difference between the sample mean and the population mean would occur naturally more or less than 5% (the chosen level of significance that will be used as the test criteria)?

The cut-off point of 1.96 standard deviations (SEMs) would correspond to 1.96 × 2.5 = 4.9 points above or below the mean. Thus a sample mean IQ of less than 95.1 or greater than 104.9 would be considered significant and the sample not representative of the population. Therefore, in the example, the group with a mean IQ of 106 would be considered statistically significant and the group not typical, and it is unlikely that they are a representative sample of the whole population, for IQ.

Some researchers present results that are supported by an even lower level of probability, usually designated by the Greek letter α, to support their argument, such as 1% ($\alpha = 0.01$), 0.5% ($\alpha = 0.005$), or even 0.1% ($\alpha = 0.001$). Two problems arise with such a practice. First, for the test to be legitimate, one school of thought says the level of significance should be set *before* the test (or even the study) is conducted. Remember that the hypothesis is a statement of expectation, one that should include what will be expected in terms of statistical outcome. It is not fair to write the rules after the game has begun. Second, there is a feeling that a lower significance level than 5% ($p < 0.05$), such as 1% ($p < 0.01$), provides greater support for the results. In other words, if the probability of the relationship existing is only 1 in 100, that must be a stronger statement than if it were only 1 in 20. This supposition will be challenged in Chapter 14 when the concept of the power of a statistical test is introduced.

ŘEHÁK, Jan. 1971. „Poznámky k analýze sociologických dat." Sociologický časopis, 27 (4), str. 425-433.

Poznámky k analýze sociologických dat

JAN ŘEHÁK
Ústav pro filosofii a sociologii ČSAV, Praha

Poznámky shrnuté v této stati nepředstavují úplnou inventarizaci úloh a problémů, s kterými se v běžné praxi analýzy sociologických dat setkáváme; týkají se několika vybraných aspektů statistické analýzy a některých aspektů, které se statistickou analýzou a statistickou inferencí úzce souvisí.

Jde hlavně o to, ukázat na omezení aplikačních možností statistických metod v sociologické praxi. Tato omezení vyplývají jednak z principů metod samých, jednak z kvality dat, s nimiž pracujeme. Statistika je vědní disciplína, která má svou vlastní teorii a metodologii a z těch vyplývají její procedury a přístupy. Teorii samu lze studovat pouze ze speciálních učebnic a se značnými matematickými znalostmi. Určitý pohled na přístupy ke statistické teorii a alespoň letmé a velmi stručné seznámení se s principy může však snad přispět k fundovanějšímu aplikačnímu postřehu.

Statistika stojí v postavení metody analýzy dat; kvalifikovaná aplikace nutně vyžaduje znalosti teorie statistiky na jedné straně a teoretické znalosti aplikační vědy na straně druhé a proto je taková aplikace většinou obtížným problémem. Kromě toho jsou tu stálé vtíravé otázky o kvalitě dat, s nimiž pracujeme. Proto závěry z dat se vytvářejí součinností inference metodologické, sociologické a statistické. Už tato jednota ukazuje, že analýza dat není vůbec jen rutinní záležitostí, ale součástí tvořivé inference vědecké.

1. Typy statistických závěrů

Statistické závěry činíme o *základním souboru, populaci,* tj. o definované množině statistických jednotek (statistických objektů), která je předmětem našeho zájmu. Základní soubor může být dvojího typu:

a) jednak je to přesně vymezený *konečný* soubor jednotek, který je jednoznačně určitelný například tím, že pořídíme jeho seznam (teoreticky je možno vytvořit seznam všech jednotek konečného souboru vždycky, prakticky to ale nebývá možné u větších souborů);

b) jednak je to soubor určený svými vlastnostmi, ale neomezený na daný konečný počet jednotek; je to soubor, jehož velikost je neznáma a někdy ani známa být nemůže — takový základní soubor můžeme nazvat *otevřeným,* respektive hypotetickým nebo *neurčitým.*

Příkladem souboru a) může být soubor všech školních dětí v určitém obvodu Prahy, příkladem b) může být soubor žáků experimentální školy v Bratislavě a všech žáků, kteří se v podobných podmínkách kdy ocitali, ocitají nebo budou ocitat v budoucnu; soubor je tu definován teoreticky a není možno jej určit jednoznačně seznamem, protože nevíme kolik dětí a které děti se budou učit za stejných podmínek v budoucnu.

Statistické údaje však zřídka získáváme od celého základního souboru. Většinou jsou data k dispozici pouze od jeho vybrané části: každou takovou část základního souboru nazveme *výběrovým souborem, respektive výběrem.* Tak výběrovým souborem v případě a) je například soubor dětí ze tří vybraných pražských tříd; v případě b) je výběrem ona bratislavská experimentální třída.

Data, která získáváme, mohou být klasifikována také z hlediska jejich proměnlivosti u statistické jednotky. Některé znaky jsou *konstantní,* nemění se podle nálady, okamžitého stavu respondenta, nepodléhají chybám při zjišťování, nemění se v čase. Jiné znaky (a těch je většina) jsou *proměnlivé.* Můžeme je považovat za náhodné proměnné, neboť jejich zjištění je podmíněno spoustou nejrůznějších faktorů — subjektivní pocity, chyba při počítání, nepozornost, špatná interpretace otázky či výběr nežádoucího obsahového prvku otázky. Tato data měříme s chybou. Zatímco pohlaví je neproměnné, věk v průběhu

krátkého časového období se nemění, zaměstnání se nemění apod. Otázka „Jak se cítíte spokojen v zaměstnání?" bude zodpovězena v závislosti na mnoha okolnostech a odpověď se může měnit den ze dne či někdy z hodiny na hodinu, obzvláště tam, kde odpovědi jsou sémanticky neurčité: velmi spokojen, spokojen atd.

Při typologii situací statistických závěrů můžeme vzít v úvahu obě hlediska současně a naznačit je v tabulce 1.

Tabulka 1.

	Data mají konstantní charakter	Data jsou realizacemi náhodných proměnných — měření s chybou
Data se získávají z celého základního souboru (ZS)	1. není problém zobecnění, úlohy se redukují na vhodné vytváření přehledných měr vlastností souborů	2. eliminujeme chyby měření a náhodnost při získávání dat
Data jsou získávána pouze z výběrového souboru (VS)	3. úlohou je tu zobecnit data z VS na ZS a vyjádřit přesnost zobecnění	4. současně 1. a 3.: tato situace je nejobtížnější a většinou se redukuje na 2. nebo 3.

Otevřený základní soubor je věcí modelu a přijaté konvence. Výběr z takového základního souboru (například experimentální třídu) můžeme považovat za základní soubor s tím, že nás nezajímá jen současný stav, ale proces, resp. výsledek procesu, který vede k současnému stavu. Dostáváme se tím z pole 3 do pole 2. Podle toho, co bylo řečeno o základních souborech, musíme rozlišovat v poli 3 dvě situace — zobecnění na konečný (3a) a na neurčitý (3b) základní soubor. V každém případě pak musíme k úloze přistupovat jinak, tj. tvořit jiné statistické, resp. pravděpodobnostní modely.

Většinou se při analýze dat dostáváme do pole 2 resp. do pole 4 a po redukci do pole 2. To je proto, že na složité otázky sociologického výzkumu odpovídá respondent většinou s možností chyby a subjektivního hodnocení obsahu kategorií odpovědi. Nebude třeba snad argumentovat pro to, že poznání možného zdroje nepřesných odpovědí a informací je velice důležité pro konečnou interpretaci. Některé pomocné „klasifikace" znaků uvádím v dalším paragrafu. (Slovo klasifikace je v uvozovkách, protože jde jen o pomocná třídění, která jsou spíše výhodná než nutná.)

2. Některé pohledy na znaky z hlediska chyb zjišťování

Většina znaků tedy podléhá různým chybám při zjišťování jejich hodnot pro jednotlivce nebo jiné statistické jednotky. Pomíjím zde problémy validity a reliability jako takové a uvádím typy znaků z hlediska chyb, které lze a priori předpokládat a tím i vhodnou metodou měření (či zjišťování) potlačovat na nejmenší míru. Přitom nemám na mysli chyby náhodné, tj. náhodné odchylky od skutečné hodnoty znaku; ty jsou relativně málo nebezpečné, neboť jsou identifikovatelné. (Příklad takové chyby je zjišťování skóre I.Q. testu: při každém zjišťování I.Q. skóre se dá předpokládat, že respondent bude mít poněkud jiný výsledek, přičemž nelze identifikovat žádný faktor, který tuto variabilitu způsobuje.)

Kritéria typů:

a) *Formální vztahy hodnot znaku:* zde rozeznáváme znaky kardinální, ordinální a nominální. Informace čerpaná z těchto vztahů mezi hodnotami znaku je podstatná pro vytváření nejrůznějších populačních měr. Obzvlášť obtížné je zpracování ordinálních znaků, tj. znaků, jejichž hodnoty jsou uspořádané kategorie (obvyklé „kvantifikace" pořadovým číslem odpovědi v uspořádání nejsou považovány vždy za vyhovující); to plyne jednak z faktu, že není známa vzdálenost kategorií na předpokládané škále a dále, že odpovědi bývají často sémanticky neurčité a subjektivně interpretovatelné respondentem.

b) *Přímé zjišťování* × *konstrukce znaků* (složené znaky, škály). Je třeba vyjádřit

ŘEHÁK, Jan. 1971. „Poznámky k analýze sociologických dat." Sociologický časopis, 27 (4), str. 425-433.

vytvářet tak, že jej rozprostřeme symetricky kolem bodového odhadu $\bar{p}$:

$$\bar{p} - d \qquad \bar{p} \qquad \bar{p} + d$$

Šířka d bude zřejmě závislá na přesnosti měření a na spolehlivosti, s jakou chceme náš *konfidenční interval* (interval spolehlivosti) konstruovat. Předpokládejme, že spolehlivost má být 95%; pak

$$d = 1{,}96 \, . \, s_{\bar{p}} \, ,$$

kde $S_{\bar{p}}$ je výběrová chyba. (Kdybychom chtěli spolehlivost např. 99 %, pak by se koeficient změnil z 1,96 na 2,58 apod.)

Výběrová chyba se vyjádří vzorcem:

$$S_{\bar{p}} = \sqrt{\left(1 - \frac{n}{N}\right) \frac{\bar{p}\,(100 - \bar{p})}{n - 1}}$$

kde N je velikost populace (N = 1000)
n je velikost výběru (n = 200)
$\bar{p}$ je procento ve výběru (postupně 62,5; 78,0; 56,3).

Pro naše tři výzkumníky dostaneme postupně výpočtem (a po zaokrouhlení) tabulku 2.

Tabulka 2

$\bar{p}$	62,5%	78,0%	56,3%
$1 - \frac{n}{N}$	0,8	0,8	0,8
$\frac{\bar{p}(100-\bar{p})}{n-1}$	11,78	8,62	12,93
$s_{\bar{p}}^2$	9,42	6,90	10,34
$s_{\bar{p}}^2$	3,07	2,63	3,22
$1{,}96\ s_{\bar{p}}$	6,02	5,15	6,30
interval spolehlivosti	(56,48;68,52)	(72,85;83,15)	(50,00;62,60)

Každý výzkumník dostává zcela odlišné výsledky, pouze první (aniž o tom ví) pokrývá skutečnou hodnotu. Všichni však doufají, že jejich výsledky jsou správné a mají za sebou spolehlivost 95%.

Jak interpretovat slovo *spolehlivost?* Jestliže vybereme náhodně jeden ze souborů o 200 statích, máme 95% šanci, že neuděláme tímto statistickým postupem chybu; v 95 % všech možných výběrů o velikosti 200 vede tento postup k pokrytí skutečného neznámého parametru. A to je též klíč k volbě tohoto čísla. Ve společenských vědách se většinou rozhodujeme pro 95 %, někdy pro 99 %. Podle problému však ve statistických úlohách může být hladina spolehlivosti volena až na 99,9 %, což je případ některých medicínských úloh nebo situací, kdy se na základě měření rozhodujeme pro velké investice.

Máme přirozeně zájem na tom, aby interval spolehlivosti byl co nejkratší. Platí ovšem, že čím větší spolehlivost požadujeme, tím širší interval dostaneme a naopak. Je tu tedy konflikt mezi oběma hledisky. Jednou krajností je bodový odhad, který má spolehlivost 0 % a na druhé straně je tu 100% spolehlivost, té však odpovídá celá škála možných hodnot, takže takový výsledek je prakticky bezcenný.

Výhodou postupu je to, že je v něm už přímo zabudována přesnost — čím užší interval při dané spolehlivosti, tím méně entropičnosti v rozhodování a tím přesnější informace. Nevýhodou však je právě to, že jde o interval, s nímž nemůžeme snadno pracovat při vytváření dalších měr a při srovnávání různých výsledků. Kromě toho je tu stále základní fakt statistických závěrů — jsou to závěry za neurčitosti. Pouze doufáme, že nám je náhoda při měření příznivá. V dlouhé řadě konstrukcí intervalů spolehlivosti očekáváme 5 % případů nepokrytí. Je-li zvolená spolehlivost dostatečně vysoká, musí každý výzkumník rozhodnout sám.

6. Testování hypotéz

Nejčastější úlohy formulované ve společenskovědním výzkumu jsou úlohy testování hypotéz. To jsou případy, kdy nás nezajímá odhadování hodnoty parametru přímo, ale rozhodnutí, zda lze přijmout ten či jiný výrok o parametrech nebo o rozložení veličin vůbec.

Existuje celá řada typů hypotéz. Tak např. výzkumník uvedený ad 4. si mohl postavit hypotézu, že skutečné procento p = 60 %. Test hypotézy pak mohl být ekvivalentní s konstrukcí intervalů spolehlivosti; jestliže interval spolehlivosti pokryje hodnotu 60 %, pak není důvodů hypotézu odmítnout, je-li pak 60 % vně intervalu, považujeme hypotézu za nesprávnou (málo pravděpodobnou). I zde je vidět omezení statistických aplikací. Předpokládejme, že naše tři osoby, každá nezávisle, vyslovily tři různé hypotézy: p = 60 %, p = 65 %, p = 55 %. První a třetí osoba přijímají své hypotézy, které jsou ale odlišné pro tentýž soubor, tj. pro tutéž realitu; druhá osoba odmítá hypotézu (ač hypotetická hodnota je tatáž jako skutečná v souboru).

Další příklady hypotéz:

a) průměrný výsledek bodovaného testu z matematiky je větší u studentů sociologie než u studentů psychologie;

b) existuje průkazná závislost mezi znakem „informovanost o daném objektu" a znakem „postoj k objektu";

c) na základě minulých měření předpokládáme, že se v průměru zlepší počasí a budou létat letadlové linky mezi Prahou a Bratislavou;

d) na základě zjištěných symptomů přijímáme diagnózu D.

Mluvíme tu o statistických hypotézách, tj. o hypotézách týkajících se pravděpodobnostních vlastností náhodných veličin, které měříme a zjišťujeme. Mluvíme-li o statistice, předpokládáme, že překlad z jazyka sociologie do jazyka statistiky, tj. statistická operacionalizace, je už provedena. Konečný výsledek rozhodnutí o veličinách musí být opět převeden zpět do jazyka sociologie.

Mnohdy při testování hypotéz nás výsledky překvapují svou neočekávatelností a zdají se zcela nemožnými — odmítáme hypotézy, které by měly být podle teoretického východiska jednoznačně potvrzeny. Interpretace takovýchto výsledků musí být krajně zdrženlivá a opatrná. Je pochopitelně nutné takový výsledek vysvětlit a na bázi teorie se s ním vyrovnat. Nesmíme však zapomínat, že to může být způsobeno nejrůznějšími faktory a ukvapené závěry o revizi východisek nelze přijmout bez prověření všech kroků ve výzkumu. Uveďme některé možné příčiny:

1. chyby v operacionalizaci teoretických pojmů;
2. špatná konstrukce modelů — tj. špatný překlad ze sociologie do matematiky;
3. chybný plán a realizace sběru dat;
4. špatná volba znaků vypovídajících o objektech; špatná volba *referenčního systému znaků;*
5. nevhodná formulace otázek;
6. sémantická neurčitost otázek — respondent je chápe jinak než výzkumník;
7. špatný zpětný překlad statistického rozhodnutí do pojmového rámce výchozí teorie;
8. je k dispozici příliš málo dat, aby mohly být prokázány formulované hypotézy;
9. data nejsou analyzována podle modelu, který byl aplikován;
10. mohl nastat případ, že právě u této konkrétní hypotézy se projevilo pravděpodobnostní riziko, se kterým pracujeme ve statistice vždy.

Není možné tedy zamítnout škrtem pera východiska, aniž bychom nesledovali celou dlouhou řadu vlivů (z nichž některé zde byly jmenovány), které mohou ovlivňovat celkové závěry a celý proces vědecké inference — nejen její statistickou část.

Statistické testování hypotéz, jako všechny úlohy řešené matematikou, vycházejí z modelu, který je abstrakcí skutečnosti. Z toho plynou další omezení aplikace metody — vždy je zachycen jen některý z aspektů reality.

Shrňme jen stručně některé z možných přístupů k testování hypotéz.

A) Formulujeme hypotézu H a za předpokladu, že platí, odvodíme teoretické pravděpodobnostní chování sledovaných veličin a jevů. Jestliže data, která získáme jsou za této hypotézy málo pravděpodobná, pak hypotézu zamítáme. Tento přístup nazveme přístupem *R. A. Fishera.* Bylo by možno jej charakterizovat jako zeslabenou analogii logického postupu

$$(H \Rightarrow D) \Leftrightarrow (\text{non}\, D \Rightarrow \text{non}\, H)$$

(z hypotézy (H) plyne chování dat (D); nechovají-li se data podle předpokladu, hypotéza neplatí).

Uvažujeme vždy riziko α (odpovídající riziku nepokrytí skutečné hodnoty při intervalovém odhadu, např. 5% resp. v jednotkách pravděpodobnosti 0,05), které je stanoveno jako horní pravděpodobnostní hranice proti hypotéze, jestliže je správná. Nastane-li málo pravděpodobný jev, pak buď hypotéza neplatí, nebo nastal zázrak (R. A. Fisher).

ŘEHÁK, Jan. 1971. „Poznámky k analýze sociologických dat." Sociologický časopis, 27 (4), str. 425-433.

B) Přístup *Neyman-Pearsonův* lze charakterizovat dvojicí hypotéz, které jsou postaveny proti sobě. Základní nulová hypotéza (H_0) je testována proti nějaké jiné možnosti — alterativní hypotéze (H_1). Předpokládáme teoreticky dva možné stavy skutečnosti — zde tedy vstupuje teorie ještě o krok dále do statistické procedury — vymezuje možnou alternativu. To umožňuje lepší výběr mezi rozhodovacími pravidly. Je tu možné formulovat hledisko optimálnosti pro takový výběr.

Rozhodovací situace může být naznačena tabulkou

Tabulka 3

		Rozhodnutí pro	
		H_0	H_1
Ve skutečnosti platí	H_0	správné rozhodnutí	chyba 1. druhu
	H_1	chyba 2. druhu	správné rozhodnutí

Rozhodovací pravidla jsou volena tak, aby pravděpodobnost chyby 1. druhu nepřevýšila dané číslo (např. 0,05; 0,01; 0,001; značíme ji obvykle α) a přitom aby pravděpodobnost chyby 2. druhu byla co nejmenší (značíme obvykle β). Rozhodnutí je tu závislé na schopnosti přesně formulovat oba modely odpovídající oběma hypotézám, na počtu pozorování a (jako vždy) na náhodě.

C) *Bayesovský přístup* je založen na Bayesově větě (publikované v roce 1763 Thomasem Bayesem). Tento přístup formuluje hypotézy $H_1 \ldots H_k$ a předpokládá, že jsou a priori (před sběrem dat) známy pravděpodobnosti těchto hypotéz; tyto apriorní pravděpodobnosti mívají nejrůznější interpretace a celý tento přístup má mnoho různých škol podle názoru na apriorní pravděpodobnosti. Můžeme snad souhrnně říci, že apriorní pravděpodobnosti odrážejí stupeň našich znalostí a zkušeností o možných jevech, ať už je vyjadřujeme empiricky a na základě dřívějších zkoumání, nebo jako subjektivní názor; mohou reprezentovat také subjektivní důvěru v platnost hypotézy.

Pomocí Bayesovy formule pak opravujeme apriorní pravděpodobnosti na základě evidence ze sebraných dat. Takto získané aposteriorní pravděpodobnosti jsou základem pro statistické rozhodování o hypotézách. Hypotézu, která má aposteriori dostatečně vysokou pravděpodobnost, můžeme přijmout za správnou. Bayesovská analýza dat je technicky značně náročná. Závěry závisí na apriorních pravděpodobnostech a na modelu, který byl zvolen.

D) V případě, že nejsme s to, či nechceme, nebo se neodvažujeme konstruovat matematický model, který umožňuje inferenci o parametrech, můžeme volit *neparametrické techniky*, v nichž jsou předpoklady daleko volnější a nezavádí se parametry, kromě velice obecných (jako je posunutí počátku, změna měřítka apod.). Výhodou těchto technik je, že se obejdou bez modelu; jedna technika zahrnuje daleko více aplikačních situací. Jsou většinou velice jednoduché, snadno pochopitelné a jejich heuristické odvození je jasné. Přesto že jsou „do šíře" aplikabilnější, nelze doporučit jejich použití tam, kde lze sestrojit model. Do modelu totiž vkládáme už apriorní informaci a ta nám umožňuje analyzovat data efektivněji; pro stejnou přesnost závěru je třeba méně pozorování. Chyby druhého druhu jsou menší. Přínos informace skrze model umožňuje snížit entropičnost rozhodovací situace. Také formální podoba neparametrických technik tento fakt potvrzuje. Jsou to většinou metody založené na pořadí nebo na konstatování výskytu nějakých jevů. Kardinální znaky pak pro použití takových testů ordinalizujeme, což je pochopitelně jistou ztrátou informace. Formální stránka, která vede k využití pořadí pozorování však předurčuje neparametrickým technikám velice široké použití ve společenských vědách — mnohdy totiž nejsme schopni získávat kardinální znaky a naše měření jsou pouze pořadového charakteru. (Např. nejsme schopni měřit schopnost žáků, ale jsme schopni je podle schopnosti uspořádat.) Formulace úlohy je obdobná jako u Neyman-Pearsonova přístupu; zavádíme též nulovou a alternativní hypotézu a chyby 1. a 2. druhu. Rozdíl spočívá v tom, že zde nezavádíme parametrické modely.

7. Typy statistických hypotéz

Pro lepší orientaci v různých úlohách testování hypotéz je možné rozčlenit je do

tří nejčastěji se vyskytujících typů. Toto členění je pomocné a má úlohu pouze orientační (z hlediska formálního i terminologického mu lze leccos vytknout). Velká většina hypotéz je v současné době založena na Neyman-Pearsonově přístupu a neparametrických technikách, které jsou jeho rozšířením.

Ve výzkumné praxi se vyskytují nejčastěji tři typy statistických hypotéz:

1. *Hypotézy o stavu populace — testy dobré shody*

H_0 je hypotéza, která říká, že pro danou populaci platí určitý model, resp. určitý výrok o neznámých parametrech. Termín testy dobré shody je tu chápán šířeji než ve statistické literatuře. Jde obecně o tvrzení, že populace je v nějakém daném hypotetickém stavu. Jako příklady mohou sloužit:

a) veličina má normální rozložení;

b) víme, či předpokládáme, že veličina má normální rozložení a H_0 specifikuje, že má průměr 38,5;

c) průměrné skóre I.Q. testu v dané konečné populaci je $\bar{X} = 70$ bodů (ze 100 možných).

V případě c) např. můžeme formulovat řadu alterativních hypotéz:

$H_1 : \bar{X} \neq 70$ $\quad$ $H_4 : \bar{X} > 70$
$H_2 : \bar{X} = 80$ $\quad$ $H_5 : \bar{X} < 70$
$H_3 : \bar{X} = 62$ $\quad$ $H_6 : \bar{X} < 50$

a mnoho dalších; výběr z nich záleží na situaci a teorii, kterou vkládáme do procesu inference.

Nulová hypotéza může být specifikována na rozložení četností, průměr, procento, rozptyl, medián, symetrii apod.

2. *Hypotézy o srovnání dvou nebo více populací — testy homogenity*

V těchto testech se promítá do statistiky srovnávací metoda. Těmito testy se srovnávají dvě nebo více populací, resp. některé rysy populací. Tak např. je možno porovnávat rozložení četností, aritmetické průměry apod.

Nulová hypotéza bývá formulována obvykle jako hypotéza homogenity, tj. hypotéza, že soubory jsou z daného hlediska statisticky podobné (je možno je smíchat a výsledné rozložení, či zkoumaná vlastnost je stejná pro celý soubor právě tak, jako pro jeho původní části). Alternativní hypotézy specifikují rozdílnost: například rozdílnost průměrů nebo jejich uspořádání atd.

3. *Testy o struktuře vztahu mezi proměnnými — testy závislosti*

V tomto případě (který by bylo možno obecně zahrnout do první kategorie) je předmětem rozhodovacího procesu pouze jedna populace, ale více proměnných, jejichž vztah nás zajímá.

Nulová hypotéza je většinou formulována jako nezávislost znaků. Alternativní hypotéza pak je specifikována kauzálním modelem, který nás zajímá, tj. strukturou vztahů, které chceme prokázat statisticky.

Je přirozené, že existuje řada dalších typů hypotéz a testů pro ně a že toto dělení je jen zcela hrubé a pomocné. Přesto může být snad užitečné při formulaci statistických úloh a při následné interpretaci.

8. Závěr

Tyto poznámky neměly být receptem jak analyzovat data, ale měly čtenáře seznámit s tím, jaké úlohy před námi ve statistice stojí, jaké požadavky můžeme formulovat a hlavně naznačit, s jakými omezeními při použití musíme počítat, jaká omezení máme. Jsou do jisté míry reakcí na současný stav aplikace statistických metod v sociologickém výzkumu . Musím tu varovat před přespřílišným a hlavně nekvalifikovaným používáním testů a jiných procedur matematické statistiky. Nekvalifikovaná aplikace udělá obvykle více škod než užitku. Domnívám se, že aplikace matematické statistiky a matematiky vůbec má velký význam pro analýzu dat i pro modelové účely a výstavbu dílčích teorií, že se bez nich společenská věda, která chce kvalifikovaně hodnotit jednotlivé měřené aspekty skutečnosti, neobejde; soudím však, že je třeba postupovat opatrně a odborně. Chtěl bych varovat před magií čísel a před absolutizováním četnosti jako důkazového materiálu. Doufám, že jsem uvedl dost různých okolností, které takovýto přístup zpochybňují.

SWOBODA Helmut. 1977. *Moderní statistika*. Praha, Svoboda.

5.5 Spolehlivost výběrových souborů — intervaly spolehlivosti

Pravděpodobnost jistoty, spolehlivost, vypovídací schopnost, přesnost, interval spolehlivosti — je to vše totožné? Pokusíme se ještě zřetelněji ukázat souvislosti a rozdíly a především se nad těmito pojmy poněkud zamyslíme.

Co je například *spolehlivost*? Zpravidla jsme tím mysleli onu spolehlivost, která říká, že výpověď je z 90 nebo 95 % správná. To je pravděpodobnost jistoty, která vymezuje také interval spolehlivosti a kterou se budeme ještě zabývat podrobněji.

„Spolehlivost" se však může vztahovat také na všeobecnou kvalitu statistické výpovědi. V tomto smyslu např. systematická chyba narušuje spolehlivost.

Další otázkou je, zda vytváření výběrových souborů je vůbec spolehlivou vědeckou metodou. Kromě toho experiment musí být zásadně opakovatelný. Specifikem statistických experimentů však je, že nelze téměř nikdy docílit přesně stejných výsledků, nýbrž jen velmi podobných. Zkoumání výběrových souborů má být opakovatelné, jeho výsledky však mohou být pouze navzájem porovnatelné, nikoli totožné. Další požadavek, který se klade na všechny vědecké pokusy, je *požadavek významnosti, vypovídací schopnosti a důležitosti*: musí se skutečně dokázat to, co má být dokázáno, nikoliv jen něco podobného.

Vraťme se však zpět k 95% *pravděpodobnosti* jistoty, ke stupni jistoty. Co si pod tím máme představit? Můžeme opět pomyslit na rozdělení četnosti pod normální křivkou a předpokládat soubor všech možných výběrových souborů v rozsahu *n* jako normálně rozdělený. Pak 95% pravděpodobnost znamená, že výběrový soubor, který jsme právě náhodně vytvořili, je právě v toleranci 95 %, které dávají střední hodnotu $\bar{x}$, ležící maximálně 2σ od skutečného průměru μ.

Zda postačí 95 % jistoty nebo nikoliv, nelze říci všeobecně. O tom je třeba rozhodnout v každém jednotlivém případě samostatně. Jestliže se nemá vydat moc peněz a postačí-li přibližný přehled (jako např. u mnohých otázek průzkumu trhu), je postačující 90% nebo ještě nižší jistota. Jde-li o zvlášť významné, velmi závažné rozhodnutí, bude snaha dosáhnout 99% nebo ještě vyšší pravděpodobnosti. Za postačující se však zpravidla pokládá 95% pravděpodobnost.

Jestliže chceme dosáhnout vyšší spolehlivosti, je nutno zkoumat větší výběrové soubory. Není-li to z nějakých důvodů možné, pak také nelze dosáhnout žádaného stupně jistoty — a dost! Nepomohou žádné početní triky či okliky a ani jiné komplikované kejkle neposkytnou východisko. Rozsah výběrového souboru a spolehlivost výsledku výběrového souboru jsou také v nejužší logické a matematické souvislosti s úkazem, který jsme pozorovali při prvním výkladu binomického rozdělení a Pascalova trojúhelníku: v protikladu k naivní představě o „zákonu velkých čísel" neexistuje sice přesné ustálení výkyvů na očekávané hodnotě, přesto však při rostoucím rozsahu výběrového souboru se budou proporce výběrového souboru relativně (ne absolutně!) méně odchylovat od skutečné proporce souboru základního.

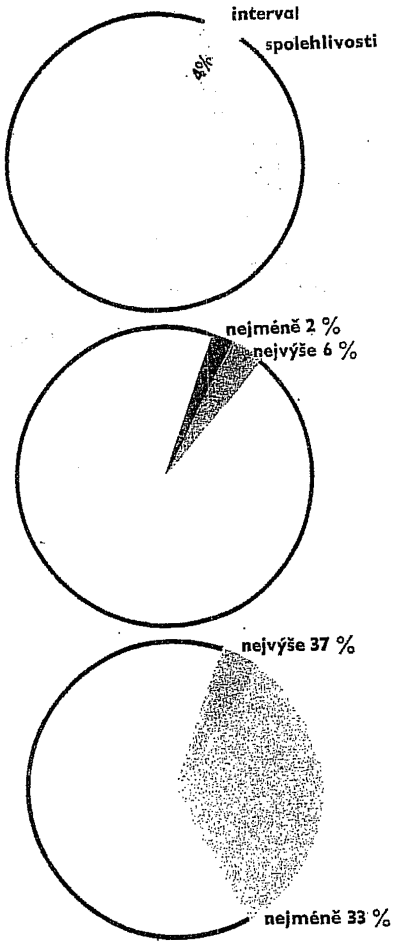
V ruletě to například znamená: je zcela dobře myslitelné, že ve „výběrovém souboru" dvaceti po sobě jdoucích her je 16 červených a jen čtyři černé, to znamená odchylka o 6 od očekávané hodnoty 10. Naproti tomu je zcela vyloučeno, aby z 2000 her bylo 1600 červených a jen 400 černých. I poměr 1100 ku 900 by byl velmi nepravděpodobný. S rostoucím rozsahem výběrového souboru bude tedy stále přesněji zaměřován skutečný průměr základního souboru (který v případě rulety známe). Při rozsahu výběrového souboru $n = 20$ se mohu velice zklamat (např. jako u neuvěřitelného výsledku 16 červených, 4 černé), při $n = 200$ se však mohu spolehnout, že neskončím příliš daleko od pravdy. To je „zákon velkých čísel" v teorii a praxi výběrových souborů.

Pravděpodobnost výběrového souboru mohu také definovat ze záporného hlediska a obdržím „riziko chyb"; 95% pravděpodobnost znamená 5% riziko chyb. Máme-li pro „pozitivní" a „negativní" dva různé výrazy, stane se situace ještě zmatenější, když se začne mluvit o „hladině významnosti". Není to zvlášť šťastně zvolený výraz, ale přesto je velmi rozšířen. „*Významnost*" je prakticky totéž co pravděpodobnost: výsledek je „statisticky významný na úrovni 5 %", protože by jen čirou náhodou nenastal v 95 % případů. Celá rozsáhlá oblast testování hypotéz (viz kap. 6), která nás bude ještě velmi zaměstnávat, se zakládá na takovýchto úvahách a výpočtech: je nebo není dosažený výsledek slučitelný s tou nebo onou hypotézou?

„Významnost" není tedy nic jiného než dohoda mezi těmi, kdo statistické metody používají. A zcela stejně jako u pravděpodobnosti lze různě posuzovat úroveň významnosti v závislosti na kladení otázek. *Zpravidla se všeobecně pokládá úroveň 5 %* (nazývaná také často úroveň 95 %) *za „významné", úroveň 1 % (99 %) za „vysoce významné".*

Pravděpodobností jistoty je také zároveň určen *interval spolehlivosti*. Tento interval se měří jako směrodatná odchylka normovaného normálního rozdělení nebo také i v absolutních hodnotách. Měli jsme již příklady pro oba způsoby vyjádření: Úspory mají rozptyl $s = 60$ kolem průměru $\bar{x} = 480$, interval spolehlivosti (95 %) sahá od 360 do 600 DM, totiž od $(480 — 2\sigma)$ až do $(480 + 2\sigma)$. Mám 95 % jistoty, že náhodně vybraný rolník má více než 360 a méně než 600 DM úspor.

Poznámka: Pozor na záměnu *pravděpodobnost jevu a pravděpodobnost statistického rozboru*. Pravděpodobnost četnosti nějakého znaku zjištěná ve vzorku (proporce výběrového souboru p) je nezávislá na pravděpodobnosti jistoty tohoto vzorku. Mohu vytvořit tři vzorky o rozsahu $n_1 = 20$, $n_2 = 100$ a $n_3 = 500$, které (náhodou) všechny mají četnost výběrového souboru $p = 0{,}7$ (tedy 70 %) znaku. Avšak kvalita, vypovídací schopnost, pravděpodobnost jistoty malého výběrového souboru budou podstatně nižší než u většího souboru.



Interval spolehlivosti ve výši 4 % má různou váhu u malých a velkých podílů. Jestliže podíl vzorku činí např. 4 %, sahá interval spolehlivosti od 2 do 6 %, tedy až do trojnásobku. Činí-li podíl vzorku naproti tomu 35 %, je rozpětí „mezi 33 a 37 %" dostatečně informativní a přesné; 37 % je totiž jen asi o desetinu větší než 33 %.

STATISTICAL INFERENCE

Now that we have established a background in probability and have seen what a normal distribution can reveal, what can a statistical test tell a researcher? It *cannot* prove that a change in one variable caused a change in another, but it can tell whether the difference in mean scores observed between those experiencing one treatment and the usual population that did not could have occurred as a random event. If the test says that it is unlikely that the difference occurred by chance alone, it is still up to the researcher to prove that the one variable was the only possible cause. Statistical tests are like the 'idiot lights' on the dashboard of your car: they only tell you that *something* has happened, but not exactly why. For example, if the oil light comes on, we assume something is not right. It could mean the engine is low on oil, the engine bearings have worn out, the oil pump has perished, the signal-sending device on the engine is broken, or a wire has shorted out to the light. The motorist obviously checks the oil level first, but if that is adequate, then it is time to call the mechanic, who will try to find the reason for the light being on. In the social sciences, the researcher should plan a study such that when the light comes on (the statistics indicate that something probably happened), then there is only one predicted, defensible link or potential cause. As seen in Chapters 1–7, designing a study to resolve such issues is not trivial.

The term *inferential statistics* refers to the process of using data collected from samples to make inferences about a larger population or populations. The research process introduces complications since:

- most research involves samples (which are *probably* representative);
- the traits usually result in distributions of scores, thus the group characteristics or tendencies are best described as measures of central tendency, such as means;
- the natural variability (with hopefully little error due to low reliability of the instrument) is best indicated by standard deviations.

Using this information, there is a desire to compare groups to determine relationships that will ultimately extend back to the original population(s). Thus any comparisons will require the nature of the distribution to be considered as well as the central tendency of the group. All of this depends heavily upon probability, and it is never possible to speak about relationships with absolute certainty, a fact that causes a distinct amount of mental anguish for most people who feel that events should have some degree of certainty.

There is a need to state the expected outcomes of inferential statistical research in terms of the *null hypothesis*: that there will *not* be any statistically significant difference. In other words, it is expected that any differences or changes or relationships found will be attributable to chance alone. Even if the null hypothesis is rejected, it only means that the difference or occurrence witnessed *probably* did not occur by chance alone. This probability level traditionally has been set at a critical level of 5%, which basically means that if a statistical test says that the probability of this event occurring by chance alone is less than 5%, then it probably did *not* occur as a random event. At this level, there is something probably influencing the event(s), or at least the event(s) has/have occurred as the result of some external influence other than natural random fluctuation. Exactly what this influence is, is not made clear by the statistical test. As noted before, it is still up to the researcher to justify that what he or she did or the variables identified, were the only possible source of influence.

This section will bring together earlier concepts from probability and combine them with research questions and hypotheses, and apply them to the cases where the variables are normally distributed. Before the actual choice of statistical tests can be considered, it is necessary to take a brief mathematical look at what underlies statistical inference and significance. This will be done graphically as much as possible, since most decisions are made on the basis of where the means of sets of data are in a normal distribution. It will provide a basis for later chapters that will continue the review of inferential statistics by considering a variety of specific tests which can be used as part of experimental, quasi-experimental and ex post facto designs to decide the acceptability of stated hypotheses.

Linking probability to statistical inference

Just as individual scores for a trait vary around a mean to form a normal distribution, the means of samples themselves will vary if a number of representative samples are taken from a population. Thus, if the frequencies of these means are plotted on a graph it is not surprising that we find yet another normal distribution. This *distribution of sampling means* will be quite useful in making inferences about the population. This was introduced earlier in Chapter 5 with reference to sampling error. Figure 13.6 shows all three types distributions for IQ scores: (a) an exemplar population distribution with parameters provided; (b) a single-sample distribution with its statistics; and (c) a distribution of sampling means.

The IQ score is used here simply because it is one distribution for which the population parameters are known, since the tests are designed to produce a mean of 100 and a standard deviation of 15. As will be seen later, this is the exception, since we rarely know what the population mean is. The situation where the population mean is known is used here primarily because it is the simplest and easiest to use to illustrate the principles behind statistical significance. Once this foundation is laid, all the others are basically variations on this ideal.

Remember that when the term *population* is used, it refers to a group sharing a limited set of common characteristics. In social sciences, these are often not obvious to the casual observer and require some form of detailed observation, measurement or questioning of the subjects. So initially, the issue is whether or not a sample as a group is similar enough to the population for the trait or characteristic in question to be considered representative. A statistical test should be able to resolve what is enough.

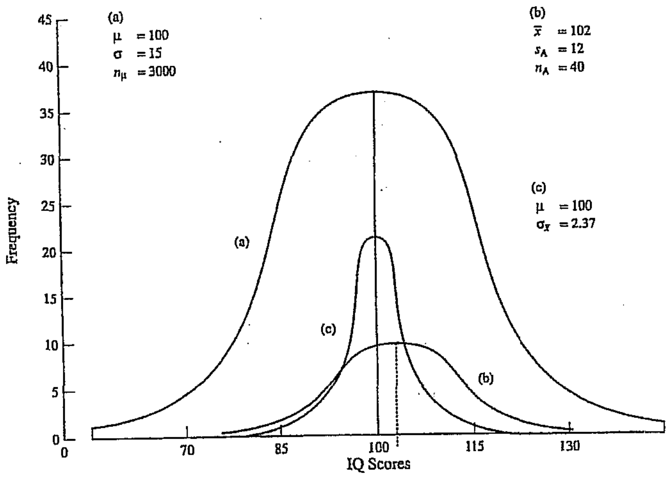


FIGURE 13.6
(a) The population distribution of IQ scores for all 3000 11-year-olds in a local education authority (LEA); (b) a single exemplar sample distribution of IQ scores of a random selection of 40 11-year-olds in the LEA; and (c) the distribution of sample means for a number of such random samples of 40 students

The first thing to notice in Figure 13.6 is that the standard deviation (and width of the bell-shaped curve) for the distribution of sample means is relatively small compared to the standard deviations for the population and any single sample. Thus it is very unlikely that a truly representative sample will have a mean very different from that of the population. This fact is used in the most basic of inferential statistical tests, deciding whether a sample is to be considered part of a defined population, or part of some other population. To distinguish this standard deviation from that of a sample of the population, the standard deviation of the distribution of sampling means is used, which is known as the *standard error of the mean* (SEM). This will be designated by $\sigma_{\bar{x}}$ if it is calculated from the population parameter and is found by

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad (13.2)$$

where σ is the population standard deviation (equation (12.2)), and n is the sample size. Obviously, the standard error of the mean depends on the sample size: for a very large sample size the standard error of the mean, and consequently the width of the curve for the sampling distribution, will be very small.

It is illustrative to consider an example: in order to carry out a study, a researcher selects a sample of 40 students from the LEA population of 11-year-olds described in Figure 13.6. They are given an IQ test: the group mean is found to be 106. Is this group typical? Let us first state this question as a null hypothesis:

> H_0: There is no significant difference between the IQ of the sample group and that of the population.

In everyday English, we would say that we expect that the sample *is* representative of the population for this trait. Here the sample mean will be used to resolve the issue. To make the decision, it is necessary to zoom in on distribution (c) in Figure 13.6, the sample means, shown enlarged in Figure 13.7. The question now becomes one that is stated in terms of probabilities:

> What is the probability that a sample with a mean of 106 would be randomly chosen from the population?

Recall that the area under the distribution for a range of scores represents the percentage of people having scores within that range (see Figure 12.21). In this situation, we are considering a distribution of sample means. Using Table B.1 in Appendix B, the number of standard deviations of sampling means (SEMs) can be used to determine what percentage of sample means one would expect below this group's. Here, a sample mean of 106 is 2.40 standard deviations (SEMs) above the population mean, as marked on Figure 13.7. From Table B.1, this

BLACK, Thomas, R. 1999. *Doing Quantitative Research in the Social Sciences. An Integrated Approach to Research Design, Measurement and Statistics*. London: Sage.

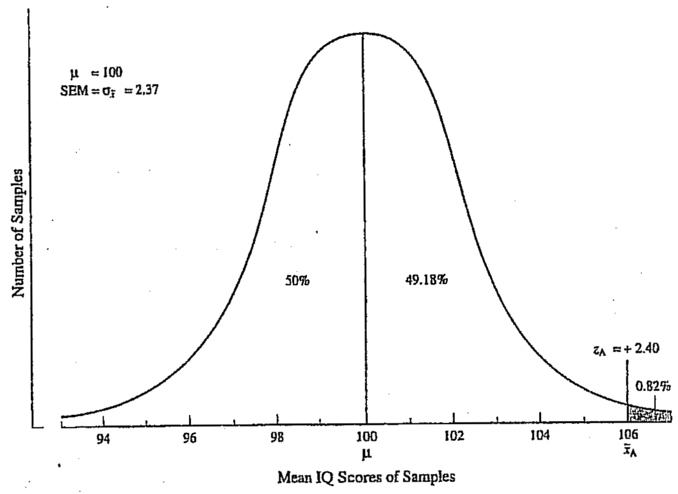


FIGURE 13.7
Distribution of sampling means (each sample size = 40), showing the position of a single sample mean, $\bar{x}_A$

tells us that 49.18% of the sample means would be expected to be between this score and the population mean. Add to this the 50% below the population mean and we find that 99.18% of the sample means should be below this, as shown in Figure 13.7. To put it another way, the probability of this event or any one beyond it occurring as a random event is 100% − 99.18% = 0.82%, or 0.82 of a chance in 100 or 8.2 chances in 1000. Thus this sample mean does seem to be a highly unlikely outcome for a random sample, but what is *unlikely enough* for researchers?

6. lekce

SROVNÁVÁNÍ SKUPIN NA ZÁKLADĚ STŘEDNÍCH HODNOT JEJICH KARDINÁLNÍCH CHARAKTERISTIK (modul Analyze: procedura means). HYPOTÉZA O SHODĚ DVOU PRŮMĚRŮ PRO NEZÁVISLÁ A PÁROVANÁ DATA: T-TESTY A EKVIVALENTNÍ NONPARAMETRICKÉ TESTY (modul ANALYZE: procedury Compare means: one-sample t-test; independent-samples t-test, paired-samples t-test).

Comparing Groups

5

How can you determine if the values of the summary statistics for a variable differ for subgroups of cases?

- What are subgroups of cases?
- What can you learn from calculating summary statistics separately for subgroups of cases?
- How can you graph means for subgroups of cases?

In Chapter 3 and Chapter 4, you used the Frequencies and Descriptives procedures to calculate summary statistics for all of the cases in your study. Often, however, you are interested in comparing summary statistics for different groups of cases. For example, you want to compare hours studied per week for college freshmen, sophomores, juniors, and seniors. Or you want to find the average income for people living in different geographical areas. There's no easy way with the Frequencies or Descriptives procedure to produce such information. In this chapter, you'll use the Means procedure to calculate simple summary statistics for subgroups of cases. You'll see if you can find a relationship between the average years of education and job satisfaction. You'll also see whether the relationship appears to be similar for men and women. (The Explore procedure described in Chapter 6 lets you examine the values of a variable for subgroups of cases in much greater detail.)

▶ This chapter uses the *gssft.sav* data file, which contains some of the variables in the *gss.sav* file, but for full-time workers only. (See "Case Selection" on p. 541 in Appendix B if you want to know how this smaller file was created.) For instructions on how to obtain the SPSS output discussed in this chapter, see "How to Obtain Subgroup Means" on p. 83.

Education and Job Satisfaction

In Figure 5.1, you see the average years of education and the standard deviation for people in each of four job satisfaction categories. To make comparisons easier to interpret, only people employed full time are included in the analysis.

Figure 5.1 Pivoted Means output for education and job satisfaction

To obtain this output, from the menus choose:

Statistics
Compare Means ▶
Means...

Select the variables educ and satjob, as shown in Figure 5.6.

The Pivot Table Editor is used to put statistics into the columns.

Highest Year of School Completed

Job Satisfaction	Mean	N	Std. Deviation
Very satisfied	14.25	327	2.79
Mod satisfied	13.80	320	2.60
A little dissatisfied	13.92	74	2.75
Very dissatisfied	14.65	26	2.37
Total	14.04	747	2.70

Statistics for individual subgroups

Statistics for all cases analyzed

Looking at the last row of Figure 5.1, you see that the 747 people who are employed full time have 14.04 years of education on average, with a standard deviation of 2.70 years. These 747 people are assigned to one of four subgroups, based on their job satisfaction. The first subgroup, *very satisfied* employees, are somewhat more educated than the group as a whole, while the fourth subgroup, *very dissatisfied* respondents, are the most educated of all. People in the two subgroups in the middle—*moderately satisfied* and *a little dissatisfied* employees—have somewhat less education than the group as a whole.

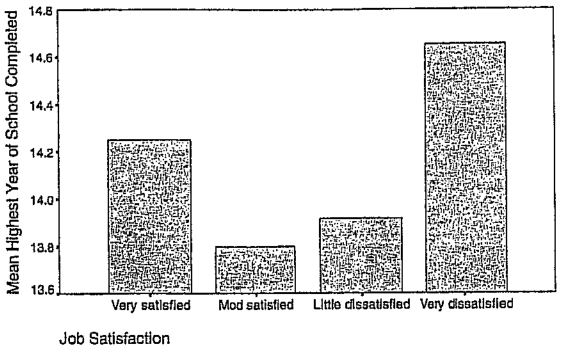
Plotting Mean and Standard Deviation

A plot of the mean years of education for the four subgroups is shown in Figure 5.2. There is a bar for each of the subgroups. The height of the bar depends on the average years of education. You can easily see that the *very dissatisfied* people have the largest mean years of education. The mean years of education for the middle two satisfaction groups appear to be similar. Note, however, that the scale for the axis on which mean education is plotted doesn't start at 0. That makes even small differences in years of education look large on the plot.

Figure 5.2 Bar chart of education by job satisfaction

You can obtain bar charts using the Graphs menu, as described in Appendix A.

In the Define Simple Bar Chart Summaries for Groups of Cases dialog box, select Other summary function and select the variable educ. Select satjob for Category axis.



Layers: Defining Subgroups by More than One Variable

In Figure 5.1, all of the people who are employed full time are subdivided into four groups based only on their answer to the job satisfaction question. If you want to see whether the relationship between education and job satisfaction is similar for males and females, you must subdivide each of the rows of Figure 5.1 further. Figure 5.3 shows summary statistics for full-time workers, subdivided first by job satisfaction and then by gender.

Figure 5.3 Pivoted means output for job satisfaction and gender subgroups

To obtain this output, select the variable sex for layer 2 in the Means dialog box, as shown in Figure 5.7.

The Pivot Table Editor is used to move the statistics to the columns.

Highest Year of School Completed

Job Satisfaction	Respondent's Sex	Mean	N	Std. Deviation
Very satisfied	Male	14.20	179	3.00
	Female	14.31	148	2.54
	Total	14.25	327	2.79
Mod satisfied	Male	13.71	173	2.68
	Female	13.90	147	2.50
	Total	13.80	320	2.60
A little dissatisfied	Male	14.02	41	2.95
	Female	13.79	33	2.53
	Total	13.92	74	2.75
Very dissatisfied	Male	15.27	15	2.40
	Female	13.82	11	2.14
	Total	14.65	26	2.37
Total	Male	14.01	408	2.85
	Female	14.06	339	2.51
	Total	14.04	747	2.70

The total line in each category is the same as Figure 5.1.

Within each category, separate statistics are shown for men and women.

You see that 327 people rate themselves as *very satisfied* with their job. They have an average of 14.25 years of education. Of the 327 *very satisfied* people, 179 are men and 148 are women. The men have an average of 14.20 years of education, while the women have 14.31 years of education. That's not much of a difference. Looking at the *moderately satisfied* males, you see that their average years of education is about half a year less than that of the *very satisfied* males.

One of the more interesting observations gleaned from Figure 5.3 is that the *very satisfied* women have the highest average years of education of all the women. Women in the remaining three satisfaction categories have very similar average years of education. In contrast, the *very dissatisfied* males have the highest average years of education, 15.27. (However, there are few cases in the *very dissatisfied* group, so your conclusions are necessarily tentative.) The *moderately satisfied* males have the smallest average years of education.

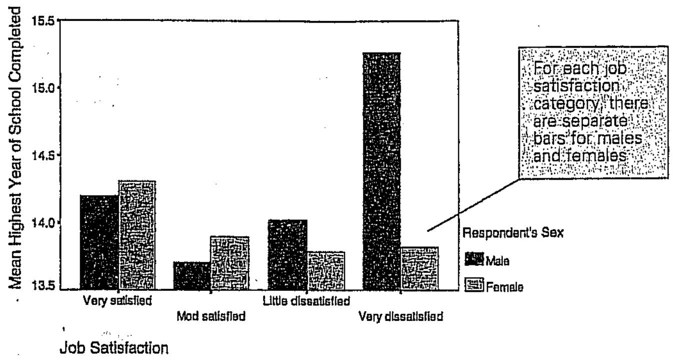
Figure 5.4 is a bar chart that displays the results of Figure 5.3. There are four sets of bars corresponding to the job satisfaction categories. Each

NORUŠIS, Marija J. 1998. *Guide to Data Analysis*. Upper Saddle River, Prentice Hall.

set of bars has separate bars for males and for females. The conclusions we reached based on the summary table are easier to see from this display. By looking at a corresponding pair of bars, you can see if the average years of education are similar for men and women within each category of job satisfaction. (Boxplots, which are a better way of comparing summary statistics for groups of cases, are described in Chapter 6.)

You can obtain this clustered bar chart using the Graphs menu, as described on p. 500. Select the variables educ, satjob, and sex in the Define Clustered Bar Summaries for Groups of Cases dialog box.

Figure 5.4 Bar chart of education by job satisfaction and sex



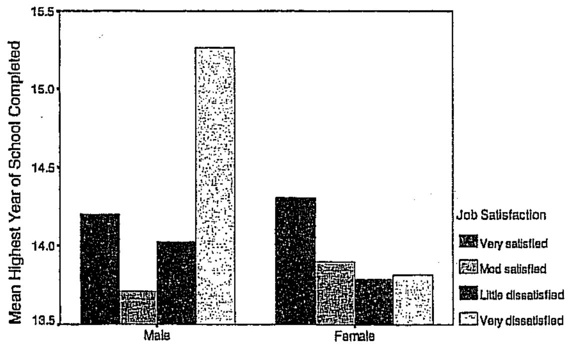
For each job satisfaction category, there are separate bars for males and females

You can also group all of the bars for men and all of the bars for women together, as shown in Figure 5.5.

You can obtain this chart by modifying Figure 5.4, as described in "Bar Charts" on p. 520 in Appendix A. Activate the chart into a chart editor window and from the menus choose:

*Series
  Transpose Data*

Figure 5.5 Bar chart of education by sex and job satisfaction



Now you have two subgroups: men and women. Within each subgroup, you see the four categories of job satisfaction. This plot makes it easy to see that the relationship between job satisfaction and education is not the same for men and women.

Means and standard deviations for groups of cases can be displayed with error bar charts. See "Error Bar Charts" on p. 523 in Appendix A.

What problems are associated with calculating statistics for subgroups of cases? As the number of subgroups you want to compare increases, the sample size in each of the subgroups diminishes. When your means are based on a small number of cases, they are not very reliable. That is, the subgroup means can change substantially if you select another sample from the same population. You'll learn more about the variability of sample means in Part 3.

Summary

How can you determine if the values of the summary statistics for a variable differ for subgroups of cases?

- Subgroups are formed when cases are subdivided into groups based on the values of one or more variables.
- By calculating summary statistics separately for subgroups of cases, you can see if there is a relationship between the summary statistics and the subgroups.
- You can make bar charts of the means of a variable for different subgroups.

NORUŠIS, Marija J. 1998. *Guide to Data Analysis*. Upper Saddle River, Prentice Hall.

13

Testing a Hypothesis about Two Independent Means

How can you test the null hypothesis that two population means are equal, based on the results observed in two independent samples?

- Why can't you use a one-sample t test?
- What assumptions are needed for the two independent-samples t test?
- Can you prove the null hypothesis is true?
- What is power, and why is it important?

You know how to test whether a single sample of data comes from a population with a known mean. You've tested whether the average cholesterol level for CEO's is the same as the average for the general population, whether college graduates work a 40-hour week on average, and whether the average change in β-endorphin values is 0 during a half-marathon run. In Chapter 12, although you had pairs of observations, you analyzed the differences between the two values and tested the hypothesis that these differences come from a population with a mean of 0.

In this chapter, you'll learn how to test whether two population means are equal based on the results observed in two independent samples—one from each of the populations of interest. You'll use a statistical technique called the two independent-samples t test. You can use the two independent-samples t test to see if, in the population, men and women have the same scores on a test of physical dexterity or if two treatments for high cholesterol result in the same mean cholesterol levels.

▶ This chapter uses the *gssft.sav* data file, which includes only cases for people holding full-time jobs. For instructions on how to obtain the independent-samples t test output shown in this chapter, see "How to Obtain an Independent-Samples T Test" on p. 250.

Looking at Age Differences

In Part 2 of this book, you examined the relationship between job satisfaction, age, and education for full-time employees. You saw that the average values of age and education vary among the different job satisfaction groups. That isn't surprising, since you know that even if the average ages and educational levels in the population are the same for all job satisfaction groups, the sample means will not be equal. Different samples from the same population result in different sample means and standard deviations. To determine if any of the observed sample differences among groups might be real, that is, not simply the result of the usual variability of sample means from a single population, you need to determine if the observed sample means would be unusual when the population means are equal.

Let's consider what happens if you form two independent groups of people—those who are very satisfied with their jobs and those who are not. You want to determine whether the population values for average age and average education are the same for the two groups. First we'll look at age.

What do you mean by independent groups? Samples from different groups are called **independent** if there is no relationship between the people or objects in the different groups. For example, if you select a random sample of males and a random sample of females from a population, the two samples are independent. That's because selecting a person for one group in no way influences the selection of a person for another group. The two groups in a paired design are not independent, since either the same people or closely matched people are in both groups.

Since you have means from two independent groups, you can't use the one-sample t test to test the null hypothesis that two population means are equal. That's because you now have to cope with the variability of two sample means: the mean for *very satisfied* people and the mean for the *not very satisfied* people. When you test whether a single sample comes from a population with a known mean, you only have to worry about how much individual means from the same population vary. The population value to which you compare your sample mean is a fixed, known number. It doesn't vary from sample to sample. You assumed that the value of 205 mg/dL for the cholesterol of the general population is an established norm based on large-scale studies. Similarly, the value of 40 hours for a work week is a commonly held belief.

NORUŠIS, Marija J. 1998. *Guide to Data Analysis.* Upper Saddle River, Prentice Hall.

The two independent-samples *t* test is basically a modification of the one-sample *t* test that incorporates information about the variability of the two independent-sample means. The standard error of the mean difference is no longer estimated from the variance and number of cases in a single group. Instead, it is estimated from the variances and sample sizes of the two independent groups.

Descriptive Statistics

Look at Figure 13.1, which shows descriptive statistics for the age variable, when full-time workers are classified into one of two distinct groups—the *very satisfied* and the *not very satisfied.*

Figure 13.1 Descriptive statistics for age by job satisfaction category

Age of Respondent

Job Satisfaction	Mean	Std. Deviation	N
Very satisfied	41.50	11.54	325
Not very satisfied	39.57	10.79	419
Total	40.41	11.16	744

You can obtain these statistics using the Means procedure, which is described in Chapter 5. Select the variables age and satjob2 in the Means dialog box.

You see that the average age of the *very satisfied* group is 41.5 years, while the average age of the *not very satisfied* group is 39.6. The standard deviation of the *very satisfied* group, 11.5 years, is slightly larger than the standard deviation of the *not very satisfied* group, 10.8 years. In the General Social Survey sample, the *very satisfied* people are on average 1.9 years older than those less content with their jobs. Based on these sample results, what can you reasonably conclude about the population of American adults who are employed full time? Can you conclude that there is a difference in average ages between the two groups?

Distribution of Differences

To answer this question, you have to determine if your observed age difference would be unusual if the two populations have the same average age. In the previous chapters, you answered similar questions by looking at the distribution of all possible means from a population. Now you'll look at the distribution of all possible *differences* between sample means from two independent groups.

Fortunately, the Central Limit Theorem works for differences of sample means as well as for individual means. So if your data are samples from approximately normal populations, or your sample size is large enough so that the Central Limit Theorem holds, the distribution of differences between two sample means is also normal. It's always a good idea to obtain stem-and-leaf plots or histograms for each of the two groups. From these, you can tell what the distribution of values looks like.

Standard Error of the Mean Difference

If two samples come from populations with the same mean, the mean of the distribution of differences is 0. However, that's not enough information to determine if the observed sample results are unusual. You also need to know how much the sample differences vary. The standard deviation of the difference between two sample means, the **standard error of the mean difference**, tells you that. When you have two independent groups, you must estimate the standard error of the mean difference from the standard deviations and the sample sizes in each of the two groups.

How do I estimate the standard error of the difference? The formula is

$$S_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$$

where S_1^2 is the variance for the first sample and S_2^2 is the variance for the second sample. The sample sizes for the two samples are n_1 and n_2. If you look carefully at the formula, you'll see that the standard error of the mean difference depends on the standard errors of the two sample means. You square the standard error of the mean for each of the two groups. Next you sum them, and then take the square root. ■■■

Computing the T Statistic

Once you've estimated the standard error of the mean difference, you can compute the *t* statistic the same way as in the previous chapters. You divide the observed mean difference by the standard error of the difference. This tells you how many standard error units from the population mean of 0 your observed difference falls. That is,

NORUŠIS, Marija J. 1998. *Guide to Data Analysis.* Upper Saddle River, Prentice Hall.

$$t = \frac{(\bar{X}_1-\bar{X}_2)-0}{S_{\bar{X}_1-\bar{X}_2}} \qquad \textbf{Equation 13.1}$$

If your observed difference is unlikely when the null hypothesis is true, you can reject the null hypothesis.

How is this different from the one-sample t *test?* The idea is exactly the same. What differs is that you now have two independent-sample means, not one. So you estimate the standard error of the mean difference based on two sample variances and two sample sizes. ■■■

Output from the Two Independent-Samples T Test

Look at Figure 13.2, which shows the results from SPSS of testing the null hypothesis that in the population the average age of *very satisfied* full-time workers and *not very satisfied* full-time workers is the same.

Figure 13.2 Independent-samples t test of age by job satisfaction

			Age of Respondent	
			Equal variances assumed	Equal variances not assumed
Levene's Test for Equality of Variances	F		.377	
	Sig.		.540	
t-test for Equality of Means	t		2.347	2.327
	df		742	672.439
	Sig. (2-tailed)		.019	.020
	Mean Difference		1.93	1.93
	Std. Error Difference		.82	.83
	95% Confidence Interval of the Mean	Lower	.32	.30
		Upper	3.54	3.56

There is only a 1.9% chance of observing a mean difference at least this large if the null hypothesis is true.

The mean age for the two samples differs by 1.93 years.

To obtain this output, from the menus choose:

Statistics
Compare Means ▸
Independent Samples T Test...

Select the variables age and satjob2, as shown in Figure 13.8.

In the output, there are two slightly different versions of the *t* test. One makes the assumption that the variances in the two populations are equal; the other does not. This assumption affects how the standard error of the mean difference is calculated. You'll learn more about this distinction later in this chapter.

Consider the column labeled *Equal variances assumed.* You see that for the observed difference of 1.93 years, the *t* statistic is 2.35. (To calculate the *t* statistic, divide the observed difference of 1.93 by 0.82, the standard error of the difference estimate when the two population variances are assumed to be equal.) The degrees of freedom for the *t* statistic are 742, the sum of the sample sizes in the two groups minus 2.

The observed two-tailed significance level is 0.019. This tells you that only 1.9% of the time would you expect to see a sample difference of 1.93 years or larger, when the two population means are equal. Since 1.9% is less than 5%, you reject the null hypothesis that the two groups of workers come from populations with the same average age. Your observed results are unusual if the null hypothesis is true.

Confidence Intervals for the Mean Difference

Take another look at Figure 13.2. The 95% confidence interval for the true difference is from 0.32 years to 3.54 years. This tells you it's likely that the true mean difference is anywhere from a third of a year to slightly more than three and one-half years. Since your observed significance level for the test that the two population means are equal was less than 5%, you know that the 95% confidence interval will not contain the value of 0. (Remember, only likely values are included in a confidence interval. Since you found 0 to be an unlikely value, it won't be included in the confidence interval.)

If I compute a 99% confidence interval for the true mean difference, will it also not include 0? The 99% confidence interval for the mean difference extends from –0.194 to 4.053. This interval does include the value of 0. That's because your observed significance level is greater than 1%. If your criterion for unusual is 1 in a 100 or less, you cannot reject the null hypothesis based on the *t* test or on the corresponding 99% confidence interval for the mean difference. ■■■

To calculate a 99% confidence interval, specify 99 in the T Test Options dialog box (see Figure 13.10).

NORUŠIS, Marija J. 1998. *Guide to Data Analysis*. Upper Saddle River, Prentice Hall.

Another Way of Looking at It

You found a small, but statistically significant, age difference between people who are *very satisfied* with their jobs and those who aren't. Since the observed sample difference is less than two years, it's tempting to dismiss this finding as not particularly interesting. However, there are many different ways you can look at the relationships between the two variables. Sometimes uninteresting information can become more interesting when looked at in another way.

You can obtain crosstabulations using the Crosstabs procedure, as discussed in Chapter 7.

Figure 13.3 Crosstabulation of job satisfaction and age

			Job Satisfaction		
			Very satisfied	Not very satisfied	Total
	18-29	Count	46	91	137
		% within age in four categories	**33.6%**	66.4%	100.0%
	30-39	Count	112	131	243
		% within age in four categories	46.1%	53.9%	100.0%
	40-49	Count	89	121	210
		% within age in four categories	42.4%	57.6%	100.0%
	50+	Count	78	76	154
		% within age in four categories	**50.6%**	49.4%	100.0%
Total		Count	325	419	744
		% within age in four categories	43.7%	56.3%	100.0%

33.6% of people less than 30 are very satisfied with their jobs

50.6% of people age 50 and over are very satisfied

Look at Figure 13.3, which is a crosstabulation of the two categories of job satisfaction and four categories of age. From the row percentages, you see that only 33.6% of people less than 30 are *very satisfied* with their jobs, while over 50% of people age 50 and over are *very satisfied*. Overall, 43.7% of full-time workers claim to be *very satisfied* with their jobs. The crosstabulation provides findings that are more interesting and easier to interpret. (In Chapter 16, you'll learn how to test hypotheses that the two variables in a crosstabulation are independent.)

Testing the Equality of Variances

You saw that there are two different *t* values in Figure 13.2. That's because there are two different ways to estimate the standard error of the difference. One of them assumes that the variances are equal in the two populations from which you are taking samples, the other one does not.

In Figure 13.1, you see that the observed standard deviations in the two samples are fairly similar. You can test the null hypothesis that the two samples come from populations with the same variances using the Levene test, which is shown in Figure 13.4. If the observed significance level for the Levene test is small, you can reject the null hypothesis that the two population variances are equal.

For this example, you can't reject the equal variances hypothesis, since the observed significance level for the Levene test is 0.54. That means you can use the results labeled *equal variances* in Figure 13.4.

To obtain this output, select the variables age and satjob2 in the Independent-Samples T Test dialog box, as shown in Figure 13.8.

Figure 13.4 Levene test for equality of variances

			Age of Respondent	
			Equal variances assumed	Equal variances not assumed
Levene's Test for Equality of Variances	F		.377	
	Sig.		.540	
t-test for Equality of Means	t		2.347	2.327
	df		742	672.439
	Sig. (2-tailed)		.019	.020
	Mean Difference		1.93	1.93
	Std. Error Difference		.82	.83
	95% Confidence Interval of the Mean	Lower	.32	.30
		Upper	3.54	3.56

You don't reject the hypothesis that the two population variances are equal based on the Levene test

If the Levene test leads you to reject the null hypothesis that the two population variances are equal, or if you are unsure, you should use the results from the column labeled *Equal variances not assumed* in Figure 13.4. Notice that the estimate of the standard error of the difference is not the same in the two columns. This affects the *t* value and confidence interval. When you use the estimate of the standard error of the difference that does not assume that the two variances are equal, the degrees of freedom for the *t* statistic are no longer the sum of the two sample sizes minus two. They are calculated based on both the sample sizes and the standard deviations in each of the groups. In this example, both *t* tests give very similar results, but that's not always the case.

Why do you get different numbers for the standard error of the mean difference depending on the assumptions you make about the population variances? If you assume that the two population variances are equal, you can compute what's called a pooled estimate of the variance. The idea is similar to that of averaging the variances in the two groups, taking into account the sample size. The formula for the pooled variance is

$$s^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1-1)+(n_2-1)}$$

It is this pooled value that is substituted for both S_1^2 and S_2^2 in the equation on p. 236. If you do not assume that the two population variances are equal, the individual sample variances are used in the equation on p. 236.

Comparing Education

From the previous analysis, you concluded that there appears to be a difference in average ages between those who are *very satisfied* with their jobs and those who are not. Younger people tend to be less satisfied with their jobs than older people. Now consider education. As always, your first step should be to look at the data values in the two groups. Look at the distributions, and try to see if there's anything unusual going on. For small sample sizes, see if the distribution of data values is approximately normal.

Figure 13.5 contains the descriptive statistics for the two groups. You see that the very satisfied people are somewhat better educated (or at least went to school longer) than those who are *not very satisfied*. The difference is slightly more than a third of a year. Figure 13.6 shows the distribution of education graphically. From the boxplot, you see that the variability of the two groups is similar. Since the median for the *very satisfied* group is in the middle of the box, the distribution of values for the group is more or less symmetric. One case stands out in the plot. That's the person who claims no formal education. For the *not very satisfied* group, the median is close to the bottom edge, indicating that there is a tail toward higher educational levels. If your sample is small, and the departures from normality are severe, you may want to substitute one of the nonparametric tests described in Chapter 17 for the independent-samples *t* test. In this case, the sample sizes are large, so the independent-samples *t* test should work just fine.

You can obtain these statistics using the Means procedure, as discussed in Chapter 5.

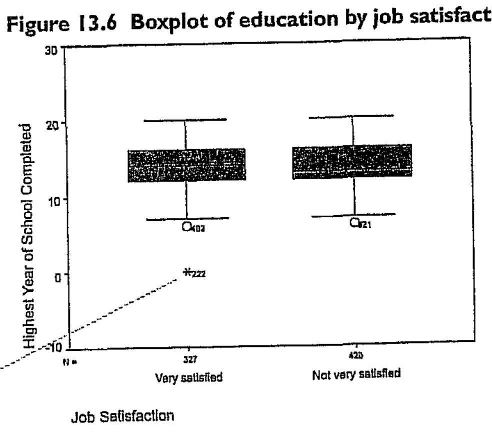
Figure 13.5 Descriptive statistics for education by job satisfaction

Highest Year of School Completed

Job Satisfaction	Mean	N	Std. Deviation
Very satisfied	14.25	327	2.79
Not very satisfied	13.87	420	2.62
Total	14.04	747	2.70

You can obtain this boxplot using the Explore procedure, as discussed in Chapter 6.

Figure 13.6 Boxplot of education by job satisfaction



Respondent with no formal education

Based on the Levene test in Figure 13.7, there is no reason to doubt that the population variances are equal, so you can use the *t* value in the column labeled *Equal variances assumed* to test the null hypothesis that in the population, the average years of education are the same for those who are *very satisfied* with their jobs and those who are not. The two-tailed significance level is 0.057, so you don't reject the null hypothesis. As expected, the 95% confidence interval for the mean difference includes the

NORUŠIS, Marija J. 1998. *Guide to Data Analysis*. Upper Saddle River, Prentice Hall.

value of 0. (The lower bound of the 95% confidence interval is given in scientific notation. The lower limit is –0.011 years.)

What kind of number has an E in it? When SPSS displays a very small or very large number, it uses scientific notation. The number that follows the letter E tells you how many places the decimal must be moved. If the number following E is negative, move the decimal to the left. If the number following E is positive, move the decimal point to the right. For example, –1.1E–02 is –0.011; –1.1E02 is –110.

If you don't like the format SPSS uses to display a number, activate the pivot table, select the cell, and from the menus choose:

Format
    Cell properties...

Select the format you prefer in the Value tab of the Cell properties dialog box.

To obtain this output, select the variables educ and satjob2 in the Independent-Samples T Test dialog box. (See Figure 13.8.) Then activate the pivot table and from the menus choose:

Pivot
  Transpose Rows and Columns

Figure 13.7 Independent-samples t test of education by job satisfaction

			Highest Year of School Completed	
			Equal variances assumed	Equal variances not assumed
Levene's Test for Equality of Variances	F		.261	
	Sig.		.609	
t-test for Equality of Means	t		1.908	1.892
	df		746	677.434
	Sig. (2-tailed)		.057	.059
	Mean Difference		.38	.38
	Std. Error Difference		.20	.20
	95% Confidence Interval of the Mean	Lower	-1.10E-02	-1.42E-02
		Upper	.77	.77

7. lekce

JAK TESTOVAT NULOVOU HYPOTÉZU O SHODĚ NĚKOLIKA POPULAČNÍCH PRŮMĚRŮ.

One-Way Analysis of Variance

14

How can you test the null hypothesis that several population means are equal?

- What is analysis of variance?
- What assumptions about the data are needed to use analysis-of-variance techniques?
- How is the *F* ratio computed, and what does it tell you?
- Why do you need multiple comparison procedures?

You've already learned how to test hypotheses about two population means using the paired-samples *t* test and the independent-samples *t* test. Often, however, you want to compare more than two population means. For example, if you are studying four methods for teaching mathematics, you want to compare average test scores for all four groups. Or, if you are testing seven different treatments for lowering cholesterol, you may want to compare the average final cholesterol levels for all seven methods. In this chapter, you'll learn how to test the null hypothesis that several independent population means are equal. The technique you'll use is called **analysis of variance**, usually abbreviated as ANOVA.

▶ This chapter uses the *gssft.sav* data file, which includes only people holding full-time jobs. For instructions on how to obtain the One-Way ANOVA output shown in the chapter, see "How to Obtain a One-Way Analysis of Variance" on p. 274.

Hours in a Work Week

In Chapter 11, you looked at the average number of hours worked in a week by college graduates. Based on the results from the General Social Survey, you rejected the null hypothesis that the average work week is 40 hours. You found the 95% confidence interval for the population value for the average number of hours worked to be from 46.16 hours to 49.30 hours. So it's not inconceivable that the average college graduate works almost an extra 8-hour day each week.

An obvious question that arises is whether it's just college graduates who suffer from the expansion of the work week, or is everyone, regardless of educational background, working more? Using the General Social Survey, you can look at the average number of hours worked by full-time employees of various educational backgrounds.

Describing the Data

You see in Figure 14.1 that the average work week for all full-time employees is 46.29 hours. (It's the entry in the column labeled *Total*.) The average work week ranges from a low of 43.69 hours for people without a high school diploma to a high of 50.27 hours for people with graduate degrees.

Figure 14.1 Descriptive statistics for hours worked

To obtain this output, from the menus choose:

Statistics
Compare Means ▶
One-Way ANOVA...

Select the variables hrs1 and degree, as shown in Figure 14.5.

In the One-Way ANOVA Options dialog box, select Descriptive, as shown in Figure 14.7.

		HRS1 Number of Hours Worked Last Week					
		DEGREE RS Highest Degree					
		Less than HS	High school	Junior college	Bachelor	Graduate	Total
N		52	387	54	162	86	741
Mean		43.69	45.77	45.87	46.38	50.27	46.29
Std. Deviation		8.72	10.58	11.66	12.89	11.44	11.27
Std. Error		1.21	.54	1.59	1.01	1.23	.41
95% Confidence Interval for Mean	Lower Bound	41.26	44.72	42.69	44.38	47.82	45.48
	Upper Bound	46.12	46.83	49.05	48.38	52.72	47.10
Minimum		20	8	25	5	34	5
Maximum		70	80	80	89	89	89

The average work week ranges from 43.69 hours to 50.27 hours

In the row labeled *Std. Deviation*, you see that the smallest variability in hours worked is for people with less than a high school diploma, while the largest is for people with bachelor's degrees. The next column, labeled *Standard Error*, tells you how much the sample means vary in repeated samples from the same population. For each group, it's the standard deviation divided by the square root of the sample size. The smallest standard error is for high school graduates, since they are the largest group.

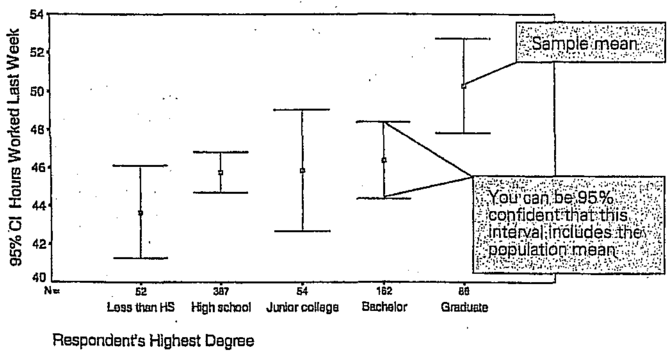
Confidence Intervals for the Group Means

In the last two columns of Figure 14.1, you see for each group the 95% confidence interval for the population value of the average hours worked per week. You are 95% confident that the true work week for those with less than a high school diploma is between 41.26 and 46.12 hours. For those with a graduate degree, you are 95% confident that the true average work week is between 47.82 and 52.72 hours.

Figure 14.2 Plot of sample means and 95% confidence intervals

You can obtain this error bar chart using the Graphs menu, as described in "Error Bar Charts" on p. 523 in Appendix A.

In the Define Simple Error Bar Summaries for Groups of Cases dialog box, select the variables hrs1 and degree.



Sample mean

You can be 95% confident that this interval includes the population mean

Plots of the means and confidence intervals are shown in Figure 14.2. You see that the 95% confidence interval for high school graduates is the narrowest. That's because there are so many of them in the sample. Many of the confidence intervals in Figure 14.2 overlap. That tells you that some of the values that are plausible for the true work week in one group are also plausible for the true work week in the others. The exception is the confidence interval for those with graduate degrees. It doesn't overlap the confidence interval for those with less than a high school education, nor the interval for those with a high school education.

Can you tell from the plot if the 40-hour work week is a reasonable guess for the true hours worked per week? Sure. Remember, if a value doesn't fall in the 95% confidence interval for the mean, you can reject the hypothesis that it's a plausible population value. You see in Figure 14.2 that the value 40 is not included in any of the confidence intervals. That means you can reject the hypothesis that it's a reasonable value for any of the groups. It appears that the 40-hour work week may be a thing of the past, regardless of your education level. ■■■

Testing the Null Hypothesis

The descriptive statistics and plots suggest that there are differences in the average work week among the five education groups. Now you need to figure out whether the observed differences in the samples may be attributed to just the natural variability among sample means or whether there's reason to believe that some of the five groups have different values in the population for average hours worked.

The null hypothesis says that the population means for all five groups are the same. That is, there is no difference in the average hours worked for people in the five education categories. The alternative hypothesis is that there is a difference. The alternative hypothesis doesn't say which groups differ from one another. It just says that the groups means are not all the same in the population; at least one of the groups differs from the others.

The statistical technique you'll use to test the null hypothesis is called analysis of variance (abbreviated ANOVA). It's called analysis of variance because it examines the variability of the sample values. You look at how much the observations within each group vary as well as how much the group means vary. Based on these two estimates of variability, you can draw conclusions about the population means. If the sample means vary more than you expect based on the variability of the observations in the groups, you can conclude that the population means are not all equal.

SPSS contains several different procedures that perform analysis of variance. In this chapter, you'll use the One-Way ANOVA procedure. It's called one-way analysis of variance because cases are assigned to different groups based on their values for one variable. In this example, you form the groups based on the values of the *degree* variable. The variable used to form groups is called a factor. In Chapter 15, you'll learn how to test hypotheses when cases are classified into groups based on their values for two factors.

Assumptions Needed for Analysis of Variance

Analysis of variance requires the following assumptions:

- Independent random samples have been taken from each population.
- The populations are normal.
- The population variances are all equal.

The Kruskal-Wallis test, described in Chapter 17, requires more limited assumptions about the data.

Independence. The independence assumption means that there is no relationship between the observations in the different groups and between the observations in the same group. For example, if you administer four different treatments to each individual, you cannot use the one-way analysis-of-variance procedure to analyze the data. Observations from the same individual appear in each of the groups, so they are not independent. (In this situation, you must use an extension of the paired-samples *t* test. It's called repeated measures analysis of variance, a topic not covered in this book.) Observations within a group are also not independent if conditions are changing with time. For example, if you are explaining a task to subjects and your instructions get better with time, early subjects may not perform as well as later subjects. In this situation, the response of the subject depends on the point in time he or she entered into the study. Consecutive subjects will be similar to each other.

Normality. The normality assumption in analysis of variance can be checked by making histograms or normal probability plots for each of the groups. In practice, the analysis of variance is not heavily dependent on the normality assumption. As long as the data are not extremely non-normal, you do not have to worry. (If your sample sizes in the groups are small, you should be aware of the impact of unusual observations, which can have a big effect on the mean and standard deviation. You can rerun the analysis without the unusual point to make sure that you reach the same conclusions.)

Equality of Variance. The equality of variance assumption can be checked by examining the spread of the observations in the box plot. You can also compute the Levene test for equality of variance, which is available in the Explore and One-Way procedures. In practice, if the number of cases in each of the groups is similar, the equality of variance assumption is not too important.

What should I do if I suspect that my data violate the necessary assumptions? Well, it depends on which assumption is being violated. For example, if you're worried about the normality or equal-variance assumptions, sometimes you can transform your data so that the distribution of values is more normal or the variances in the groups are more similar. Taking logarithms or square roots of the data values is often helpful. If this fails, you can use a statistical test that makes fewer assumptions about the data. In particular, you may want to use the Kruskal-Wallis test described in Chapter 17.

The situation is considerably more complicated if you're worried about whether the groups are somehow biased. That is, you're concerned that one or more of your samples differs in some important way from the population of interest. For example, if you want to compare four medical treatments, and the participating physicians have assigned the sickest patients to a particular group, you've got a real problem. You may not be able to draw any correct conclusions from your data. That's why it's very important when comparing several treatments or conditions, to make sure that the subjects are randomly assigned to the different groups. *Randomly* doesn't mean *haphazardly.* It means that you must have a well organized system for random assignment of cases. ■■■

Analyzing the Variability

In analysis of variance, the observed variability in the sample is divided (partitioned, in statistical lingo) into two parts: variability of the observations *within* a group about the group mean, and variability *between* the group means.

Why are we talking about variability? Aren't we testing hypotheses about means? Yes, we're testing hypotheses about population means; but as you've seen in previous chapters, your conclusions about population means are always based on looking at the variability of sample means. You have to determine if your sample mean is outside the usual range of variability of sample means from the population.

In analysis of variance, you'll look at how much your observed sample means vary. You'll compare this observed variability to the expected variability if the null hypothesis that all population means are the same is true. If the sample means vary more than you'd expect, you have reason to believe that this extra variability is because some of groups don't have the same population mean. (If you have two independent groups, you'll get the same results using ANOVA or the equal variance *t* test.) ■■■

Let's look a little more closely now at the two types of variability and how they are used to test the null hypothesis that the population values for average hours worked per week are the same for people in the five education categories. The game plan is as follows: You want to know whether your sample means vary more than you would expect if the null hypothesis is true. First, you'll see how much the observations in a group vary, and then you'll see how much the sample means vary. If the sample means vary more than you expect, you'll reject the null hypothesis.

Within-Groups Variability

The within-groups estimate of variability, as its name suggests, tells you how much the observations within a group vary. The sample variance of each group estimates within-groups variability. One of the assumptions of analysis of variance is that all groups come from populations with the same variance. That makes it possible for you to average the variances in each of the groups to come up with a single number, which is the within-groups variance. (You'll see later how this averaging is done. You can't just add up the sample variances and divide by the number of groups.)

You might wonder why you can't just put all of your observations together and compute the variance. The reason is that you don't know if all of the groups have the same population mean. If they don't, pooling all the values together will give you the wrong answer. For example, suppose that all people without a high school diploma work exactly 40 hours a week; all people with a high school diploma work exactly 43 hours a week; and all people with a college degree work exactly 45 hours. The variance in each of the groups is 0, since the values within a group don't vary at all. The correct estimate of the within-groups variance is also 0. If you compute the variance for all cases together, it wouldn't be close to 0. The observed variability would be the result of differences in the means of the three groups.

Between-Groups Variability

You have a sample mean for each of the groups in your study. If all of the groups have the same number of cases, you can find the standard deviation of the sample means. What would that tell you? If all the groups come from populations with the same mean and variance, the standard deviation of the sample means tells you how much sample means from the same population vary. The standard deviation of the sample means is an estimate of the standard error of the mean.

From the standard error of the mean, you can estimate the standard deviation of the observations. You do this by multiplying the standard error of the mean by the square root of the number of cases in a group.

Where did that come from? The standard error of the mean is the standard deviation of the observations divided by the square root of the sample size. So, using simple algebra, the standard deviation is the standard error of the mean multiplied by the square root of the sample size. Thus,

$$\text{standard error} = \frac{\text{standard deviation}}{\sqrt{\text{sample size}}}$$

and

$$\text{standard deviation} = \text{standard error} \times \sqrt{\text{sample size}}$$ ■■■

If you square the estimate of the standard deviation, you have a quantity that's called the **between-groups estimate of variability**. It's called the between-groups estimate of variability because it's based on how much sample means vary *between* the groups.

Comparing the Two Estimates of Variability

You now have two estimates of how much the observations within a group vary: the within-groups estimate and the between-groups estimate. These two estimates differ in a very important way: the between-groups estimate of variance will be correct only if the null hypothesis is true. If the null hypothesis is false, the between-groups estimate of variance will be too large. The observed variability of the sample means will be the result of two factors: the variability of the observations within a group and the variability of the population means. The within-groups estimate of variability doesn't depend on the null hypothesis being true. It's always a good estimate.

Your decision about the null hypothesis will be based on comparing the between-groups and the within-groups estimates of variability. You'll see how much the number of hours worked varies for individuals in the same education group. This will give you the within-groups estimate of variability. Then you'll see how much the means of the five groups vary. Based on this, you'll calculate the between-groups estimate of variability. If the between-groups estimate is sufficiently larger than the within-groups estimate, you'll reject the null hypothesis that all of the means are equal in the population.

The Analysis-of-Variance Table

The estimates of variability that we've been talking about are usually displayed in what's called an **analysis-of-variance table**. Figure 14.3 is the analysis-of-variance table for the test of the null hypothesis that the population value for average hours worked per week is the same for people in five categories of education. By looking at this table, you'll be able to tell whether you have enough evidence to reject the null hypothesis.

tain this output,
: the variables
ınd degree in the
Nay ANOVA
ı box, as shown
ure 14.5.

Figure 14.3 Analysis-of-variance table

		Sum of Squares	df	Mean Square	F	Sig.
Number of Hours Worked Last Week	Between Groups	1825.917	4	456.479	**3.646**	**.006**
	Within Groups	92148.280	736	125.201		
	Total	93974.197	740			

Ratio of mean squares

Probability of obtaining F ratio at least this large when null hypothesis is true

The two estimates of variability are shown in the column labeled *Mean Square*. Their ratio is in the column labeled *F*. If the null hypothesis is true, you expect the ratio of the between-groups mean square to the within-groups mean square to be close to 1, since they are both estimates of the population variance. Large values for the *F* ratio indicate that the sample means vary more than you would expect if the null hypothesis were true.

You can tell if your observed *F* ratio of 3.65 is large enough for you to reject the null hypothesis by looking at the observed significance level, which is labeled *Sig.* You see that the probability of obtaining an *F* ratio of 3.65 or larger when the null hypothesis is true is 0.006. Only 6 times in 1000, when the null hypothesis is true, would you expect to see a ratio this large or larger. So you can reject the null hypothesis. It's unlikely that the number of hours worked per week is the same for the five groups in the population.

Now that you know the punch line, let's see where all the numbers are coming from.

Estimating Within-Groups Variability

You need three steps to compute the within-groups estimate of variability:

1. First, you must compute what's called the **within-groups sum of squares**. Take all of the standard deviations in Figure 14.1 and square them to obtain variances. Then multiply each variance by one less than the number of cases in the group. Finally, add up the values for all of the groups. The within-groups sum of squares is:

$$(8.72^2 \times 51) + (10.58^2 \times 386) + (11.66^2 \times 53) + (12.89^2 \times 161) + (11.44^2 \times 85) = 92148.28 \qquad \textbf{Equation 14.1}$$

You see this number in the second row of Figure 14.3 in the column labeled *Sum of Squares*. (You have to use more decimal places for the standard deviation than shown above to get exactly the answer given.)

2. Next, you must compute the degrees of freedom. That's easy to do. For each group, you compute the number of cases minus 1, and then add up these numbers for all of the groups. In this example, the degrees of freedom are:

$$\text{degrees of freedom} = 51 + 386 + 53 + 161 + 85 = 736 \qquad \textbf{Equation 14.2}$$

This number is shown in the *Within Groups* row of Figure 14.3, in the column labeled *df* (for degrees of freedom).

3. Finally, divide the sum of squares by its degrees of freedom, to get what's called a **mean square**. This is the estimate of the average variability in the groups. It's really nothing more than an average of the variances in each of the groups, adjusted for the fact that the number of observations in the groups differs. Your estimate of the variance for the number of hours worked, based on the variability of the observations within each of the groups, is 125.20.

Estimating Between-Groups Variability

You also need three steps to calculate the between-groups estimate of variability.

1. First, you compute the **between-groups sum of squares**. Subtract the overall mean (the mean of all of the observations) from each group mean. Then square each difference, and multiply the square by the

number of observations in its group. Finally, add up all the results. For this example, the between-groups sum of squares is:

$$\begin{aligned} & 52 \times (43.69 - 46.29)^2 \\ & + 387 \times (45.77 - 46.29)^2 \\ & + 54 \times (45.87 - 46.29)^2 \qquad \textbf{Equation 14.3} \\ & + 162 \times (46.38 - 46.29)^2 \\ & + 86 \times (50.27 - 46.29)^2 = 1825.92 \end{aligned}$$

2. Next, you must compute the degrees of freedom. The degrees of freedom for the between-groups sum of squares is just the number of groups minus 1. In this example, there are five education groups, so the degrees of freedom for the between-groups sum of squares is 4.

3. Finally, calculate the between-groups mean square by dividing the between-groups sum of squares by its degrees of freedom. The between-groups mean square is 456.48.

Calculating the F Ratio

You now have the two estimates of the variability in the population: the within-groups mean square and the between-groups mean square. The *F* ratio is simply the ratio of these two estimates. Take the between-groups mean square and divide it by the within-groups mean square:

$$F = \frac{\text{between-groups mean square}}{\text{within-groups mean square}} = \frac{456.48}{125.20} = 3.65 \qquad \textbf{Equation 14.4}$$

(Remember, the within-groups mean square is based on how much the observations within each of the groups vary. The between-groups mean square is based on how much the group means vary among themselves.) If the null hypothesis that the average hours worked per week is the same for the five groups is true, the two numbers should be close to each other. If you divide one by the other, the result should be close to 1.

As you see, the ratio of the two estimates is not 1. Does that mean you automatically reject the null hypothesis? No. You know that your sample ratio will not be exactly 1, even if the null hypothesis is true. You need to figure out how often you would expect to see a sample value of 3.65 or greater when the null hypothesis is true. That is, you need to determine whether your sample results are unlikely if the null hypothesis is true.

The observed significance level is calculated by comparing your observed *F* ratio to values of the *F* distribution. The observed significance level depends on both the observed *F* ratio and the degrees of freedom for the two mean squares.

What's the F *distribution?* Like the normal and *t* distributions, the *F* distribution is defined mathematically. It's used when you want to test hypotheses about population variances. The Central Limit Theorem doesn't work for variances. Their distributions are not normal. The ratio of two sample variances from normal populations has an *F* distribution. The *F* distribution is indexed by two values for the degrees of freedom, one for the numerator and one for the denominator. The degrees of freedom depend on the number of observations used to calculate the two variances.

In Figure 14.3, you see that the observed significance level for this example is 0.006. Since the value is small, you can reject the null hypothesis that the average hours worked per week in the population is the same for the five groups. The observed sample results are not likely to occur when the null hypothesis is true.

Multiple Comparison Procedures

A statistically significant *F* ratio tells you only that it appears unlikely that all population means are equal. It doesn't tell you which groups are different from each other. You can reject the null hypothesis that all population means are equal in a variety of situations. For example, it may be that the average hours worked differs for all of the five groups. Or it may be that only one or two of the groups differ from the rest. Usually when you've rejected the null hypothesis, you want to pinpoint exactly where the differences are. To do this, you must use multiple comparison procedures.

Why do you need yet another statistical technique? Why can't you just compare all possible pairs of means using t *tests?* The reason for not using many *t* tests is that when you make many comparisons involving the same means, the probability increases that one or more comparisons will turn out to be statistically significant, even when all the population means are equal. This is known as the **multiple comparison problem.**

For example, if you have 5 groups and compare all pairs of means, you're making 10 comparisons. When the null hypothesis is true, the probability that at least 1 of the 10 observed significance levels is less than 0.05 is about 0.29. With 10 means (45 comparisons), the probability of finding at least one significant difference is about 0.63. The more comparisons you make, the more likely it is that you'll find 1 or more pairs to be statistically different, even if all population means really are equal.

Multiple comparison procedures protect you from calling differences *significant* when they really aren't. This is accomplished by adjusting the observed significance level for the number of comparisons that you are making, since each comparison provides another opportunity to reject the null hypothesis. The more comparisons you make, the larger the difference between pairs of means must be for a multiple comparison procedure to call it statistically significant. That's why you should look only at differences between pairs of means that you are interested in. When you use a multiple comparison procedure, you can be more confident that you are finding true differences.

Many multiple comparison procedures are available. They differ in how they adjust the observed significance level. One of the simplest is the **Bonferroni procedure.** It adjusts the observed significance level by multiplying it by the number of comparisons being made. For example, if you are making five comparisons, the observed significance level for each comparison must be less than 0.05/5, or 0.01, for the difference to be significant at the 0.05 level.

To obtain the Bonferroni test, select Post Hoc in the One-Way ANOVA dialog box. Then select Bonferroni, as shown in Figure 14.6.

Figure 14.4 Bonferroni multiple comparison test on hours worked

Dependent Variable: Number of Hours Worked Last Week

Bonferroni

(I) DEGREE RS Highest Degree	(J) DEGREE RS Highest Degree	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Less than HS	High school	-2.08	1.653	1.000	-6.73	2.57
	Junior college	-2.18	2.174	1.000	-8.30	3.94
	Bachelor	-2.69	1.783	1.000	-7.71	2.33
	Graduate	-6.58*	1.986	.009	-12.11	-1.04
High school	Less than HS	2.08	1.653	1.000	-2.57	6.73
	Junior college	-9.78E-02	1.625	1.000	-4.67	4.48
	Bachelor	-.61	1.047	1.000	-3.56	2.34
	Graduate	-4.49*	1.334	.008	-8.25	-.74
Junior college	Less than HS	2.18	2.174	1.000	-3.94	8.30
	High school	9.78E-02	1.625	1.000	-4.48	4.67
	Bachelor	-.51	1.758	1.000	-5.46	4.44
	Graduate	-4.40	1.943	.239	-9.87	1.07
Bachelor	Less than HS	2.69	1.783	1.000	-2.33	7.71
	High school	.61	1.047	1.000	-2.34	3.56
	Junior college	.51	1.758	1.000	-4.44	5.46
	Graduate	-3.88	1.493	.094	-8.09	.32
Graduate	Less than HS	6.58*	1.986	.009	1.04	12.11
	High school	4.49*	1.334	.008	.74	8.25
	Junior college	4.40	1.943	.239	-1.07	9.87
	Bachelor	3.88	1.493	.094	-.32	8.09

*. The mean difference is significant at the .05 level.

If you want to compare all five education groups to one another, you can form 10 unique pairs of groups. Statistics for all pairs of group comparisons using the Bonferroni multiple comparison procedure are shown in Figure 14.4. There are a lot of numbers, but they're not hard to understand. Each row corresponds to a comparison of two groups. The first row is for the comparison of the less than high school group to the high school group. The last row is for the comparison of the graduate group to the bachelor's degree group. The difference in hours worked between the two groups is shown in the column labeled *Mean Difference*. Pairs of means that are significantly different from each other are marked with an asterisk. You see that people with graduate degrees work significantly longer than people with less than a high school education and people with graduate degrees work significantly longer than people with just a high school education. No two other groups are significantly different from one another. The table shows all possible pairs of groups twice. There is a row for the comparison of bachelor to graduate and another row for the comparison of graduate to bachelor. These two rows are identical, except for the sign of the mean difference.

The column labeled *Std. Error* (of the difference) is calculated from the within-groups estimate of the standard deviation and the sample sizes in each of the two groups. The observed significance level for the test of the null hypothesis that the two groups come from populations with the same mean is shown in the column labeled *Sig.* Looking down the column of observed significance levels, you see that four of them are less than 0.05. (Note that the mean differences for these pairs are marked with an asterisk.) The 95% confidence interval for the mean difference gives you a range of values that you expect would include the true population difference between the two groups. For example, it's possible that the true difference between the hours worked by people with graduate degrees and people with less than a high school education is anywhere between 1 and 12 hours. Note that the confidence intervals for the pairs that are significantly different from one another do not include the value 0. The confidence intervals are also modified to take into account the fact that 10 pairs of means are being compared. They are wider than they would be if only one pair of means was being compared.

How come the graduate degree group isn't different from the junior college group too? Whether a difference between two groups is statistically significant depends on how big the difference is between the two groups and how many cases there are in each of the groups. (The same estimate of variance is used for all groups.) The average hours worked for junior college grads is very similar to the average hours worked for high school grads. However, there are only 54 people with junior college degrees. It's possible that you'll find in the pairwise table of differences that smaller differences between two groups may be significant, while larger differences between other groups are not. That's the result of differences in the sample sizes between the groups.

Summary

How can you test the null hypothesis that several population means are equal?

- Analysis of variance is a statistical technique that is used to test hypotheses about two or more population means.
- To use analysis of variance, your groups should be random samples from normal populations with the same variance.
- The *F* ratio is the ratio of two estimates of the population variance: the between-groups and the within-groups mean squares.
- The analysis-of-variance *F* test does not pinpoint which means are significantly different from each other. That's why multiple comparison procedures, which protect you against calling too many differences significant, are used to identify groups that appear to be different from each other.

8. lekce

ZÁKLADY BIVARIAČNÍ ANALÝZY: ROZLOŽENÍ DAT V KONTINGENČNÍ TABULCE - POVAHA VZTAHU MEZI HODNOTAMI PROMĚNNÝCH A POROVNÁVÁNÍ POZOROVANÝCH S OČEKÁVANÝMI ČETNOSTMI. MĚŘENÍ (SÍLY) ASOCIACE MEZI DVĚMA KATEGORIZOVANÝMI PROMĚNNÝMI: KOEFICIENTY ASOCIACE (modul ANALYZE: procedura Crosstabs).

LOETHER Herman J. and Donald G. McTAVISH. 1988. *Descriptive and Inferencial Statistics. An Introduction.* Boston, Allyn and Bacon.

6.3.1 Creating a Bivariate Frequency Distribution

To illustrate how a table is constructed, Table 6.4 shows a tolerance score and a gender score for each of 13 fictitious individuals. The bivariate distribution is set up as shown, with the categories of tolerance (here two categories, high and low) down the **stub** or side of the table, and the categories of gender across the **heading** at the top.

This table is a 2 by 2 table, or **fourfold table**, because it has two rows and two columns (or four cells in the **body** of the table where the rows and columns intersect). Tables can, of course, have any number of rows and column (in general we refer to an *r* by *c* table where *r* refers to the number of rows and *c* to the number of columns); *r* and *c* depend upon the number of categories that are distinguished for row and column variables.

The problem now is to count the number of cases that have various possible combinations of values on the two variables, and to enter these totals into the table to form a bivariate frequency distribution. Notice that 3 of the 13 cases in this sample are males who are "high" on tolerance, 3 are males who are "low" on tolerance, 5 are females who are "high" on tolerance, and 2 are females who are "low" on tolerance. These numbers are written in the boxes or cells in the body of the table corresponding to the appropriate row and column labels shown in Table 6.4. Sometimes it is helpful to create a tally within each cell as a workmanlike way to assure accuracy.

Each of the boxes in the table is called a **cell** and the frequency in a cell is called a **cell frequency**. Cell frequencies are sometimes symbolized by the small letter n_{ij}, where the first subscript (i) indicates the number of the row and the second subscript (j) indicates the number of the column, as follows:

	Column 1	*Column 2*	*Row Totals*
Row 1	n_{11}	n_{12}	$\sum_j n_{1j}$
Row 2	n_{21}	n_{22}	$\sum_j n_{2j}$
Column totals	$\sum_i n_{i1}$	$\sum_i n_{i2}$	N

They indicate the number of cases in the total sample that fall in a certain category of the row and column variables as indicated by the row and column labels.* The cell frequencies indicate the number of cases with two characteristics simultaneously. Row and column totals each add up to 13, the total number of cases there were. The cell frequencies constitute the conditional distributions, and the row and column totals reflect the marginals or univariate distribution of each variable.

6.3.2 Traditions of Table Layout

As mentioned above (Section 6.3) tables usually are set up so that the dependent variable is the one with categories listed down the *stub*, or left side of the table, and the independent variable is listed across the top in the *heading*. This convention, of course, is not always kept, but it does tend to aid the examination of conditional distributions in each column to have it set up this way. Table 6.3 illustrates proper labeling of a table. Notice that low categories of the independent variable, where there are low categories on that variable, are listed at the left and the high categories at the right. For the dependent variable the high categories are at the top of the table and the low categories are at the bottom. This is similar to the labeling of other graphs, although in the case of tables the convention is not as rigidly adhered to, and the investigator would do well to double check the table layout before proceeding to make any interpretation.

A table usually has a title that lists the dependent variable, whether the table contains frequencies or percentages (or some other measure), the independent variable(s), and the kind of case upon which the measurements were taken. Table 6.3 contains data on 7,714 individuals. If the table is a percentaged table, it is important to indicate the base upon which the percentage was computed in brackets, at the bottom by the column total percentages*, when this is done, cell frequencies may be omitted from a percentaged table. The source of data is indicated, typically, in a footnote to the table, and both the stub and heading are clearly marked with the variable and the name of each of the categories of each variable.

Table 6.4 is a frequency table. It has categories of the tolerance variable down the side (the stub), and categories of the variable, gender, across the top. In this case, tolerance played the role of dependent variable.

Notice that cell frequencies in columns are summed, and the sums are put at the bottom of the table. Rows are also summed, and the totals put at the right-hand side. These row and column totals are called **marginals**, or simply row totals and column totals, and they are merely the univariate distribution of each variable separately.

If the table shows percentages it is called a bivariate percentage distribution, and if frequencies are shown, it is called a bivariate frequency distribution.

6.3.3 Percentaged Tables

Probably the most often used type of table is the percentaged table. Its value lies in the way it helps one to make comparisons across the conditional distributions one wants to compare. The basic rule for computing percentages in a table is as follows:

Compute percentages in the direction of the independent variable.

This means that percentages should sum up to 100% for each category of the independent variable. For tables set up such as Table 6.3, the percentaging rule

* Some authors use a different set of symbols for row and column totals, where a dot is used in place of a subscript for totals. Thus $n_{.1}$ would be the sum of column 1 over all of the rows in the table (the row subscript is replaced by a dot) instead of $\sum_j n_{i1}$, and $n_{1.}$ would symbolize the total of row one, instead of $\sum_j n_{1j}$. One could use n rather than N or $\Sigma\Sigma n_{ij}$ to indicate the grand total number of cases.

* This is true if column totals are the bases of percentages. If rows sum to 100% then row total frequencies are given.

leads to computation with column totals as the base of the percentage: thus column percentages add up to 100% for each column. If the independent variable and the dependent variable were switched around, the percentages would have to be run in the other direction. There are three ways that a table can be percentaged, as shown in Table 6.5, using the hypothetical data from Table 6.4. Tables could be percentaged with *column totals* as the base of percentages, with *row totals* as the base of percentages, and with the *grand total* as the base of percentages. Since the dependent variable is down the stub of Table 6.4, the proper table to examine to see what differences there may be between categories of the gender variable would be to percentage with column totals (the number of males or the number of females) as the base of the percentages. One wants to contrast the distribution of the dependent variable between men and women, and the only way to do this is to take out the effect of different numbers of men and women by percentaging down (in the direction of the independent variable). This type of operation permits one to make comparisons in the *other direction. Comparisons are made in the opposite direction from the way percentages are run.*

TABLE 6.5 Illustration of Different Ways Percentages Can Be Computed on Tables

Original Frequency Distribution from Table 6.4

	Gender		
Tolerance Level	*Male*	*Female*	*Total*
High	3	5	8
Low	3	2	5
Total	6	7	13

A. Percentaging to Column Totals as the Base

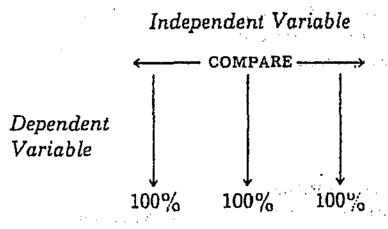
	Gender		
Tolerance Level	*Male*	*Female*	*Total*
High	50%	71%	62%
Low	50	29	38
Total	100%	100%	100%

B. Percentaging to Row Totals as the Base

	Gender		
Tolerance Level	*Male*	*Female*	*Total*
High	38%	62	100%
Low	60%	40	100%
Total	46%	54	100%

C. Percentaging to Overall Grand Total as the Base

	Gender		
Tolerance Level	*Male*	*Female*	*Total*
High	23%	39%	62%
Low	23%	15%	38
Total	46%	54	100%



Comparisons are made in a percentaged table by examining differences between percentages. In Table 6.5A, for example, the difference between percentage "high" on tolerance among men and women is 21% (71% − 50% = 21%). This value is called **epsilon**, the percentage difference in a table, and it is symbolized by the Greek letter ϵ. For tables larger than a 2 by 2 table, there are a number of percentage contrasts or epsilons that may be computed and used in interpretation. Epsilon will be discussed further later on in this chapter.

Sometimes an investigator will compute percentages, as in Table 6.5C, with the total number of cases (N) as *the base for all cell percentages.* Where this is done, we no longer can compare conditional distributions, but we can express the percentage of cases that have each of the different combinations of characteristics labeled by the rows and columns.

If it is not clear which variable is dependent or independent, or if we could think of the data in both ways, we might compute percentages to *both row and column totals* (as in Tables 6.5A and 6.5B) and *examine each table.* Table 6.5A would permit us to say that females are more likely to be higher on tolerance than are males. Table 6.5B would permit us to say that high-tolerance people are more likely to be female than are low-tolerance people—a subtle shift with worlds of import, as we shall soon see.

As shown in Table 6.5, percentaging *down* permits an examination of any influence gender may have on the distribution of tolerance; percentaging *across* shows the possible recruitment pattern into tolerance levels from each gender, and percentaging to the *grand total* permits us to examine the joint percentage distribution of tolerance levels and gender.

6.4 FOUR CHARACTERISTICS OF AN ASSOCIATION

Going back to a bivariate distribution such as that shown in Table 6.3, we can think of that distribution as a relationship between two variables. Suppose we want to know how the distribution of the dependent variable varies as we move from category to category of the other variable. The way two variables relate to each other is called an **association** between the variables. In Table 6.3, as city size increased, the percentage of individuals showing higher tolerance increased. The two variables were associated in that particular fashion.

We can speak of the association of any two variables and describe that association in terms of a percentaged table, as we have shown. There are other ways to summarize the association, however, and, in fact, there are four characteristics of an association that we will single out for summary, just as there are three characteristics of an univariate distribution that we summarized in terms of different index numbers (*i.e.*, central tendency, variation, and form). The four aspects of a bivariate association are:

1. Whether or not an association *exists.*
2. The *strength* of that association.
3. The *direction* of the association.
4. The *nature* of the association.

Each of these characteristics will be discussed in turn, and in the next chapter we will develop several alternative measures of them. In fact, we will create a single number that will be used to describe the first three features of an association listed above and in some cases a simple formula can be used as an efficient description of the last.

6.4.1 The Existence of an Association

An association is said to exist between two variables if the distribution of one variable differs in some respect between at least some of the categories of the other variable. This rather general statement can be pinned down in a number of ways, the first of which we have already discussed. If, after computing percentages in the appropriate direction in a table, there is *any* difference between percentage distributions, we would say that an association exists in these data. In the table, below, the distribution of education is slightly different for men compared with women. We know this by percentaging in the direction of the independent variable and comparing across.

Education	*Men*	*Women*	*Total*
High	40%	38%	38%
Low	60	62	62
Total	100%	100%	100%
	(43)	(56)	(99)

In the table below, however, there is *no* association between "toenail length" and "education," and this is shown by the fact that there is no difference in the percentage distribution of education (the dependent variable) regardless of the category of the independent variable within which we examine the dependent variable.

	Toenail Length		
Education	*Short*	*Long*	*Total*
High	33%	33%	33%
Low	67	67	67
Total	100%	100%	100%
	(521)	(1756)	(2277)

In the following table it is clear that there *is* an association between social class and the number of arrests, because the percentage distributions, comparing across the way percentages were run, are different.

	Social Class		
Number of Arrests	*Low*	*Medium*	*High*
None	16%	28%	45%
Few	18	18	35
Many	66	54	20
Total	100%	100%	100%
	(129)	(129)	(73)

Recall that there is a name for these comparisons: **epsilon** (ϵ), which is the percentage difference computed across the way percentages were run in a table. In a table where *all* of the epsilons are 0, there is *no* association. If any epsilon is non-0, there is an association in the data even though we may not choose to consider the very small differences important enough to talk about.

The second way to tell whether or not there is an association in a table is to compare the **actual observed table frequencies** with the frequencies we would expect if there were no association, or **expected frequencies**. If the match between actual data and our model of no association is perfect, then there is no association in the actual data between the two variables that were cross-tabulated in the table.

6.4.1a No-Association Models

A **model of no association** can be set up for a specific table as follows. Usually in setting up a model of the way frequencies in a table should look if there were no association, we assume that the marginal distribution of each variable is the way it is in the observed data table, and that the total number of cases is the same. The problem is to specify the pattern of cell frequencies in the body of the table in a way that shows no association. As an example, suppose the marginals for variables X and Y are as follows:

	(X)		
(Y)	*Low*	*High*	*Total*
High	a	b	57
Low	c	d	50
Total	34	73	107

The problem is to find a pattern of frequencies for cells a, b, c, and d such that they exhibit no association between X and Y. The reasoning goes like this. If there is no association in the table, then the ratio of "high" cell frequencies for variable Y as related to the corresponding column totals should be the same throughout the table, as it is in the overall distribution of Y itself, namely 57 to 107. In the table above. we would expect 57/107ths of the 34 cases in the "low" category of X to be in the "high" category of Y. Furthermore, we would expect the same ratio, 57/107ths of the 73 cases in the high column of X to be in the top row. This would mean that, relatively speaking, there is no difference between the proportion of cases in the top row for any column of the table.

$$\frac{57}{107}(34) = .533(34) = 18.1 \text{ cases } \textit{expected} \text{ in cell } a$$

$$\frac{57}{107}(73) = .533(73) = 38.9 \text{ cases } \textit{expected} \text{ in cell } b$$

Given that one of the above cell frequencies in a 2 by 2 table is computed, the other expected cell frequencies could be determined by subtraction. The resulting table of expected cell frequencies (expected if there were no association between the two variables X and Y for these 107 cases) is shown below.

"Expected" Cell Frequencies

	(X)		
(Y)	*Low*	*High*	*Total*
High	18.1	38.9	57.0
Low	15.9	34.1	50.0
Total	34.0	73.0	107.0

This is a hypothetical tabulation showing no association and thus fractional frequencies are acceptable.

Expected cell frequencies (f_e) can be computed for a given cell by multiplying the row total for that cell by the column total for that cell and dividing by N, which is the operation explained above.

$$f_{e_{ij}} = \frac{(n_{i.})(n_{.j})}{N} \tag{6.1}$$

where $f_{e_{ij}}$ refers here to the expected cell frequency for the cell in the ith row and jth column of the table; $n_{i.}$ is the total for the ith row and $n_{.j}$ is the total for the jth column; and N is the total number of cases. An expected cell frequency is computed (or found by subtraction) for each cell in the table.

Now the difference between the table of observed data and the model we could construct of how this table would look if there were no association can be compared. This comparison is made by subtracting an expected cell frequency, f_e, from the corresponding observed cell frequency, f_o. The difference is called **delta**, and in this text we will symbolize delta with the upper case Greek letter delta (Δ). For a given cell,

$$\Delta = f_o - f_e \tag{6.2}$$

A delta value can be computed for each cell in a table, regardless of the size of the table. If any of the deltas are *not* 0, then there is at least some association shown in the table. Whenever all deltas are 0, all epsilons will also be 0. Later we will discuss summary measures of association based on these ideas.

In summary, whether or not an association exists in a table of observed frequencies can be exactly determined in two ways that yield the same conclusions. One way is to compute percentages in one direction and compare across in the other direction, using epsilon. The other way is to create a table of expected cell frequencies and compare the observed and expected cell frequencies, cell by corresponding cell, using delta. If all of the epsilons that can be computed in a table, or if all of the delta values for a table, amount to 0, then there is no association between the two variables cross-tabulated in the bivariate distribution. This is called **statistical independence**. If, on the other hand, there is any epsilon or any delta that is not 0, then there is an association in the observed frequency table, however slight or large that association might be.

6.4.2 Degree (Strength) of Association

Where the differences between percentages (epsilons) are large, or where the deltas are large, we speak of a strong **degree of association** between the two variables; that is, the dependent variable is distributed quite differently within the different conditional distributions defined by the independent variable. This can be contrasted with a weak association where there is very little difference or where the epsilons and deltas are very small, approaching or equaling 0.

Often investigators use epsilon (or delta) as a crude first indicator of the strength of association. The problem with both delta and epsilon is that it is difficult to determine what a given-sized delta or epsilon means, other than that there is some association in the table. The reason for this is that both delta and epsilon values for any cell(s) can vary from 0 or near-0 up to a magnitude that is not, in general, fixed. They are not "normed" or standardized. Later, in this chapter and in the next, the problem of creating good standardized measures of the strength of association will be discussed and several alternative measures will be described. Suffice it to say here that some tables show a strong relationship between independent and dependent variables, and some show a weak association or no association at all.

6.4.3 Direction of Association

Where the dependent and independent variables in a table are at least ordinal variables, it makes sense to speak about the **direction** of an association that may exist. If the tendency in the table as shown by the percentage distribution is for the higher values of one variable to be associated with the higher values of the other variable (and the lower values of each variable also tend to go together), then the association is called a *positive* association. Height and weight tend to have a positive association, since taller persons tend to be heavier; in general, across the people in a general population.

On the other hand, if the higher values of one variable are associated with lower values of the other (and the lower values of the first with higher values of the second), the association is said to be *negative*. Sociologists generally expect that the higher the educational level of people, the lower their degree of normlessness will be—a negative association.

The association between city size and tolerance scores (see Table 6.3) is positive because the larger the city, in general, the higher the tolerance level becomes (*i.e.*, the higher the percentage of people who have high tolerance scores). The older a person's age, in general, the fewer the years left until retirement, a negative association.

6.4.4 The Nature of Association

Finally, the **nature** of an association is a feature of a bivariate distribution referring to the general *pattern* of the data in the table. This is often discovered by examining the pattern of percentages in a properly percentaged table. Often the pattern is irregular, and an investigator would cite many epsilons in describing where the various concentrations of cases are in the different categories of the independent variable. Sometimes there is a rather uniform progression in concentration of cases on the dependent variable as we move toward higher values of the independent variable. If, with an increase of one step in one variable, cases tend to move up (or down) a certain number of steps on the other variable we might call the nature of the association "linear." That is, the concentrations of cases on the dependent variable (the mode, for example) tend to fall along a straight line that could be drawn through the table.

The nature of association will be discussed at length in the next chapter. Simple linear associations have an intrinsic interest to investigators as one of the simplest natures of association, but some associations are curvilinear in nature, or of some more complex patterning. In most cases the nature of association will be determined from a percentage table or a scatter plot, but in some cases nature can be described in terms of an equation.

At this point we should pause to examine several tables and describe them in terms of these four features of an association. Table 6.9 presents a series of examples together with brief summary statements.

TABLE 6.9A Percentage Distribution of Total Money Income for Families by Ethnic Background of Householder, 1983

Income	*White*	*Spanish*	*Black*	*Total*
\$35,000 and over	$\underline{31.5}$	15.1	13.2	29.6
\$25,000 to 34,999	$\underline{20.3}$	14.3	14.0	19.5
\$15,000 to 24,999	23.7	$\underline{27.0}$	21.5	23.4
\$ 7,500 to 14,999	16.0	$\underline{23.6}$	$\underline{23.7}$	16.8
Under \$7,500	8.5	20.0	$\underline{27.6}$	10.7
Total	100.0	100.0	100.0	100.0
N (millions)	(53.9)	(3.6)	(6.7)	(62.0)

Source: U.S. Bureau of the Census. (1985).

1. *Existence of Association.* This table is percentaged down in the direction of the independent variable, ethnicity, so that comparisons may be made across. The percentages across are different so that there is an association evident in the table. Compare any row, say the row for incomes of \$35,000 or more; percentages range from a high of 31.5% down to 13.2%, all different from 29.6%, the total for that income category.

2. *Strength of Association.* Overall, if ethnicity makes any difference in the distribution of family income, we should find fairly substantial epsilon's. Here, the white-black epsilon for \$35,000 and over incomes is 18.3. That for Spanish-black for the same income category is 1.9 and for white-Spanish is 16.4. These are all less than 100% but substantially larger than 0.

3. *Direction of Association.* Here, ethnicity is a nominal variable so that it is impossible to talk about direction of association.

4. *Nature of Association.* To see the pattern of association most clearly, it is helpful to underline the highest percentages in each row-wise comparison. Here, for the highest income category, 31.5% is clearly the largest percentage and is underlined. In the next row, 20.3% is largest and is underlined; 27.0% is underlined in the third row. We will underline both the 23.6% and 23.7% in the fourth row because they are essentially equivalent in magnitude. Finally, 27.6% is underlined in the bottom row. Notice that we are underlining the percentages from the body of the table, not the marginal distribution. The nature of association is the pattern of white's having higher percentages in the highest two income groups, compared with the other two ethnic groups; Spanish have higher percentages in the next group and are tied in percentages in the \$7,500 to 14,999 income category. Finally, the black group has highest percentages in the lowest income category. The pattern of high and low percentages from comparisons across the way the percentages were run is the nature of the association of ethnicity and income for families in 1983.

TABLE 6.9B PERCENTAGE DISTRIBUTION OF FATHER'S OCCUPATION BY OCCUPATION OF 30–59-YEAR-OLD CHILD, SWEDEN, 1977

Occupation of Father	Occupation of Child				
	Farmer	Worker	Entrepreneur	Middle Class	Not Known
Middle Class	2	10	13	29	14
Entrepreneur	6	10	23	15	10
Worker	14	52	38	39	39
Farmer	77	24	21	14	21
Not Known	1	4	5	3	16
Total	100.	100.	100.	100.	100.
N	(241)	(2964)	(525)	(2557)	(166)

Source: Adapted from Sundstrom (1986:369).

1. *Existence of Association.* This table presents the results of a Swedish survey of occupations of adult (30–59-year-old) children and their fathers. The table is percentaged in the direction of the child's occupation to show the distribution of fathers' occupation. Comparing across, the percentages are different, thus there is an association in the table. Note that this way of percentaging the table permits one to make statements about background occupational experience of children (e.g., their father's occupation). Percentaging the table the other way would permit one to say something about the distribution of children's occupations for fathers in different occupations. Percentaging to the total would permit statements about the percentage of people who, for example, stayed in the same occupation that their father had. The way percentages are run permits quite different kinds of comparisons.

2. *Strength of Association.* In this table, the highest epsilon ought to be seen in comparing extreme categories of child's occupation for the highest (or lowest) category of father's occupation. Taking the middle class fathers, the farmer to middle-class epsilon is 27%. For fathers who are farmers, the same epsilon comparison is 63%. These are both very substantial epsilons. The association is quite strong.

3. *Direction of Association.* As an aid in finding the direction of an association between variables each of which is at least ordinal, a useful procedure is to make comparisons across the way percentages are run, underlining the highest percentage for each comparison. In this table we could make four comparisons (aside from the not known category that the author provides here). For the middle class row the highest percentage is "middle class." For the entrepreneur row, it is "entrepreneur"; it is "worker" for the worker row, and "farmer" for the farmer row. Notice that the highest percentage is for father having the same occupation as the child, or no social mobility between generations. One could draw a diagonal line through the underlined percentages in the table. In this case, the line would extend from the "farmer"–"farmer" or lower status occupational category to the highest category combination, "middle class"–"middle class." This indicates a positive association: the higher the occupational status of the father, the higher the child's occupational status tends to be. In this "occupational mobility" table, people in the upper left-hand area above the diagonal are downwardly mobile (fathers had higher status occupations than the child) and people in the lower right-hand area under the diagonal are upwardly mobile (child has a higher status occupation than father).

4. *Nature of Association.* In this table, the nature of association (i.e., the pattern of concentration in the table) tends to be almost linear. There is a relatively uniform shift toward higher status child's occupation with shifts upward in the category of father's occupation. There are no "reversals" in this general trend of concentration. Because the variables are ordinal, it would be more appropriate to speak of this nature of association as "monotonic" rather than linear. If distances were defined then one could determine whether in fact there is a constant amount of shift in values of one variable, given a fixed amount of difference in the other. In ordinal variables one can only say that the value of one variable remained the same or shifted in a fixed direction with increases in the other variable—a monotonic nature of association. Contrasted with this type of nature are those such as the one shown in Table 6.9C.

TABLE 6.9C PERCENTAGE DISTRIBUTION OF BODY WEIGHT BY AGE FOR PERSONS 20 YEARS AND OLDER

Percentage Above or Below Desired Weight	Age					
	20–34	35–44	45–54	55–64	65–74	75+
30% or more above	10.7	16.7	21.8	21.3	21.5	11.6
20–29.9% above	7.8	11.0	13.3	13.9	13.3	11.2
10–19.9% above	16.5	21.5	23.1	23.3	22.0	20.1
5–9.9% above	12.6	13.1	13.1	12.9	12.3	15.2
Plus or minus 4.9%	27.6	22.5	18.1	18.9	18.1	*21.5*
5–9.9% below	11.5	8.0	5.7	5.0	6.2	*8.4*
10% or more below	13.3	7.2	4.9	4.7	6.6	*12.0*
Total	100.0	100.0	100.0	100.0	100.0	100.0
N (1000's)	(59.9)	(28.9)	(22.2)	(22.0)	(16.4)	(9.5)

Source: National Center for Health Statistics (1986).

1. *Existence of Association.* In this table, age is treated as the independent variable and the amount by which one's weight is above or below the desired weight is the dependent variable. Comparing across, there is a percentage difference, thus there is an association shown in the table.

2. *Strength of Association.* Among all the possible percentage comparisons, the strongest percentage difference should be evident in comparing the extreme categories of age for the extreme categories of percentage above or below desired weight. Taking the top row, the overall epsilon is only 1.1 and in the bottom row it is only 1.3. These are indeed small percentage differences. Yet, there are larger percentage differences in the table; for example, the difference between the 20–34 and 45–54 age categories for the top row of the table. As we shall see, the pattern of association in this table is irregular, making the assessment of strength of association more complex. Even at its best, however, the percentage differences are rather small, suggesting a weak association between the two variables.

3. *Direction of Association.* As before, we will underline the highest percentages in each comparison, underlining more than one where percentages are very close in magnitude. Table 6.9C shows a pattern that moves generally from upper right to lower left, between overweight associated with older ages and underweight with younger ages, a generally "positive" association.

4. *Nature of Association.* Although the overall pattern of association indicated by drawing a diagonal line through the underlined percentages (or the middle of several underlined percentages in a given comparison) is linear, there are other patterns that need to be examined. Notice that percentages are essentially tied from 45–74 years old for the top two rows but there is a clearer concentration of high percentage in one column for the bottom three rows. There is a broader area of high percentages for overweight rows than for underweight rows. Notice too that the percentages along the bottom three or four rows drop down as one moves across from low to high age categories and then the percentages begin to rise again. If the second "high" is italicized for each of the bottom three rows one can see a more complex pattern emerging from the table. The nature of association begins to appear "curvilinear." Underweight is concentrated in the lowest age bracket and to some extent in the oldest age bracket while overweight is more likely found in middle age categories. It was this curved nature of association that made an assessment of strength of association difficult to determine if we made epsilon comparisons as if we expected a monotonic relationship. We will have more to say about this later but suffice it to say that one needs to be aware of the nature of association in selecting measures of strength of association.

7 Counting Responses for Combinations of Variables

How can you study the relationship between two or more variables that have a small number of possible values?

- Why is a frequency table not enough?
- What is a crosstabulation?
- What kinds of percentages can you compute for a crosstabulation, and how do you choose among them?
- What's a dependent variable? An independent variable?
- What if you want to examine more than two variables together?
- How can you use a chart to display a crosstabulation?

The Means and Explore procedures described in Chapter 5 and Chapter 6 are useful only when statistics such as the mean and standard deviation are appropriate measures for the variable whose values you want to summarize. You can't use Means or Explore to look for relationships between color of car driven and region of the country, since it doesn't make sense to compute an average color or region. When you want to look at the relationship between two variables that have a small number of values or categories (sometimes called **categorical variables**), you may want to use a **crosstabulation**, a table that contains counts of the number of times various combinations of values of two variables occur. For example, you can count how many men and how many women are in each of the job satisfaction categories, or you can see the distribution of car colors for various regions of the country.

In this chapter, you'll use a crosstabulation to look at the relationship between job satisfaction and total family income, measured on a four-point scale.

▶ This chapter continues to use the *gssft.sav* file. For instructions on how to obtain the crosstabulation output shown in this chapter, see "How to Obtain a Crosstabulation" on p. 122.

Income and Job Satisfaction

In the General Social Survey, respondents are asked to select the range of values into which their annual family income falls. There are 21 categories, ranging from under $1,000 (assigned a code of 1) to $75,000 and over (assigned a code of 21). To look at the relationship between income and job satisfaction for full-time employees, you'll use four income groups with roughly the same number of cases. That is, you will use quartiles of income. (The variable *income4* contains the income data recoded into quartile categories.) You see from Figure 7.1, which shows a frequency distribution of the four categories of income, that as expected, roughly 25% of the people fall into each of the income groupings.

To obtain this frequency table, select the variable income4 in the Frequencies dialog box. See Chapter 3 for information on frequency tables.

Figure 7.1 Frequency table for income quartiles

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	24,999 or less	174	23.3	23.3	23.3
	25,000 to 39,999	194	26.0	26.0	49.3
	40,000 to 59,999	156	20.9	20.9	70.1
	60,000 or more	223	29.9	29.9	100.0
	Total	747	100.0	100.0	

To examine the relationship between income and job satisfaction, you want to count how many *very satisfied, moderately satisfied, a little dissatisfied*, and *very dissatisfied* people there are in each of the income categories. Figure 7.2 contains this information. The income groups make up the columns of the table. The rows are the job satisfaction categories. A cell appears in the table for each combination of values of the two variables. The first cell, at the top left of the table, is for *very satisfied* people in the lowest income group. You see that 53 people fall into this cell. The cell in the second row of the first column is for *moderately satisfied* people in the lowest income category. There are 93 people in this cell. Similarly, the cell in the fourth row of the fourth column tells you that there are 7 *very dissatisfied* people in the highest income group.

Figure 7.2 Crosstabulation of job satisfaction by income

To obtain this crosstabulation, from the menus choose:

Statistics ▶ Summarize ▶ Crosstabs...

In the Crosstabs dialog box, select the variables satjob and income4, as shown in Figure 7.9.

Count

		Total Family Income in quartiles				
		24,999 or less	25,000 to 39,999	40,000 to 59,999	60,000 or more	Total
Job Satisfaction	Very satisfied	53	90	74	110	327
	Mod satisfied	93	79	61	87	320
	A little dissatisfied	24	17	14	19	74
	Very dissatisfied	4	8	7	7	26
Total		174	194	156	223	747

194 people are in the second income group

110 people with incomes of $60,000 or more are very satisfied

To the right and at the bottom of the table are totals—often called **marginal totals** because they are in the table's margin. The margins on the table show the same information as frequency tables for each of the two variables. In the right margin, labeled *Total*, you have the total number of people who gave each of the job satisfaction answers. Similarly, the first column total of 174 is the number of people in the lowest income category. The very last number, 747, is the total number of people in the table.

Will the marginal totals that I get in a crosstabulation table always be the same as those I would get from frequency tables for the variables individually? Not if you have missing values for either of the two variables in the crosstabulation. For example, the crosstabulation in Figure 7.2 includes only cases that have nonmissing values both for job satisfaction and for income. The marginal totals for income are therefore based on cases that have nonmissing values for both income and job satisfaction. When you make a frequency table for income, the only cases excluded from the valid percentages are those with missing values for income.

If you look at the counts in the crosstabulation, you see that 53 people from the lowest income category said they are *very satisfied* with their jobs, 90 from the second income category, 74 from the third income category, and 110 from the highest income category. Can you tell from the counts just what the relationship is between income and job satisfaction? Of course not, since you can't just compare the counts when there are different numbers of people in the four income groups. To compare the groups you must look at percentages instead of counts. That is, you must look at the percentage of people in each of the income groups who gave each of the job satisfaction responses.

Row and Column Percentages

Figure 7.3 contains both the counts and the column percentages. From the totals for each of the rows, you see that, overall, 43.8% of the sample are *very satisfied* with their jobs. You also see that 30.5% of the lowest income group, 46.4% of the second income group, 47.4% of the third income group, and 49.3% of the highest income group are *very satisfied* with their jobs. It appears that the lowest income people are less likely than average to be *very satisfied*, while the high income people are more likely than average to be *very satisfied*.

To obtain column percentages, select Cells in the Crosstabs dialog box. Then select Column, as shown in Figure 7.11.

Use the Pivot Table Editor to specify labels of your choice.

Figure 7.3 Crosstabulation showing column percentages

			Total Family Income in quartiles				
			24,999 or less	25,000 to 39,999	40,000 to 59,999	60,000 or more	Total
Job Satisfaction	Very satisfied	Count	53	90	74	110	327
		% within Total Family Income in quartiles	30.5%	46.4%	47.4%	49.3%	43.8%
	Mod satisfied	Count	93	79	61	87	320
		% within Total Family Income in quartiles	53.4%	40.7%	39.1%	39.0%	42.8%
	A little dissatisfied	Count	24	17	14	19	74
		% within Total Family Income in quartiles	13.8%	8.8%	9.0%	8.5%	9.9%
	Very dissatisfied	Count	4	8	7	7	26
		% within Total Family Income in quartiles	2.3%	4.1%	4.5%	3.1%	3.5%
Total		Count	174	194	156	223	747
		% within Total Family Income in quartiles	100.0%	100.0%	100.0%	100.0%	100.0%

You can change the label to indicate that the column percentages are shown

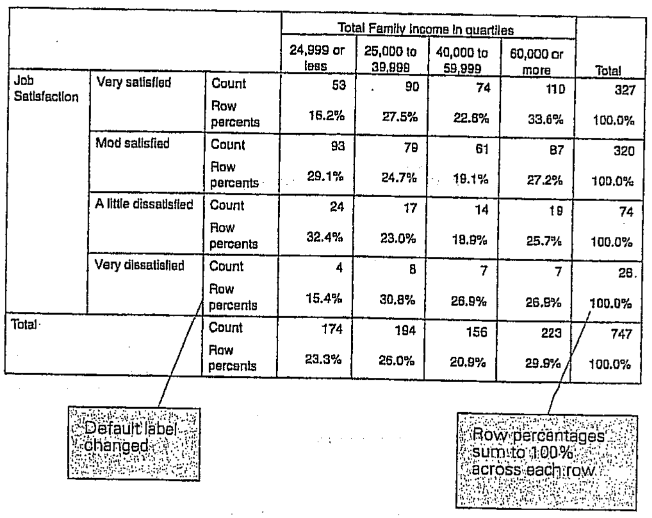
Column percentages sum to 100% in each column

The percentages you used to make comparisons are known as column percentages, since they express the number of cases in each cell of the table as a percentage of the column total. That is, for each income group, they tell you the distribution of job satisfaction. The column percentages sum up to 100% for each of the columns. In Figure 7.3, you can see that 30.5% of people with incomes less than $25,000 are very satisfied with their jobs, while 49.3% of people with incomes of $60,000 and up are very satisfied. The percentage of people who are very dissatisfied or a little dissatisfied is largest in the lowest income category (16.1%).

You can also calculate row percentages for the table. Row percentages tell you what percentage of the total cases of a row fall into each of the columns. For each job satisfaction category, they tell you the percentage of cases in each income category. (You can also compute what are called total percentages. The count in each cell of the table is expressed as a percentage of the total number of cases in the table.) Figure 7.4 contains counts and row percentages for our example.

To obtain row percentages, select Cells in the Crosstabs dialog box. Then select Row. (See Figure 7.11.)

Figure 7.4



			Total Family Income in quartiles				
			24,999 or less	25,000 to 39,999	40,000 to 59,999	60,000 or more	Total
Job Satisfaction	Very satisfied	Count	53	90	74	110	327
		Row percents	16.2%	27.5%	22.6%	33.6%	100.0%
	Mod satisfied	Count	93	79	61	87	320
		Row percents	29.1%	24.7%	19.1%	27.2%	100.0%
	A little dissatisfied	Count	24	17	14	19	74
		Row percents	32.4%	23.0%	18.9%	25.7%	100.0%
	Very dissatisfied	Count	4	8	7	7	26
		Row percents	15.4%	30.8%	26.9%	26.9%	100.0%
Total		Count	174	194	156	223	747
		Row percents	23.3%	26.0%	20.9%	29.9%	100.0%

Default label changed

Row percentages sum to 100% across each row

From the row percentages, you see that 16.2% of the *very satisfied* respondents are in the lowest income group, 27.5% are in the second income group, 22.6% are in the third income group, and 33.6% are in the fourth income group. The four row percentage values sum to 100 for each of the rows. In this example, the row percentages aren't very helpful, since you can't make much sense of them without taking into account the overall percentages of cases in each of the income categories. That is, you can't tell whether the percentage of high income cases in the *very satisfied* category is due to a large number of high income cases in your sample or to high satisfaction rates in that category.

How can I tell whether a table contains row or column percentages? If the column labeled *Total* shows all 100%, the table contains row percentages, which necessarily sum to a 100 for each row. If the row labeled *Total* contains 100%, the tables contains column percentages. ■■■

For a particular table, you must determine whether the row or column percentages answer the question of interest. This can be done easily if one of the variables can be thought of as an independent variable and the other as a dependent variable. An **independent variable** is a variable that is thought to influence another variable, the **dependent variable**. For example, if you are studying the incidence of lung cancer in smokers and nonsmokers, smoking is the independent variable. Smoking influences whether people get cancer, the dependent variable. Similarly, if you are studying the income categories of men and women, gender is the independent variable since it might influence how much you get paid.

If you can identify one of your variables as independent and the other as dependent, then you should compute percentages so that they sum to 100 for each category of the independent variable. In other words, what you want to see is the same number of people in each of the categories of the independent variable. Having the percentages sum to 100 for each category of the independent variable is the equivalent of having 100 cases in each category. For example, you want 100 smokers and 100 nonsmokers. Then you can compare the incidence of lung cancer in the two groups. In the current example, income category is the independent variable and job satisfaction is the dependent variable. That means you'd like to see 100 people in each of the income categories. Since income is the column variable in Figure 7.3, you use column percentages that sum to 100 for each category of income.

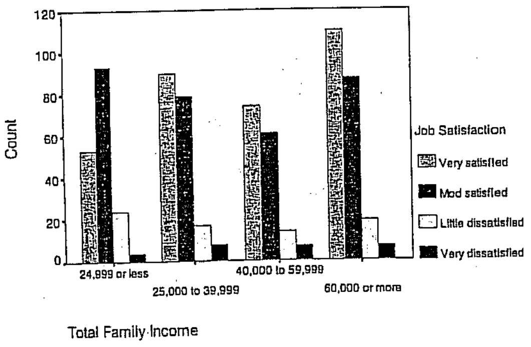
Can't you analyze these data using the Means procedure? The General Social Survey codes income in unequal intervals. For example, the interval from $8,000 to $9,999 is coded 8, but the interval $60,000 to $74,999 is coded 20. So you don't want to compute means for these codes. Instead, if you want to compute average family income, you must change the coding scheme so that the code for a case is the midpoint of the appropriate income interval. For example, an income anywhere in the range of $8,000 to $9,999 would be assigned a code of $9,000. Similarly, incomes in the range of $60,000 to $74,999 would be assigned a code of $67,500, the midpoint of the interval. You can then compute descriptive statistics for the recoded incomes. Of course, the means won't be the same as those you would get if you had the exact income for each person, but they're the best you can do given the limitations of the data. ■■■

Bar Charts

You can display the results of a crosstabulation in a clustered bar chart. Consider Figure 7.5, which is a bar chart of family income by job satisfaction. The length of a bar tells you the number of cases in a category.

You can obtain this bar chart using the Graphs menu, as described in "Bar Charts" on p. 520 in Appendix A. In the Clustered Bar Summaries for Groups of Cases dialog box, select the variables income4 and satjob.

Figure 7.5 Bar chart of income by job satisfaction



There is a cluster of bars for each of the four income categories. Within each cluster, there is a bar for each of the job satisfaction categories.

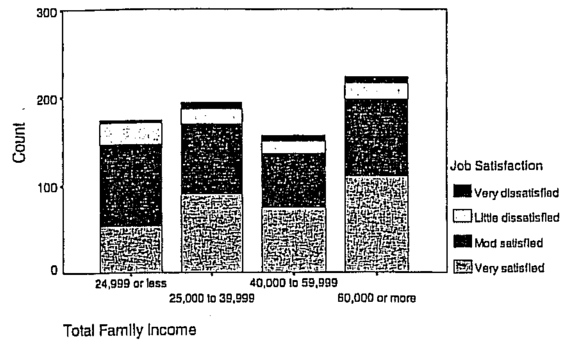
Since there are unequal numbers of people in the income categories, comparing bar lengths across income categories presents the same problem as looking at simple counts in a crosstabulation. All you can really do with this bar chart is compare bar lengths within a cluster and see whether the patterns are the same across clusters.

You can obtain this chart by modifying Figure 7.5, as described in "Bar Charts" on p. 520 in Appendix A. From the Chart Editor menus choose:

Gallery
Bar...

In the Bar Charts dialog box, select Stacked.

Figure 7.6 Stacked bar chart



Stacked Bar Charts

You can stack the bars in a clustered bar chart one on top of the other. The result is the stacked bar chart in Figure 7.6. Now it's easier to see for each income category the proportion of people in each of the job satisfaction categories. However, the lengths of the bars aren't equal for the four income categories, so that still gets in the way.

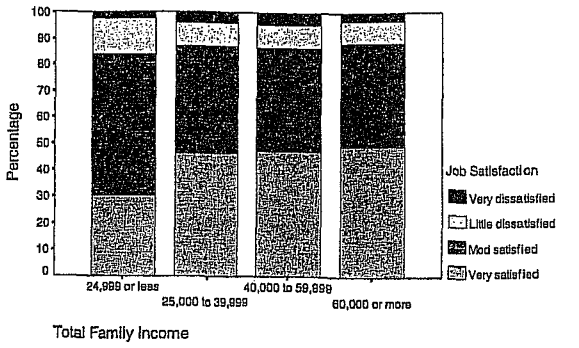
Ideally, you want each of the bars to be of the same length, so you can easily compare the areas across bars. What you'd really like to see is a plot of the column percentages from Figure 7.3. You can do this by turning the counts in each bar into percentages, as shown in Figure 7.7. Now each of the bars has the same length—100%—and you can easily compare the job satisfaction distributions across bars. You see that people in the lowest income group are least likely to be *very satisfied* with their jobs. They are also least likely to be *very dissatisfied*. The distribution of job satisfaction categories seems to be very similar for the other three income groups. The proportion of *very satisfied* people doesn't increase with income for these three groups. You can also see that the sum of the percentages for *very satisfied* and *moderately satisfied* is very similar for the four groups.

You can obtain this bar chart using the Graphs menu, as described in "Bar Charts" on p. 520 in Appendix A. In the Clustered Bar Summaries for Groups of Cases dialog box, select the variables income4 and satjob. Or you can select Display clustered bar charts in the Crosstabs dialog box. From the Chart Editor menus choose:

Series
Transpose Data

to cluster bars within income categories.

Figure 7.7 Stacked bar chart with percentage scale



Summary

How can you study the relationship between two or more variables that have a small number of possible values?

- A crosstabulation shows the numbers of cases that have particular combinations of values for two or more variables.
- The number of cases in each cell of a crosstabulation can be expressed as the percentage of all cases in that row (the row percentage) or the percentage of all cases in that column (the column percentage).
- A variable that is thought to influence the values of another variable is called an independent variable.
- The variable that is influenced is called the dependent variable.
- If there is an independent variable, percentages should be calculated so that they sum to 100% for each category of the independent variable.
- When you have more than two variables, you can make separate crosstabulations for each of the combinations of values of the other variables.
- Bar charts can be used to display a crosstabulation graphically.

16 Comparing Observed and Expected Counts

How can you test the null hypothesis that two variables are independent?

- What are observed and expected counts?
- How do you compute the chi-square statistic?
- What assumptions are needed for the chi-square test of independence?
- What is a one-sample chi-square test?
- Why is sample size important?

You know how to test a variety of hypotheses about population means. However, these tests are useful only when it makes sense to compute a mean for a variable. If you want to look at the relationship between preference among car colors and region of the country, or between type of treatment and remission of symptoms, you can't use a *t* test because it doesn't make sense to compare means. Rather, such variables are best summarized by a crosstabulation. In this chapter, you'll use the **chi-square test** to examine hypotheses about data that are best summarized by a crosstabulation.

▶ This chapter uses the *gss.sav* data file. The chi-square test output shown can be obtained using the SPSS Crosstabs procedure. (For more information on Crosstabs, see Chapter 7.)

Education and Anomia

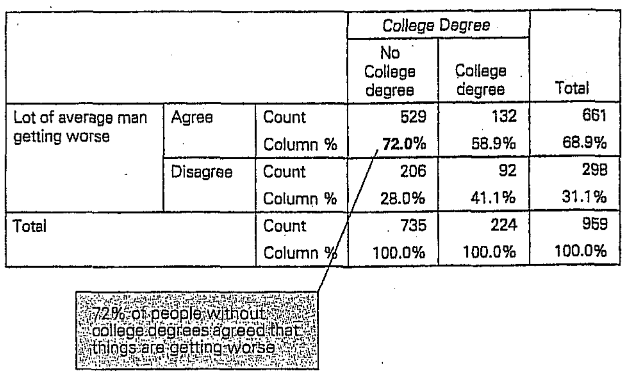
The French sociologist Emile Durkheim introduced the concept of anomie to represent the feelings of alienation and rootlessness common in the modern world. The General Social Survey attempts to measure such feelings with a scale called *anomia*. One item on this scale asks respondents whether they agree or disagree with the following statement:

"In spite of what some people say, the lot of the average man is getting worse, not better." Let's consider whether education is related to the likelihood of agreeing with this statement.

You can obtain a crosstabulation using the Crosstabs procedure, as discussed in Chapter 7.

Select anomia5 and degree2 in the Crosstabs dialog box.

Figure 16.1 Crosstabulation of anomia and education



			College Degree		
			No College degree	College degree	Total
Lot of average man getting worse	Agree	Count	529	132	661
		Column %	**72.0%**	58.9%	68.9%
	Disagree	Count	206	92	298
		Column %	28.0%	41.1%	31.1%
Total		Count	735	224	959
		Column %	100.0%	100.0%	100.0%

72% of people without college degrees agreed that things are getting worse

Figure 16.1 is a crosstabulation of responses to the statement for those with and without college degrees. You see that 72% of respondents who have not completed college agree with the statement, while 58.9% of respondents with college degrees agree with this statement. Based on these results, do you think that, in the population, there is a difference between college graduates and non-college graduates in the perception of the lot of the average man? Certainly in this sample, college graduates are less pessimistic than nongraduates. But as usual, the sample results are not what you're interested in. You want to know what you can conclude about the population based on the observed sample results. You want to know whether you have enough evidence to reject the null hypothesis that, in the population, the same percentage of college graduates and nongraduates agree with the statement.

Observed and Expected Counts

The basic element of a crosstabulation table is the count of the number of cases in each cell of the table. The statistical procedure you'll use to test the null hypothesis is based on comparing the observed count in each of the cells to the expected count. The expected count is simply the number of cases you would expect to find in a cell if the null hypothesis is true. Here's how the expected counts are calculated.

Calculating Expected Counts

If the null hypothesis is true, you expect college graduates and nongraduates to answer the question in the same way. That is, you expect the *percentage* agreeing with the statement to be the same for the two groups of cases. You don't expect the same *number* of graduates and nongraduates to agree with the statement, since you don't have the same number of people in the two education categories.

From the row marginals in Figure 16.1, you see that in the sample, 68.9% of the respondents agreed with the statement and 31.1% disagreed. If the null hypothesis is true, these are the best estimates for the *percentages* you would expect for both graduates and nongraduates. To convert the percentages to the actual number of cases in each of the cells, multiply the expected percentages by the numbers of graduates and nongraduates. For example, the expected number of nongraduates agreeing with the statement is

$$68.93\% \times 735 = 506.6 \qquad \textbf{Equation 16.1}$$

Similarly, the expected number of nongraduates disagreeing with the statement is

$$31.07\% \times 735 = 228.4 \qquad \textbf{Equation 16.2}$$

For college graduates, the expected values are calculated in the same way, substituting the number of college graduates (224) for the number of nongraduates (735) in the above two equations.

Is there a simple way I can remember how to calculate expected values? Sure. The following rule is equivalent to what you've just done: To calculate the expected number of cases in any cell of a crosstabulation, multiply the number of cases in the cell's row by the number of cases in the cell's column and divide by the total number of cases in the table. Try it. You'll see it always works. ■ ■ ■

Figure 16.2 Observed and expected counts

To obtain observed and expected counts, select Cells in the Crosstabs dialog box.

See Chapter 7 for more information.

			College Degree		
			No College degree	College degree	Total
Lot of average man getting worse	Agree	Count	529	132	661
		Expected Count	506.6	154.4	661.0
		Residual	22.4	-22.4	
	Disagree	Count	206	92	298
		Expected Count	228.4	69.6	298.0
		Residual	-22.4	22.4	
Total		Count	735	224	959
		Expected Count	735.0	224.0	959.0

Residual is the difference between observed and expected counts

You see in Figure 16.2 the observed and expected counts for all four cells. The last entry in a cell is the **residual**, the difference between the observed and expected counts. A positive residual means that you observed more cases in a cell than you would expect if the null hypothesis were true. A negative residual indicates that you observed fewer cases than you would expect if the null hypothesis were true.

The sum of the expected counts for any row or column is the same as the observed count for that row or column. For example, the expected counts for college graduates add up to the observed number of college graduates. Similarly, the expected counts for the number agreeing add up to the observed number of cases agreeing. Another way of saying this is that the residuals add up to 0 across any row and any column.

The Chi-Square Statistic

When you test the null hypothesis that two population means are equal, you compute the t statistic, and then, using the t distribution, calculate how unusual the observed value is if the null hypothesis is true. To test hypotheses about data that are counts, you compute what's called a chi-square statistic and compare its value to the chi- square distribution to see how unlikely the observed value is if the null hypothesis is true.

What assumptions are needed to use the chi-square test? All of your observations must be independent. That implies that an individual can appear only once in a table. You can't let a person choose two favorite car colors and then make a table of color preference by gender. (Each person would appear twice in such a table.) It also means that the categories of a variable can't overlap. (For example, you can't use the age groups less than 30, 25–40, 35–90.) Also, most of the expected counts must be greater than 5, and none less than 1. ■ ■ ■

To compute the Pearson chi-square statistic, do the following:

1. For each cell, calculate the expected count by multiplying the number of cases in the cell's row by the number of cases in the cell's column and dividing the result by the total count.
2. Find the difference between the observed and expected counts.
3. Square the difference.
4. Divide the squared difference by the expected count for the cell.
5. Add up the results of the previous step for all of the cells.

In the current example, the value for the Pearson chi-square statistic is

$$\frac{(529-506.6)^2}{506.6} + \frac{(132-154.4)^2}{154.4} + \frac{(206-228.4)^2}{228.4} + \frac{(92-69.6)^2}{69.6} = 13.64 \qquad \textbf{Equation 16.3}$$

If the null hypothesis is true, the observed and expected values should be similar. Of course, even if the null hypothesis is true, the observed and expected values won't be identical, since the results you observe in a sample vary somewhat around the true population value. As before, you have to determine how often to expect a chi-square value at least as large as the one you've calculated, if the null hypothesis is true.

To determine whether a chi-square value of 13.64 is unusual, you compare it to the chi-square distribution. Like the t distribution, the chi-square distribution depends on the parameter called the degrees of freedom. The degrees of freedom for the chi-square statistic depend *not* on the number of cases in your sample, as they did for the t statistic, but on

the number of rows and columns in your crosstabulation. The degrees of freedom for the chi-square statistic are

$$(\text{number of rows in the table} - 1) \times (\text{number of columns in the table} - 1) \qquad \textbf{Equation 16.4}$$

For this example, there is one degree of freedom, since there are two rows and two columns.

What's the logic behind the calculation of the degrees of freedom? For any row or column of a crosstabulation, the residuals sum to 0. That means that you can tell what the expected values must be for the last row and last column of a table without doing any calculations other than summing the expected values in the preceding rows or columns. The number of cells for which you have to calculate expected values is equal to the number of cells when you remove the last row and the last column from your table. The number of cells in a table when one row and one column are removed is the number of rows minus 1 multiplied by the number of columns minus 1, which is the formula for the degrees of freedom. ■ ■ ■

Figure 16.3 Pearson chi-square test for anomia by education

To obtain the Pearson chi-square test along with a crosstabulation, select Statistics in the Crosstabs dialog box.

See Chapter 7 for more information.

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	**13.639**	1	.000		
Continuity Correction	13.036	1	.000		
Likelihood Ratio	13.201	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	13.624	1	.000		
N of Valid Cases	959				

Pearson chi-square (from Equation 16.3)

In Figure 16.3, you see that the observed significance level for the Pearson chi-square value of 13.64 is less than 0.0005. This means that, if the null hypothesis is true, you expect to see a chi-square value at least as large as 13.64 less than five times out of 10,000. Since the observed significance level is small, you can reject the null hypothesis that college graduates and those who did not graduate from college give the same responses to the question. It appears that college graduates are more optimistic about the lot of the common man than high school graduates.

What's all that other stuff in Figure 16.3 along with the Pearson chi-square? The continuity-corrected chi-square is a modification of the Pearson chi-square for two-by-two tables. Most statisticians agree that the modification is unnecessary, so you can ignore it. The likelihood-ratio chi-square is a statistic very similar to the Pearson chi-square. For large sample sizes, the two statistics are close in value. The *Linear-by-Linear Association test* is a measure of the linear association between the row and column variables. It's useful only if both the row and column variables are ordered from smallest to largest. Ignore it in other situations.

If you have a table with two rows and two columns, you'll also find something labeled *Fisher's Exact Test* on your output. The advantage of Fisher's exact test is that it is appropriate for 2 × 2 tables in which the expected value in one or more cells is small. The disadvantage is that it requires a very restrictive assumption about the data: that you know in advance the number of cases in the margins. There's controversy among statisticians about the appropriateness of Fisher's exact test when this assumption is not met. In general, Fisher's exact test is less likely to find true differences than it should. Statistically, a test like this is called conservative.

College Degrees and Perception of Life

In the previous example, you tested whether college graduates and those who are not college graduates respond in the same way to the question about the lot of the average man. The null hypothesis can be stated in several equivalent ways. You can say the null hypothesis is that the percentage agreeing with the statement is the same for the two categories of education. Another way of stating the null hypothesis is that educational status and response are independent.

Independence means that knowing the value of one of the variables for a case tells you nothing about the value of the other variable. For example, if marital status and happiness with life are independent, knowing a person's marital status gives you no information about how happy they are with life. College education and perception of the lot of man, on the other hand, don't seem to be independent. If you know that a person is a college graduate, you know that he or she is less likely to agree with the pessimistic statement about the lot of the average man than is a person who is not a college graduate.

A Larger Table

The chi-square test can be used to test the hypothesis of independence for a table with any number of rows and columns. The idea is the same as for the two-row and two-column table. As an example, let's look at the relationship between highest degree earned and whether life is perceived as exciting, routine, or dull.

Figure 16.4 is a crosstabulation of highest degree earned and the response to the perception of life question.

Figure 16.4 Crosstabulation of education and life

			Is life exciting or dull			
			Dull	Routine	Exciting	Total
RS Highest Degree	Less than HS	Count	24	96	66	186
		Expected Count	12.0	85.8	88.2	186.0
		Row %	12.9%	51.6%	35.5%	100.0%
		Residual	12.0	10.2	-22.2	
	High school	Count	35	251	231	517
		Expected Count	33.3	238.5	245.3	517.0
		Row %	6.8%	48.5%	44.7%	100.0%
		Residual	1.7	12.5	-14.3	
	Junior college	Count	2	33	27	62
		Expected Count	4.0	28.6	29.4	62.0
		Row %	3.2%	53.2%	43.5%	100.0%
		Residual	-2.0	4.4	-2.4	
	Bachelor	Count	2	58	97	157
		Expected Count	10.1	72.4	74.5	157.0
		Row %	1.3%	36.9%	61.8%	100.0%
		Residual	-8.1	-14.4	**22.5**	
	Graduate	Count	1	21	51	73
		Expected Count	4.7	33.7	34.6	73.0
		Row %	1.4%	28.8%	69.9%	100.0%
		Residual	-3.7	-12.7	**16.4**	
Total		Count	64	459	472	995
		Expected Count	64.0	459.0	472.0	995.0
		Row %	6.4%	46.1%	47.4%	100.0%

College graduates have large positive residuals in the Exciting column

From the row percentages, you see that almost 70% of people with graduate degrees find life exciting. (They probably don't read or write statistics books!) Only 36% of people with less than a high school diploma find life exciting. In fact, as education increases, so does the likelihood of finding life exciting. (Don't be alarmed by the large number of missing observations. Not all people in the General Social Survey were asked the question.)

To test the null hypothesis that highest degree and perception of life are independent, you compute a chi-square statistic for this table the same way you did for a 2 × 2 table. For example, if the null hypothesis is true, the expected number of people without high school diplomas who find life exciting is 88.2. (That can be calculated by multiplying the overall percentage of people who find life exciting, 47.4%, by the number of people without high school diplomas, 186.)

The Pearson chi-square value for the table is shown in Figure 16.5. You see that the observed significance level is less than 0.0005, which leads you to reject the null hypothesis that degree and perception of life are independent. By looking at the residuals in Figure 16.4, you see that college graduates have large positive residuals for the response *Exciting*. That means that the observed number of college graduates in those cells is larger than that predicted by the independence hypothesis. By examining the residuals in a crosstabulation, you can tell where the departures from independence are.

Figure 16.5 Pearson chi-square for crosstabulation of education and life

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	53.962[1]	8	.000
Likelihood Ratio	55.874	8	.000
Linear-by-Linear Association	47.633	1	.000
N of Valid Cases	995		

[1] 2 cells (13.3%) have expected count less than 5. The minimum expected count is 3.99.

Check these values to be sure your test is valid

After the chi-square statistics are printed, SPSS tells you what the small est expected count is in any cell of the table. In this example, the *Minimum Expected Frequency* is 3.99. This is important because, if too many of the expected values in a table are less than 5, the observed significance level based on the chi-square distribution may not be correct. As a general rule, you should not use the chi-square test if more than 20% of the cells have expected values less than 5, or if the minimum expected frequency is less than 1.

What should I do if one of these conditions is not satisfied? If your table has more than two rows and two columns, you can see if it makes sense to combine some of the rows or columns. For example, if you have few people with graduate degrees, you can combine them into a single category with bachelor's degrees. Similarly, if necessary, you can combine the junior college graduates with the high school graduates, since their responses appear to be similar.

A One-Sample Chi-Square Test

So far, you've used the chi-square test to test for independence in a crosstabulation of two variables. You can also use the chi- square test to test null hypotheses about the distribution of values of a single variable. That is, you can see whether the distribution of observed counts in a frequency table is compatible with a set of expected counts. The expected counts are specified by the null hypothesis that you want to test. For example, you can test the hypothesis that people are equally likely to find life exciting, routine, or dull. Or you can test the null hypothesis that there are twice as many people without college degrees as there are with college degrees.

You can obtain this output using the Chi-Square Test procedure, as described in "Chi-Square Test" on p. 337 in Chapter 17. Select the variable life and All categories equal in the Chi-Square dialog box.

Figure 16.6 Chi-square test for life

	Observed N	Expected N	Residual
Dull	65	332.3	-267.3
Routine	459	332.3	126.7
Exciting	473	332.3	140.7
Total	997		

Expected counts if null hypothesis is true

	Is life exciting or dull
Chi-Square	322.865
df	2
Asymp. Sig.	.000

Look at Figure 16.6, which shows counts of the number of people who find life exciting, routine, and dull. Before you looked at the data, you might have thought that people were equally likely to find life exciting, routine, or dull. To test the null hypothesis that the three responses are equally likely in the population, you have to determine the expected counts for each of the categories. That's easy to do. For this hypothesis, the expected count for each category is just the total number of cases divided by 3.

You calculate the chi-square statistic the same way as before. Square each of the residuals (difference between observed and expected), divide by the expected count, and sum up for all of the cells. In Figure 16.6, you see that the chi-square value is a whopping 322.86. Its degrees of freedom are 2, one less than the number of categories in the table. Based on the observed significance level, you can handily reject the null hypothesis.

Let's try another test, this time specifying unequal numbers of expected counts for the categories. You want to test the null hypothesis that there are twice as many people in the population without college degrees as there are people with college degrees. That means you expect two-thirds of the people not to have college degrees and one-third to have college degrees. The expected counts for the two cells are 997.3 and 498.7.

Figure 16.7 Chi-square test for degree

	Observed N	Expected N	Residual
No College degree	1149	997.3	151.7
College degree	347	498.7	-151.7
Total	1496		

	College Degree
Chi-Square	69.193
df	1
Asymp. Sig.	.000

Observed results are highly unlikely if null hypothesis is true

The results of this test are shown in Figure 16.7. You see that the expected count for people without a degree is twice as large as it is for people with a college degree. From the residuals, you see that the two-to-one-ratio hypothesis predicts more college graduates than you observe. In the sample, the ratio is slightly larger than three to one. Again the chi-square statistic is large and the observed significance level is small, so you reject the null hypothesis that in the population, non-college graduates are twice as common as college graduates.

Power Concerns

You know that your ability to reject the null hypothesis when it's false, the power of a test, depends not only on the size of the discrepancy from the null hypothesis, but also on the sample size. The same is true, of course, for chi-square tests. The value of the chi-square statistic depends on the number of observations in the sample. For example, if you leave the table percentages unchanged but multiply the number of cases in each cell by 10, the chi-square value will be multiplied by 10 as well. This means that if you have small sample sizes, you may not be able to reject the null hypothesis even when it's false. Similarly, for large sample sizes, you will find yourself rejecting the null hypothesis even when the departures from independence are quite small.

When one or both of the variables in your crosstabulation is measured on an ordinal scale (for example, good/better/best), the chi-square test is not as powerful as some other statistics for detecting departures from independence. These other statistics make use of the additional information available for ordinal variables to measure both the strength and the direction of the relationship between two variables. If examination of the residuals in such a table leads you to suspect that there are departures from independence, you should use one of the measures described in Chapter 18.

Summary

How can you test the null hypothesis that two variables are independent?

- In a crosstabulation, the observed count is the number of cases in a particular cell.
- An expected count is the number of cases predicted if the two variables are independent.
- The chi-square statistic is based on a comparison of observed and expected counts.
- To use the chi-square test, your observations must be independent, and most of the expected values must be at least 5.
- A one-sample chi-square test is used to test whether a sample comes from a population with specified probabilities for the occurrence of each value.

de VAUS D. A. 1990. *Survey in Social Research*. London: Unwin Hyman.

Table 11.4 The links between correlations and tests of significance

Correlation	Significance	N	Interpretation
0 35	0 27	100	Moderate association in sample but too likely to be due to sampling error Continue to assume correlation of 0 in the population
0 15	0 001	1500	Weak association but is very likely to hold in the population.
0 64	0 01	450	Strong relationship that is likely to hold in the population.
0 04	077	600	Negligible association Highly probable that the correlation differs from zero due only to sampling error Continue to assume correlation of 0 in the population.

de VAUS D. A. 1990. *Survey in Social Research*. London: Unwin Hyman.

Table 11.5 Guidelines for selecting measures of association

Level of measurement of variables		Appropriate methods	Appropriate descriptive summary statistics	Appropriate inferential statistic
1 Nominal/Nominal	'Shape' of variables 2 by 2	Crosstabulations	i Phi ii Yules Q iii Lambda iv Goodman & Kruskall's tau	chi square
2 Nominal/Nominal	3+ by 2+	Crosstabulations	i Lambda ii Goodman & Kruskall's tau iii Cramers V	chi square
3 Nominal/Ordinal	Nominal variable with 3+ categories	Crosstabulations	i Theta ii Any statistics in 2 above	Mann - Whitney U-test (dichotomous nominal independent variable) K-sample median test Kruskal · Wallis
4 Nominal/Interval	Nominal variable independent	a Crosstabulations (if interval variable has only a few categories) b Comparison of means (esp. if interval variable has many categories)	i Eta (also called correlation ratio) ii Any statistics in 2 or 3 above but not very wise. i Eta	F-test (one-way analysis of variance) chi square F-test (one-way analysis of variance)
5 Ordinal/Ordinal	Both with few categories	Crosstabulations	i Gamma ii Kendall's tau b (square tables) iii Kendall's tau c (any shape table)	Test for significance of gamma Test for significance of tau
6 Ordinal/Ordinal	One variable with many categories	Rank correlation	i Kendall's tau	Test for significance of tau
7 Ordinal/Ordinal	Both variables with many categories	Rank correlation	i Kendall's tau ii Spearman's rho	as above Test for significance of rho
8 Ordinal/Interval	Both with few categories	a Crosstabulations b Comparison of means (if dependent variable is interval)	i Eta ii Any statistics in 5 above i Eta	F-test F-test
9 Ordinal/Interval	Ordinal with low categories Interval with many	a Comparison of means b Rank order correlation	i Eta i Kendall's tau	F-test Test for significance of tau
10 Ordinal/Interval	Both with many categories	Rank correlation	i Kendall's tau ii Spearman's rho	as above Test for significance of rho

de VAUS D. A. 1990. *Survey in Social Research*. London: Unwin Hyman.

Table 11.5 Guidelines for selecting measures of association

Level of measurement of variables			Appropriate methods	Appropriate descriptive summary statistics	Appropriate inferential statistic
11	Interval/Interval	Both variable with small number of categories	Crosstabulations	i Pearson's r	Test for significance of r
12	Interval/Interval	At least one variable with many categories	Scattergram	i Pearson's r ii Regression	

Table 10.15 Characteristics of various measures of association

	Appropriate table size	Range	Directional	Symmetric	Linear only	Other features
Phi	2 × 2	0 - 1[2]	no	yes	no	Lower co-efficients than Yule's Q
Cramer's V	larger than 2 × 2	0 - 1	no	yes	no	More sensitive to a wider range of relationships than lambda
Yule's Q	2 × 2	0 - 1	no	yes	no	1. Higher co-efficients than phi 2. Same as gamma 2 by 2 case 3 Always 1 00 if an empty cell
Lambda	any size[1]	0 - 1	no	yes[4]	no	Insensitive and therefore not recommended
Goodman and Kruskal's tau	any size	0 - 1[3]	no	no	no	More sensitive than lambda but not available on SPSS
Gamma	any size	0 - 1	yes	yes	yes	Gives higher co-efficients than Kendall's Tau_b or Tau_c
Kendall's Tau_b	square tables only	0 - 1	yes	yes	yes	
Kendall's Tau_c	any size	0 - 1	yes	yes	yes	
Eta	any size	0 - 1	no	no	no	
Pearson's r	any size	0 1	yes	yes	yes	

Notes: (1) i.e. given the qualifications in section 10 1.5
(2) Under certain conditions the maximum may be less than 1 (see Guilford, 1965 336)
(3) Will only be if there is perfect association and if the independent variable has the same number of categories as the dependent variable
(4) There is both a symmetric and asymmetric version

9. lekce

MĚŘENÍ (SÍLY) ASOCIACE MEZI DVĚMA SPOJITÝMI PROMĚNNÝMI: KORELAČNÍ KOEFICIENTY A GRAFY - SCATTERPLOTS (modul GRAF: procedura Scatter) A KORELAČNÍ MATICE (modul CORRELATE: procedura bivariate) A.

WILLIAMSON, J. B., David A. KARP, John R. DALPHIN, Paul S. GRAY. 1982. *The Research Craft. An Introduction to Social Research Methods.* Boston: Little, Brown and Company.

The scattergrams in Figure 16.5 illustrate several alternative relationships between the independent variable X and the dependent variable Y. The following observations can be made on the basis of the information in Figure 16.5:

(A) the data for this scattergram illustrate a moderately strong positive correlation that would be approximately .60. You will note that in this scattergram, as in the others, the X values increase from left to right, that is, from L (low) to H (high); and the Y values increase from bottom to top (also from low to high). As with all positive correlations, there is a tendency for the Y values to increase as the X values increase.

(B) Here all the data points fall along a straight line; this is what happens when there is a perfect positive correlation between X and Y ($r = 1.00$). The correlation is perfect only in the sense that it represents the upper limit for the correlation coefficient. In actual social research applications we do not get correlations of 1.00 unless we have somehow managed to correlate a variable with itself.

(C) Here there is no relationship between X and Y ($r = .00$).

(D) Here there is a weak positive correlation ($r = +.20$) between X and Y.

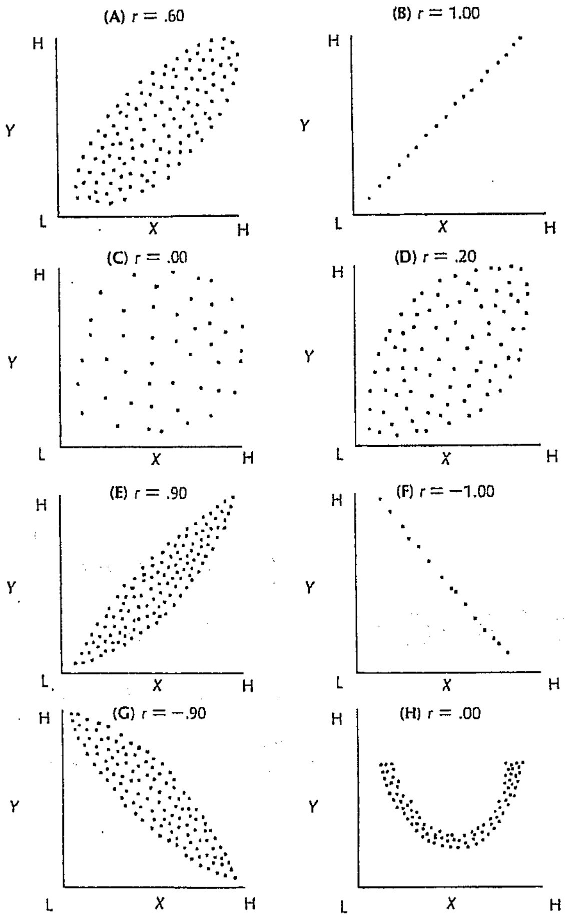
(E) Here there is a very strong positive correlation ($r = +.90$).

(F) Here there is a perfect negative correlation ($r = -1.00$). Note that for a negative correlation Y decreases as X increases.

(G) Here there is a strong negative correlation ($r = -.90$). An example of a negative correlation would be the relationship between cigarette consumption (X) and life expectancy (Y). As cigarette consumption increases, life expectancy decreases. (Undoubtedly, the actual correlation between these two variables is weaker than $-.90$.)

(H) Here there is a strong NONLINEAR RELATIONSHIP between X and Y ($r = .00$). It is not appropriate to use the correlation coefficient to summarize this relationship. The low correlation masks the evidence of its pronounced nonlinear shape.

FIGURE 16.5 Scattergrams for Alternative Correlations between X and Y



MORGAN George A. and Griego V. ORLANDO. 1998. *Easy Use and Interpretation of SPSS for Windows: Answering Research Questions With Statistics*. Mahwah: Lawrence Erlbaum Associates.

CHAPTER 9

Correlation and Scatterplots

In this assignment, you will learn about how to compute the basic associational statistics. The Pearson correlation is a parametric statistic used when both variables are at least interval scale. When you have ranked data or when other assumptions (such as normality of the data) are markedly violated, one should use a nonparametric equivalent of the Pearson correlation coefficient (such as Spearman's rho or Kendall's tau). The Kendall's tau is said to deal with ties in a better way than the Spearman rho. Here we ask you to compute all three correlations and compare them.

Chapter 7 is important background because it will help you understand when to compute/choose associational statistics, and it will remind you about what the significance test means and how to interpret it.

Problems/Research Questions

1. What is the association between grades in high school and math achievement? You will compute three bivariate (2 variable) **correlations** (Pearson, Spearman, and Kendall's tau-b) of *grades* and *mathach.*

2. What are the correlations among all of the variables, *mathach, visual, mosaic, mathcrs, pleasure, comptnc,* and *motivatn,* using Pearson correlations.

3. In this problem, you will compare **pairwise** and **listwise** exclusion of missing data.

4. Using the Graphs menu, you will request **Scatterplots** with the linear, quadratic, and cubic regression lines and r^2 printed on the scatterplot for *grades* and *mathach* and for some of the other correlations.

Lab Assignment E

Logon and Get Data

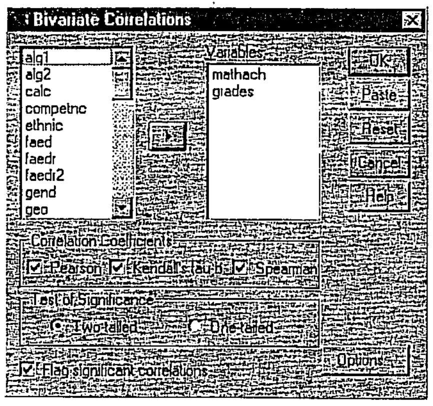
- Retrieve **hsbdataD** from your **Data file.**

Problem 1: Correlate Grades and Math Achievement

To do Pearson, Kendall, and Spearman correlations follow these commands:

- **Statistics => Correlate => Bivariate.**
- Move *mathach* and *grades* to the **Variables** box.
- Next, ensure that the **Pearson, Kendall's tau-b, and Spearman** boxes are checked.

- Make sure that the **Two-tailed** (under **Test of Significance**) and **Flag significant correlations** are checked (see Fig. 9.1).
- Now click on **Options** to get that dialog box.
- Click on **Means and standard deviations** and note that **Exclude cases pairwise** is checked. Does your screen look like Fig. 9.2?
- Click on **Continue** then on **OK.** What does your output file look like? Compare Output 9.1 to your output and syntax.


Bivariate Correlations

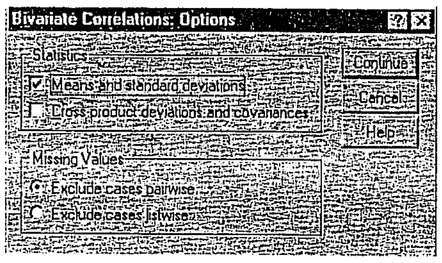
Variables: mathach, grades

Correlation Coefficients: ☑ Pearson ☑ Kendall's tau-b ☑ Spearman

Test of Significance: ● Two-tailed ○ One-tailed

☑ Flag significant correlations

Fig. 9.1. Bivariate correlations.


Bivariate Correlations: Options

Statistics: ☑ Means and standard deviations ☐ Cross-product deviations and covariances

Missing Values: ● Exclude cases pairwise ○ Exclude cases listwise

Fig. 9.2. Bivariate correlations: Options.

Problem 2: Correlation Matrixes for Several Interval Scale Variables

Now, on your own, compute a **Pearson** correlation among all the following variables: *mathach, visual; mosaic, mathcrs, pleasure, competnc,* and *motivatn.* Follow similar procedures outlined previously except:

- Click *off* **Kendall's tau-b** and **Spearman** (under **Correlation Coefficients**).

It is usually best, except in exploratory research with small samples, to use two-tailed tests. The 'flag" puts an asterisk beside the correlation coefficients that are statistically significant so that

MORGAN George A. and Griego V. ORLANDO. 1998. Easy Use and Interpretation of SPSS for Windows: *Answering Research Questions With Statistics*. Mahwah: Lawrence Erlbaum Associates.

they can be identified quickly. The output also prints the exact significance level (p) which is redundant with the asterisk so you wouldn't report both in a thesis or paper.

- For **Options**, obtain **Means and standard deviations**, and **Exclude cases pairwise**.

This will produce Output 9.2, which was reduced in size to fit on the page. To see if you are doing the work right, compare your own syntax file and output to Output 9.2.

Problem 3: Correlations With Pairwise Exclusions

Next, rerun the same analysis, except:

- Click *off* **Means and standard deviation** (in the **Options** window).
- Change **Exclude cases pairwise** to **Exclude cases listwise** (under **Missing Values)**.

Now, compare the correlations in Output 9.3 (listwise exclusion of participants with any missing data) to the Pearson correlations in Output 9.2 (pairwise deletion). Are they the same?

Problem 4: Scatterplots - Mathach With Grades

Let's now work on developing a scatterplot of the correlations of *mathach* with *grades*. Follow these commands:

- **Graphs => Scatter**. This will give you Fig. 9.3.
- Click on **Simple** then **Define** which will bring you to Fig. 9.4.

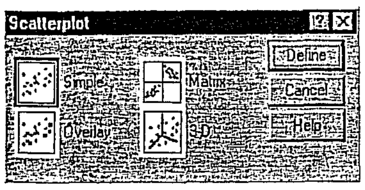
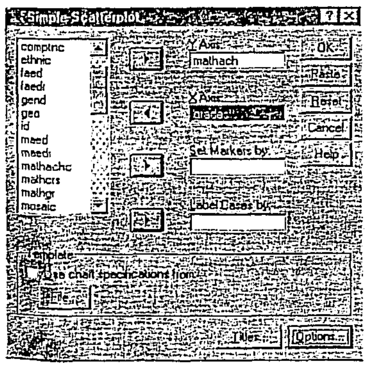


Fig. 9.3. Scatterplot.



Y Axis: mathach

X Axis: grades

Fig. 9.4. Simple scatterplot.

- Now, move *mathach* to the **Y** axis and *grades* to the **X** axis (*the dependent variable goes on the Y axis*).
- Click on **Options** and make sure **Exclude cases listwise** is highlighted (see Fig. 9.5).
- Click on **Continue.**
- Next, click on **Titles** (in Fig. 9.4) and type **"Correlation of math achievement with high school grades"** (see Fig. 9.6).
- Click on **Continue** then on **OK**. You will get an output chart which looks like Fig. 9.7.

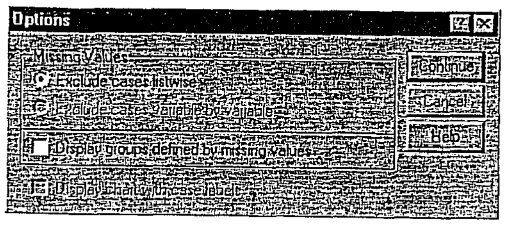
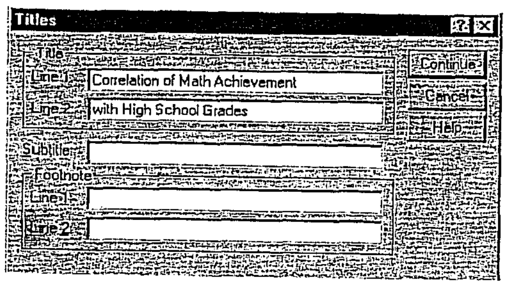

Options
Missing Values
Exclude cases listwise
Exclude cases variable by variable
Display groups defined by missing values
Display chart with case labels
Continue
Cancel
Help

Fig. 9.5. Options.



Title Line 1: Correlation of Math Achievement

Title Line 2: with High School Grades

Fig. 9.6. Titles.

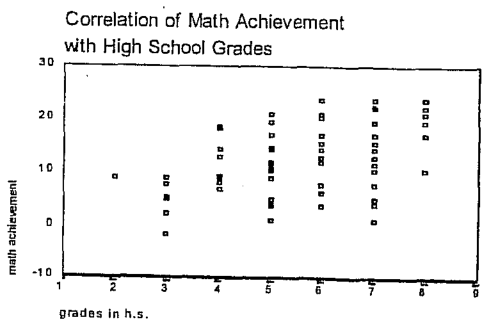


Fig. 9.7. Scatterplot output.

MORGAN George A. and Griego V. ORLANDO. 1998. Easy Use and Interpretation of SPSS for Windows: *Answering Research Questions With Statistics*. Mahwah: Lawrence Erlbaum Associates.

Now let's put the regression lines on the scatterplot so we can get a better sense of the correlation and how much scatter or deviation from the line there is.

- *Double click* on the chart in the output file. You will see a dialog box like Fig. 9.8.
- Select **Chart => Options** until you see Fig. 9.9.
- Click on **Total** in the **Fit Line** box and **Show sunflowers**; there is no need to change the **Sunflower Options**. The sunflowers indicate, by the number of petals, how many participants had essentially the same point on the scatterplot.
- Next, click on the **Fit Options** button, which will give you Fig. 9.10.
- Ensure that the **Linear Regression** box is highlighted.
- Then check the **Individual** box and **Display R-Square in legend** box. Check to be sure your window is like Fig. 9.10.
- Click on **Continue** then **OK**.

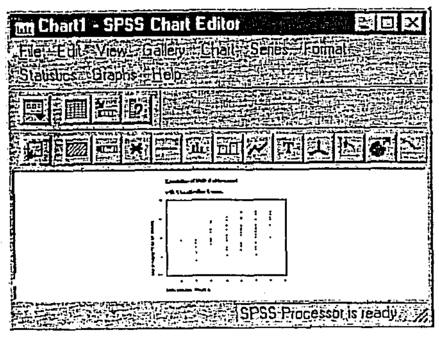
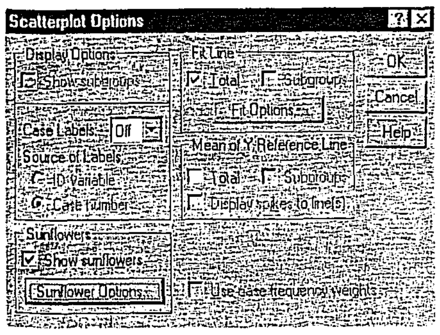


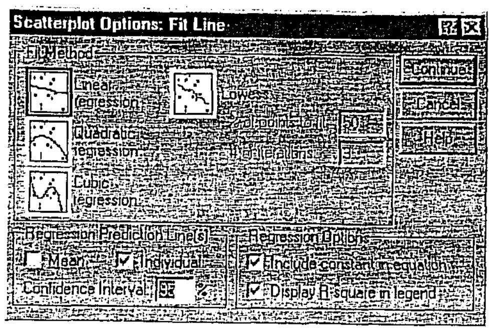
Fig. 9.8. SPSS chart editor.



Fit Line: ☑ Total □ Subgroups

Sunflowers: ☑ Show sunflowers

Fig. 9.9. Scatterplot options.



□ Mean ☑ Individual

☑ Include constant in equation

☑ Display R-square in legend

Fig. 9.10. Chart: Scatterplot.

Now, if the points on the scatterplot do not lie close to the regression line, it could be that the data were curvilinear (better fit a curved line). If so, you could (in Fig. 9.10) click on **Quadratic** and possibly the **Cubic regression** boxes (one at a time) to and see what the fit and r^2 look like. If the quadratic and/or cubic r^2 are quite a bit higher, a linear Pearson correlation is not the best statistic to use. Output 9.4 shows the quadratic and cubic regression lines as well as the linear chart. Check your syntax and output against Output 9.4.

Now try the following scatterplots by doing the same steps as Problem 3. Don't forget to *change the title* before you run each scatterplot.

1. *Mosaic* (X) with *mathach* (Y).
2. *Mathcrs* (X) with *mathach* (Y).

Do your syntax and output look like the ones in Output 9.5 and 9.6?

Print, Save, and Exit

- Print your output if you want.
- Save your data file as **hsbdataE** (**File => Save As**).
- Save the SPSS log files as **hsblogE**.
- Exit SPSS.

Interpretation Questions

1. In Output 9.1: a) What do the correlation coefficients tell us? b) What is r^2 for the Pearson correlation? What does it mean? c) Compare the Pearson, Kendall, and Spearman correlations on both correlation size and significance level. d) When should you use which type?

2. In Output 9.2, how many of the Pearson correlation coefficients are significant?

3. In Output 9.3: a) How many Pearson correlations are there? b) How many are significant?

4. Write an interpretation of a) one of the significant and b) one of the nonsignificant correlations in Output 9.3. Include whether or not the correlation is significant, your decision about the null hypothesis, *and* a sentence or two describing the correlations in nontechnical terms.

5. What is the difference between the pairwise and listwise correlation matrixes?

6. Using Outputs 9.5, and 9.6, inspect the scatterplots. a) What is r^2? b) Is the linear relationship as good as a curvilinear (quadratic) one? c) Why should one do scatterplots?

Outputs and Interpretations

```
GET
FILE='A:\hsbdataD.sav'.
EXECUTE .
```

Output 9.1: Pearson, Spearman, and Kendall's Tau-b Correlations

Syntax for Pearson correlation of math achievement with grades in h.s.

```
CORRELATIONS
  /VARIABLES=mathach grades
  /PRINT=TWOTAIL SIG
  /STATISTICS DESCRIPTIVES
  /MISSING=PAIRWISE .
```

Interpretation of Output 9.1

The first table provides **descriptive statistics** for the variables to be correlated, in this case math achievement and grades. The two **correlations** tables are the key. Each has three parts, with the information in matrix form which, unfortunately, means that every number is presented twice. We have provided callout boxes to help you.

The Pearson correlation coefficient is .504; the significance level or *p* is .000 and the number of participants with both variables (*mathach* and *grades*) is 75. In a report, this would usually be written as: *r* (73) = .50, *p* < .001. Note that the degrees of freedom (*N*-2 for correlations) is put in parentheses after the statistic (*r* for Pearson correlation) which is usually rounded to two decimal places. The significance or *p* value follows and is stated as less than .001 rather than .000. Note that the correlation values for Kendall's tau-b and Spearman's rho are different from *r*, but in this case they have the same significance level (*p* < .001).

This correlation is significant, because the "sig" is less than .05, (*p* < .05) so we can reject the null hypothesis of no association and state that there *is an association* between grades and math achievement. Because the correlation is positive, students who have high grades generally have high math achievement scores and vice versa. This means that high grades generally are *associated* with high achievement, medium with medium, and low with low. If the correlation is significant and *negative* (e.g., -.50), high grades would be associated with low achievement and vice versa. If the correlation was not significant, there would be *no* systematic association between a student's grades and achievement. In that case you could not predict anything about math achievement from knowing someone's grades.

Descriptive Statistics

	Mean	Std. Deviation	N
math achievement	12.5645	6.6703	75
grades in h.s.	5.68	1.57	75

Correlations

		math achievement	grades in h.s.
Pearson Correlation	math achievement	1.000	.504
	grades in h.s.	.504	1.000
Sig. (2-tailed)	math achievement	.	.000
	grades in h.s.	.000	.
N	math achievement	75	75
	grades in h.s.	75	75

Note the diagonal is always 1.00.

Correlation between math achievement and grades in high school.

The statistical level of significance or *p*.

Number of participants with both variables.

These numbers are redundant. Ignore them.

Syntax for Kendall's Tau-b and Spearman Rho correlations of math achievement with grades in h.s.

```
NONPAR CORR
  /VARIABLES=mathach grades
  /PRINT=BOTH TWOTAIL NOSIG
  /MISSING=PAIRWISE .
```

Nonparametric Correlations

Correlations

			math achievement	grades in h.s.
Kendall's tau-b	Correlation Coefficient	math achievement	1.000	.370**
		grades in h.s.	.370**	1.000
	Sig. (2-tailed)	math achievement	.	.000
		grades in h.s.	.000	.
	N	math achievement	75	75
		grades in h.s.	75	75
Spearman's rho	Correlation Coefficient	math achievement	1.000	.481**
		grades in h.s.	.481**	1.000
	Sig. (2-tailed)	math achievement	.	.000
		grades in h.s.	.000	.
	N	math achievement	75	75
		grades in h.s.	75	75

**. Correlation is significant at the .01 level (2-tailed).

Output 9.2: Pearson Correlation Matrix (Pairwise Exclusion)

Syntax for Pearson correlation matrixes (pairwise exclusion of missing data)

```
CORRELATIONS
  /VARIABLES=mathach visual mosaic mathcrs pleasure competnc motivatn
  /PRINT=TWOTAIL NOSIG
  /STATISTICS DESCRIPTIVES
  /MISSING=PAIRWISE .
```

Interpretation of Output 9.2

Notice that after the descriptive statistics table, there is a large **correlations** table divided into three sections: Pearson correlation coefficients, significance, and *N*. These numbers are, as in Output 9.1, each given twice so you have to be careful in reading them. It is a good idea to look only at the numbers below the diagonal (1.00 as in the coefficients section, dots in the significance section, and 75s in the *N* section). There are 21 different correlations in the table. In the first column, there is the correlation of each of the other six variables with math achievement. In the second column, each of the other six variables is correlated with visualization score, but note that the .423 for *visual* and *mathach* is the same as the correlation of *mathach* and *visual* in the first column, so ignore it. The Pearson correlations on this table are interpreted similarly to the one in Output 9.1. However, because there are 21 correlations, the odds are that at least one could be statistically significant by chance (i.e., .05= 1/20). Thus, it would be prudent to use the .01 level of significance. The Bonferroni correction (.05/21= .002) would be a conservative approach designed to keep the significance level at .05 for the whole study.

Descriptive Statistics

	Mean	Std. Deviation	N
math achievement	12.5645	6.6703	75
visualization score	5.2433	3.9120	75
mosaic, pattern test	27.413	9.574	75
Math course taken	2.11	1.67	75
Pleasure scale	3.2267	.6300	75
Competence scale	3.2945	.6645	73
Motivation scale	2.8744	.6382	73

Correlations with math achievement.

Correlations with visualization score.

Use correlations above or below line.

Correlations

		math achievement	visualization score	mosaic, pattern test	Math course taken	Pleasure scale	Competence scale	Motivation scale
Pearson Correlation	math achievement	1.000	.423**	.213	.794**	.094	.332**	.316**
	visualization score	.423**	1.000	.030	.399**	-.160	.007	.047
	mosaic, pattern test	.213	.030	1.000	-.059	.085	.111	.083
	Math course taken	.794**	.399**	-.059	1.000	-.006	.309**	.298*
	Pleasure scale	.094	-.160	.085	-.006	1.000	.431**	.305**
	Competence scale	.332**	.007	.111	.309**	.431**	1.000	.570**
	Motivation scale	.316**	.047	.083	.298*	.305**	.570**	1.000
Sig. (2-tailed)	math achievement	.	.000	.067	.000	.421	.004	.006
	visualization score	.000	.	.798	.000	.171	.954	.695
	mosaic, pattern test	.067	.798	.	.616	.466	.349	.487
	Math course taken	.000	.000	.616	.	.958	.008	.010
	Pleasure scale	.421	.171	.466	.958	.	.000	.009
	Competence scale	.004	.954	.349	.008	.000	.	.000
	Motivation scale	.006	.695	.487	.010	.009	.000	.
N	math achievement	75	75	75	75	75	73	73
	visualization score	75	75	75	75	75	73	73
	mosaic, pattern test	75	75	75	75	75	73	73
	Math course taken	75	75	75	75	75	73	73
	Pleasure scale	75	75	75	75	75	73	73
	Competence scale	73	73	73	73	73	73	71
	Motivation scale	73	73	73	73	73	71	73

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Output 9.3: Pearson Correlation Matrix (Listwise Exclusion)

Syntax for Pearson correlation matrix

```
CORRELATIONS
  /VARIABLES=mathach visual mosaic mathcrs pleasure competnc motivatn
  /PRINT=TWOTAIL NOSIG
  /MISSING=LISTWISE .
```

Interpretation of Output 9.3

In this table there is not a separate section for *N* because, with **listwise** exclusion, only the same 71 subjects who have scores on all seven variables are used for all correlations. Note that the correlations are slightly different from those in Output 9.2 where the *N*'s varied depending on how many subjects had each pair of variables. Factor analysis, Cronbach's alpha, and multiple regression (Assignments F, G, and H) all use listwise deletion, so if you have one or more variables with quite a bit of missing data the *N* may be dramatically reduced.

MORGAN George A. and Griego V. ORLANDO. 1998. *Easy Use and Interpretation of SPSS for Windows: Answering Research Questions With Statistics*. Mahwah: Lawrence Erlbaum Associates.

Correlations[a]

		math achievement	visualization score	mosaic, pattern test	Math course taken	Pleasure scale	Competence scale	Motivation scale
Pearson Correlation	math achievement	1.000	.434**	.246*	.804**	.101	.335**	.316**
	visualization score	.434**	1.000	.035	.429**	-.191	.010	.047
	mosaic, pattern test	.246*	.035	1.000	.000	.067	.104	.083
	Math course taken	.804**	.429**	.000	1.000	.005	.318**	.301*
	Pleasure scale	.101	-.191	.067	.005	1.000	.443**	.309**
	Competence scale	.335**	.010	.104	.318**	.443**	1.000	.570**
	Motivation scale	.316**	.047	.083	.301*	.309**	.570**	1.000
Sig. (2-tailed)	math achievement	.	.000	.038	.000	.400	.004	.007
	visualization score	.000	.	.773	.000	.111	.931	.695
	mosaic, pattern test	.038	.773	.	.999	.576	.386	.489
	Math course taken	.000	.000	.999	.	.964	.007	.011
	Pleasure scale	.400	.111	.576	.964	.	.000	.009
	Competence scale	.004	.931	.386	.007	.000	.	.000
	Motivation scale	.007	.695	.489	.011	.009	.000	.

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

a. Listwise *N*=71

Cases with missing values for any variable are excluded from all correlations using listwise.

Output 9.4: Scatterplots, Grades With Math Achievement

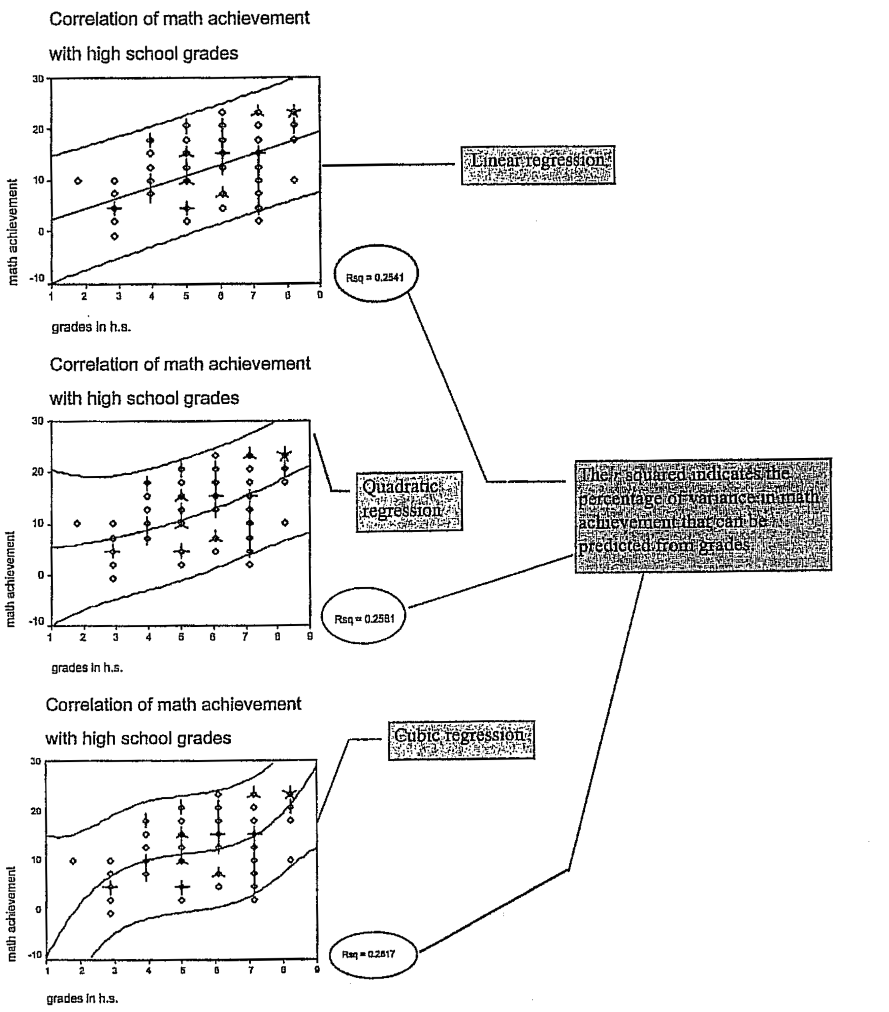
Syntax for Scatterplot of grades with math achievement with linear, quadratic (curved), and cubic (2 bend) regression

```
GRAPH
  /SCATTERPLOT(BIVAR)=grades WITH mathach
  /MISSING=LISTWISE
  /TITLE= 'Correlation of math achievement' 'with high school grades'.
```

Interpretation of Output 9.4

The three scatterplots shown in Output 9.4 are the same except that the first shows the best fit for a straight (linear) regression line (i.e., it minimizes the squared differences between the points and the line), and the second is the best fit for a line with one curve (quadratic) and the third is the best fit for a line with two bends or curves (cubic). To the right of each scatterplot is the r^2 which indicates the percentage of variance in the dependent variable (*mathach*) that can be predicted from the independent variable (*grades*). Note that the r^2 (.2541) for the linear (first) scatterplot means that $r = .504$, which is what Output 9.1 had for the same correlation.

Note also that the r^2 for the quadratic (.2581) and the r^2 for the cubic regression (.2817) are quite similar to that (.2541) for the linear regression line. Thus, there is little improvement in fit over the linear regression, which is what is used in computing the Pearson and Kendall correlations.



Linear regression

Quadratic regression

Cubic regression

The r squared indicates the percentage of variance in math achievement that can be predicted from grades.

10. lekce

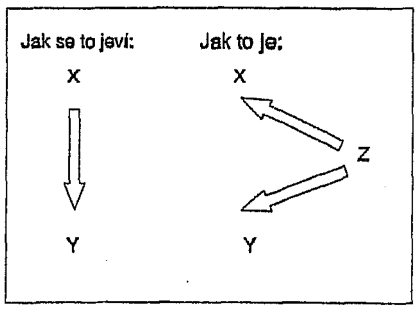
JAK ODHALIT VLIV TŘETÍ PROMĚNNÉ: PRÁCE S PODSOUBORY NEBOLI TŘÍDĚNÍ VYŠŠÍCH STUPŇŮ A PARCIÁLNÍ KOEFICIENTY.

Nepravá korelace

V řadě evropských regionů bylo zjištěno, že čím více čápů žije v určité krajině, tím vyšší je tam porodnost. Korelační koeficienty byly tak významné, že je velice nepravděpodobné, že zjištěná souvislost je náhodná. Jsme tedy ochotni přijmout hypotézu, že čápi přece jen nosí děti? Asi sotva. Ale pak je naší povinností navrhnout hypotézu, která by uspokojivě vysvětlovala naměřenou souvislost.

Graf 1.2.

Nepravá korelace



Toto je klasický příklad nepravé korelace ("spurious correlation"). Zkreslení vzniká tehdy, když třetí nepozorovaná nebo neanalyzovaná proměnná ovlivňuje nějak obě proměnné X a Y, které studujeme.

Cvičení 1.3.

Podívejte se pečlivě na graf 1.2., popisující nepravou korelaci. Navrhněte, co může být to tajemné Z.

Jistě nám nehrozí nebezpečí, že bychom přijali hypotézu, že čápi nosí děti. Ale představme si, že nepravá korelace se zdá potvrzovat naši oblíbenou hypotézu. Potom výzkumník musí mít objektivnost anděla a trpělivost nerostného krystalu, aby pracně zabil to, co se po měsíce pokoušel dokázat.

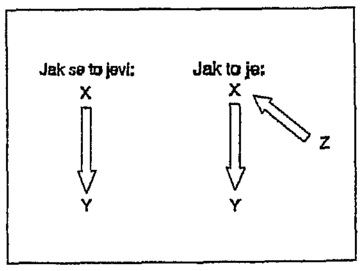
Nepravá korelace je skutečným nebezpečím ve výzkumu. Není to ani tak technický problém analýzy, ale spíše problém lidské kvality výzkumníka.

Vývojová sekvence

Tak nazýváme zkreslení, způsobené faktem, že proměnná X, která ovlivňuje Y, je určována předcházející, ale nepozorovanou proměnnou Z.

Graf 1.3.

Vývojová sekvence



Taková situace je skutečně naprosto nevyhnutelná. Každá příčina má totiž jinou příčinu, ta zase jinou příčinu, která má opět svoji příčinu, a tak bychom mohli pokračovat až k aktu stvoření, nebo k tomu, co astronomové nazývají Big Bang. To je problém velmi dobře známý filozofům, kteří ho obvykle nazývají "regresus ad infinitivum"

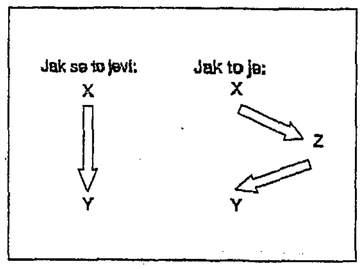
Nicméně, někdy může předčasné přerušení kauzálního řetězce vést k mylné interpretaci. Některé studie tvrdí, že četba pornografické literatury vyvolává násilné chování mužů k ženám. Nelze však vyloučit, že je zde nějaký předcházející činitel, jako kupř. autoritativní metoda socializace respondenta v dětství, který vyvolal silný zájem jedince o pornografii. Takový omyl je závažný zejména tehdy, když cílem výzkumu je sociální intervence.

Chybějící střední člen

Tak je označována situace, kde mezi nezávisle proměnnou X a závislou Y je ještě proměnná Z, kterou jsme nezahrnuli do analýzy. Graf 1.3. tuto situaci jasně popisuje. Je to opět konfigurace, která je téměř všudypřítomná. Kdybychom se jakousi sociologickou lupou podívali, co se děje mezi nějakou příčinou a jejím následkem, existuje ještě řada mezikroků. Často můžeme tyto elementy ignorovat bez rizika zkreslení. Ne však vždycky.

Graf 1.4.

Chybějící střední člen



Řekněme, že X reprezentuje pohlaví respondentů a Y jejich skóre v testu inteligence. Je možné, že výsledky žen, a to zejména žen příslušejících k nižším sociálním třídám, by byly signifikantně nižší, než výsledky mužů.

Cvičení 1.4.

Zamyslete se, prosím, nad předchozím odstavcem a navrhněte alternativní hypotézu, ukazující mužským šovinistům, že takové výsledky nepotvrzují superioritu nás, pánů tvorstva.

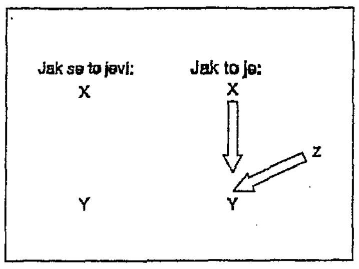
Zkreslení tohoto typu může být nebezpečné. Můžeme je často najít v quasivědeckých pracích, podporujících rasismus, v některých politických pamfletech atd. Mnohé noviny se dopouštějí tohoto hříchu z nevědomosti, když publikují výsledky statistických šetření.

Dvojí příčina

Takto můžeme označit situaci, kdy závislá proměnná Y má dvě příčiny, ale jenom jedna z nich, X, byla zahrnuta do výzkumu. Toto je asi nejčastější problém výzkumu v sociálních vědách. Pravděpodobně neexistuje žádný sociální jev, který by měl jedinou příčinu. I v našem nesmírně zmenšeném vesmíru, složeném jenom ze tří proměnných, si můžeme představit, jaké zkreslení může vyvolat, není-li tato další příčina zahrnuta do analýzy.

Graf 1.5.

Dvojí příčina



Cvičení 1.5.

Představme si třeba, že X je vzdělání jedince a že Y je jeho příjem. Pokud přepokládáme, že se vzděláním příjem poroste, mohli bychom zjistit, že souvislost je velice nízká, nebo dokonce nulová. Co mohlo vyvolat toto zkreslení?

Teoreticky by bylo možné namítnout, že v některých situacích neměřená proměnná Z může posilovat vliv příčiny X. Ale to je krajně nepravděpodobné. Jak vidíme z grafu 1.4., mezi X a Y není žádný příčinný vztah. I naše cvičení 1.5. je platné jenom uvnitř našeho nerealisticky miniaturizovaného systému tří proměnných. Je jasné, že realisticky by bylo třeba zahrnout další proměnnou "věk", která ovlivňuje vzdělání, a prostřednictvím "zkušenosti", "seniority" i příjem.

9.1. Proč tabulka nemusí být placatá

Někdy není těžké zapochybovat, že s určitou zjištěnou souvislostí není všechno v pořádku. Zejména je snadné pozastavit se nad některými nálezy, publikovanými v denním tisku. *Toronto Star* jsou dobré a seriózní noviny, jedny z nejlepších v Kanadě a možná i v celé Americe. Je proto vysoce pravděpodobné, že zpráva, ze které uvádíme výňatek, není plodem novinářovy fantazie, ale je založena na poněkud svérázné interpretaci skutečně existujícího výzkumu.

> TEMPE, Ariz. (AFP)
>
> Ve studii sponzorované US vládou se uvádí, že osoby, které rády jedí hamburgy, milují své rodiny, svoji práci a náboženství... Labužníci, kteří dávají přednost ústřicím a kaviáru, mají obecně ateistické a liberální postoje..
>
> *Toronto Star*, 9.listopadu 1981.

Nesnažili jsme se obdržet původní data, ale ta mohla mít třeba takovou distribuci, jako v tabulce 9.1.

Tabulka 9.1.

	Preferované jídlo:		Celkem:
Zbožnost:	HAMBURGY	KAVIÁR	
vysoká	78% 780	25% 125	905
nízká	22% 220	75% 375	595
Celkem: N	100% 1000	100% 500	1500

LAMBDA = .420

Tahle data ukazují, že mezi oběma proměnnými existuje povážlivá souvislost. Kdybychom tuto souvislost interpretovali jako kauzální, znamenalo by to v sociologii úplnou revoluci. Sotva by někdo z nás by byla připraven obhajovat teorii o biochemických determinantách postojů nebo experimentálně testovat možnost, jak změny v dietě změní jednotlivcovu morálku. Všichni si včas vzpomeneme na klasický příklad s čápy a porodností, na koncept nepravé souvislosti. Teď jde jen o to navrhnout, co je ten třetí faktor, který vyvolal souběžné změny v obou našich proměnných, a hlavně **dokázat**, že souvislost naměřená v naší tabulce je v nějaké podstatnější míře vyvolána tímto faktorem.

Nalézt něco, co ovlivňuje právě tak vzorce preferencí ve stravě, jako postoje k náboženství, rodině atd., nebude tak těžké navrhnout. Bude to pravděpodobně životní styl. Operacionalizovat životní styl by bylo obtížné. Tak si zjednodušíme situaci tím, že proměnnou "životní styl" necháme zastupovat proměnnou "vzdělání". Osoby s vyšším vzděláním - alespoň v severoamerické společnosti - jsou spíše ochotny přiznat, že to s jejich náboženstvím není tak žhavé. Výše vzdělané osoby mají většinou také vyšší plat, takže si mohou dovolit - alespoň občas - kaviár, nebo ústřice. A aby naše diskuse byle opravdu přehledná, předstírejme, že proměnná "vzdělání" má jen dvě kategorie.

Teď zbývá už jen jedno: dokázat, že souvislost pozorovaná v tabulce 9.1. je nepravá, že je vyvolána vlivem vzdělání na obojí, na "zbožnost" a na "preferovanou stravu". Jednoduchá technika, kterou můžeme použít, se nazývá **kontrola dalším faktorem** ("control for test factor"). Tady je návod:

Kontrola pro další faktor:

(1) Náš vzorek rozdělíme do tolika podsouborů, kolik kategorií má proměnná, jejíž vliv kontrolujeme. Všichni jedinci v každém podsouboru budou mít v této proměnné stejnou hodnotu. (V našem případě získáme jeden podsoubor, ve kterém budou všichni jedinci mít jenom "nízké" vzdělání a v druhém podsouboru budou jenom jedinci s "vysokým" vzděláním.)

(2) Pak zkonstruujeme pro každý podsoubor tabulku, která má stejnou formu jako původní tabulka popisující souvislost, o jejíž platnosti pochybujeme. (V našem případě budeme mít dvě tabulky, formou shodné s tabulkou 9.1. V jedné budou jen osoby s nízkým vzděláním, v druhé jen s vysokým.)

(3) Porovnáme intenzitu souvislosti v původní tabulce se souvislostí zjištěnou v nových tabulkách. Je-li souvislost v původní tabulce funkcí třetího faktoru, v nových tabulkách souvislost mezi původními daty zmizí, nebo je alespoň podstatně oslabena.

A teď to můžeme vyzkoušet. Tabulka 9.2. obsahuje data o osobách s "nízkým" vzděláním, tabulka 9.3. data o osobách s vysokým vzděláním:

Tabulka 9.2.

Osoby s nízkým vzděláním

	Preferované jídlo:		Celkem:
Zbožnost:	HAMBURGY	KAVIÁR	
vysoká	78% 624	78% 156	780
nízká	22% 176	22% 44	220
Celkem: N	100% 800	100% 200	1000

LAMBDA = 0

Tabulka 9.3.

Osoby s vysokým vzděláním

	Preferované jídlo:		Celkem:
Zbožnost:	HAMBURGY	KAVIÁR	
vysoká	25% 50	25% 75	125
nízká	75% 150	75% 325	475
Celkem: N	100% 200	100% 400	600

LAMBDA = 0

Výsledky jsou docela jasné: souvislost mezi původními dvěma proměnnými úplně zmizela. Lambda v obou nových tabulkách klesla na nulu. Vidíme to i bez jakéhokoliv měření síly souvislosti: procento zbožných je v obou zcela shodné pro ty, kteří dávají přednost kaviáru, jako pro ty, kteří mají raději hamburgy. Můžeme tedy uzavřít, že původní souvislost mezi zbožností a jídlem byla jen a jen funkcí třetí proměnné, funkcí vzdělání.

Jistě už víte, že to tak pěkně vyšlo jen proto, že jsme připravili data tak, aby bylo všechno jasné a průhledné. Ve skutečnosti by nám kontrola dalších faktorů v nových tabulkách nedala nulovou souvislost. Životní styl může být nadto ovlivněn dalšími faktory, třeba tím, zda respondent žije ve velkém městě nebo na venkově. Ale i takové situace může kontrolu dalšího faktoru zvládnout. Tady je návod, jak kontrolovat souvislost mezi stravou a zbožností pro dva další faktory (vzdělání a místo bydliště) současně:

Kontrola dalších dvou faktorů

(1) Nejdříve rozdělíme náš vzorek na dvě skupiny. V jedné budou jenom respondenti žijící ve městě, ve druhé jenom ti, kteří žijí na venkově.

(2) Pak každou skupinu rozdělíme do dvou podsouborů podle vzdělání respondentů. Výsledkem budou čtyři skupiny dat.

(3) Pro každou skupinu zkonstruujeme tabulku shodnou v její formě s původní tabulkou 9.1. Budeme mít tedy čtyři tabulky. V jedné budou data o jedincích, žijících ve městě a majících vysoké vzdělání. Ve druhé budou data o těch, kdo žijí ve městě, ale mají nízké vzdělání. Ve třetí tabulce budou data o respondentech žijících na venkově a majících vysoké vzdělání. v poslední tabulce pak budou respondenti, kteří žijí na venkově a mají nízké vzdělání.

(4) Poslední krok bude stejný jako předtím: porovnáme souvislosti v nových tabulkách se souvislostí zjištěnou v původní tabulce.

A nyní by nám už měla být jasná **logika testování dalšího faktoru:**

> **Vytvořením nových tabulek je testovaný faktor udržován na konstantní hodnotě. Tím je souvislost mezi původními proměnnými očištěna od zkreslujícího vlivu této další proměnné.**

Teoreticky nemáme důvod, proč omezit tuto kontrolu na jednu nebo dvě další proměnné. Popsaná logika může být aplikována i na vyšší počet testovaných faktorů.

Dr.Watson:

Výborně! Mě vždycky zajímalo, jak vzdělání ovlivňuje rozhodnutí, pro koho budou lidé hlasovat ve volbách. a teď to mohu zjistit mnohem jasněji. Budu kontrolovat typ vzdělání, povolání, příjem a velikost obce. a také to, jak respondent hlasoval v minulých volbách a ovšem pohlaví a věk.

Teoreticky má dr. Watson pravdu. Prakticky je v tom háček. Podívejme se, co všechno by náš přítel musel pro navrženou kontrolu udělat. Řekněme, že by proměnná "volební preference" měla jenom 6 kategorii, a proměnná "vzdělání" jenom pět; původní tabulka by měla tedy 30 polí. Abychom mohli kontrolovat pro "typ vzdělání" museli bychom původní tabulku opakovat pro každou, řekněme ze 4 kategorií této proměnné. Máme teď 4 tabulky se 120 poli. Pak musíme tuto sérii 4 tabulek opakovat pro každou z kategorií proměnné "povolání". Ale teď se to už stává trochu nepřehledné. Shrneme si to tedy do tabelární formy:

Proměnná:	**Počet kategorií:**	**Počet polí:**
preferovaná strana	6	6
vzdělání	5	30
typ vzdělání	4	120
povolání	5	600
příjem	5	3.000
velikost obce	4	12.000
strana volená v minulých volbách	6	72.000
věk	3	216.000
pohlaví	2	432.000

Tak tohle by opravdu nešlo. Co bychom si počali s 7.200 tabulek? Takové množství tabulek by vůbec nebylo možné interpretovat. Ale hlavně, pro takovéhle cvičení **nemáme dost lidí!** I kdyby dr. Watson měl hodně štědrého sponzora a mohl si dovolit vzorek s dvěma tisíci jedinci, **více než 99% polí v jeho tabulkách by bylo prázdných.**

Zkusme tedy podstatně skromnější přístup. Budeme kontrolovat jen ty proměnné, které jsou snad nejdůležitější: povolání a pohlaví. Teď bychom skončili s 300 poli v deseti tabulkách. I zde by vzorek 2.000 jedinců sotva stačil. Teoreticky by na každé pole v tabulkách připadlo o něco méně než 7 pozorování. To by nemuselo být dost. **Prázdná pole v tabulce, jakož i pole s velice nízkým počtem pozorování, mohou podstatně zkreslit význam koeficientů, měřících souvislost.**

Počet faktorů, které si můžeme dovolit kontrolovat nezávisí ovšem jen na počtu proměnných, ale i na počtu kategorií každé z nich. Tak kdybychom kontrolovali povolání a pohlaví naši tabulku 9.1. a dostali bychom ve výsledných deseti tabulkách jen 80 polí a při stejné velikosti vzorku by na každé pole připadalo v průměru 25 pozorování. To už je podstatně lepší, ale kdo si kdy může dovolit dvoutisícový vzorek?

Někdy nám však i kontrola jediného dalšího faktoru může podstatně prospět. Například jsme zapojeni do výzkumu trhu a studujeme, zda balení typu A je atraktivnější, než balení B. První výsledky jsou v tabulce 9.4.:

Tabulka 9.4.

	Balení A	Balení B	Celkem:
asi by koupil	40% 80	40% 160	240
asi ne	60% 120	60% 240	360
Celkem:	100% 200	100% 400	600

LAMBDA = 0

Data zřejmě ukazují, že typy balení nemá vliv na úmysl zakoupit výrobek. Přesně stejné procento respondentů vyjádřilo úmysl koupit, ať již byl výrobek uveden v balení a nebo B. Ale je tomu opravdu tak? Podívejme se, co se stane, když budeme kontrolovat pohlaví respondentů.

Tabulka 9.5. shrnuje údaje pro muže. Na prvý pohled se zdá, že se nic nezměnilo: proměnná "typ balení" a proměnná "úmysl koupit" jsou vzájemně naprosto nezávislé:

Tabulka 9.5.

Muži:

	Balení A	Balení B	Celkem:
asi by koupil	40% 40	40% 40	80
asi ne	60% 60	60% 60	120
Celkem:	100% 100	100% 100	100

LAMBDA = 0

Ale tabulka 9.6. nám podává docela jiný obraz:

Tabulka 9.6.

Ženy:

	Balení A	Balení B	Celkem:
asi by koupil	100% 100	20% 60	160
asi ne	0% 0	80% 240	240
Celkem:	100% 100	100% 300	400

LAMBDA = .625

DISMAN Miroslav. 1993. *Jak se vyrábí sociologická znalost.* Praha: Karolinum.

Všechny ženy ve vzorku vyjádřily úmysl zakoupit výrobek v balení A a jen pětina z nich v balení B. Pro marketing to je jistě velice užitečná informace, která byla zcela neviditelná v původní tabulce 9.4.

Pro nás je tenhle poznatek také pěkně důležitý: ukazuje nám, jak statistická analýza více proměnných **současně** může odhalit nejen nepravou souvislost, produkovanou nějakým faktorem, ale může i odkrýt nepravou nezávislost mezi proměnnými. Příčinou tohoto typu zkreslení může být fakt, že souvislost existuje pouze v určité části vzorku, v našem specifickém případě jen mezi ženami. Zde může být kontrola dalších faktorů velice účinným nástrojem.

Nicméně, musíme-li pracovat najednou s mnoha proměnnými, kontrola dalšího faktoru brzy ztrácí dech. To jsme si už demonstrovali. Musíme se tedy porozhlédnout po jiných postupech, které by umožnily dr. Watsonovi realizovat jeho volební projekt.

9.2. Výprava do čtyřrozměrného prostoru

> Z poselství Vogona Jetze, člena Plánovacího výboru pro galaktický nadprostor:
> "Bohužel, vaše planeta je jednou z těch, které byly určeny pro demolici. Tento proces započne za necelé dvě vaše pozemské minuty. **Nepodléhejte panice!** Děkuji vám.
>
> Douglas Adams, *The Hitch-Hikers Guide to the Gallaxy*, 1979, p.30

On to bude víc než čtyřrozměrný prostor, ale nepanikařte. i když operace, o kterých budeme hovořit, mají vznešená a lehce hrozivá jména, jako vícerozměrná regresní analýza, "path analysis", faktorová analýza atd. Jejich **logika a jejich interpretace** není složitá. Složitá je jen logika jejich výpočtu a logika zdůvodnění těchto výpočtů. Ale tím se zde nebudeme zabývat. s tím se setkáte, až budete studovat skutečnou statistiku.

V podstatě značnou část toho, o čem tu budeme mluvit, už znáte. Zde to jen trošku rozšíříme. Např. už dovedeme pomocí regrese odhadnout jednotlivcův příjem, když známe jeho vzdělání. Ale příjem nezávisí jenom na vzdělání, ale i na délce odborné praxe, povolání atd. Zkusme, zda je možné aplikovat postup, který už známe, i na více proměnných.

1.
Pro každého jedince jsou pozorované hodnoty násobeny koeficienty odpovídajícími koeficienty prvé diskriminantní funkce. (To jsou ta čísla v prvém sloupci tabulky 9.9.) Součet těchto násobků reprezentuje jednotlivcovu pozici na prvé diskriminační funkci.

2.
Stejnou operaci opakujeme s koeficienty druhé funkce a obdržíme jeho pozici na této funkci.

3.
Teď máme pro jedince dvě souřadnice a můžeme ho zakreslit do mapy, odpovídající mapě v grafu 9.7.

4.
Pak už zbývá jen jedno, rozhodnout ke kterému z centroidů je jednotlivcova pozice nejblíže a zařadit ho do té skupiny. (Ve skutečnosti jsou i obě poslední operace prováděny matematicky.)

Tabulka 9.11. shrnuje výsledky takové klasifikace:

Tabulka 9.11.
Výsledky klasifikace

Skutečné členství	Odhadnuté členství			
	Amerika	Střední Evropa	Itálie	N
Amerika	66%	14%	20%	616
Střední Evropa	13%	69%	18%	1145
Itálie	26%	21%	53%	512

Z celkového počtu 2273 jedinců bylo správně klasifikováno 65%. Uvážíme-li charakter použitých prediktorů, je to opravdu docela pozoruhodný výsledek. Sociálně psychologické rozdíly podmíněné rozdílnou kulturou a rozdílným politicko-ekonomickým systémem jsou daleko silnější, než jsme očekávali. Ale tím se opět dostáváme do obsahové oblasti, a ta, bohužel, nepatří do naší knížky.

Tak nezbývá než poznamenat, že diskriminační analýza je nejen zajímavá, ale i velice užitečná hračka a přejít k poslední ze statistických technik, které zde budeme probírat.

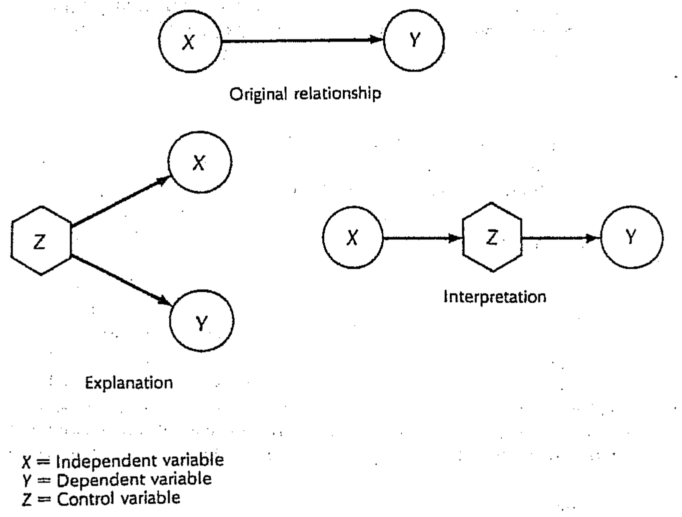
WILLIAMSONN, J. B., David A. KARP, John R. DALPHIN, Paul S. GRAY. 1982. *The Research Craft. An Introduction to Social Research Methods.* Boston: Little, Brown and Company.

FIGURE 16.8 The Bivariate Relationship between Abortion Attitude and Size of Birth Place

		Size of Birthplace	
		Town	City
"Should it be possible for a woman to obtain an abortion on demand?"	No	82% (410)	18% (90)
	Yes	18% (90)	82% (410)
	Total	100% (500)	100% (500)

INTERPRETATION. Figure 16.8, a cross-tabulation of the relationship between *attitudes toward abortion* and *size of one's birthplace*, shows persons from cities much more likely (82 percent) to endorse the right of women to obtain an abortion than are persons from towns (18 percent). Suppose, as in the previous example, we try to explain away the relationship but fail to discover any control variable that meets both requirements (i.e., associated with and causally prior to both original variables). When explanation fails to reduce such a nonobvious relationship between two variables, there still exists the possibility that we can uncover a third factor to help clarify the chain of circumstances that connects the two variables to one another.

FIGURE 16.9 Models Illustrating the Distinction Between Explanation and Interpretation



INTERPRETATION, the second part of the elaboration paradigm, is the search for a control variable (Z) that *causally intervenes* between the independent variable (X) and the dependent variable (Y). Figure 16.9 diagrams the differences between explanation and interpretation as they modify the original relationship between the independent and dependent variables. An INTERVENING VARIABLE must be related to both the independent and the dependent variable, and it must be plausible to think of it as somehow a result of the independent variable that, in turn, affects the dependent variable. Figure 16.10 illustrates the effects of an intervening variable.

Searching for an intervening variable that might explain the relationship between abortion attitude and size of birthplace (Figure 16.8), one might hypothesize that towns and cities promote very different kinds of political and social ideologies, which in turn might account for the city/town differences in abortion attitudes. In effect, people born in towns are more likely to be conservative than are people born in cities, and conservatives are more likely than liberals to oppose abortion. Note in Figure 16.10 that there are no longer any differences in abortion attitudes between town people and city people in either subtable; all town/city differences have been accounted for by subdividing the sample into *conservatives* and *liberals*. Hence we have successfully interpreted the relationship by locating an intervening variable.

Compare Figure 16.7 with Figure 16.10. Notice that the results have the same statistical form; that is, the introduction of a control variable makes the original relationship disappear. Hence the difference between explanation and interpretation rests in the underlying logic, not in the statistics. We now turn to a third form of elaboration, referred to as SPECIFICATION, in which the objective is not to make the original relationship disappear, but rather to specify the conditions under which the strength of the original relationship varies in intensity.

SPECIFICATION. Figure 16.11 reexamines the relationship between size of birthplace and attitudes toward abortion while controlling for a third variable, the *region of the country in which a person was born*. Here the original relationship changes (compare with Figure 16.8) but does not disappear; instead, it takes on a different form from one subtable to the next. The original relationship disappears for persons born in the South, where town and city people show identical attitudes

FIGURE 16.10 The Relationship Between Abortion Attitude and Size of Birthplace Controlling for Political Ideology: an Example of Interpretation

		Political Ideology			
		Conservative		Liberal	
		Size of Birthplace		*Size of Birthplace*	
		Town	City	Town	City
"Should it be possible for a woman to obtain an abortion on demand?"	No	90% (405)	90% (45)	10% (5)	10% (45)
	Yes	10% (45)	10% (5)	90% (45)	90% (405)
	Total	100% (450)	100% (50)	100% (50)	100% (450)

WILLIAMSON, J. B., David A. KARP, John R. DALPHIN, Paul S. GRAY. 1982. *The Research Craft. An Introduction to Social Research Methods.* Boston: Little, Brown and Company.

toward abortion; it remains strong in the West, where town people are more likely than city people to oppose abortion (86 percent versus 21 percent); and it intensifies in the North, where differences between town and city people regarding abortion attitudes are most pronounced (89 percent versus 0 percent oppose abortion). Introducing a control variable has enabled us to analyze the relationship between size of birthplace and attitude toward abortion more precisely, pinpointing the circumstances under which the association holds. This is an example of specification.

It is entirely possible that the use of a control variable for specification of a relationship, as in Figure 16.11, may produce fundamentally different relationships in different subtables. It is conceivable that town persons might favor abortion more than city persons in one region, and yet the opposite might be true in another area. When this occurs, there is good reason to suspect that other, undiscovered factors are affecting the relationship. A specification that results in such markedly different subtables is an invitation to pursue the analysis further, as the following case illustrates.

SUPPRESSOR VARIABLES. Suppose we have a table in which no relationship appears, even though we had good reason to expect to find an association. In Figure 16.11 the data for the West and the North indicate a strong association between size of birthplace and abortion attitude; yet the association disappears in data for the South. Why? It is possible that some hidden third factor is *suppressing* the true relationship between the two original variables. Such a factor is referred to as a SUPPRESSOR VARIABLE, because it hides the actual relationship until it is controlled.

Figure 16.12 reanalyzes this data for the South, controlling for another variable, *percentage of persons in the community who are black.* Whereas the original data showed no relationship between size of birthplace and abortion attitude, these two subtables each show strong (but opposite) associations. Subtable 2 shows data that are consistent with the overall findings presented in Figure 16.11, while subtable 1 isolates the deviant cases. When the two subtables are combined, as they were in Figure 16.11, the relationship is no longer discernible.

FIGURE 16.11 The Relationship Between Abortion Attitudes and Size of Birthplace, Controlling for Region of Birthplace: Example of Specification

Region of Birthplace

		South		West		North	
		Size of Birthplace		*Size of Birthplace*		*Size of Birthplace*	
		Town	City	Town	City	Town	City
"Should it be possible for a woman to obtain an abortion on demand?"	No	50% (40)	50% (40)	86% (160)	21% (50)	89% (210)	0% (0)
	Yes	50% (40)	50% (40)	14% (25)	79% (190)	11% (25)	100% (180)
	Total	100% (80)	100% (80)	100% (185)	100% (240)	100% (235)	100% (180)
		(Subtable 1)		(Subtable 2)		(Subtable 3)	

FIGURE 16.12 A Three-Way Table Illustrating the Effect of Introducing a Suppressor Variable

Percent Black in Community of Birth for Respondents Born in South

		High		Low	
		Size of Birthplace		*Size of Birthplace*	
		Town	City	Town	City
"Should it be possible for a woman to obtain an abortion on demand?"	No	100% (40)	0% (0)	0% (0)	100% (40)
	Yes	0% (0)	100% (40)	100% (40)	0% (0)
	Total	100% (40)	100% (40)	100% (40)	100% (40)
		(Subtable 1)		(Subtable 2)	

The data we have presented in this discussion of various methods of elaboration (Figures 16.6 to 16.8 and 16.10 to 16.12) are hypothetical and exaggerated to illustrate points of analysis. In actual research, relationships are seldom so strong, nor are distinctions between types of elaboration so clear. However, the logic that underlies these idealized examples embodies the range of possibilities for analysis that you will encounter in real research, and a thorough knowledge of these classifications will serve as a useful guide.

For the sake of simplicity we have developed elaborations around *dichotomies*—variables with only two values. The same logic applies to more complex variables, but when tables get larger, the elaborations soon become unwieldy. Indeed, it is often desirable to control for the effects of more than one variable, but we find ourselves confronted with the same practical difficulty. Just as correlation analysis was introduced to solve the analogous problem for two-variable tables with many cells, a technique called partial correlation exists to aid in the analysis of more complex elaborations.

NORUŠIS, Marija J. 1998. *Guide to Data Analysis.* Upper Saddle River, Prentice Hall.

Adding Control Variables

So far, you've considered the relationship between income and job satisfaction for the entire sample. It's possible that if you consider additional variables, the relationship you've seen between the two variables may change. For example, it may be that the relationship between income and job satisfaction is different for men and women. To test this, you can make separate tables of income and job satisfaction for men and for women. Gender is then called a control variable, since its effect is removed, or "controlled" for, in each of the separate tables. Figure 7.8 shows separate crosstabulation tables for men and women.

To obtain these separate—or layered—crosstabulations, select sex as a layer variable, as shown in Figure 7.10.

Figure 7.8 Job satisfaction by income for men and women

Respondent's Sex			Total Family Income in quartiles				Total
			24,999 or less	25,000 to 39,999	40,000 to 59,999	60,000 or more	
Male	Very satisfied	Count	30	51	41	57	179
		Column %	34.9%	47.2%	45.6%	46.0%	43.9%
	Mod satisfied	Count	44	44	36	49	173
		Column %	51.2%	40.7%	40.0%	39.5%	42.4%
	A little dissatisfied	Count	10	10	7	14	41
		Column %	11.6%	9.3%	7.8%	11.3%	10.0%
	Very dissatisfied	Count	2	3	6	4	15
		Column %	2.3%	2.8%	6.7%	3.2%	3.7%
	Total	Count	86	108	90	124	408
		Column %	100.0%	100.0%	100.0%	100.0%	100.0%
Female	Very satisfied	Count	23	39	33	53	148
		Column %	26.1%	45.3%	50.0%	53.5%	43.7%
	Mod satisfied	Count	49	35	25	38	147
		Column %	55.7%	40.7%	37.9%	38.4%	43.4%
	A little dissatisfied	Count	14	7	7	5	33
		Column %	15.9%	8.1%	10.6%	5.1%	9.7%
	Very dissatisfied	Count	2	5	1	3	11
		Column %	2.3%	5.8%	1.5%	3.0%	3.2%
	Total	Count	88	86	66	99	339
		Column %	100.0%	100.0%	100.0%	100.0%	100.0%

Each cell contains counts and column percentages. An interesting difference emerges between the two tables. For women, job satisfaction seems to increase with income. Almost twice as many women in the highest income category (53.5%) are *very satisfied* compared to women in the lowest income category (26.1%). For men, the difference is not as striking. Almost 35% of men in the lowest income category are *very satisfied*, compared to 46% of the men in the highest income category.

Case Selection

Sometimes you want to analyze only part of your cases. For example, some of the analyses described in this book look only at full-time workers or only at college graduates.

The Select Cases dialog box allows you to restrict your analysis to a specific group of cases. There are a number of options for selecting cases:

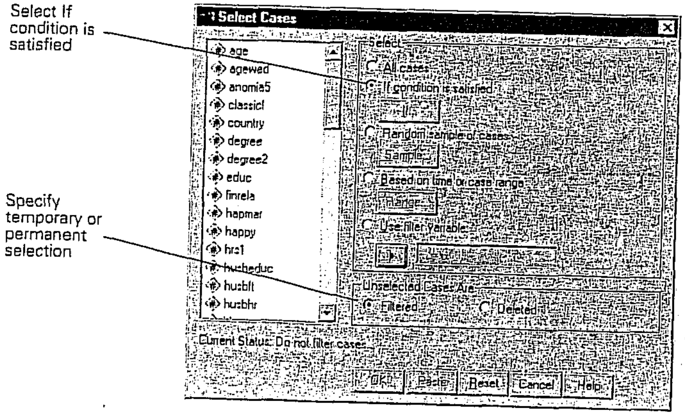
- You can choose cases according to a logical condition based on their data values.
- You can select a random sample of the cases in your file.
- You can select a range of cases according to their order in the file.
- You can select those cases that are marked with a non-zero value for a "filter variable."

▶ From the menus choose:

Data
 Select Cases...

This opens the Select Cases dialog box, as shown in Figure B.12.

Figure B.12 Select Cases dialog box



Temporary or Permanent Selection

The Select Cases dialog box offers a choice between filtering cases (selecting temporarily) or deleting cases (selecting permanently). The distinction between temporary and permanent case selection is important to understand.

- When you filter cases, or select a temporary subset, the unselected cases remain in the working data file. You can regain all the original cases at any time.
- When you delete cases, or select a permanent subset, SPSS deletes them forever from your working data file. If you save the working data file, replacing the copy on your disk, the deleted cases are gone forever from that file, too. This can be useful because it allows you to save a smaller data file.

If you haven't saved the working data file, you can often "undo" a permanent case selection by reopening the original data file. If you have saved the working data file there is no way to get back cases that have been deleted, unless you have a backup copy of the data file.

Example: Selecting Full-Time Employees

Many of the analyses in this book that use the GSS data restrict the analysis to full-time workers. In the *gss.sav* file, respondents who are employed full time are coded 1 for the variable *wrkstat*. To select full-time employees:

▶ From the Data Editor menus choose:

Data
 Select Cases...

▶ Select If condition is satisfied alternative in the Select Cases dialog box (see Figure B.12).

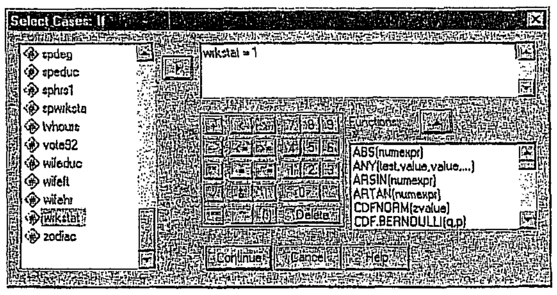
▶ Select Filtered in the Unselected Cases Are group.

This assures that the unselected cases will remain in the working data file if you want to use them in future analyses.

▶ Click If.

This opens the Select Cases If dialog box, as shown in Figure B.13. This dialog box, which strongly resembles the Compute Variable If Cases dialog box shown in Figure B.9, allows you to specify a conditional expression.

Figure B.13 Select Cases If dialog box


wkstat = 1

▶ Enter the expression wrkstat = 1.

▶ Click Continue to return to the Select Cases dialog box.

▶ Click OK.

Cases for people who work full time are now selected. In the Data Editor, unselected cases are indicated by a slash mark over the row number.

▶ To turn off case selection, open the Select Cases dialog box again, select All cases, and click OK.

Example: Selecting College Graduates

In the *gss.sav* and *gssft.sav* data files, the variable *degree* indicates the highest degree earned by each respondent. Four-year college graduates are coded 3 (for bachelor's degree) or 4 (for advanced degree). To select people with bachelor's or advanced degrees:

▶ Open the Select Cases dialog box as described above.

▶ Select Filtered in the Unselected Cases Are group.

▶ Select If condition is satisfied and click If.

▶ Enter degree >= 3 in the Numeric Expression box.

This expression specifies that cases should be selected "if degree is greater than or equal to 3."

▶ Click Continue to return to the Select Cases dialog box and click OK.

Other Selection Methods

Other options available in the Select Cases dialog box include:

Random sample. Sometimes you want a random subset of cases. You have no particular criterion for choosing which cases to process, but you don't want the whole data file.

Based on time or case range. Under some circumstances, it is desirable to select a range of cases according to the order of cases, as displayed in the Data Editor. This can be useful for time series data files.

Use filter variable. A filter variable is simply a variable that indicates whether or not a particular case should be selected. Cases for which the specified filter variable has a valid non-zero value are retained. Cases for which it is 0 or missing are dropped.

Testing for intervening variables

The quest for intervening variables is different from the search for potentially spurious relationships. An intervening variable is one that is both a product of the independent variable and a cause of the dependent variable. Taking the data examined in Table 10.1, the sequence depicted in Figure 10.2 might be imagined. The analysis presented in Table 10.4 strongly suggests that the level of people's interest in their work is an intervening variable. As with Tables 10.2 and 10.3, we partition the sample into two groups (this time those who report that they are interested and those reporting no interest in their work) and examine the relationship between work variety and job satisfaction for each group. Again, we can compare d_1 in Table 10.1 with d_1 and d_2 in Table 10.4. In Table 10.1 d_1 is 56 per cent, but in Table 10.4 d_1 and d_2 are 13 per cent and 20 per cent respectively. Clearly, d_1 and d_2 in Table 10.3 have not been reduced to zero (which would suggest that the whole of the relationship was through interest in work), but they are also much lower than the 56 per cent difference in Table 10.1. If d_1 and d_2 in Table 10.4 had remained at or around 56 per cent, we would conclude that interest in work is not an intervening variable.

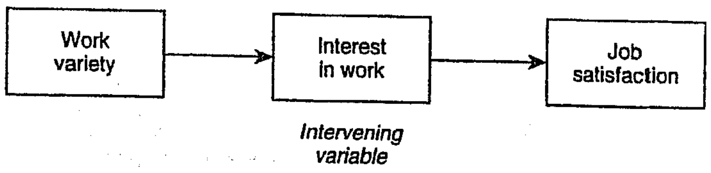


Figure 10.2 Is the relationship between work variety and job satisfaction affected by an intervening variable?

The sequence in Figure 10.2 suggests that variety in work affects the degree of interest in work that people experience, which in turn affects their level of job satisfaction. This pattern differs from that depicted in Figure 10.1 in that if the analysis supported the hypothesized sequence, it suggests that there is a relationship between amount of variety in work and job satisfaction, but the relationship is not direct. The search for intervening variables is often referred to as *explanation* and it is easy to see why. If we find that a test variable acts as an intervening variable, we are able to gain some explanatory leverage on the bivariate relationship. Thus, we find that there is a relationship between amount of variety in work and job satisfaction and then ask why that relationship might exist. We speculate that it may be because those who have varied work become more interested in their work, which heightens their job satisfaction.

It should be apparent that the computation of a test for an intervening variable is identical to a test for spuriousness. How, then, do we know which is which? If we carry out an analysis like those shown in Tables 10.2, 10.3 and 10.4, how can we be sure that what we are taking to be an intervening variable is not in fact an indication that the relationship is spurious? The answer is that there should be only one logical possibility, that is, only one that makes sense. If we take the trio of variables in Figure 10.1, to argue that the test variable – size of firm – could be an intervening variable would mean that we would have to suggest that a person's level of work variety affects the size of the firm in which he or she works – an unlikely scenario. Similarly, to argue that the trio in Figure 10.2 could point to a test for spuriousness, would mean that we would have to accept that the test variable – interest in work – can affect the amount of variety in a person's work. This too makes much less sense than to perceive it as an intervening variable.

One further point should be registered. It is clear that controlling for interest in work in Table 10.4 has not totally eliminated the difference between those reporting varied work and those whose work is not varied, in terms of job satisfaction. It would seem, therefore, that there are aspects of the relationship between amount of variety in work and job satisfaction that are not totally explained by the test variable, interest in work.

Testing for moderated relationships

A moderated relationship occurs when a relationship is found to hold for some categories of a sample but not others. Diagrammatically this can be displayed as in Figure 10.3. We may even find the character of a relationship can differ for categories of the test variable. We might find that for one category those who report varied work exhibit greater job satisfaction, but for another category of people the reverse may be true (i.e. varied work seems to engender lower levels of job satisfaction than work that is not varied).

Table 10.3 A non-spurious relationship: the relationship between work variety and job satisfaction, controlling for size of firm (imaginary data)

		Large firms		Small firms	
		Varied work	Not varied work	Varied work	Not varied work
Job satisfaction	Satisfied	1 (166) 83% $d_1 =$	2 (14) 28% 55%	3 (34) 68% $d_2 =$	4 (46) 23% 45%
	Not satisfied	5 (34) 17% $d_3 =$	6 (36) 72% 55%	7 (16) 32% $d_4 =$	8 (154) 77% 45%

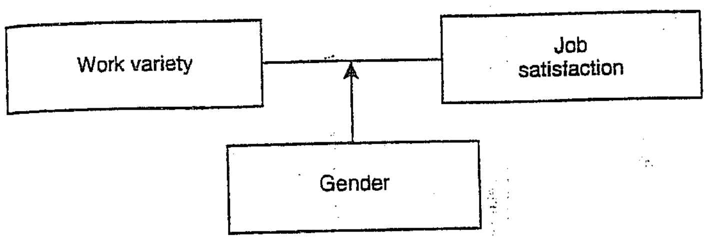


Figure 10.3 Is the relationship between work variety and job satisfaction moderated by gender?

Table 10.5 looks at the relationship between variety in work and job satisfaction for men and women. Once again, we can compare d_1 (56 per cent) in Table 10.1 with d_1 and d_2 in Table 10.5, which are 85 per cent and 12 per cent respectively. The bulk of the 56 percentage point difference between those reporting varied work and those reporting that work is not varied in Table 10.1 appears to derive from the relationship between variety in work and job satisfaction being far stronger for men than women and there being more men (300) than women (200) in the sample. Table 10.5 demonstrates the importance of searching for moderated relationships in that they allow the researcher to avoid inferring that a set of findings pertains to a sample as a whole, when in fact it only really applies to a portion of that sample. The term *interaction effect* is often employed to refer to the situation in which a relationship between two variables differs substantially for categories of the test variable. This kind of occurrence was also addressed in Chapter 9. The discovery of such an effect often inaugurates a new line of inquiry in that it stimulates reflection about the likely reasons for such variations.

The discovery of moderated relationships can occur by design or by chance. When they occur by design, the researcher has usually anticipated the possibility that a relationship may be moderated (though he or she may be wrong of course). They can occur by chance when the researcher conducts a test for an intervening variable or a test for spuriousness and finds a marked contrast in findings for different categories of the test variable.

Multiple causation

Dependent variables in the social sciences are rarely determined by one variable alone, so that two or more potential independent variables can usefully be considered in conjunction. Figure 10.4 suggests that whether someone is allowed participation in decision-making at work also affects their level of job satisfaction. It is misleading to refer to participation in decision-making as a test variable in this context, since it is really a second independent variable. What, then, is the impact of amount of variety in work on job satisfaction when we control the effects of participation?

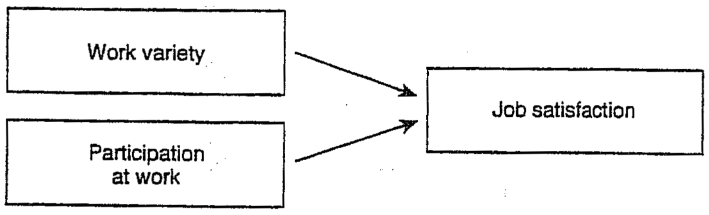


Figure 10.4 Work variety and participation at work

Again, we compare d_1 in Table 10.1 (56 per cent) with d_1 and d_2 in Table 10.6. The latter are 19 and 18 per cent respectively. This suggests that although the effect of amount of variety in work has not been reduced to zero or nearly zero, its impact has been reduced considerably. Participation in decision-making appears to be a more important cause of variation in job satisfaction. For example, compare the percentages in cells 1 and 3 in Table 10.6: among those respondents who report that they perform varied work, 93 per cent of those who experience participation exhibit job satisfaction, whereas only 30 per cent of those who do not experience participation are satisfied.

One reason for this pattern of findings is that most people who experience participation in decision-making also have varied jobs, that is (cell1 + cell5) − (cell2 + cell6). Likewise, most people who do not experience participation have work which is not varied, that is (cell4 + cell8) − (cell3 + cell7). Could this mean that the relationship between variety in work and job satisfaction is really spurious, when participation in decision-making is employed as the test variable? The answer is that this is unlikely, since it would mean that participation in decision-making would have to cause variation in the amount of variety in work, which is a less likely possibility (since technological conditions tend to be the major influence on variables like work variety). Once again, we have to resort to a combination of intuitive logic and theoretical reflection in order to discount such a possibility. We will return to this kind of issue in the context of an examination of the use of multivariate analysis through correlation and regression.

Table 10.4 An intervening variable: the relationship between work variety and job satisfaction, controlling for interest in work (imaginary data)

		Interested		Not interested	
		Varied work	*Not varied work*	*Varied work*	*Not varied work*
Job satisfaction	*Satisfied*	1 (185) 93% $d_1=$	2 (40) 80% 13%	3 (15) 30% $d_2=$	4 (20) 10% 20%
	Not satisfied	5 (15) 7% $d_3=$	6 (10) 20% 13%	7 (35) 70% $d_4=$	8 (180) 90% 20%

Table 10.5 A moderated relationship: the relationship between work variety and job satisfaction, controlling for gender (imaginary data)

		Men		Women	
		Varied work	*Not varied work*	*Varied work*	*Not varied work*
Job satisfaction	*Satisfied*	1 (143) 95% $d_1=$	2 (15) 10% 85%	3 (57) 57% $d_2=$	4 (45) 45% 12%
	Not satisfied	5 (7) 5% $d_3=$	6 (135) 90% 85%	7 (43) 43% $d_4=$	8 (55) 55% 12%

Table 10.6 Two independent variables: the relationship between work variety and job satisfaction, controlling for participation at work (imaginary data)

		Participation		Little or no participation	
		Varied work	*Not varied work*	*Varied work*	*Not varied work*
Job satisfaction	*Satisfied*	1 (185) 93% $d_1=$	2 (37) 74% 19%	3 (15) 30% $d_2=$	4 (23) 12% 18%
	Not satisfied	5 (15) 7% $d_3=$	6 (13) 26% 19%	7 (35) 70% $d_4=$	8 (177) 88% 18%

3.2.4. Partial Correlation

3.2.4.1. The Theory behind Part and Partial Correlation

I mentioned earlier that there is a type of correlation that can be done that allows you to look at the relationship between two variables when the effects of a third variable are held constant. For example, analyses of the exam anxiety data (in the file **ExamAnx.sav**) showed that exam performance was negatively related to exam anxiety, but positively related to revision time, and revision time itself was negatively related to exam anxiety. This scenario is complex, but given that we know that revision time is related to both exam anxiety and exam performance, then if we want a pure measure of the relationship between exam anxiety and exam performance we need to take account of the influence of revision time. Using the values of R^2 for these relationships, we know that exam anxiety accounts for 19.4% of the variance in exam performance, that revision time accounts for 15.7% of the variance in exam performance, and that revision time accounts for 50.2% of the variance in exam anxiety. If revision time accounts for half of the variance in exam anxiety, then it seems feasible that at least some of the 19.4% of variance in exam performance that is accounted for by anxiety is the same variance that is accounted for by revision time. As such, some of the variance in exam performance explained by exam anxiety is not *unique* and can be accounted for by revision time. A correlation between two variables in which the effects of other variables are held constant is known as a partial correlation.

Figure 3.13 illustrates the principle behind partial correlation. In part 1 of the diagram there is a box labelled exam performance that represents the total variation in exam scores (this value would be the variance of exam performance). There is also a box that represents the variation in exam anxiety (again, this is the variance of that variable). We know already that exam anxiety and exam performance share 19.4% of their variation (this value is the correlation coefficient squared). Therefore, the variations of these two variables overlap (because they share variance) creating a third box (the one with diagonal lines). The overlap of the boxes representing exam performance and exam anxiety is the common variance. Likewise, in part 2 of the diagram the shared

variation between exam performance and revision time is illustrated. Revision time shares 15.7% of the variation in exam scores. This shared variation is represented by the area of overlap (filled with diagonal lines). We know that revision time and exam anxiety also share 50% of their variation: therefore, it is very probable that some of the variation in exam performance shared by exam anxiety is the same as the variance shared by revision time.

Part 3 of the diagram shows the complete picture. The first thing to note is that the boxes representing exam anxiety and revision time have a large overlap (this is because they share 50% of their variation). More important, when we look at how revision time and anxiety contribute to exam performance we see that there is a portion of exam performance that is shared by both anxiety and revision time (the dotted area). However, there are still small chunks of the variance in exam performance that are unique to the other two variables. So, although in part 1 exam anxiety shared a large chunk of variation in exam performance, some of this overlap is also shared by revision time. If we remove the portion of variation that is also shared by revision time, we get a measure of the unique relationship between exam performance and exam anxiety. We use partial correlations to find out the size of the unique portion of variance. Therefore, we could conduct a partial correlation between exam anxiety and exam performance while 'controlling' the effect of revision time. Likewise, we could carry out a partial correlation between revision time and exam performance 'controlling' for the effects of exam anxiety.

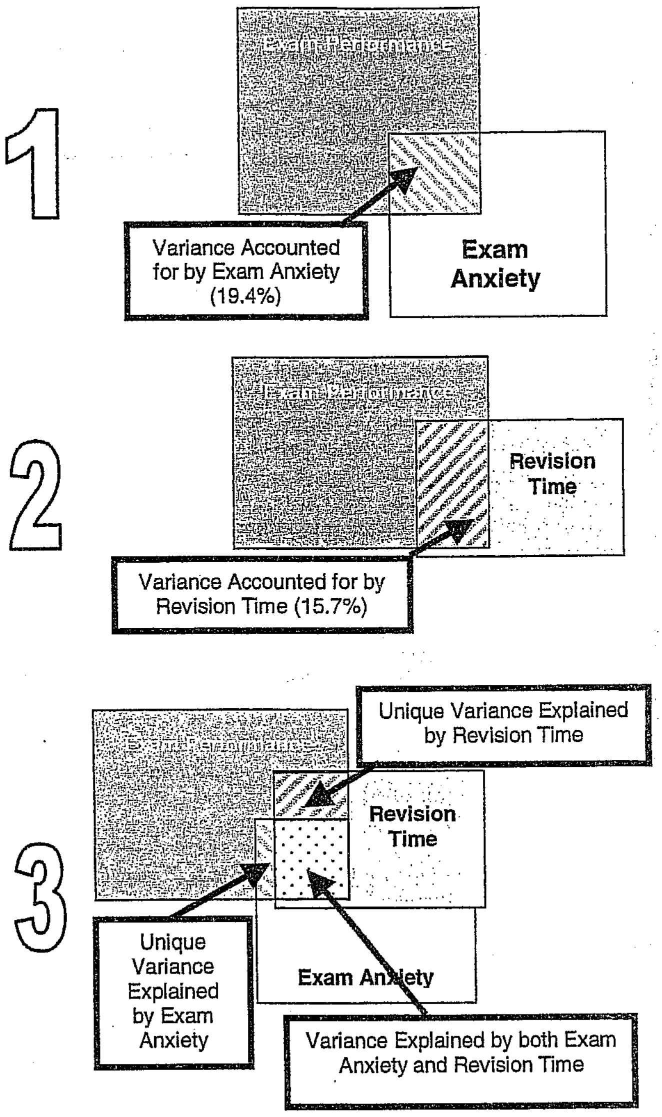
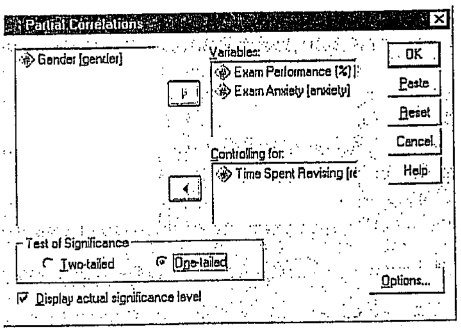

Exam Performance
1
Variance Accounted for by Exam Anxiety (19.4%)
Exam Anxiety
Exam Performance
2
Revision Time
Variance Accounted for by Revision Time (15.7%)
Unique Variance Explained by Revision Time
Exam Performance
Revision Time
3
Unique Variance Explained by Exam Anxiety
Exam Anxiety
Variance Explained by both Exam Anxiety and Revision Time

Figure 3.13: Diagram showing the principle of partial correlation

3.2.4.2. Partial Correlation Using SPSS

To conduct a partial correlation on the exam performance data select the *Correlate* option from the *Analyze* menu and then select *Partial* (**Analyze⇒Correlate⇒Partial**) and the dialog box in Figure 3.14 will be activated. This dialog box lists all of the variables in the data editor on the left-hand side and there are two empty spaces on the right-hand side. The first space is for listing the variables that you want to correlate and the second is for declaring any variables the effects of which you want to control. In the example I have described, we want to look at the unique effect of exam anxiety on exam performance and so we want to correlate the variables **exam** and **anxiety**, while controlling for **revise**. Figure 3.14 shows the completed dialog box. If you click on Options... then another dialog box appears as shown in Figure 3.15.


Partial Correlations

Gender [gender]

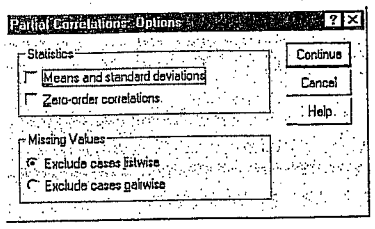
Variables:
Exam Performance [%]
Exam Anxiety [anxiety]

Controlling for:
Time Spent Revising [re

Test of Significance
○ Two-tailed ● One-tailed

☑ Display actual significance level

Figure 3.14: Main dialog box for conducting a partial correlation


Partial Correlations: Options

Statistics
□ Means and standard deviations
□ Zero-order correlations

Missing Values
● Exclude cases listwise
○ Exclude cases pairwise

Figure 3.15: Options for partial correlation

These further options are similar to those in bivariate correlation except that you can choose to display zero-order correlations. Zero-order correlations are the bivariate correlation coefficients without controlling for any other variables. So, in our example, if we select the tick-box for zero-order correlations SPSS will produce a correlation matrix of

anxiety, **exam** and **revise**. If you haven't conducted bivariate correlations before the partial correlation then this is a useful way to compare the correlations that haven't been controlled against those that have. This comparison gives you some insight into the contribution of different variables. Tick the box for zero-order correlations but leave the rest of the options as they are.

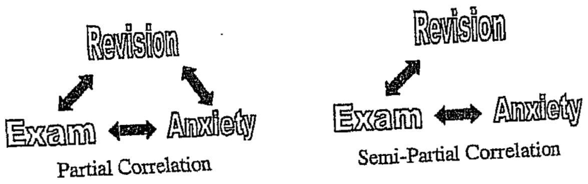
```
PARTIAL CORRELATION COEFFICIENTS

Zero Order Partials

                EXAM        ANXIETY      REVISE

EXAM           1.0000       -.4410        .3967
              (     0)      (  101)      (  101)
              P= .          P= .000      P= .000

ANXIETY        -.4410       1.0000       -.7092
              (  101)      (     0)      (  101)
              P= .000       P= .         P= .000

REVISE          .3967       -.7092       1.0000
              (  101)      (  101)      (     0)
              P= .000       P= .000      P= .

(Coefficient / (D.F.) / 1-tailed Significance)

PARTIAL CORRELATION COEFFICIENTS

Controlling for..     REVISE

                EXAM        ANXIETY

EXAM           1.0000       -.2467
              (     0)      (  100)
              P= .          P= .006

ANXIETY        -.2467       1.0000
              (  100)      (     0)
              P= .006       P= .

(Coefficient / (D.F.) / 1-tailed Significance)

" . " is printed if a coefficient cannot be computed
```

SPSS Output 3.6: Output from a partial correlation

SPSS Output 3.6 shows the output for the partial correlation of exam anxiety and exam performance controlling for revision time. The first thing to notice is the matrix of zero-order correlations, which we asked for using the *options* dialog box. The correlations displayed here are identical to those obtained from the Pearson correlation procedure (compare this matrix with the one in SPSS Output 3.2). Underneath the zero-order correlations is a matrix of correlations for the variables **anxiety** and **exam** but controlling for the effect of revision. In this instance we have controlled for one variable and so this is known as a first-order partial correlation. It is possible to control for the effects of two variables at the same time (a second-order partial correlation) or control three variables (a third-order partial correlation) and so on. First, notice that the partial correlation between exam performance and exam anxiety is −0.2467, which is considerably less than the correlation when the effect of revision time is not controlled for (r = −0.4410). In fact, the correlation coefficient is nearly half what it was before. Although this correlation is still statistically significant (its p value is still below 0.05), the relationship is diminished. In terms of variance, the value of R^2 for the partial correlation is 0.06, which means that exam anxiety can now account for only 6% of the variance in exam performance. When the effects of revision time were not controlled for, exam anxiety shared 19.4% of the variation in exam scores and so the inclusion of revision time has severely diminished the amount of variation in exam scores shared by anxiety. As such, a truer measure of the role of exam anxiety has been obtained. Running this analysis has shown us that exam anxiety alone does explain much of the variation in exam scores, and we have discovered a complex relationship between anxiety and revision that might otherwise have been ignored. Although causality is still not certain, because relevant variables are being included, the third variable problem is, at least, being addressed in some form.

3.2.4.3. Semi-Partial (or Part) Correlations

In the next chapter, we come across another form of correlation known as a semi-partial correlation (also referred to as a part correlation). While I'm babbling on about partial correlations it is worth me explaining the difference between this type of correlation and a semi-partial correlation. When we do a partial correlation between two variables, we control for the effects of a third variable. Specifically, the effect that the third variable has on *both* variables in the correlation is controlled. In a semi-partial correlation we control for the effect that the third variable has on only one of the variables in the correlation. Figure 3.16 illustrates this principle for the exam performance data. The partial correlation that we calculated took account not only of the effect of revision on exam performance, but also of the effect of revision on anxiety. If we were to calculate the semi-partial correlation for the same data, then this would control for only the effect of revision on exam performance (the effect of revision on exam anxiety is ignored). Partial correlations are most useful for looking at the unique relationship between two variables when other variables are ruled out. Semi-partial correlations are, therefore, useful when trying to explain the variance in one particular variable (an outcome) from a set of predictor variables. This idea leads us nicely toward Chapter 4 ...



Partial Correlation

Semi-Partial Correlation

Figure 3.16: The difference between a partial and a semi-partial correlation

11. lekce

ZÁKLADY LINEÁRNÍ REGRESE - VZTAH SPOJITÝCH PROMĚNNÝCH

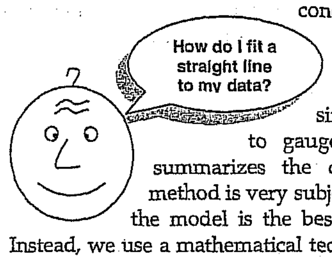
FIELD Andy. 2000. *Discovering Statistics Using SPSS for Windows. Advanced Techniques for the Beginner*. London: Sage.

4 Regression

4.1. An Introduction to Regression

Correlations can be a very useful research tool but they tell us nothing about the predictive power of variables. In regression analysis we fit a predictive model to our data and use that model to predict values of the dependent variable (DV) from one or more independent variables (IVs).[1] Simple regression seeks to predict an outcome from a single predictor whereas multiple regression seeks to predict an outcome from several predictors. This is an incredibly useful tool because it allows us to go a step beyond the data that we actually possess. The model that we fit to our data is a linear one and can be imagined by trying to summarize a data set with a straight line (think back to Figure 1.5).

With any data set there are a number of lines that could be used to summarize the general trend and so we need a way to decide which of many possible lines to chose. For the sake of drawing accurate conclusions we want to fit a model that *best* describes the data. There are several ways to fit a straight line to the data you have collected. The simplest way would be to use your eye to gauge a line that looks as though it summarizes the data well. However, the 'eyeball' method is very subjective and so offers no assurance that the model is the best one that could have been chosen. Instead, we use a mathematical technique to establish the line that best describes the data collected. This method is called the *method of least squares.*



[1] Unfortunately, you will come across people (and SPSS for that matter) referring to regression variables as dependent and independent variables (as in controlled experiments). However, correlational research by its nature seldom controls the independent variables to measure the effect on a dependent variable. Instead, variables are measured simultaneously and without strict control. It is, therefore, inaccurate to label regression variables in this way. For this reason I label 'independent variables' as *predictors*, and the 'dependent variable' as the *outcome*.

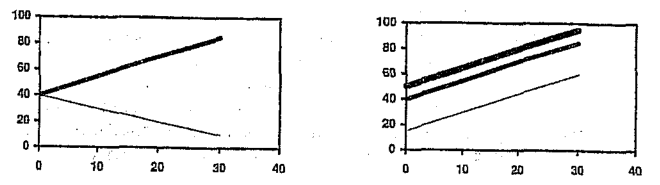
4.1.1. Some Important Information about Straight Lines

To use linear regression it is important that you know a few algebraic details of straight lines. Any straight line can be drawn if you know two things: (1) the slope (or gradient) of the line, and (2) the point at which the line crosses the vertical axis of the graph (known as the *intercept* of the line). The equation of a straight line is defined in equation (4.1), in which *Y* is the outcome variable that we want to predict and X_i is the *i*th subject's score on the predictor variable. β_1 is the gradient of the straight line fitted to the data and β_0 is the intercept of that line. There is a residual term, ε_i, which represents the difference between the score predicted by the line for subject *i* and the score that subject *i* actually obtained. The equation is often conceptualized without this residual term (so, ignore it if it's upsetting you); however, it is worth knowing that this term represents the fact our model will not fit perfectly the data collected.

$$Y = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (4.1)$$

A particular line has a specific intercept and gradient. Figure 4.1 shows a set of lines that have the same intercept but different gradients, and a set of lines that have the same gradient but different intercepts. Figure 4.1 also illustrates another useful point: that the gradient of the line tells us something about the nature of the relationship being described. In Chapter 3 we saw how relationships can be either positive or negative (and I don't mean the difference between getting on well with your girlfriend and arguing all the time!). A line that has a gradient with a positive value describes a positive relationship, whereas a line with a negative gradient describes a negative relationship. So, if you look at the graph in Figure 4.1 in which the gradients differ but the intercepts are the same, then the thicker line describes a positive relationship whereas the thinner line describes a negative relationship.

If it is possible to describe a line knowing only the gradient and the intercept of that line, then we can use these values to describe our model (because in linear regression the model we use is a straight line). So, the model that we fit to our data in linear regression can be conceptualized as a straight line that can be described mathematically by equation (4.1). With regression we strive to find the line that best describes the data collected, then estimate the gradient and intercept of that line. Having defined these values, we can insert different values of our predictor variable into the model to estimate the value of the outcome variable.



Same intercept, different slopes

Same slope, different intercepts

Figure 4.1: Shows lines with the same gradients but different intercepts, and lines that share the same intercept but have different gradients

4.1.2. *The Method of Least Squares*

I have already mentioned that the method of least squares is a way of finding the line that best fits the data (i.e. finding a line that goes through, or as close to, as many of the data points as possible). This 'line of best fit' is found by ascertaining which line, of all of the possible lines that could be drawn, results in the least amount of difference between the observed data points and the line. Figure 4.2 shows that when any line is fitted to a set of data, there will be small differences between the values predicted by the line, and the data that were actually observed. We are interested in the vertical differences between the line and the actual data because we are using the line to predict values of Y from values of the X variable. Although some data points fall exactly on the line, others lie above and below the line, indicating that there is a difference between the model fitted to these data and the data collected. Some of these differences are positive (they are above the line, indicating that the model underestimates their value) and some are negative (they are below the line, indicating that the model overestimates their value). These differences are usually called *residuals*. In the discussion of variance in section 1.1.3.1 I explained that if we sum positive and negative differences then they tend to cancel each other out. To avoid this problem we square the differences before adding them up. These squared differences provide a gauge of how well a particular line fits the data: if the squared differences are large, the line is not representative of the data; if the squared differences are small then the line is representative. The sum of squared differences (or sum of squares for short) can be calculated for any line that is fitted to some data; the 'goodness-of-fit' of each line can then be compared by looking at the sum of squares for each. The method of least squares works by selecting the line that has the lowest sum of squared differences (so it chooses the line that best represents the observed data). One way to select this optimal line would be to fit every possible line to a set of data, calculate the sum of squared differences for each line, and then choose the line for which this value is smallest. This would take quite a long time to do! Fortunately, there is a mathematical technique for finding maxima and minima and this technique (calculus) is used to find the line that minimizes the sum of squared differences. The end result is that the value of the slope and intercept of the 'line of best fit' can be estimated. Social scientists generally refer to this line of best fit as a regression line.

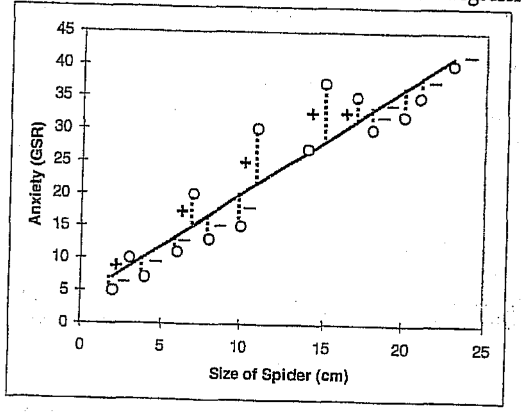
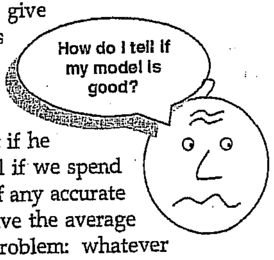


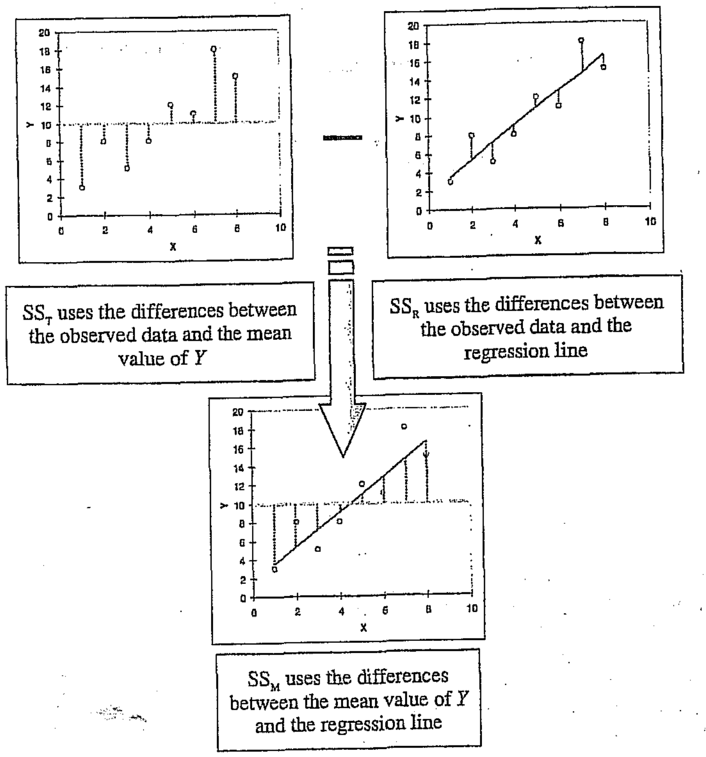
Figure 4.2: This graph shows a scatterplot of some data with a line representing the general trend. The vertical lines (dotted) represent the differences (or residuals) between the line and the actual data

4.1.3. *Assessing the Goodness-of-Fit: Sums of Squares, R and R^2*

Once we have found the line of best fit it is important that we assess how well this line fits the actual data (we assess the *goodness-of-fit* of the model). In section 1.1.3.1 we saw that one measure of the adequacy of a model is the sum of squared differences. Sticking with this theme, there are several sums of squares that can be calculated to help us gauge the contribution of our model to predicting the outcome. Imagine that I was interested in predicting record sales (Y) from the amount of money spent advertising that record (X). One day my boss came in to my office and said 'Andy, how many records will we sell if we spend £100,000 on advertising?' If I didn't have an accurate model of the relationship between record sales and advertising, what would my best guess be? Well, probably the best answer I could give would be the mean number of record sales (say, 200,000) because on average that's how many records we expect to sell. This response might well satisfy a brainless record company executive. However, what if he had asked 'How many records will we sell if we spend £1 on advertising?' Again, in the absence of any accurate information, my best guess would be to give the average number of sales (200,000). There is a problem: whatever amount of money is spent on advertising I always predict the same levels of sales. It should be pretty clear then that the mean is fairly useless as a model of a relationship between two variables—but it is the simplest model available.



So, as a basic strategy for predicting the outcome, we might choose to use the mean, because on average (*sic*) it will be a fairly good guess of an outcome. Using the mean as a model, we can calculate the difference between the observed values, and the values predicted by the mean. We saw in section 1.1.3.1 that we square all of these differences to give us the sum of squared differences. This sum of squared differences is known as the *total sum of squares* (denoted SS_T) because it is the total amount of differences present when the most basic model is applied to the data. This value represents how good the mean is as a model of the observed data. Now, if we fit the more sophisticated model to the data, such as a line of best fit, we can again work out the differences between this new model and the observed data. In the previous section we saw that the method of least squares finds the best possible line to describe a set of data by minimizing the difference between the model fitted to the data and the data themselves. However, even with this optimal model there is still some inaccuracy, which is represented by the differences between each observed data point and the value predicted by the regression line. As before, these differences are squared before they are added up so that the directions of the differences do not cancel out. The result is known as the *sum of squared residuals* (SS_R). This value represents the degree of inaccuracy when the best model is fitted to the data. We can use these two values to calculate how much better the regression line (the line of best fit) is than just using the mean as a model (i.e. how much better is the best possible model than the worst model?). The improvement in prediction resulting from using the regression model rather than the mean is calculated by calculating the difference between SS_T and SS_R. This difference shows us the reduction in the inaccuracy of the model resulting from fitting the regression model to the data. This improvement is the *model sum of squares* (SS_M). Figure 4.3 shows each sum of squares graphically.



SS_T uses the differences between the observed data and the mean value of Y

SS_R uses the differences between the observed data and the regression line

SS_M uses the differences between the mean value of Y and the regression line

Figure 4.3: Diagram showing from where the regression sums of squares derive.

If the value of SS_M is large then the regression model is very different from using the mean to predict the outcome variable. This implies that the regression model has made a big improvement to how well the outcome variable can be predicted. However, if SS_M is small then using the regression model is little better than using the mean (i.e. the regression model is no better than taking our 'best guess'). A useful measure arising from these sums of squares is the proportion of improvement due to the model. This is easily calculated by dividing the sum of squares for the model by the total sum of squares. The resulting value is called R^2 and to express this value as a percentage you should multiply it by 100. So, R^2 represents the amount of variance in the outcome explained by the model (SS_M) relative to how much variation there was to explain in the first place (SS_T). Therefore, as a percentage, it

FIELD Andy. 2000. Discovering Statistics Using SPSS for Windows. Advanced Techniques for the Beginner. London: Sage.

represents the percentage of the variation in the outcome that can be explained by the model.

$$R^2 = \frac{SS_M}{SS_T} \qquad (4.2)$$

Interestingly, this value is the same as the R^2 we met in Chapter 3 (section 3.2.3.3) and you'll notice that it is interpreted in the same way. Therefore, in simple regression we can take the square root of this value to obtain the Pearson correlation coefficient. As such, the correlation coefficient provides us with a good estimate of the overall fit of the regression model, and R^2 provides us with a good gauge of the substantive size of the relationship.

A second use of the sums of squares in assessing the model is through the *F*-test. The *F*-test is something we will cover in greater depth in Chapter 7, but briefly this test is based upon the ratio of the improvement due to the model (SS_M) and the difference between the model and the observed data (SS_R). In fact, rather than using the sums of squares themselves, we take the mean sums of squares (referred to as the *mean squares* or MS). To work out the mean sums of squares it is necessary to divide by the degrees of freedom (this is comparable to calculating the variance from the sums of squares—see section 1.1.3.1). For SS_M the degrees of freedom are simply the number of variables in the model, and for SS_R they are the number of observations minus the number of parameters being estimated (i.e. the number of beta coefficients including the constant). The result is the mean squares for the model (MS_M) and the residual mean squares (MS_R). At this stage it isn't essential that you understand how the mean squares are derived (it is explained in Chapter 7). However, it is important that you understand that the *F*-ratio (equation (4.3)) is a measure of how much the model has improved the prediction of the outcome compared to the level of inaccuracy of the model.

$$F = \frac{MS_M}{MS_R} \qquad (4.3)$$

If a model is good, then we expect the improvement in prediction due to the model to be large (so, MS_M will be large) and the difference between the model and the observed data to be small (so, MS_R will be small). In short, a good model should have a large *F*-ratio (greater than one at least) because the top half of equation (4.3) will be bigger than the bottom. The exact magnitude of this *F*-ratio can be assessed using critical values for the corresponding degrees of freedom.

4.1.4. Simple Regression on SPSS

So far, we have seen a little of the theory behind regression, albeit restricted to the situation in which there is only one predictor. To help clarify what we have learnt so far, we will go through an example of a simple regression on SPSS. Earlier on I asked you to imagine that I worked for a record company and that my boss was interested in predicting record sales from advertising. There are some data for this example in the file **Record1.sav**. This data file has 200 rows, each one representing a different record. There are also two columns, one representing the sales of each record in the week after release and the other representing the amount (in pounds) spent promoting the record before release. This is the format for entering regression data: the outcome variable and any predictors should be entered in different columns, and each row should represent independent values of those variables. The pattern of the data is shown in Figure 4.4 and it should be clear that a positive relationship exists: so, the more money spent advertising the record, the more it is likely to sell. Of course there are some records that sell well regardless of advertising (top left of scatterplot), but there are none that sell badly when advertising levels are high (bottom right of scatterplot). The scatterplot also shows the line of best fit for these data: bearing in mind that the mean would be represented by a flat line at around the 200,000 sales mark, the regression line is noticeable different.

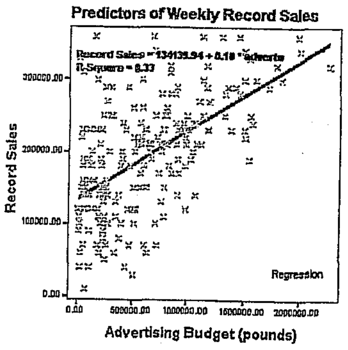
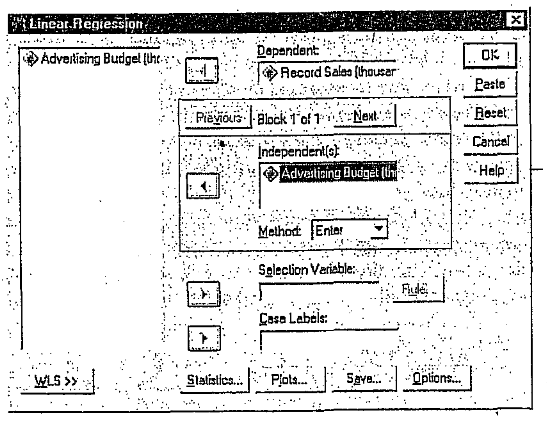


Figure 4.4: Scatterplot showing the relationship between record sales and the amount spent promoting the record

To find out the parameters that describe the regression line, and to see whether this line is a useful model, we need to run a regression analysis.

FIELD Andy. 2000. Discovering Statistics Using SPSS for Windows. Advanced Techniques for the Beginner. London: Sage.

To do the analysis you need to access the main dialog box by using the Analyze⇒Regression⇒Linear... menu path. Figure 4.5 shows the resulting dialog box. There is a space labelled *Dependent* in which you should place the outcome variable (in this example **sales**). So, select **sales** from the list on the left-hand side, and transfer it by clicking on ▸. There is another space labelled *Independent(s)* in which any predictor variable should be placed. In simple regression we use only one predictor (in this example **adverts**) and so you should select **adverts** from the list and click on ▸ to transfer it to the list of predictors. There are a variety of options available, but these will be explored within the context of multiple regression (see section 4.2). For the time being just click on OK to run the basic analysis.



Dependent: Record Sales (thousar
Independent(s): Advertising Budget (th
Method: Enter

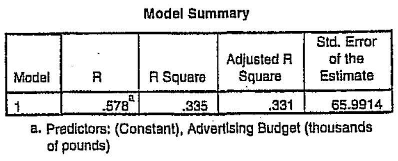
Figure 4.5: Main dialog box for regression

4.1.5. Output from SPSS

4.1.5.1. Overall Fit of the Model

The first table provided by SPSS is a summary of the model. This summary table provides the value of *R* and R^2 for the model that has been derived. For these data, *R* has a value of 0.578 and because there is only one predictor, this value represents the simple

correlation between advertising and record sales (you can confirm this by running a correlation using what you were taught in Chapter 3). The value of R^2 is 0.335, which tells us that advertising expenditure can account for 33.5% of the variation in record sales. In other words, if we are trying to explain why some records sell more than others, we can look at the variation in sales of different records. There might be many factors that can explain this variation, but our model, which includes only advertising expenditure, can explain 33% of it. This means that 66% of the variation in record sales cannot be explained by advertising alone. Therefore, there must be other variables that have an influence also.



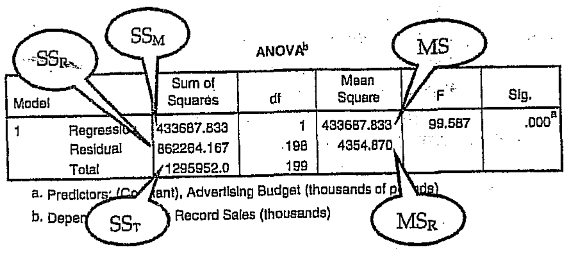
Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.578[a]	.335	.331	65.9914

a. Predictors: (Constant), Advertising Budget (thousands of pounds)

SPSS Output 4.1

The next part of the output reports an analysis of variance (ANOVA—see Chapter 7). The summary table shows the various sums of squares described in Figure 4.3 and the degrees of freedom associated with each. From these two values, the average sums of squares (the mean squares) can be calculated by dividing the sums of squares by the associated degrees of freedom. The most important part of the table is the *F*-ratio, which is calculated using equation (4.3), and the associated significance value of that *F*-ratio. For these data, *F* is 99.59, which is significant at $p < 0.001$ (because the value in the column labelled *Sig.* is less than 0.001). This result tells us that there is less than a 0.1% chance that an *F*-ratio this large would happen by chance alone. Therefore, we can conclude that our regression model results in significantly better prediction of record sales than if we used the mean value of record sales. In short, the regression model overall predicts record sales significantly well.


SS_R
SS_M
MS
SS_T
MS_R

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	433687.833	1	433687.833	99.587	.000[a]
	Residual	862264.167	198	4354.870		
	Total	1295952.0	199			

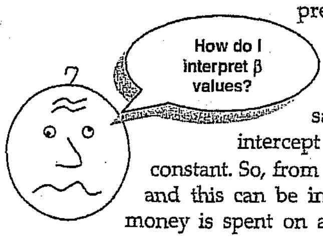
a. Predictors: (Constant), Advertising Budget (thousands of pounds)

b. Dependent Variable: Record Sales (thousands)

SPSS Output 4.2

4.1.5.2. Model Parameters

The ANOVA tells us whether the model, overall, results in a significantly good degree of prediction of the outcome variable. However, the ANOVA doesn't tell us about the individual contribution of variables in the model (although in this simple case there is only one variable in the model and so we can infer that this variable is a good



predictor). The table in SPSS Output 4.3 provides details of the model parameters (the beta values) and the significance of these values. We saw in equation (4.1) that β_0 was the Y intercept and this value is the value B for the constant. So, from the table, we can say that β_0 is 134.14, and this can be interpreted as meaning that when no money is spent on advertising (when $X = 0$), the model predicts that 134,140 records will be sold (remember that our unit of measurement was thousands of records). We can also read off the value of β_1 from the table and this value represents the gradient of the regression line. It is 9.612 E–02, which in unabbreviated form is 0.09612.[2] Although this value is the slope of the regression line, it is more useful to think of this value as representing *the change in the outcome associated with a unit change in the predictor*. Therefore, if our predictor variable is increased by 1 unit (if the advertising budget is increased by 1), then our model predicts that 0.096 extra records will be sold. Our units of measurement were thousands of pounds and thousands of records sold, so we can say that for an increase in advertising of £1000 the model predicts 96 (0.096 × 1000 = 96) extra record sales. As you might imagine, this investment is pretty bad for the record company: they invest £1000 and get only 96 extra sales! Fortunately, as we already know, advertising accounts for only one-third of record sales!

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	134.140	7.537		17.799	.000
	Advertising Budget (thousands of pounds)	9.612E-02	.010	.578	9.979	.000

a. Dependent Variable: Record Sales (thousands)

SPSS Output 4.3

[2] You might have noticed that this value is reported by SPSS as 9.612 E–02 and many students find this notation confusing. Well, this notation simply means 9.61×10^{-2} (which might be a more familiar notation). OK, some of you are still confused. Well think of E–02 as meaning 'move the decimal place 2 steps to the left', so 9.612 E–02 becomes 0.09612. If the notation read 9.612 E–01, then that would be 0.9612, and if it read 9.612 E–03, that would be 0.009612. Likewise, think of E+02 (notice the minus sign has changed) as meaning 'move the decimal place 2 places to the right'. So 9.612 E+02 becomes 961.

The values of β represent the change in the outcome resulting from a unit change in the predictor. If the model was useless at predicting the outcome, then if the value of the predictor changes, what might we expect the change in the outcome to be? Well, if the model is very bad then we would expect the change in the outcome to be zero. Think back to Figure 4.3 (see the panel representing SS_T) in which we saw that using the mean was a very bad way of predicting the outcome. In fact, the line representing the mean is flat, which means that as the predictor variable changes, the value of the outcome does *not* change (because for each level of the predictor variable, we predict that the outcome will equal the mean value). The important point here is that a bad model (such as the mean) will have regression coefficients of zero for the predictors. A regression coefficient of zero means: (a) a unit change in the predictor variable results in no change in the predicted value of the outcome (the predicted value of the outcome does not change at all), and (b) the gradient of the regression line is zero, meaning that the regression line is flat. Hopefully, what should be clear at this stage is that if a variable significantly predicts an outcome, then it should have a β value significantly different from zero. This hypothesis is tested using a *t*-test (see Chapter 6). The *t*-statistic tests the null hypothesis that the value of β is zero: therefore, if it is significant we accept the hypothesis that the β value is significantly different from zero and that the predictor variable contributes significantly to our ability to estimate values of the outcome.

One problem with testing whether the β values are different from zero is that their magnitude depends on the units of measurement (for example advertising budget has a very small β value, yet it seems to have a strong relationship to record sales). Therefore, the *t*-test is calculated by taking account of the standard error. The standard error tells us something about how different β values would be if we took lots and lots of samples of data regarding record sales and advertising budgets and calculated the β values for each sample. We could plot a frequency distribution of these samples to discover whether the β values from all samples would be relatively similar, or whether they would be very different. We can use the standard deviation of this distribution (known as the *standard error*) as a measure of the similarity of β values across samples. If the standard error is very small, then it means that most samples are likely have a β value similar to the one in the sample collected (because there is little variation across samples). The *t*-test tells us whether the β value is different from zero relative to the variation in β values for similar samples. When the standard error is small even a small deviation from zero can reflect a meaningful difference because β is representative of the majority of possible samples.

Equation (4.4) shows how the *t*-test is calculated and you'll find a general version of this equation in Chapter 6 (equation (6.1)). The β_{expected} is simply the value of β that we would expect to obtain if the null hypothesis were true. I mentioned earlier that the null hypothesis is that β is zero and so this value can be replaced by zero. The equation simplifies to become the observed value of β divided by the standard error with which it is associated.[3]

$$\begin{aligned} t &= \frac{\beta_{\text{observed}} - \beta_{\text{expected}}}{\text{SE}_\beta} \\ &= \frac{\beta}{\text{SE}_\beta} \end{aligned} \tag{4.4}$$

The values of t can then be compared to the values that we would expect to find by chance alone: if t is very large then it is unlikely to have occurred by chance. SPSS provides the exact probability that the observed value of t is a chance result, and as a general rule, if this observed significance is less than 0.05, then social scientists agree that the result reflects a genuine effect. For these two values, the probabilities are 0.000 (zero to 3 decimal places) and so we can say that the probability of these t values occurring by chance is less than 0.001. Therefore, they reflect genuine effects. We can, therefore, conclude that advertising budget makes a significant contribution ($p < 0.001$) to predicting record sales.

4.1.5.3. Using the Model

So far, we have discovered that we have a useful model, one that significantly improves our ability to predict record sales. However, the next stage is often to use that model to make some predictions. The first stage is to define the model by replacing the β values in equation (4.1) with the values from SPSS Output 4.3. In addition, we can replace the X and Y with the variable names so that the model becomes:

$$\begin{aligned} \text{Record Sales} &= \beta_0 + \beta_1 \text{Advertising Budget}_i \\ &= 134.14 + (0.09612 \times \text{Advertising Budget}_i) \end{aligned} \tag{4.5}$$

It is now possible to make a prediction about record sales, by replacing the advertising budget with a value of interest. For example, imagine a record executive wanted to spend £100,000 on advertising a new record. Remembering that our units are already in thousands of pounds, we can simply replace the advertising budget with 100. He would discover that record sales should be around 144,000 for the first week of sales.

[3] To see that this is true you can use the values from SPSS Output 4.3 to calculate t for the constant. For advertising budget, the standard error has been rounded to 3 decimal places, so to verify how t is calculated you should use the unrounded value. This value is obtained by double-clicking the table in the SPSS output and then double-clicking the value that you wish to see in full. You should find that $t = 0.096124 / 0.009632 = 9.979$.

$$\begin{aligned} \text{Record Sales} &= 134.14 + (0.09612 \times \text{Advertising Budget}_i) \\ &= 134.14 + (0.09612 \times 100) \\ &= 143.75 \end{aligned} \tag{4.6}$$

12. lekce

FAKTOROVÁ ANALÝZA - REDUKCE DAT A VSTUP DO MULTIVARIAČNÍ ANALÝZY (Modul ANALYZE: pocedura: Data reduction-factor analysis).

DISMAN Miroslav. 1993. *Jak se vyrábí sociologická znalost.* Praha: Karolinum.

Faktorová analýza

To je technika, která byla a někde ještě je přijímána s určitou nedůvěrou. Domníváme se, že *neprávem*. Důvodem pro tuto nedůvěru je zřejmě fakt, že funkce faktorové analýzy je velice odlišná od ostatních statistických technik. Většina statistických operací v sociologickém výzkumu je používána pro **testování hypotéz**, faktorová analýza je používána pro tento účel spíše výjimečně. Faktorová analýza je spíše nástrojem pro explorativní výzkum. Většinou netestuje hypotézy, ale je nástrojem pro jejich formulování a upřesňování. Je ovšem i neobyčejně účinným nástrojem zjednodušování dat.

Funkci faktorové analýzy si můžeme vysvětlit - velice nespisovně, ale jasně a v podstatě správně - asi takto: faktorová analýza je schopna **nalézt seskupení proměnných, které patří <u>nějakým</u> způsobem k sobě.** Jaký je to způsob, to nám faktorová analýza neřekne. Na to musí odpověď nalézt výzkumník sám, na základě své odborné znalosti.

A teď si ukažme, až v neslušně zjednodušené a zkrácené formě, jak tato analýza funguje. Vstup do faktorové analýzy je korelační matice, tabulka korelačních koeficientů mezi všemi proměnnými, které hodláme analyzovat. Výstupem z faktorové analýzy jsou sloupce čísel, z nichž každý představuje jeden extrahovaný **faktor**. Faktor je ono dosud nepojmenované "něco" co sdružuje proměnné s vysokými čísly v daném sloupci. Čísla v sloupcích jsou nazývána "factor loadings", **faktorová zátěž,** míra spojení proměnné v daném řádku s tímto faktorem. Je možné říci, že faktorová zátěž je korelace mezi proměnnou a faktorem.

Ideální by bylo, kdyby skupiny proměnných měly vysokou zátěž v jednom faktoru a téměř nulovou ve všech ostatních faktorech. Faktorová analýza se snaží dosáhnout tohoto cíle *značně složitými matematickými postupy, které bychom snad mohli nejlépe popsat* jako rotaci souřadnic v mnohodimenzionálním prostoru. Tyto postupy nejsou bez problémů a pouštět se do faktorové analýzy bez spolupráce statistika může být někdy velice riskantní. Nicméně ani tyto velice chytré postupy nemohou izolovat zcela čisté faktory. Nezapomeňte na první Dismanův zákon ("data jsou potvory") nebo jinými slovy, na rozsáhlé sítě souvislostí v společenských vědách. Tak v tabulce 9.12. vidíme, že proměnná (5), strava, má vysokou zátěž v prvém faktoru, který, jak uvidíme, se zdá zachycovat etnickou symboliku stravy, ale i nezanedbatelnou náložv druhém faktoru, odrážejícím materiální utilitu stravy.

A jak interpretovat obsahový význam faktorů? Na to nemáme jednoduchý recept kromě jednoho. Použít svou odbornou znalost a zdravý rozum. Někdy to může být velice jednoduché. Řekněme, že jsme třeba připravili sérii otázek, o nichž se domníváme, že měří týž koncept. Faktorová analýza nám dává možnost testovat tento předpoklad. V ideálním případě by měla být schopna extrahovat jediný faktor, t.j. koeficienty by měly být nejsilnější v prvém faktoru. Proměnné, které by měly vysoké koeficienty v jiném než prvém faktoru, měří zřejmě něco jiného a měly by být ze souboru vyloučeny. To je ovšem případ, kdy je faktorová analýza použita pro testování hypotéz. Tato analýza je také výhodným nástrojem na zjednodušení souboru proměnných.

Rozsáhlá faktorová analýza, používající data z desítek velice rozdílných populací byla s to redukovat stovky stimulů, používaných v technice sémantického diferenciálu. V běžné praxi se nyní dosti standardně používá kolem deseti párů podnětů, a to bez jakékoliv podstatné ztráty informace. Faktorová analýza jako nástroj pomáhající ustavit validitu jiných výzkumných technik má značně široké pole použití. Slyšeli jsme *dokonce o tom*, jak tato technika byla použita pro odhalení tazatele, který falšoval rozhovory; nenavštěvoval respondenty, ale *vymýšlel si odpovědi* u *svého* psacího stolu. Byl dosti chytrý, aby odpovědi v jednotlivých rozhovorech byly konzistentní, ale nebyl - a nemohl být - chytrý natolik, aby se faktory vyvozené z jeho dat shodovaly s faktory extrahovanými z dat ostatních tazatelů.

Ovšem důležitou oblastí aplikace této techniky je explorativní výzkum a zde se interpretace významů faktorů stává základním úkolem. Klíčem pro identifikaci faktoru je v podstatě odpověď na tuto otázku: "Co mají proměnné s vysokými koeficienty v daném faktoru společného?" Forslund (1980) studoval delikvenci mládeže v malém městě ve Wyomingu. (Citováno v Babbie, 1989.) Rozdal středoškolským studentům dotazník, ve kterém se dotazoval na řadu lehce delikventních aktivit. Ve *faktorové analýze* pak identifikoval 4 zřetelné faktory.

Kupř. následující proměnné měly vysoké koeficienty ve faktoru 1:

- rozbíjení pouličních světel
- rozbíjení oken
- vypouštění vzduchu z pneumatik automobilů
- drobné krádeže atd.

Tento faktor byl identifikován jako přestupky proti majetku.

Druhý faktor byl identifikován jako nezvládnutelnost a měl vysokou zátěž kupř. v následujících proměnných:

- neuposlechnutí rodičů
- psaní a malování po stěnách a lavicích
- odmlouvání učitelům
- ánonymní telefonické rozhovory

Poslední faktor byl zcela jednoznačný, měl vysokou zátěž jen ve dvou proměnných, týkajících se rvaček a bitek.

Třetí faktor jsme si nechali na konec. Je totiž zajímavý z hlediska metodologie faktorové analýzy a proto zde uvedeme všechny proměnné, ve kterých měl významnou zátěž, i spolu s koeficienty:

-	kouření marihuany	.755
-	používání jiných drog	.669
-	falšování podpisů na omluvenkách	.395
-	pití alkoholu v nepřítomnosti rodičů	.358
-	chození za školu	.319

Vidíme zřetelně, že tento faktor shrnuje dva různé problémy: drogy a chození za školu. Forslund také použil pro tento faktor toto dvojité jméno. Pro nás je tato podvojnost faktoru ilustrací jedné důležité a nebezpečné vlastnosti faktorové analýzy. V naší interpretaci jsou obě složky faktoru sdruženy jinak, než tomu bylo u zbývajících faktorů. U nich proměnné byly sdruženy tím způsobem, že všechny proměnné byly ukazateli shodného podloženého faktoru. V případě faktoru 3. jeho obě složky mohou patřit k dvěma poněkud odlišným podloženým konceptům, "chození za školu" a "drogy". Faktorová analýza je prezentovala jako jediný faktor prostě proto, že tyto dvě složky jsou v nějakém, patrně silném, kauzálním vztahu. Je jedno, zda nedostatek dohledu v nekontrolované situaci "za školou" zvyšuje pravděpodobnost přístupu k drogám, nebo zda snaha získat "marjanku" vede k chození za školu.

Tohle není útok na validitu Forslundova výzkumu, ta byla dobře potvrzena i tím, že faktorová analýza byla opakována odděleně pro subpopulace chlapců a děvčat a v obou případech extrahovala faktory srovnatelné s výsledky původní analýzy. Jde nám jen o to zdůraznit technické aspekty, které jsou spojeny právě s jedinečnou schopností faktorové analýzy nalézt struktury v datech, která "nějak" spolu souvisejí. Ono "nějak" může být opravdu odrazem obsahu proměnných, a tak umožní identifikovat struktury, které mohou být

teoreticky významné a které jsou neviditelné jiným analytickým postupům. Ale to "nějak" může mít někdy velice prozaický a mechanický charakter. Faktorová analýza může prostě identifikovat jako samostatný faktor skupinu proměnných jenom proto, že informace byla získána shodnou technikou sběru dat. Řekněme, že existují určité sociální typy osobnosti, charakterizované jednak postoji jedince, jednak jeho demografickými charakteristikami. Kdybychom chtěli "objevit" tyto typy s pomocí faktorové analýzy, dočkali bychom se pravděpodobně značného zklamání. Analýza by asi objevila jenom (nebo hlavně) dva faktory: jeden, obsahující demografické proměnné, druhý postojové proměnné. Jsou cesty, jak takové problémy překonat, ale obecně bychom se měli vyvarovat toho, používat jako vstup do faktorové analýzy proměnné, které jsou formálně různorodé.

Pokud není faktorová analýza použita jako pouhé testování hypotéz, představuje jenom prvý krok v poznávacím procesu. Jejím výsledkem může být třeba jen nová formulace problému, navržení nových hypotéz. Ale to není málo.

V našem torontském výzkumu o postojích starých Italů, Portugalců a anglicky mluvících osob, narozených v Kanadě jsme se zajímali o struktury obsahu strachu, který staří lidé pociťují téměř univerzálně, uvažují-li o možnosti, že budou nucení vstoupit do domova důchodců. Zjišťovali jsme mimo jiné, jak by pro respondenta byly důležité - v případě institucionalizace - následující body:

1. být odříznut od sousedství, na které je respondent zvyklý;
2. možnost pokračovat v koníčcích, které nyní má;
3. mít pokoj pro sebe;
4. přinést si do instituce nábytek a jiné věci;
5. dostávat stravu, na kterou je respondent zvyklý;
6. bydlet s někým, kdo je s respondentem sociálně srovnatelný a
7. bydlet s osobou mluvící respondentovým jazykem.

Výsledky byly podrobeny faktorové analýze a zde jsou její výsledky pro italskou subpopulaci:

Tabulka 9.12.

FAKTOR:	(1)	(2)	(3)
Proměnná:			
(6) osoba jako respondent	.901*	-.071	-.047
(7) jazyk	.896*	.015	.007
(5) strava	.604*	.335	.100
(4) přinést věci	.026	.813*	-.123
(3) pokoj pro sebe	.067	.772*	.323
(2) koníček	-.258	.180	.818*
(1) sousedství	.397	-.070	.677*

Zdá se, že prvý faktor reprezentuje **etnickou identitu** respondentů: jazyk jako symbol národní identity, potřeba partnera "jako já" jako symbol kulturní a třídní sounáležitosti, a konečně strava, která je v literatuře zcela shodně považována za jeden z nejdéle přežívajících prvků etnické kultury.

Druhý faktor se zdá reprezentovat jedincovo **pragmatické, materiální okolí** a konečně třetí faktor může reprezentovat **sféru soukromých, osobních zájmů.** Plausibilita této interpretace zdá se být podporována i následujícím faktem. Zcela shodné faktory byly extrahovány i pro portugalský vzorek, ne však pro vzorek rozených, anglicky mluvících Kanaďanů. Pro ně, jako členy hlavního kulturního proudu není etnická identita vůbec problémem. Zde faktorová analýza izolovala pouze dva slabé faktory, odpovídající přibližně faktorům 2. a 3. ve zbývajících etnických skupinách. Další teoretickou podporou pro naši interpretaci jsou výsledky analýzy jiného souboru proměnných ze stejného výzkumu. Tyto proměnné měřily míru obav asociovaných s různými prvky spojenými se vstupem do instituce pro staré. Faktorová analýza těchto dat extrahovala opět faktory významem podobné těm,které jsme zde právě představili.

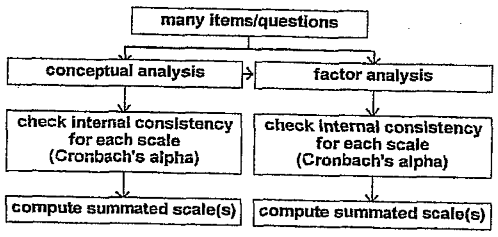
Tyto nálezy samy o sobě neznamenají mnoho. Ale otevřely pro nás docela zajímavou cestu do studia obsahu strachu a jeho sociálních a zejména kulturních determinant. Ale to je už zase jiná povídka, které musí být teprve napsána.

CHAPTER 10

Factor Analysis: Data Reduction With Principal Components Analysis

Factor analysis, a complex associational technique, is used for several purposes, but the main one is data reduction.[1] When you have a number of questions about the same general topic (e.g., attitudes about mathematics), you may want to ask whether the questions could be grouped into a smaller number of composite variables. Table 10.1 shows that *either* factor analysis or conceptual analysis (i.e., thinking based on theory and/or literature) can be used to reduce the number of variables to a more manageable and meaningful number of summated scales. The table also shows that you should check the internal consistency reliability of these new scales with Cronbach's alpha (see Assignment G) before actually computing the scales. You may want to check your conceptual analysis with factor analysis as well as Cronbach alphas. That is what we are doing here for our conceptualization of the three math attitude scales.

Table 10.1. ***Two Strategies for Reducing Many Related Items to Fewer Composite Variables***


many items/questions
conceptual analysis
factor analysis
check internal consistency for each scale (Cronbach's alpha)
check internal consistency for each scale (Cronbach's alpha)
compute summated scale(s)
compute summated scale(s)

After doing your conceptual or factor analysis, you have several decisions:

1. If you identify only one conceptual scale or one factor and it is supported by an alpha above .70, compute one overall summary scale score. However,
2. If there is more than one conceptual scale or more than one factor and they have good alphas, compute the several summated scale scores. Also, compute an overall scale score *if* it makes sense conceptually. However,
3. If the factor analysis results do not make good conceptual sense, do not use them. In this case, use the conceptual factors, rethink the conceptualization, or use each item separately

[1] Statisticians call what we have done in this chapter principal components analysis (PCA) rather than exploratory factor analysis (EFA). In SPSS, principal components analysis is done with the factor analysis program using the **principal components extraction** method. This is consistent with common usage, but there are technical differences between PCA and EFA (see Grimm & Yarnold, 1994).

Problems/ Research Questions

1. Can variables *q01* to *q13* be grouped into a smaller number of composite variables called components or factors? Using the principal components extraction method of the factor analysis program, you will have the computer sort the variables and suppress printing of values if the factor loading is less than .30. You will use a Varimax rotation and allow all factors with eigenvalues over 1.00 to be computed.
2. Rerun the factor analysis but specify that you want the number of factors to be three because our conceptualization is that there are three math attitude scales or factors: motivation, competence, and pleasure.
3. Run a factor analysis to see how the four "achievement" variables, *mathach, visual, mosaic,* and *grades*, cluster or factor.

Lab Assignment F

Logon and Get Data

- Retrieve your most recent data file: **hsbdataE.**

Problem 1: Factor Analysis on Math Attitude Variables

To begin factor analysis use these commands:
- **Statistics => Data Reduction => Factor** to get Fig. 10.1.
- Next select the variables *q01* through *q13*. *Do not* include q04r or any of the other reversed questions.

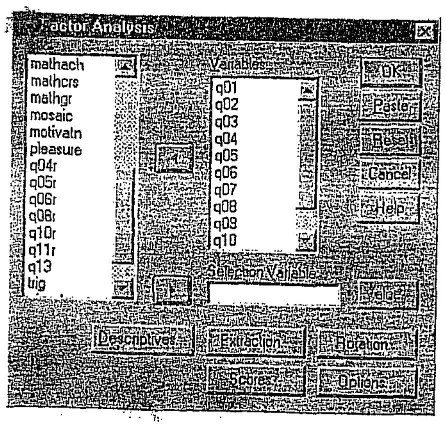

Factor Analysis
mathach
mathcrs
mathgr
mosaic
motivatn
pleasure
q04r
q05r
q06r
q08r
q10r
q11r
q13
trig
Variables:
q01
q02
q03
q04
q05
q06
q07
q08
q09
q10
Selection Variable:

Fig. 10.1. Factor analysis.

Now click on **Descriptives** to produce Fig. 10.2.
- Then click on the following: **Initial solution** (under **Statistics**), **Coefficients, Determinant, KMO and Bartlett's test of sphericity** (under **Correlation Matrix**).
- Click on **Continue.**

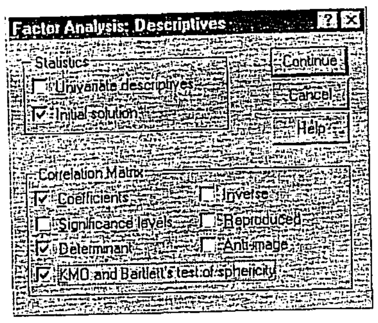

Factor Analysis: Descriptives
Statistics
☐ Univariate descriptives
☑ Initial solution
Correlation Matrix
☑ Coefficients
☐ Inverse
☐ Significance levels
☐ Reproduced
☑ Determinant
☐ Anti-image
☑ KMO and Bartlett's test of sphericity

Fig. 10.2. Factor analysis: Descriptives.

- Next, click on **Extraction** at the bottom of Figure 10.1. This will give you Fig. 10.3.
- Make sure **Eigenvalues over 1** is checked.

This default setting will allow the *computer to decide* how many math attitude factors to compute; i.e., as many as have eigenvalues (a measure of variability explained) greater than 1.0. If you have a clear theory or conceptualization about how many factors or scales there should be, you can set the **number of factors** to that number, as we will in Problem 2.

- *Unclick* display **Unrotated factor solution.**
- Click on **Continue.**

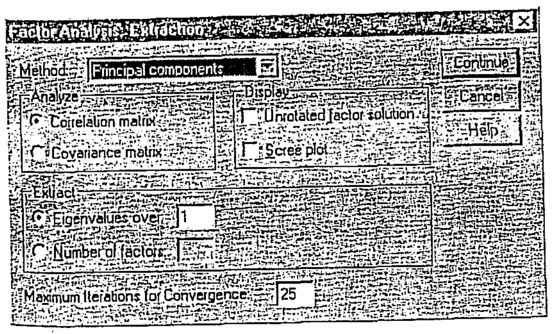

Factor Analysis: Extraction
Method: Principal components
Analyze
● Correlation matrix
○ Covariance matrix
Display
☐ Unrotated factor solution
☐ Scree plot
Extract
● Eigenvalues over: 1
○ Number of factors:
Maximum Iterations for Convergence: 25

Fig. 10.3. Extraction method to produce principal components analysis (PCA).

- Now click on **Rotation,** which will give you Fig. 10.4.
- Click on **Varimax.**
- Then click on **Continue.**

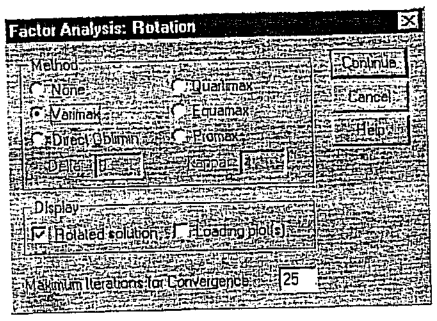

Factor Analysis: Rotation
Method
○ None
○ Quartimax
● Varimax
○ Equamax
○ Direct Oblimin
○ Promax
Delta: 0
Kappa: 4
Display
☑ Rotated solution
☐ Loading plot(s)
Maximum Iterations for Convergence: 25

Fig. 10.4. Factor analysis: Rotation.

- Next, click on **Options** which will give you Fig. 10.5.
- Then click on **Sorted by size.**
- Click on **Suppress absolute values less than** and type .3 (point 3) in the box (see Fig. 10.5). Suppressing small factor loadings makes the output easier to read.
- Click on **Continue** then **OK**. Compare Output 10.1 to your output and syntax.

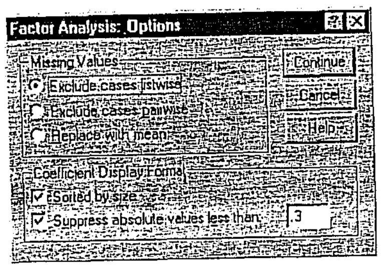

Factor Analysis: Options
Missing Values
● Exclude cases listwise
○ Exclude cases pairwise
○ Replace with mean
Coefficient Display Format
☑ Sorted by size
☑ Suppress absolute values less than: .3

Fig. 10.5. Factor analysis: Options.

Problem 2: Factor Analysis on Math Attitude Variables With Three Factors Specified

Now try doing factor analysis yourself with the same variables, rotation, and options. This time, however, *click off* everything on the descriptive screen except **Initial solution.** Also, use a different **Extraction** subcommand; click on **Extract Number of factors** and then type 3 because our conceptualization is that there are 3 factors. Compare Output 10.2 to your output and syntax. Note that the **Initial Statistics** table is the same, but the **Rotated Component Matrix** now shows three somewhat different factors.

Problem 3: Factor Analysis on Achievement Variables

Now try doing another factor analysis yourself on the "achievement variables," *mathach, visual, mosaic,* and *grades*.

- First press **Reset**.
- Use the default settings (i.e., the boxes that are already checked) for **Extraction**.
- In addition, under **Rotation**, check **Varimax**, display **Rotated solution**, and **Loading plots**.
- Under **Descriptives** check **Univariate descriptives, Initial solution, Coefficients, Determinant**, and **KMO** and **Bartlett's test of sphericity.**

We have requested a more extensive, less simplified output for contrast with the earlier ones. Compare Output 10.3 to your syntax and output.

Print, Save, and Exit

- **Print** your lab assignment results if you want.
- **Save** your data file as **hsbdataF** (**File => Save As**).
- **Save** the SPSS log files as **hsblogF.**
- **Exit** SPSS.

Interpretation Questions

1. Using Output 10.1: a) Make a table of the five highest correlations and the five lowest. Indicate whether the variables for the highest and lowest correlations are in the same or in different conceptual clusters (i.e., competence, motivation, and pleasure) as indicated on page 19 for each question. b) What might you name each component or factor in the rotated factor matrix? c) How do these statistical components differ from the three conceptual math attitude composite variables (competence, motivation, and pleasure) computed in Assignment C and shown in chapter 2 and the codebook (Appendix D)?

2. Using Output 10.2: a) How do the rotated components in Output 10.2 differ from those in Output 10.1? b) Are the factors in Output 10.2 closer to the conceptual composites in the codebook? c) How might you name the three factors in Output 10.2?

3. Using Output 10.3: a) Are the assumptions that were tested violated? Explain. b) How many components or factors are there with eigenvalues greater than 1.0, and what total/cumulative percent of variance is accounted for by them? c) Describe the main aspects of the correlation matrix, rotated component matrix, and plot in Output 10.3.

```
GET
  FILE='A:\hsbdataE.sav'.
EXECUTE .
```

Output 10.1: Factor Analysis for Math Attitude Questions

```
Syntax for factor analysis of math attitude questions

FACTOR
  /VARIABLES q01 q02 q03 q04 q05 q06 q07 q08 q09 q10 q11 q12 q13  /MISSING
 LISTWISE /ANALYSIS q01 q02 q03 q04 q05 q06 q07 q08 q09 q10 q11 q12 q13
  /PRINT INITIAL CORRELATION DET KMO ROTATION
  /FORMAT SORT BLANK(.3)
  /CRITERIA MINEIGEN(1) ITERATE(25)
  /EXTRACTION PC
  /CRITERIA ITERATE(25)
  /ROTATION VARIMAX
  /METHOD=CORRELATION .
```

Interpret Output 10.1

The factor analysis program generates a number of tables depending on which options you have chosen. The first table in Output 10.1 is a **correlation matrix** showing how each of the 13 questions is associated with each of the other 12. Note some of the correlations are high (e.g., + or - .60 or greater) and some are low (i.e., near zero). The high correlations indicate that two items are associated and will probably be grouped together by the factor analysis.

Next, several assumptions are tested. The **determinant** (located under the correlation matrix) should be more than .00001. For instance, 2.316E-3 is the same as .002316 so this assumption is met. The **KMO** should be greater than .70, and is inadequate if less than .50. The **Bartlett** test should be significant (i.e., significance less than .05); these assumptions also are met.

The **Total Variance Explained** table shows how the variance is divided among the 13 possible components/factors. Note that four factors have **eigenvalues** (a measure of explained variance) greater than 1.0, which is a common criterion for a factor to be useful. Thus, unless you specify otherwise, as we will in Problem 2, the computer will look for the best four-factor solution.

In this case, the computer tried seven iterations before converging on the solution shown in the **Rotated Component Matrix** table. This table is the key one for understanding the results of the analysis. Note that the computer has sorted the 13 math attitude questions (Q01 to Q13) into four groups of 5, 3, 3, and 2 items, respectively. Within each component, the items are sorted from the one with the highest factor weight or loading (i.e., Q05 for factor 1, with a loading of -.88) to the one with the lowest (q02) that was still loaded *the most* on that factor. We have enclosed these items in circles for easy identification. Loadings are correlation coefficients of each item with the component so they range from -1.0 through 0 to + 1.0. A negative loading just means that the question needs to be reversed when interpreting that factor.

The investigator should examine the content of the items that load high on each factor to see if they fit together conceptually and can be named. Items 5, 3, and 11 were intended to reflect an attitude or perception of competence at math (see page 19). Item 1 was intended to measure motivation for doing math, but in retrospect one can imagine that the phrase "until I can do them well" could be interpreted as competence. Likewise, Item 2, " I feel happy after solving a hard problem," although intended to measure pleasure at doing math, might also reflect competence at doing math. Every item has a weight or loading on every factor, but in a "clean" factor analysis almost all of the loadings that are not in the circles that we have drawn will be quite low (less than .40). We asked the computer to print only loadings of .30 or above so all the blanks in the table are low loadings. Note that Item 11 and, especially, Item 2 load above .40 on both Components 1 and 4. The latter component could be labeled pleasure at math, which was conceptually composed of Items 2, 6 and 10.

For our purposes, we will ignore the Factor Transformation Matrix; it was used to convert the initial factor matrix into the rotated factor matrix.

Factor Analysis

High correlations mean these items probably will be in the same factor

Indicates how each question is associated with each other

Correlation Matrix[a]

		question 1	question 2	question 3	question 4	question 5	question 6	question 7	question 8	question 9	question 10	question 11	question 12	question 13
Correlation	question 1	1.000	.484	.626	-.305	-.745	-.165	.461	-.340	.209	.071	-.441	.166	.187
	question 2	.484	1.000	.389	-.166	-.547	-.312	.361	-.176	.219	-.369	-.401	.116	.028
	question 3	.626	.389	1.000	-.348	-.743	-.209	.423	-.248	.328	.027	-.513	.165	.170
	question 4	-.305	-.166	-.348	1.000	.363	.323	-.596	.576	-.120	.102	.398	-.391	-.334
	question 5	-.745	-.547	-.743	.363	1.000	.260	-.536	.276	-.351	.130	.605	-.187	-.169
	question 6	-.165	-.312	-.209	.323	.260	1.000	-.268	.192	-.131	.217	.418	-.044	.001
	question 7	.461	.361	.423	-.596	-.536	-.268	1.000	-.608	.226	-.169	-.331	.347	.361
	question 8	-.340	-.176	-.248	.576	.276	.192	-.608	1.000	-.243	.067	.370	-.392	-.308
	question 9	.209	.219	.328	-.120	-.351	-.131	.226	-.243	1.000	-.109	-.407	.406	.286
	question 10	.071	-.369	.027	.102	.130	.217	-.169	.067	-.109	1.000	.250	-.059	-.062
	question 11	-.441	-.401	-.513	.398	.605	.418	-.331	.370	-.407	.250	1.000	-.148	-.006
	question 12	.166	.116	.165	-.391	-.187	-.044	.347	-.392	.406	-.059	-.148	1.000	.507
	question 13	.187	.028	.170	-.334	-.169	.001	.361	-.308	.286	-.062	-.006	.507	1.000

a. Determinant = 2.316E-03

Low correlations will not be in the same factor

Should be greater than .0001. If close to zero, colinearity is too high

Test of assumptions

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.787
Bartlett's Test of Sphericity	Approx. Chi-Square	393.413
	df	78
	Sig.	.000

Should be greater than .70

Should be significant (less than .05)

Communalities

	Initial
question 1	1.000
question 2	1.000
question 3	1.000
question 4	1.000
question 5	1.000
question 6	1.000
question 7	1.000
question 8	1.000
question 9	1.000
question 10	1.000
question 11	1.000
question 12	1.000
question 13	1.000

Extraction Method: Principal Component Analysis.

Notice these are greater than one. Eigenvalues refer to the variance explained or accounted for.

Total Variance Explained

	Initial Eigenvalues			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.805	36.963	36.963	3.207	24.666	24.666
2	1.826	14.049	51.011	2.327	17.898	42.564
3	1.333	10.255	61.267	1.887	14.514	57.078
4	1.133	8.718	69.985	1.678	12.907	69.985
5	.883	6.791	76.776			
6	.666	5.120	81.895			
7	.541	4.159	86.055			
8	.453	3.481	89.536			
9	.380	2.920	92.456			
10	.299	2.299	94.755			
11	.285	2.193	96.948			
12	.241	1.853	98.801			
13	.156	1.199	100.000			

Extraction Method: Principal Component Analysis.

Over 2/3 of the variance is accounted for by these four factors.

Percent of variance for each component after rotation.

Component Matrix[a]

a. 4 components extracted.

Rotated Component Matrix [a]

Factor weights or loadings are interpreted similarly to correlations.

	Component			
	1	2	3	4
question 5	-.878			
question 3	.839			
question 1	.833			
question 11	-.596			.447
question 2	.559			-.551
question 4		.838		
question 8		.763		
question 7	.371	-.699		
question 12		-.323	.805	
question 13		-.315	.768	
question 9	.368		.662	
question 10				.853
question 6		.382		.560

Items in each component are sorted from highest factor weight to lowest.

Questions 11 and 2 "load" above .40 on factor 4 as well as factor 1.

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 7 iterations.

Component Transformation Matrix

Component	1	2	3	4
1	.711	-.540	.343	-.292
2	-.495	-.410	.681	.352
3	.484	.193	.028	.853
4	.123	.709	.647	-.251

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

Output 10.2: Factor Analysis for Math Attitude Questions Limited to Three Factors

Syntax for factor analysis of math attitude questions limited to three factors

```
FACTOR
  /VARIABLES q01 q02 q03 q04 q05 q06 q07 q08 q09 q10 q11 q12 q13  /MISSING
  LISTWISE /ANALYSIS q01 q02 q03 q04 q05 q06 q07 q08 q09 q10 q11 q12 q13
  /PRINT INITIAL ROTATION
  /FORMAT SORT BLANK(.3)
  /CRITERIA FACTORS(3) ITERATE(25)
  /EXTRACTION PC
  /CRITERIA ITERATE(25)
  /ROTATION VARIMAX
  /METHOD=CORRELATION .
```

Factor Analysis

Communalities

	Initial
question 1	1.000
question 2	1.000
question 3	1.000
question 4	1.000
question 5	1.000
question 6	1.000
question 7	1.000
question 8	1.000
question 9	1.000
question 10	1.000
question 11	1.000
question 12	1.000
question 13	1.000

Extraction Method: Principal Component Analysis.

Total Variance Explained

	Initial Eigenvalues			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.805	36.963	36.963	3.289	25.300	25.300
2	1.826	14.049	51.011	2.843	21.869	47.169
3	1.333	10.255	61.267	1.833	14.097	61.267
4	1.133	8.718	69.985			
5	.883	6.791	76.776			
6	.666	5.120	81.895			
7	.541	4.159	86.055			
8	.453	3.481	89.536			
9	.380	2.920	92.456			
10	.299	2.299	94.755			
11	.285	2.193	96.948			
12	.241	1.853	98.801			
13	.156	1.199	100.000			

Extraction Method: Principal Component Analysis.

Component Matrix [a]

a. 3 components extracted.

MORGAN George A. and Griego V. ORLANDO. 1998. Easy Use and Interpretation of SPSS for Windows: Answering Research Questions With Statistics. Mahwah: Lawrence Erlbaum Associates.

Rotated Component Matrix [a]

	Component		
	1	2	3
question 5	-.878		
question 1	.851		
question 3	.845		
question 11	-.598		.484
question 12		.815	
question 13		.786	
question 8		-.682	
question 4		-.651	
question 7	.421	.611	
question 9	.303	.397	
question 10			.807
question 6			.651
question 2	.534		-.537

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization

a. Rotation converged in 5 iterations.

Component Transformation Matrix

Component	1	2	3
1	.729	.585	-.356
2	-.477	.806	.350
3	.492	-.086	.867

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

Output 10.3: Factor Analysis for Achievement Scores

Syntax for factor analysis of achievement scores

```
FACTOR
  /VARIABLES mathach visual mosaic grades  /MISSING LISTWISE /ANALYSIS
  mathach visual mosaic grades
  /PRINT UNIVARIATE INITIAL CORRELATION DET KMO EXTRACTION ROTATION
  /PLOT ROTATION
  /CRITERIA MINEIGEN(1) ITERATE(25)
  /EXTRACTION PC
  /CRITERIA ITERATE(25)
  /ROTATION VARIMAX
  /METHOD=CORRELATION .
```

Interpret Output 10.3

Compare Output 10.3 to your output and syntax in Output 10.1. Note that in addition to the tables in Output 10.1 you have: a) a table of descriptive statistics for each variable; it also provides the listwise *N* which you would not know otherwise, b) a table of communalities, c) a component matrix, which is unrotated and is used for purposes beyond the scope of this book, and d) plots of the factor loadings. Note that the default setting we used does not sort the variables by factors and does not suppress low loadings in the **rotated factor matrix**. Thus, you have to organize the table yourself, i.e., *mathach*, *grades*, and *visual* in that order are factor 1, and *mosaic* is factor 2.

Factor Analysis

Descriptive Statistics

	Mean	Std. Deviation	Analysis N
math achievement	12.5645	6.6703	75
visualization score	5.2433	3.9120	75
mosaic, pattern test	27.413	9.574	75
grades in h.s.	5.68	1.57	75

Correlation Matrix [a]

		math achievement	visualization score	mosaic, pattern test	grades in h.s.
Correlation	math achievement	1.000	.423	.213	.504
	visualization score	.423	1.000	.030	.127
	mosaic, pattern test	.213	.030	1.000	-.012
	grades in h.s.	.504	.127	-.012	1.000

a. Determinant = .562

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.468
Bartlett's Test of Sphericity	Approx. Chi-Square	41.414
	df	6
	Sig.	.000

This is not adequate because there is only one score (mosaic) to represent the second component. You should have several for each component.

Communalities

	Initial	Extraction
math achievement	1.000	.801
visualization score	1.000	.401
mosaic, pattern test	1.000	.959
grades in h.s.	1.000	.603

Extraction Method: Principal Component Analysis.

MORGAN George A. and Griego V. ORLANDO. 1998. *Easy Use and Interpretation of SPSS for Windows: Answering Research Questions With Statistics*. Mahwah: Lawrence Erlbaum Associates.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.755	43.873	43.873	1.755	43.873	43.873	1.717	42.928	42.928
2	1.009	25.237	69.110	1.009	25.237	69.110	1.047	26.183	69.110
3	.872	21.795	90.906						
4	.364	9.094	100.000						

Extraction Method: Principal Component Analysis.

Component Matrix [a]

	Component	
	1	2
math achievement	.894	3.309E-02
visualization score	.629	-6.96E-02
mosaic, pattern test	.267	.943
grades in h.s.	.699	-.339

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

Rotated Component Matrix [a]

	Component	
	1	2
math achievement	.864	.234
visualization score	.629	7.393E-02
mosaic, pattern test	4.740E-02	.978
grades in h.s.	.757	-.173

Refer to the questions at the end of the assignment.

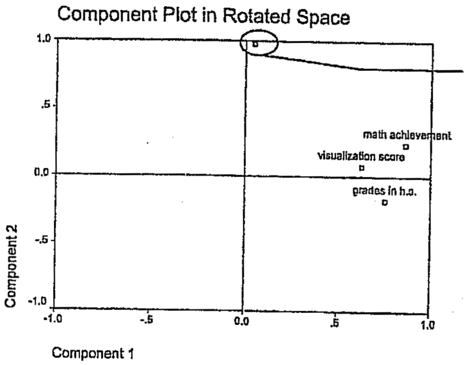
Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

Component Transformation Matrix

Component	1	2
1	.974	.225
2	-.225	.974

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.



This is mosaic.

MORGAN George A. and Griego V. ORLANDO. 1998. *Easy Use and Interpretation of SPSS for Windows: Answering Research Questions With Statistics*. Mahwah: Lawrence Erlbaum Associates.

are unequal intervals between first, second, and third place with a very small interval between second and third and a much larger one between first and second.

Interval and Ratio Scales/Variables (i.e., Equal Interval Scales)

Interval scales have not only mutually exclusive categories that are ordered from low to high, but also the categories are equally spaced (i.e., have equal intervals between them). Most physical measurements (length, weight, money, etc.) are ratio scales because they not only have equal intervals between the values/categories, but also have a true zero, which means in the above examples, no length, no weight, or no money. Few psychological scales have this property of a true zero and thus even if they are very well constructed equal interval scales, it is not possible to say that one has no intelligence or no extroversion or no attitude of a certain type. While there are differences between interval and ratio scales, the differences are not important for us because we can do all of the types of statistics that we have available with interval data. As long as the scale has equal intervals, it is not necessary to have a true zero.

Distinguishing Between Ordinal and Interval Scales

It is usually fairly easy to tell whether three categories are ordered or not, so students and researchers can distinguish between nominal and ordinal data, except perhaps when there are only two categories, and then it does not matter. The distinction between nominal and ordinal makes a lot of difference in what statistics are appropriate. However, it is considerably harder to distinguish between ordinal and interval data. While almost all *physical* measurements provide either ratio or interval data, the situation is less clear with regard to psychological measurements.

When we come to the measurement of psychological characteristics such as attitudes, often we cannot be certain about whether the intervals between the ordered categories are equal, as required for an interval level scale. Suppose we have a five-point scale on which we are to rate our attitude about a certain statement from strongly agree as 5 to strongly disagree as 1. The issue is whether the intervals between a rating of 1 and 2, 2 and 3, 3 and 4, and 4 and 5 are all equal or not. One could argue that because the numbers are equally spaced on the page, and because they are equally spaced in terms of their numerical values, the subjects will view them as equal intervals. However, especially if the in-between points are identified (e.g., strongly agree, agree, neutral, disagree, and strongly disagree), it could be argued that the difference between strongly agree and agree is not the same as between agree and neutral; this contention would be hard to disprove. Some questionnaire or survey items have response categories that are not exactly equal intervals. For example, let's take the case where the subjects are asked to identify their age as one of five categories: 21 to 30, 31 to 40, 41 to 50, 51 to 60, and 61 and above. It should be clear that the last category is larger in terms of number of years covered than the other four categories. Thus, the age intervals are not exactly equal. However, we would consider this scale and the ones above to be at least *approximately interval*.

On the other hand, an example of an ordered scale that is clearly not interval would be one that asked how frequently subjects do something. The answers go something like this: every day, once a week, once a month, once a year, once every 5 years. You can see that the categories become wider and wider and, therefore, are not equal intervals. There is clearly much more difference between 1 year and 5 years than there is between 1 day and 1 week. Most of the above information is summarized in the top of Table 3.2.

Table 3.2. *Selection of Appropriate Descriptive Statistics for One Dependent Variable*

	Level/Scale of Measurement of Variable		
	Nominal	**Ordinal**	**Interval or Ratio**
Characteristics of the Variable	- Qualitative data - Not ordered - True categories: only names, labels	- Quantitative data - Ordered data - Rank order only	- Quantitative data - Ordered data - Equal intervals between values
Examples	Gender, school, curriculum type, hair color	1st, 2nd, 3rd place, ranked preferences	Age, height, good test scores, good rating scales
Frequency Distribution	Redhead - III Blond - IIII Brunette - II	Best - II Better - III Good - III	5 - I 4 - II 3 - III 2 - III 1 - II
Frequency Polygon/ Histogram	No	Yes	Yes
Bar Graph or Chart	Yes	Yes	Yes
Central Tendency			
Mean	No	Mean Rank	Yes
Median	No	Yes	Yes
Mode	Yes	Yes	Yes
Variability			
Standard Deviation	No	of Ranks	Yes
Range	No	Yes, but[1]	Yes
How many categories	Yes	Yes	Yes
Percent in each	Yes	Yes	Yes
Shape			
Skewness	No	No	Yes
Kurtosis	No	No	Yes

[1] The range of ordinal data may well be misleading

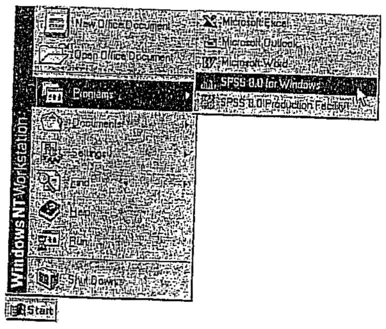
1. lekce

POVAHA HROMADNÝCH DAT A LOGIKA SURVEY. PRÁCE S HROMADNÝMI DATY PŘED JEJICH ANALÝZOU (Modul FILES: procedury), PRÁCE S PROSTŘEDÍM (Moduly Edit, View, Utilities) A VÝSTUPY Z ANALÝZY (Modul : Output).

Starting SPSS for Windows

The easiest way to run SPSS for Windows is by using the Start button. During the installation of SPSS, the Setup procedure adds SPSS to the menu that appears when you click the Start button, as shown in Figure 2.1.

Figure 2.1 SPSS on the Start menu

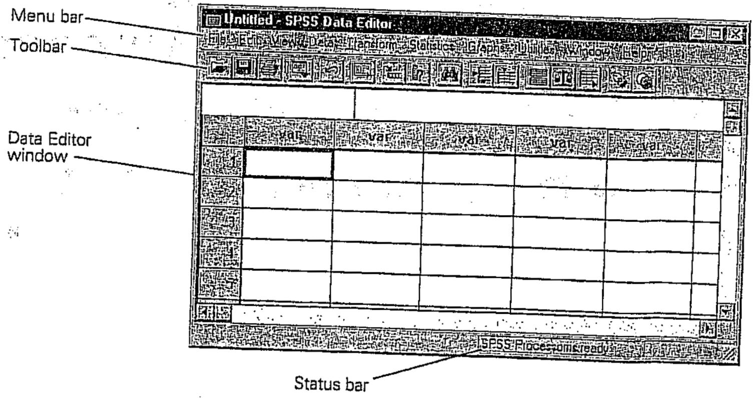


Always use the left mouse button unless the right one is specifically indicated.

▶ To start SPSS, click Start to display the Start menu, then click SPSS 8.0 for Windows.

The SPSS Data Editor window is displayed, as shown in Figure 2.2. You can move it, like any other window, by clicking and dragging its title bar, or resize it by clicking and dragging its sides or corners.

Figure 2.2 SPSS Data Editor window


Menu bar
Toolbar
Data Editor window
Status bar
Untitled - SPSS Data Editor
File Edit View Data Transform Statistics Graphs Utilities Window Help
var
var
var
var
var
SPSS Processor is ready

Opening a Data File

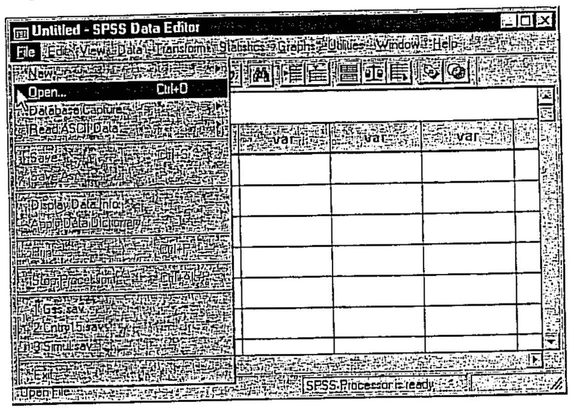
The SPSS Data Editor window displays your working data file. You don't have one yet—that's why the Data Editor is empty. If you have data of your own that are not in the computer yet, you can type the numbers right into the Data Editor. If the data are already in a spreadsheet or database file, you can probably read that file into SPSS. The data used in this book are already in the form of SPSS data files. To use them for the exercises, or just to follow along in the analysis, simply open the appropriate data file. To open a data file:

▶ Click the left mouse button on the word File on the SPSS Data Editor menu bar, as shown in Figure 2.3.

The File menu is displayed.

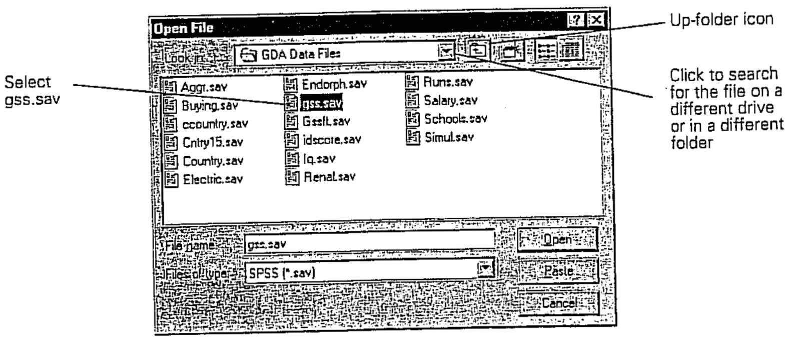
▶ On the File menu, click Open.

Figure 2.3 Opening a data file



When you click Open on the File menu, the Open File dialog box appears, as shown in Figure 2.4.

Figure 2.4 Open File dialog box



File name: gss.sav

Files of type: SPSS (*.sav)

▶ Click the *gss.sav* data file where it appears in the list.

▶ Click Open.

What if the gss.sav file doesn't appear? Only files in the current drive and directory are listed. The file you want may either be in another directory or saved on a different drive.

To look in a parent folder (one that contains the current folder), click the up-folder icon, as shown in Figure 2.4.

To look in a subfolder (one contained in the current folder), double-click it in the list.

To look on a different drive, click the up-folder icon repeatedly until you reach My Computer, then double-click the desired drive icon and continue down through the folder hierarchy on that drive. ■■■

When SPSS has finished reading the data file, it displays the data in the Data Editor, as shown in Figure 2.5. This particular data file contains selected information for 1500 people who were interviewed in the 1993 General Social Survey, which annually asks a broad range of questions to a sample of adults in the United States population.

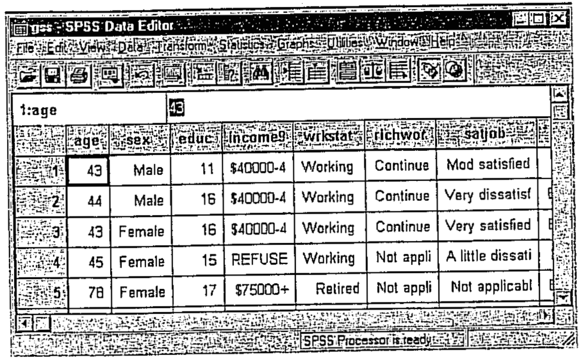
To view the data in the Data Editor, from the menus choose:

Window
gss - SPSS Data Editor

If your screen displays all numbers rather than value labels such as Male and Female in the cells, from the menus choose:

View
Value Labels

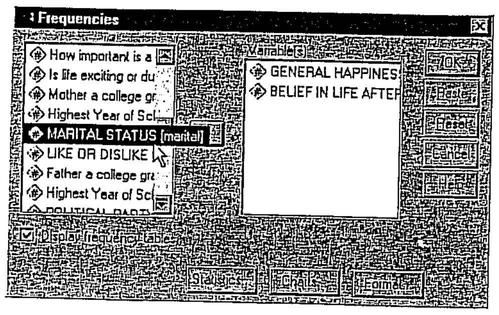
Figure 2.5 Data Editor window with GSS data


gss - SPSS Data Editor
File Edit View Data Transform Statistics Graphs Utilities Window Help
SPSS Processor is ready

1:age 43

	age	sex	educ	income9	wrkstat	richwor	satjob
1	43	Male	11	$40000-4	Working	Continue	Mod satisfied
2	44	Male	16	$40000-4	Working	Continue	Very dissatisf
3	43	Female	16	$40000-4	Working	Continue	Very satisfied
4	45	Female	15	REFUSE	Working	Not appli	A little dissati
5	78	Female	17	$75000+	Retired	Not appli	Not applicabl

Figure 2.7 Frequencies dialog box with default variable labels



To make this book easier to read, we'll use variable names instead of labels in dialog boxes, as shown in Figure 2.7. To display variable names rather than labels in your dialog boxes (so you can follow along with the text), you need to change one of SPSS's default options.

- From the menus select:

  Edit
  Options...

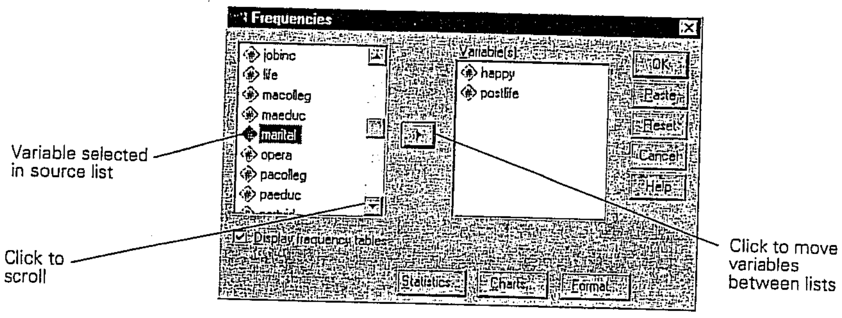
- In the Options dialog box, click the General tab.

- In the Variable Lists group box, click Display names.

- Click OK.

This change doesn't take effect until the next time you open a data file. The effect of the changed option is shown in Figure 2.8.

Figure 2.8 Frequencies dialog box



To use this dialog box:

- Click *happy* in the scroll list and then click ▶. This moves *happy* into the Variable(s) list.

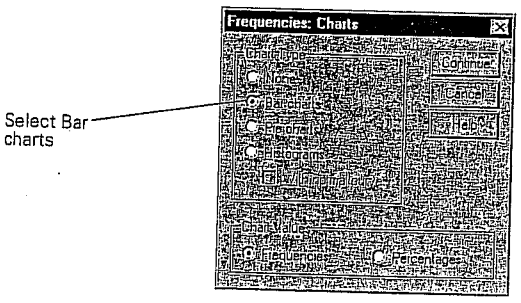
- Scroll down the source list until you see *postlife* and move it into the Variable(s) list as well.

- Click Charts.

As a shortcut to scroll the source list, click in the list and type the letter p. This scrolls to the first variable beginning with p.

This opens the Frequencies Charts dialog box, as shown in Figure 2.9. Here you can request charts along with your frequency tables.

Figure 2.9 Frequencies Charts dialog box

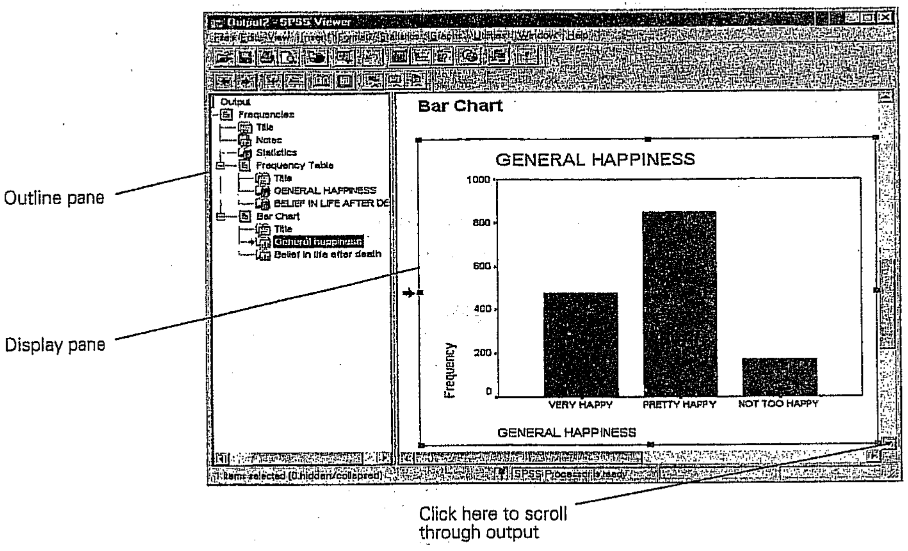


- Select Bar charts, as shown in Figure 2.9.

The Viewer Window

The Viewer window is where you see the statistics and graphics—the output—from your work in SPSS. As shown in Figure 2.10, the Viewer window is split into two parts, or **panes**. (A piece of a window is often called a pane in computer software, just as it is at your local hardware store.)

Figure 2.10 Viewer window



The left side (the **outline pane**) contains an outline view of all the different pieces of output in the Viewer, whether they are currently visible or not. The right side (the **display pane**) contains the output itself.

- To change the sizes of the two panes (for example, to make the display pane wider), just point the mouse at the line that divides them, press the left mouse button, and drag the line to the left or right.

It's possible to ignore the outline pane and simply scroll through the output displayed in the display pane on the right side of the Viewer. The outline view offers some handy tricks, however.

The Outline Pane

Individual portions of output are associated with "book" icons in the outline pane. Each icon represents a particular piece of output, such as a table of statistics or a chart.

- If you click one of these icons in the outline pane, the associated piece of output appears instantly in the display pane. (But it may be hidden! See below.)

These icons are the quickest navigational controls in the Viewer.

The book icons are also used to hide or display pieces of output temporarily. Notice that most of them in the outline pane are "open book" icons, while a few look more like closed books. A "closed book" icon represents a hidden piece of output. Hidden output doesn't appear in the display pane but can be recovered any time you want to look at it.

- To hide a single piece of output, double-click the open book icon. This closes the icon and hides the output associated with it.

- To display a hidden piece of output, double-click the closed book icon. This opens the icon and displays the output associated with it.

- To hide *all* of the output from a procedure such as Frequencies, click the little box containing a minus sign to the left of the procedure name. That whole part of the outline collapses, and the minus sign changes to a plus sign to show you that more output is hiding there. Click the plus sign to show it all again.

You will find that you can do lots of things in fairly obvious ways by playing with the outline pane. Try rearranging the output (press the left mouse button on a book icon, drag it to a different place in the outline, and then release the mouse button), or deleting part of the output (click the icon and press the Delete key). The SPSS Help system can tell you all the details.

The Display Pane

The display pane shows as much of the SPSS output as can fit in it. To see more, you can either scroll the pane or use the outline pane to jump around.

The output in the display pane includes several different kinds of objects: tables of numbers (actually a special kind of tables, called pivot tables); charts; and bits of text such as titles. You have complete control over the appearance, and even the content, of most of these objects.

- To change something about an object, double-click it in the display pane.

Double-clicking an object opens an editor that is specially designed to modify it. The appearance of the object changes to show that you are editing it. The menu bar may change. If the object is a chart, a special chart editing window opens to offer you a powerful set of tools for changing the chart's appearance.

Let's look at these objects in the Viewer.

Viewer Objects

In the outline panel, the first line is a container for the entire batch of output. It's simply called Output. There might be a line below it called Log, which isn't going to be discussed in this book. The next line, *Frequencies*, is a heading that contains all the various kinds of output produced by the Frequencies procedure that you just ran. In order, they are:

- Title. The title of the procedure, which is simply text.
- Notes. Notes are usually hidden, so this probably looks like a closed book in the outline pane.
- Statistics. This is a pivot table, which reports the number of cases, or "observations," that were processed by the Frequencies procedure. Most procedures start by producing such a table. The icon is an open book, so if you click it, the display pane will show you what it looks like.
- A frequency table for the first variable processed (*happy*). Frequency tables are discussed in Chapter 3. Note that the icon in the outline pane is labeled *GENERAL HAPPINESS*, which is a descriptive label that was assigned to the variable *happy* when the data file was set up.

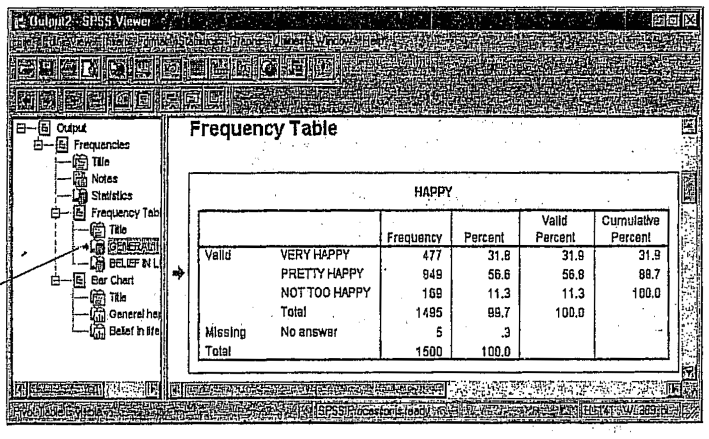
- A frequency table for the next variable, *postlife*, whose icon is labeled *BELIEF IN LIFE AFTER DEATH* in the outline pane.
- A bar chart for *happy*.
- A bar chart for *postlife*.

Let's see what these pivot tables and charts are like.

Pivot Tables

First, a pivot table. Most of SPSS's tabular and statistical output appears in the Viewer in the form of pivot tables.

Figure 2.11 Pivot tables in the Viewer



Frequency Table

HAPPY

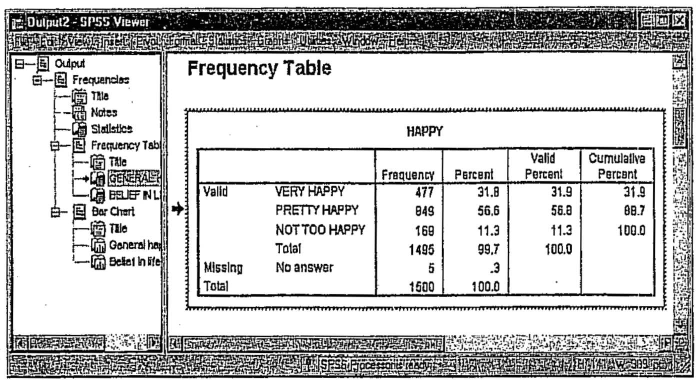
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	VERY HAPPY	477	31.8	31.9	31.9
	PRETTY HAPPY	849	56.6	56.8	88.7
	NOT TOO HAPPY	169	11.3	11.3	100.0
	Total	1495	99.7	100.0	
Missing	No answer	5	.3		
Total		1500	100.0		

Click here to display the frequencies table for GENERAL HAPPINESS

▶ In the outline, click the icon for the pivot table labeled *GENERAL HAPPINESS*. The table instantly appears in the display pane, with an arrow pointing to it, as shown in Figure 2.11.

▶ Move the mouse over to the display pane and double-click on the table itself to indicate that you want to edit it.

Figure 2.12 An activated pivot table



Frequency Table

HAPPY

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	VERY HAPPY	477	31.8	31.9	31.9
	PRETTY HAPPY	849	56.6	56.8	88.7
	NOT TOO HAPPY	169	11.3	11.3	100.0
	Total	1495	99.7	100.0	
Missing	No answer	5	.3		
Total		1500	100.0		

Not a lot seems to happen in Figure 2.12. The pivot table is now surrounded by a cross-hatched line to indicate that it is active in the Pivot Table Editor. The SPSS toolbar vanishes, and if you watch carefully, the menu bar changes—there is now a Pivot menu.

- Double-clicking a pivot table lets you edit it "in place"; that is, right where it sits in the display pane of the Viewer. If you need more room, select the pivot table by clicking once with the left mouse button, and from the menus choose:

```
Edit
  SPSS Pivot Table Object ▸
    Open...
```

This command opens the pivot table into a window of its own.

When you are editing a pivot table either way, you can change almost anything about it you want. If you don't like the label, just double-click it. It reappears as highlighted text. Type in the label the way you want it, perhaps *Happiness in General*, and click somewhere else to enter the new label. To change the font of the title or make it bold or italic, click the title and from the menus choose:

```
Format
  Font...
```

Then choose a different font or a bold or italic style.

If you don't like the way numbers are displayed in the pivot table, make sure the Pivot Table Editor is active (by double-clicking the table in the display pane), and then either make a selection from the Format menu, or *right-click* one of the numbers in the table to pop up a context menu for it. Most of the things you might want to change can be found in either of these menus under Table Properties or Cell Properties. Check out Table Looks, too, to see how you can apply consistent sets of formatting to whole tables.

Changing fonts and styles and even the text of the labels in a table can make a big difference in the way the table looks. An SPSS pivot table lets you do much more than that, however. You can change the basic organization of the data presented in the table. The Pivot menu (which appears only when you have double-clicked a pivot table in the display frame to activate it) gives you access to powerful tools for reorganizing the table. To get a feel for these tools, activate a pivot table, and from the menus choose:

```
Pivot
  Transpose Rows and Columns
```

The same information is displayed. The different codes or responses to the question, which were laid out vertically, are now laid out horizontally; the different types of statistical summaries, which were laid out horizontally, are now laid out vertically.

To see the pivot table as it was before, simply transpose the rows and columns again.

This example is a very simple pivot table. Multidimensional tables offer many more structural possibilities. You can explore those in the SPSS online Help system, or if you like, to see how things work you can build a complex table and start pivoting.

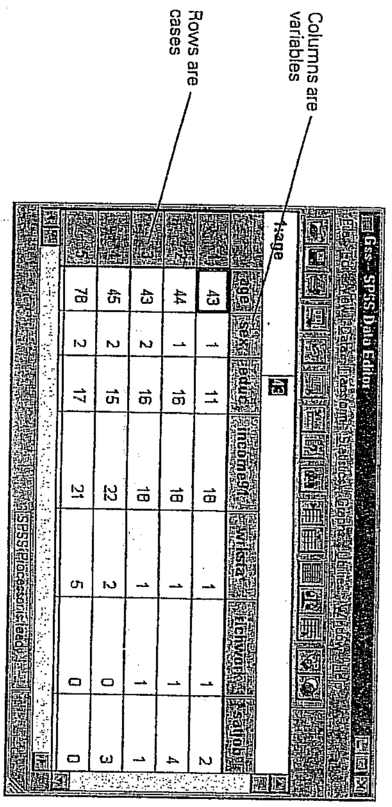
NORUŠIS, Marija J. 1998. *Guide to Data Analysis*. Upper Saddle River, Prentice Hall.

The Data Editor Window

Let's take a closer look at the Data Editor (see Figure 2.14). You can select it from the Window menu or simply click on it if any part of it is visible on your screen.

If you've ever used a spreadsheet, the Data Editor should look familiar. It's just an array of rows and columns. In the Data Editor, each row is a case, and each column is a variable. Cases and variables are fundamental concepts in data analysis. It's time we stopped to define them.

Figure 2.14 Data Editor window



	age	sex	educ	income91	wrkstat	richwork	satjob
1	43	1	11	18	1	1	2
2	44	1	16	18	1	1	4
3	43	2	16	18	1	1	1
4	45	2	15	22	2	0	3
5	78	2	17	21	5	0	0

Cases (rows) are the people who participate in a survey or experiment. (Another word often used is **observation**.) Actually, a case need not be a person. It can be anything. If you're doing experiments on rats, the case is the individual rat. If you're studying the beef content of hamburgers, each hamburger is a case. Generally speaking, the case is the unit for which you take measurements.

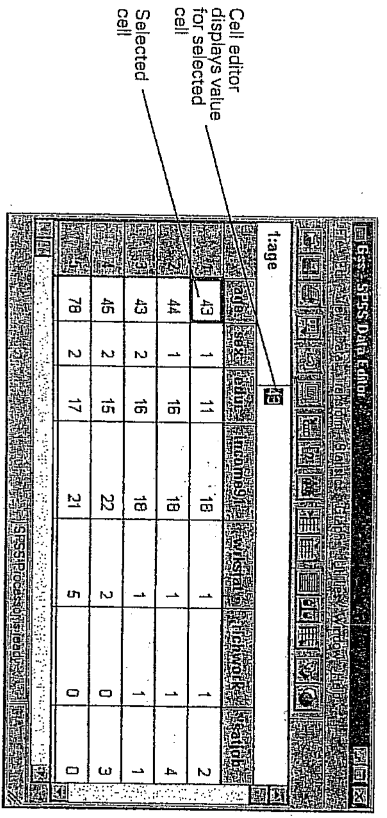
Variables (columns) are the different items of information you collect for your cases. Think about the way you conduct a survey. You ask each person for the same type of information: date of birth, sex, marital status, education, views on whatever subjects your survey is about. Each item for which you record an answer is known as a variable. The answer a particular person gives is known as the **value** for that variable. Year of birth is a variable; responses such as 1952 or 1899 are values for that variable.

The intersection of the row and the column is called a **cell**. Each cell holds the value of a particular variable for a particular case. You can edit values in the Data Editor, as follows:

- Click in one of the cells with the mouse.

The cell editor displays the value for the selected cell, as shown in Figure 2.15.

Figure 2.15 Data Editor with cell selected


Cell editor displays value for selected cell
Selected cell
Gss - SPSS Data Editor
File Edit View Data Transform Statistics Graphs Utilities Window Help
1:age
43
SPSS Processor is ready

	age	sex	educ	income91	wrkstat	richwork	satjob
1	43	1	11	18	1	1	2
2	44	1	16	18	1	1	4
3	43	2	16	18	1	1	1
4	45	2	15	22	2	0	3
5	78	2	17	21	5	0	0

- Type a number to replace the existing value and press ↵Enter.

The new value appears in the cell editor as you type it, but the value in the cell is not updated until you press ↵Enter.

- Change another value in the cell editor, but instead of pressing ↵Enter, press Esc.

When you press Esc rather than ↵Enter, the original value in the cell remains unchanged.

MORGAN George A. and Griego V. ORLANDO. 1998. *Easy Use and Interpretation of SPSS for Windows: Answering Research Questions With Statistics*. Mahwah: Lawrence Erlbaum Associates.

CHAPTER 2

Overview of the High School and Beyond (HSB) Data Set and SPSS 7.5

The Modified Hsbdata File

The file name of the data set used with this manual is hsbdata; it stands for high school and beyond data. It is based on a national sample of data from more than 28,000 high school students. The current data set is a sample of 75 students drawn randomly from the larger population. The data that we have from this sample includes school outcomes such as grades and the number of mathematics courses of different types that the students took in high school. Also there are several kinds of standardized test data and demographic data such as gender and mother's and father's education. To provide an example of questionnaire type data, we have included 13 questions about math attitudes. These data were developed for this manual and, thus, are not really the math attitudes of the 75 students in this sample. The questions, however, are based on ones used by the authors to study mastery motivation. Also we made up ethnic group data which, although somewhat realistic overall, do not represent the actual ethnic groups of the 75 students in this sample. This enables us to do some additional analyses.

We have provided you with a disk which contains the data for each of the 75 participants on 28 variables. The hsbdata file, shown in Table 2.1, has already been entered and labeled to enable you to get started on analyses quickly. In Assignments A and M, you will enter some additional data to practice entering it yourself. Also you will, in several assignments, label variables and their values so that your printouts will include the new variable names and the value labels.

The Raw HSB Data and Data Editor

Notice the short variable names at the top of the hsbdata file. (Actually we have transferred the HSB file from the SPSS data editor to Excel and reduced it so that it would fit on two pages, but in SPSS it will look very similar to Table 2.1.) Be aware that the subjects/participants are listed down the page from ID 1 to ID 75 at the bottom of the second page, and the variables are listed across the top. You will always enter data this way. If a variable is measured more than once, such as a pretest and posttest, it will be entered as two variables perhaps called Pre and Post. This method of entering data follows that suggested in chapter 7. Note that most of the values are single digits but that *visual, mosaic,* and *mathach* include some decimals and even minus numbers. Notice also that some cells like variable Q09 for participant ID 1 are blank because a datum is missing. Perhaps participant 1 did not answer question 9 and participant 2 did not answer question 4, etc. Blank is the "system missing" value that can be used for any missing data in an SPSS data file. However, other values also can be used for missing data. Notice that for father's and mother's education level we have used -1 for the missing values, and for ethnic group we have defined 9 as missing. For your purposes, however, we suggest that you leave missing data blank, but you may run across "user defined" missing data codes like -1 or 9 in other researchers' data.

Table 2.1. *Hsbdata Data Set in the SPSS Data Editor*

ID	VISUAL	MOSAIC	GRADES	MATHGR	ALG1	ALG2	GEO	TRIG	CALC	MATHACH	FAED	MAED	GEND	Q01	Q02	Q03	Q04	Q05	Q06	Q07	Q08	Q09	Q10	Q11	Q12	Q13	ETHNIC
1	8.75	31.0	4	0	0	0	0	0	0	9.00	10	10	2	10	4	3	1	4	2	2	1	.	2	1	3	3	2
2	4.75	56.0	5	0	0	0	0	0	0	10.33	2	2	2	3	40	3	.	1	2	3	3	3	1	1	4	2	3
3	4.75	25.0	6	1	0	0	0	0	0	7.67	2	2	2	1	1	1	4	4	4	1	2	4	1	1	3	2	2
4	1.00	22.0	3	0	1	0	0	0	0	5.00	3	3	1	1	1	1	2	4	3	1	3	4	3	4	4	3	1
5	2.25	17.5	3	0	0	0	0	0	0	-1.67	-1	3	2	1	4	1	4	4	4	1	4	1	1	4	2	2	2
6	1.00	23.5	5	1	0	0	0	0	0	1.00	3	2	2	1	1	1	4	4	4	1	4	1	1	4	1	2	3
7	2.50	28.5	6	0	1	1	1	0	0	12.00	9	6	1	.	1	3	1	1	3	3	1	3	1	1	3	3	1
8	3.50	29.5	4	0	1	0	0	0	0	8.00	5	3	2	1	1	3	2	4	2	1	2	3	2	1	3	3	1
9	3.50	28.0	7	1	1	0	0	0	0	13.00	3	3	2	3	1	.	1	1	2	2	2	4	2	1	3	2	1
10	3.75	27.5	5	0	0	0	0	0	0	3.67	8	2	1	1	3	1	4	4	4	1	3	3	2	4	3	4	2
11	11.00	27.0	6	0	1	1	1	1	1	21.00	3	4	1	4	4	4	3	1	4	4	1	4	4	2	4	4	4
12	4.75	26.5	8	1	1	1	1	1	1	23.67	8	9	1	3	4	4	2	1	2	3	2	4	2	1	3	3	1
13	1.00	13.0	5	0	1	0	0	0	0	4.00	2	2	2	1	4	1	3	4	2	1	1	4	1	1	4	3	3
14	1.00	18.0	2	0	1	0	0	0	0	9.00	6	3	1	3	4	3	2	1	1	4	1	4	1	1	4	3	2
15	4.75	25.0	3	0	1	1	0	0	0	5.33	2	3	2	4	4	3	2	1	3	3	1	4	1	1	3	3	1
16	13.50	33.0	7	1	1	1	1	0	0	19.67	-1	3	1	3	4	3	2	1	3	4	2	4	1	2	4	3	2
17	-.25	33.0	7	0	1	0	1	0	0	7.67	3	3	1	4	4	3	2	1	3	3	1	4	1	1	4	3	1
18	6.00	27.0	5	0	1	0	1	0	0	14.33	9	7	1	3	4	3	2	1	1	4	2	3	1	1	3	3	2
19	9.75	26.0	6	1	1	1	1	0	0	14.33	3	3	1	3	4	3	1	2	2	4	1	4	1	1	4	2	3
20	8.75	25.0	8	1	1	1	1	1	0	19.67	8	3	2	3	2	1	2	2	4	3	2	3	3	2	4	3	4
21	-.25	41.0	5	0	1	0	1	0	0	17.00	5	5	1	4	4	4	1	1	3	3	1	3	1	1	4	4	1
22	7.25	27.0	3	0	0	0	0	0	0	7.67	2	3	2	2	4	3	1	2	1	4	2	1	1	1	3	2	1
23	4.75	22.0	5	0	1	1	1	1	0	19.67	9	7	1	4	4	3	1	1	3	3	2	3	2	1	4	3	1
24	1.00	53.0	4	0	0	0	0	0	0	6.67	2	2	2	3	3	2	2	2	1	1	2	2	2	2	3	3	1
25	9.75	27.0	7	1	1	1	1	1	0	14.33	6	6	2	3	4	4	2	1	3	3	2	3	1	1	3	2	2
26	7.25	31.0	5	0	1	0	0	0	0	10.67	5	3	2	3	4	3	3	2	3	3	2	3	1	2	2	3	3
27	2.25	29.0	6	0	1	0	1	0	0	14.33	3	6	2	4	4	3	3	1	1	2	3	3	1	1	3	3	4
28	2.25	26.0	5	1	1	1	0	0	0	14.33	3	2	1	3	4	4	1	1	1	3	1	4	1	1	4	3	1
29	1.00	22.0	6	1	1	0	0	0	0	14.33	5	3	2	4	4	3	2	1	2	3	2	3	1	2	3	3	1
30	13.50	31.0	3	0	0	0	0	0	0	9.00	2	3	1	2	3	2	3	1	3	4	2	3	1	1	3	3	3
31	9.75	27.0	7	1	1	1	1	1	0	17.00	2	3	2	2	4	4	2	1	1	3	2	4	1	1	3	2	2
32	9.50	34.0	6	0	1	1	1	0	0	23.67	9	5	1	3	4	2	2	2	3	2	2	4	1	1	4	3	1
33	9.75	24.5	4	0	1	1	1	1	0	18.33	5	7	1	4	4	4	1	1	2	4	1	4	1	1	3	4	4
34	6.00	23.0	7	1	1	1	1	1	0	15.67	8	8	1	3	4	3	2	2	4	2	2	3	1	1	3	3	3
35	2.25	32.5	8	1	1	0	0	0	0	17.00	4	2	2	3	4	3	3	2	2	2	2	3	2	2	3	2	1
36	9.75	29.5	6	0	1	1	1	1	0	15.67	8	6	1	3	3	3	3	2	3	2	2	3	2	2	1	1	1
37	-.25	33.0	7	0	1	0	1	0	0	7.67	3	3	1	3	4	3	1	1	2	4	1	4	1	1	4	3	1
38	3.75	24.0	7	1	1	1	1	1	1	22.67	2	2	2	3	4	4	1	1	2	4	4	4	1	1	4	4	1
39	6.00	11.0	7	1	1	1	0	0	0	14.33	4	2	2	4	4	3	3	1	3	2	4	4	1	1	4	4	2
40	6.50	23.0	5	0	1	1	1	0	0	21.00	2	2	1	2	4	2	2	3	1	3	3	4	1	1	2	2	4
41	1.00	26.0	6	1	0	0	0	0	0	6.33	2	2	2	2	4	3	2	1	4	3	3	4	1	1	2	2	1
42	6.00	22.5	6	1	1	0	1	0	0	3.67	3	6	2	4	2	4	1	1	3	4	1	2	2	1	3	4	1
43	3.50	20.5	7	0	1	1	0	0	0	10.33	10	10	1	3	4	3	4	1	3	3	3	4	1	2	3	3	2
44	3.50	28.0	8	1	1	1	1	1	1	23.67	8	8	2	3	4	3	2	1	1	4	1	4	1	1	4	3	1

MORGAN George A. and Griego V. ORLANDO. 1998. *Easy Use and Interpretation of SPSS for Windows: Answering Research Questions With Statistics*. Mahwah: Lawrence Erlbaum Associates.

Table 2.1. ***Hsbdata Data Set in the SPSS Data Editor*** **(continued)**

ID	VISUAL	MOSAIC	GRADES	MATHGR	ALG1	ALG2	GEO	TRIG	CALC	MATHACH	FAED	MAED	GEND	Q01	Q02	Q03	Q04	Q05	Q06	Q07	Q08	Q09	Q10	Q11	Q12	Q13	ETHNIC
45	1.00	16.0	7	1	1	0	0	0	0	4.00	2	3	2	4	4	3	3	1	2	2	3	3	1	1	3	2	2
46	-.25	30.0	6	0	1	1	0	0	.	17.00	4	6	2	3	4	3	1	1	2	4	1	3	1	1	3	3	1
47	-.25	23.5	6	0	1	1	1	0	0	6.33	10	9	2	2	4	2	4	1	3	1	3	3	1	1	2	2	1
48	-.25	20.5	5	0	1	0	0	0	0	3.67	2	3	2	3	4	2	3	2	1	4	1	4	1	1	2	3	1
49	4.75	22.5	5	0	1	0	0	0	0	9.00	3	5	2	4	3	4	3	1	2	4	2	4	2	1	2	2	1
50	5.00	24.0	6	0	1	1	1	1	0	20.33	8	7	2	3	4	3	3	1	3	3	2	4	1	1	3	3	1
51	4.75	37.0	6	0	1	0	0	0	0	13.00	2	2	2	4	4	4	2	1	3	4	1	3	1	2	3	3	4
52	4.75	31.0	5	1	1	0	0	0	0	12.00	2	3	2	2	4	4	3	1	3	1	3	4	1	1	2	2	9
53	5.00	56.0	6	1	0	0	0	0	0	17.00	6	2	2	3	3	3	2	1	2	2	2	3	2	1	3	2	2
54	11.00	35.5	8	1	1	1	1	1	1	23.67	5	7	2	3	4	2	2	2	2	3	2	3	1	2	2	3	3
55	11.00	28.0	8	1	1	1	1	1	0	21.00	2	2	1	3	3	3	2	1	2	3	2	3	1	1	3	3	1
56	7.25	36.0	7	1	1	1	1	1	1	22.33	10	9	1	3	4	3	1	1	1	4	1	3	1	1	3	2	1
57	3.50	35.0	3	0	0	0	0	0	0	2.33	2	3	2	4	4	3	1	1	2	4	1	4	1	1	2	3	2
58	7.25	44.0	8	1	1	1	1	1	0	22.33	10	6	2	4	4	3	3	2	1	2	3	2	4	1	1	1	3
59	3.50	-4.0	7	1	1	0	0	0	0	5.00	2	2	2	2	3	2	2	2	3	1	3	3	1	1	1	1	1
60	6.00	24.5	5	1	0	0	0	0	0	5.00	2	3	2	3	4	2	2	2	1	3	1	3	1	1	3	3	2
61	9.75	25.0	8	1	1	1	1	0	0	10.33	2	3	2	3	3	3	2	2	3	3	2	3	2	2	3	2	3
62	2.25	51.5	3	0	0	0	0	0	0	5.00	3	3	1	4	4	3	4	1	1	1	4	4	1	1	2	1	1
63	1.00	25.0	7	0	1	0	0	0	0	1.00	10	3	2	4	4	4	1	1	2	4	1	4	1	1	4	4	1
64	3.50	30.5	4	0	1	0	1	0	0	13.00	2	3	1	3	4	2	2	2	3	2	2	3	1	1	3	3	4
65	4.75	29.0	4	0	0	0	0	0	0	9.33	3	2	2	3	4	3	2	1	1	3	2	4	1	1	4	3	1
66	-.25	20.0	7	0	1	1	0	0	0	10.33	2	3	2	4	3	4	2	1	3	3	1	4	3	1	3	3	8
67	4.75	22.0	7	1	1	1	1	1	1	23.67	9	3	1	4	4	3	2	1	3	3	1	3	1	1	3	3	1
68	6.00	7.5	3	0	0	0	0	0	0	5.00	5	5	1	3	4	2	1	1	1	3	1	3	1	1	2	2	1
69	7.25	32.0	8	1	1	1	1	1	1	23.67	9	6	1	4	4	4	3	1	3	3	1	4	1	1	4	2	1
70	1.00	23.0	5	0	1	1	0	0	0	14.67	3	3	1	4	4	3	1	1	2	4	1	4	1	1	4	4	1
71	11.00	24.5	4	0	1	1	1	1	0	18.67	7	3	1	3	4	1	2	3	4	4	1	2	1	2	3	1	1
72	8.50	23.0	7	0	1	1	1	0	0	11.67	5	4	1	4	4	4	3	1	4	1	3	3	3	1	2	1	1
73	8.50	30.5	4	0	1	0	0	0	0	14.33	3	6	1	2	2	2	2	2	3	3	2	3	3	2	3	3	1
74	14.75	13.5	5	0	1	0	1	0	0	11.67	3	3	1	3	3	2	3	2	3	2	3	3	2	2	3	2	1
75	14.75	30.0	7	1	1	1	1	1	0	14.33	6	3	1	3	3	3	1	1	3	3	1	4	2	1	3	3	1

FIELD Andy. 2000. *Discovering Statistics Using SPSS for Windows. Advanced Techniques for the Beginner*. London: Sage.

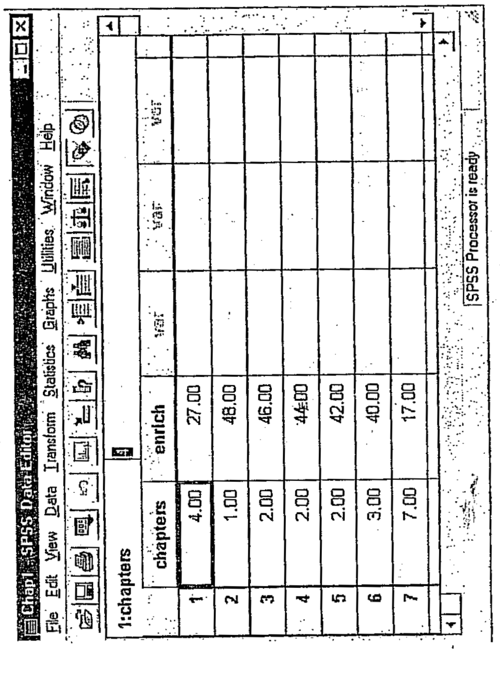
1.2. The SPSS Environment

There are several excellent texts that give introductions to the general environment within which SPSS operates. The best ones include Kinnear and Gray (1997) and Foster (1998). These texts are well worth reading if you are unfamiliar with Windows and SPSS generally because I am assuming at least some knowledge of the system. However, I appreciate the limited funds of most students and so to make this text usable for those inexperienced with SPSS I will provide a brief guide to the SPSS environment—but for a more detailed account see the previously cited texts and the SPSS manuals. This book is based primarily on version 9.0 of SPSS (at least in terms of the diagrams); however, it also caters for versions 7.0, 7.5 and 8.0 (there are few differences between versions 7.0, 8.0 and 9.0 and any obvious differences are highlighted where relevant).

Once SPSS has been activated, the program will automatically load two windows: the data editor (this is where you input your data and carry out statistical functions) and the output window (this is where the results of any analysis will appear). There are a number of additional windows that can be activated. In versions of SPSS earlier than version 7.0, graphs appear in a separate window known as the *chart carousel*; however, versions 7.0 and after include graphs in the output window, which is called the *output navigator* (version 7.0) and the *output viewer* (version 8.0 and after). Another window that is useful is the syntax window, which allows you to enter SPSS commands manually (rather than using the window-based menus). At most levels of expertise, the syntax window is redundant because you can carry out most analyses by clicking merrily with your mouse. However, there are various additional functions that can be accessed using syntax and sick individuals who enjoy statistics can find numerous uses for it! I will pretty much ignore syntax windows because those of you who want to know about them will learn by playing around and the rest of you will be put off by their inclusion (interested readers should refer to Foster, 1998, Chapter 8).

1.2.1. The Data Editor

The main SPSS window includes a data editor for entering data. This window is where most of the action happens. At the top of this screen is a menu bar similar to the ones you might have seen in other programs (such as Microsoft Word). Figure 1.6 shows this menu bar and the data editor. There are several menus at the top of the screen (e.g. *File*, *Edit* etc.) that can be activated by using the computer mouse to move the on-screen arrow onto the desired menu and then pressing the left mouse button once (pressing this button is usually known as *clicking*). When you have clicked on a menu, a menu box will appear that displays a list of options that can be activated by moving the on-screen arrow so that it is pointing at the desired option and then clicking with the mouse. Often, selecting an option from a menu makes a window appear; these windows are referred to as *dialog boxes*. When referring to selecting options in a menu I will notate the action using bold type with arrows indicating the path of the mouse (so, each arrow represents placing the on-screen arrow over a word and clicking the mouse's left button). So, for example, if I were to say that you should select the *Save As* ... option in the *File* menu, I would write this as select **File⇒Save As ...**.



1:chapters 4

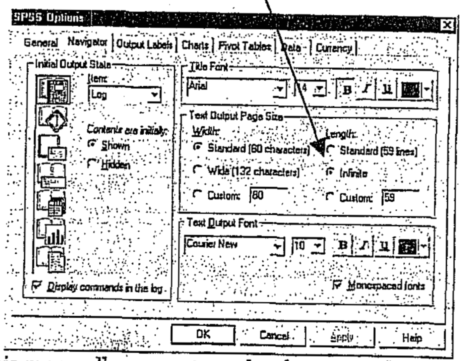
	chapters	enrich	var	var	var
1	4.00	27.00			
2	1.00	48.00			
3	2.00	46.00			
4	2.00	44.00			
5	2.00	42.00			
6	3.00	40.00			
7	7.00	17.00			

Figure 1.6: The SPSS data editor

Within these menus you will notice that some letters are underlined: these underlined letters represent the *keyboard shortcut* for accessing that function. It is possible to select many functions without using the mouse, and the experienced keyboard user may find these shortcuts faster than manoeuvring the mouse arrow to the appropriate place on the screen. The letters underlined in the menus indicate that the option can be obtained by simultaneously pressing ALT on the keyboard and the underlined letter. So, to access the *Save As*... option, using only the keyboard, you should press ALT and F on the keyboard simultaneously (which activates the *File* menu) then, keeping your finger on the ALT key, press A (which is the underlined letter).

Below is a brief reference guide to each of the menus and some of the options that they contain. This is merely a summary and we will discover the wonders of each menu as we progress through the book.

- **File**: This menu allows you to do general things such as saving data, graphs, or output. Likewise, you can open previously saved files and print graphs, data or output. In essence, it contains all of the options that are customarily found in *File* menus.
- **Edit**: This menu contains edit functions for the data editor. In SPSS for Windows it is possible to *cut* and *paste* blocks of numbers from one part of the data editor to another (which can be very handy when you realize that you've entered lots of numbers in the wrong place). You can also use the *Options* to select various preferences such as the font that is used for the output. The default preferences are fine for most purposes, the only thing you might want to change (for the sake of the environment) is to set the text output page size length of the viewer to infinite (this saves hundreds of trees when you come to print things).



Item: Log

Contents are initially: ● Shown ○ Hidden

☑ Display commands in the log

Title Font: Arial 14

Width: ● Standard (80 characters) ○ Wide (132 characters) ○ Custom: 80

Length: ○ Standard (59 lines) ● Infinite ○ Custom: 59

Text Output Font: Courier New 10

☑ Monospaced fonts

- **Data**: This menu allows you to make changes to the data editor. The important features are *insert variable*, which is used to insert a new variable into the data editor (i.e. add a column); *insert case*, which is used to add a new row of data between two existing rows of data; *split file*, which is used to split the file by a grouping variable (see section 2.4.1); and *select cases*, which is used to run analyses on only a selected sample of cases.
- **Transform**: You should use this menu if you want to manipulate one of your variables in some way. For example, you can use *recode* to change the values of certain variables (e.g. if you wanted to adopt a slightly different coding scheme for some reason). The *compute* function is also useful for transforming data (e.g. you can create a new variable that is the average of two existing variables). This function allows you to carry out any number of calculations on your variables (see section 6.2.2.1).

- **Analyze**: This menu is called **Statistics** in version 8.0 and earlier. The fun begins here, because the statistical procedures lurk in this menu. Below is a brief guide to the options in the statistics menu that will be used during the course of this book (this is only a small portion of what is available):
  - (a) **Descriptive Statistics**: This menu is called **Summarize** in version 8.0 and earlier. This menu is for conducting descriptive statistics (mean, mode, median etc.), frequencies and general data exploration. There is also a command called *crosstabs* that is useful for exploring frequency data and performing tests such as chi-square, Fisher's exact test and Cohen's kappa.
  - (b) **Compare Means**: This is where you can find *t*-tests (related and unrelated—Chapter 6) and one-way independent ANOVA (Chapter 7).
  - (c) **General Linear Model**: This is called *ANOVA Models* in version 6 of SPSS. This menu is for complex ANOVA such as two-way (unrelated, related or mixed), one-way ANOVA with repeated measures and multivariate analysis of variance (MANOVA).
  - (d) **Correlate**: It doesn't take a genius to work out that this is where the correlation techniques are kept! You can do bivariate correlations such as Pearson's *R*, Spearman's rho (ρ) and Kendall's tau (τ) as well as partial correlations (see Chapter 3).
  - (e) **Regression**: There are a variety of regression techniques available in SPSS. You can do simple linear regression, multiple linear regression (Chapter 4) and more advanced techniques such as logistic regression (Chapter 5).
  - (f) **Data Reduction**: You find factor analysis here (Chapter 11).
  - (g) **Nonparametric**: There are a variety of non-parametric statistics available such the chi-square goodness-of-fit statistic, the binomial test, the Mann-Whitney test, the Kruskal-Wallis test, Wilcoxon's test and Friedman's ANOVA (Chapter 2).
- **Graphs**: SPSS comes with its own, fairly versatile, graphing package. The types of graphs you can do include: bar charts, histograms, scatterplots, box-whisker plots, pie charts and error bar graphs to name but a few. There is also the facility to edit any graphs to make them look snazzy—which is pretty smart if you ask me.
- **View**: This menu deals with system specifications such as whether you have grid lines on the data editor, or whether you display value labels (exactly what value labels are will become clear later).
- **Window**: This allows you to switch from window to window. So, if you're looking at the output and you wish to switch back to your data sheet, you can do so using this menu. There are icons to shortcut most of the options in this menu so it isn't particularly useful.

- **Help**: This is an invaluable menu because it offers you on-line help on both the system itself and the statistical tests. Although the statistics help files are fairly useless at times (after all, the program is not supposed to teach you statistics) and certainly no substitute for acquiring a good knowledge of your own, they can sometimes get you out of a sticky situation.

As well as the menus there are also a set of *icons* at the top of the data editor window (see Figure 1.6) that are shortcuts to specific, frequently used, facilities. All of these facilities can be accessed via the menu system but using the icons will save you time. Below is a brief list of these icons and their function:

This icon gives you the option to open a previously saved file (if you are in the data editor SPSS assumes you want to open a data file, if you are in the output viewer, it will offer to open a viewer file).

This icon allows you to save files. It will save the file you are currently working on (be it data or output). If the file hasn't already been saved it will produce the *save data as* dialog box.

This icon activates a dialog box for printing whatever you are currently working on (either the data editor or the output). The exact print options will depend on the printer you use. One useful tip when printing from the output window is to highlight the text that you want to print (by holding the mouse button down and dragging the arrow over the text of interest). In version 7.0 onwards, you can also select parts of the output by clicking on branches in the viewer window (see section 1.2.4). When the *print* dialog box appears remember to click on the option to print only the selected text. Selecting parts of the output will save a lot of trees because by default SPSS will print everything in the output window.

Clicking this icon will activate a list of the last 12 dialog boxes that were used. From this list you can select any box from the list and it will appear on the screen. This icon makes it easy for you to repeat parts of an analysis.

This icon allows you to go directly to a case (i.e. a subject). This is useful if you are working on large data files. For example, if you were analysing a survey with 3000 respondents it would get pretty tedious scrolling down the data sheet to find a particular subject's responses. This icon can be used to skip directly to a case (e.g. case 2407). Clicking on this icon activates a dialog box that requires you to type in the case number required.

Clicking on this icon will give you information about a specified variable in the data editor (a dialog box allows you to choose which variable you want summary information about).

This icon allows you to search for words or numbers in your data file and output window.

Clicking on this icon inserts a new case in the data editor (so, it creates a blank row at the point that is currently highlighted in the data editor). This function is very useful if you need to add new data or if you forget to put a particular subject's data in the data editor.

Clicking this icon creates a new variable to the left of the variable that is currently active (to activate a variable simply click once on the name at the top of the column).

Clicking on this icon is a shortcut to the **Data⇒Split File ...** function (see section 2.4.1). Social scientists often conduct experiments on different groups of people. In SPSS we differentiate groups of people by using a coding variable (see section 1.2.3.1), and this function lets us divide our output by such a variable. For example, we might test males and females on their statistical ability. We can code each subject with a number that represents their gender (e.g. 1 = female, 0 = male). If we then want to know the mean statistical ability of each gender we simply ask the computer to split the file by the variable **gender**. Any subsequent analyses will be performed on the men and women separately.

This icon shortcuts to the **Data⇒Weight Cases ...** function. This function is necessary when we come to input frequency data (see section 2.8.2) and is useful for some advanced issues in survey sampling.

This icon is a shortcut to the **Data⇒Select Cases ...** function. If you want to analyze only a portion of your data, this is the option for you! This function allows you to specify what cases you want to include in the analysis.

Clicking this icon will either display, or hide, the value labels of any coding variables. We often group people together and use a coding variable to let the computer know that a certain subject belongs to a certain group. For example, if we coded gender as 1 = female, 0 = male then the computer knows that every time it comes across the value 1 in the **gender** column, that subject is a female. If you press this icon, the coding will appear on the data editor rather than the numerical values; so, you will see the words *male* and *female* in the **gender** column rather than a series of numbers. This idea will become clear in section 1.2.3.1.

1.2.2. Inputting Data

When you first load SPSS it will provide a blank data editor with the title *New Data*. When inputting a new set of data, you must input your data in a logical way. The SPSS data editor is arranged such that *each row represents data from one subject while each column represents a variable*. There is no discrimination between independent and dependent variables: both types should be placed in a separate column. The key point is that each row represents one participant's data. Therefore, any information about that case should be entered across the data editor. For example, imagine you were interested in sex differences in perceptions of pain created by hot and cold stimuli. You could place some people's hands in a bucket of very cold water for a minute and ask them to rate how painful they thought the experience was on a scale of 1 to 10. You could then ask them to hold a hot potato and again measure their perception of pain. Imagine I was a subject. You would have a single row representing my data, so there would be a different column for my name, my age, my gender, my pain perception for cold water, and my pain perception for a hot potato: Andy, 25, male, 7, 10. The column with the information about my gender is a grouping variable: I can belong to either the group of males or the group of females, but not both. As such, this variable is a between-group variable (different people belong to different groups). Therefore, between-group variables are represented by a single column in which the group to which the person belonged is defined using a number (see section 1.2.3.1). Variables that specify to which of several groups a person belongs can be used to split up data files (so, in the pain example you could run an analysis on the male and female subjects separately—see section 2.4.1). The two measures of pain are a repeated measure (all subjects were subjected to hot and cold stimuli). Therefore, levels of this variable can be entered in separate columns (one for pain to a hot stimulus and one for pain to a cold stimulus).

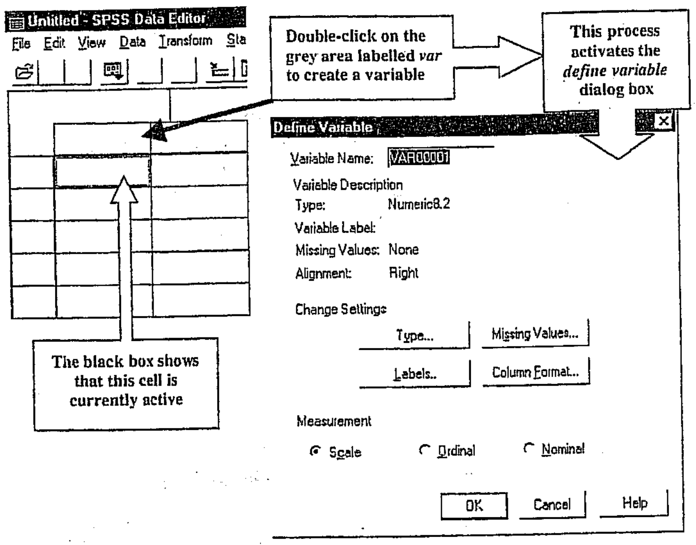
In summary, any variable measured with the same subjects (a repeated measure) should be represented by several columns (each column representing one level of the repeated measures variable). However, when a between-group design was used (e.g. different subjects were assigned to each level of the independent variable) the data will be represented by two columns: one that has the values of the dependent variable and one that is a coding variable indicating to which group the subject belonged. This idea will become clearer as you learn about how to carry out specific procedures.

The data editor is made up of lots of *cells*, which are just boxes in which data values can be placed. When a cell is active it becomes highlighted with a black surrounding box (as in Figure 1.7). You can move around the data editor, from cell to cell, using the arrow keys ← ↑ ↓ → (found on the right of the keyboard) or by clicking the mouse on the cell that you wish to activate. To enter a number into the data editor simply move to the cell in which you want to place the data value, type the value, then press the appropriate arrow button for the direction in which you wish to move. So, to enter a row of data, move to the far left of the row, type the value and then press → (this process inputs the value and then moves you into the next cell on the left).

1.2.3. Creating a Variable

There are several steps to creating a variable in the SPSS data editor (see Figure 1.7):

- Move the on-screen arrow (using the mouse) to the grey area at the top of the first column (the area labelled *var*.
- Double-click (i.e. click two times in quick succession) with the left button of the mouse.
- A dialog box should appear that is labelled *define variable* (see Figure 1.7).
- In this dialog box there will be a default variable name (something like **var00001**) that you should delete. You can then give the variable a more descriptive name. There are some general rules about variable names, such as that they must be 8 characters or less and you cannot use a blank space. If you violate any of these rules the computer will tell you that the variable name is invalid when you click on OK. Finally, the SPSS data editor is not case sensitive, so if you use capital letters in this dialog box it ignores them. However, SPSS is case sensitive to labels typed into the *Variable Label* part of the *define labels* dialog box (see section 1.2.3.1); these labels are used in the output.
- If you click on OK at this stage then a variable will be created in the data editor for you. However, there are some additional options that you might find useful.


Untitled - SPSS Data Editor
File Edit View Data Transform Sta
Double-click on the grey area labelled *var* to create a variable
This process activates the *define variable* dialog box
The black box shows that this cell is currently active
Define Variable
Change Settings
Type...
Missing Values...
Labels..
Column Format...
OK
Cancel
Help

Variable Name: VAR00001

Variable Description

Type: Numeric8.2

Variable Label:

Missing Values: None

Alignment: Right

Measurement

● Scale ○ Ordinal ○ Nominal

Figure 1.7: Creating a variable

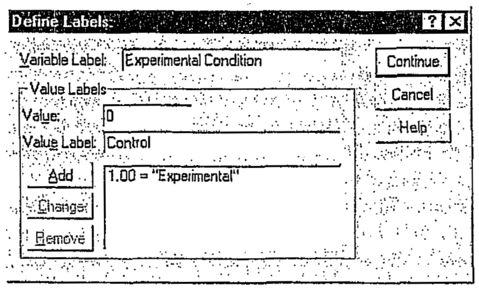
In versions 8 and 9 of SPSS, the *define variable* dialog box contains three options for selecting the level of measurement at which the variable was measured (earlier versions do not have these options). If you are using the variable as a coding variable (next section) then the data are categorical (also called *nominal*) and so you should click on the *Nominal* option. For example, if we asked people whether reading this chapter bores them they will answer *yes* or *no*. Therefore, people fall into two categories: bored and not bored. There is no indication as to exactly how bored the bored people are and therefore the data are merely labels, or categories into which people can be placed. Interval data are scores that are measured on a scale along the whole of which intervals are equal. For example, rather than asking people if they are bored we could measure boredom along a 10-point scale (0 being very interested and 10 being very bored). For data to be interval it should be true that the increase in boredom represented by a change from 3 to 4 along the scale should be the same as the change in boredom represented by a change from 9 to 10. Ratio data have this property, but in addition we should be able to say that someone who had a score of 8 was twice as bored as someone who scored only 4. These two types of data are represented by the *Scale* option. It should be obvious that in some social sciences (notably psychology) it is extremely difficult to establish whether data are interval (can we really tell whether a change on the boredom scale represents a genuine change in the experience of boredom?). A lower level of measurement is ordinal data, which does not quite have the property of interval data, but we can be confident that higher scores represent higher levels of a construct. We might not be sure that an increase in boredom of 1 on the scale represents the same change in experience between 1 and 2 as it does between 9 and 10. However, we can be confident that someone who scores 9 was, in reality, more bored than someone who scored only 8. These data would be ordinal and so you should select *Ordinal*. The *define variable* dialog box also has four buttons that you can click on to access other dialog boxes and these functions will be described in turn.

1.2.3.1. Creating Coding Variables

In the previous sections I have mentioned coding variables and this section is dedicated to a fuller description of this kind of variable (it is a type of variable that you will use a lot). A coding variable (also known as a grouping variable) is a variable consisting of a series of numbers that represent levels of a treatment variable. In experiments, coding variables are used to represent independent variables that have been measured between groups (i.e. different subjects were assigned to different groups). So, if you were to run an experiment with one group of subjects in an experimental condition and a different group of subjects in a control group, you might assign the experimental group a code of 1, and the control group a code of 0. When you come to put the data into the data editor, then you would create a variable (which you might call **group**) and type in the value 1 for any subjects in the experimental group, and 0 for any subject in the control group. These codes tell the computer that all of the cases that have been assigned the value 1 should be treated as belonging to the same group, and likewise for the subjects assigned the value 0.

There is a simple rule for how variables should be placed in the SPSS data editor: levels of the between-group variables go down the data editor whereas levels of within-subject (repeated measures) variables go across the data editor. We shall see exactly how we put this rule into operation in chapter 6.

To create a coding variable we create a variable in the usual way, but we have to tell the computer which numeric codes we are assigning to which groups. This can be done by using the Labels... button in the *define variable* dialog box (see Figure 1.7) to open the *define labels* dialog box (see Figure 1.8). In the *define labels* dialog box there is room to give your variable a more descriptive title. For the purposes of the data editor itself, I have already mentioned that variable labels have to be 8 characters or less and that they have to be lower case. However, for the purposes of the output, it is possible to give our variable a more meaningful title (and this label can also have capital letters and space characters too—great!). If you want to give a variable a more descriptive title then simply click with the mouse in the white space next to where it says *Variable Label* in the dialog box. This will place the cursor in that space, and you can type a title: in Figure 1.8 I have chosen the title *Experimental Condition*. The more important use of this dialog box is to specify group codings. This can be done in three easy steps. First, click with the mouse in the white space next to where it says *Value* (or press ALT and U at the same time) and type in a code (e.g. 1). These codes are completely arbitrary: for the sake of convention people usually use 1, 2 and 3 etc., but in practice you could have a code of 495 if you were feeling particularly arbitrary. The second step is to click the mouse in the white space below, next to where it says *Value Label* (or press ALT and E at the same time) and type in an appropriate label for that group. In Figure 1.8 I have typed in 0 as my code and given this a label of *Control*. The third step is to add this coding to the list by clicking on Add. In Figure 1.8 I have already defined my code for the experimental group, to add the coding for the control group I must click on Add. When you have defined all of your coding values simply click on OK; if you click on OK and have forgotten to add your final coding to the list, SPSS will display a message warning you that any pending changes will be lost. In plain English this simply tells you to go back and click on Add.



Define Labels:

Variable Label: Experimental Condition

Value Labels

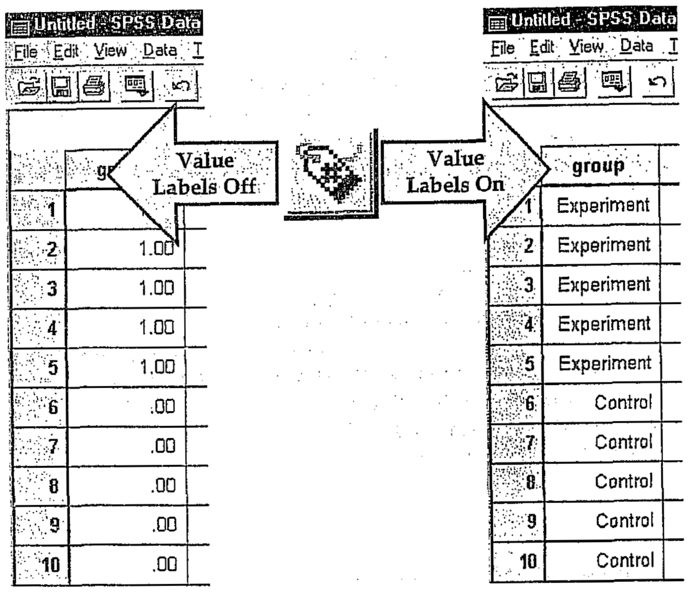
Value: 0

Value Label: Control

1.00 = "Experimental"

Figure 1.8: Defining coding values in SPSS

Having defined your codings, you can then go to the data editor and type these numerical values into the appropriate column. What is really groovy is that you can get the computer to display the codings themselves, or the value labels that you gave them by clicking on (see Figure 1.9). Figure 1.9 shows how the data should be arranged for a coding variable. Now remember that each row of the data editor represents one subject's data and so in this example it is clear that the first five subjects were in the experimental condition whereas subjects 6–10 were in the control group. This example also demonstrates why grouping variables are used for variables that have been measured between subjects: because by using a coding variable it is impossible for a subject to belong to more than one group. This situation should occur in a between-group design (i.e. a subject should not be tested in both the experimental and the control group). However, in repeated measures designs (within subjects) each subject is tested in every condition and so we would not use this sort of coding variable (because each subject does take part in every experimental condition).



	group
1	
2	1.00
3	1.00
4	1.00
5	1.00
6	.00
7	.00
8	.00
9	.00
10	.00

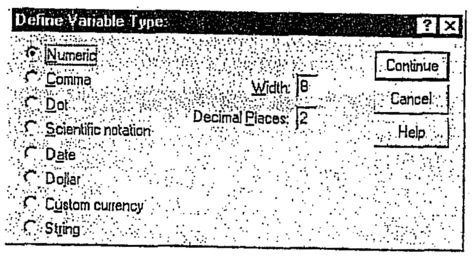
	group
1	Experiment
2	Experiment
3	Experiment
4	Experiment
5	Experiment
6	Control
7	Control
8	Control
9	Control
10	Control

Figure 1.9: Coding values in the data editor with the value labels switched off and on

1.2.3.2. Types of Variables

There are different types of variables that can be used in SPSS. In the majority of cases you will find yourself using numeric variables. These variables are ones that contain numbers and include the type of coding variables that have just been described. However, one of the other options when you create a variable is to specify the type of variable and this is done by clicking on Type... in the *define variable* dialog box. Clicking this button will activate the dialog box in Figure 1.10, which shows the default settings. By default, a variable is set up to store 8 digits, but you can change this value by typing a new number in the space labelled *Width* in the dialog box. Under normal circumstances you wouldn't require SPSS to retain any more than 8 characters unless you were doing calculations that need to be particularly precise. Another default setting is to have 2 decimal places displayed (in fact, you'll notice by default that when you type in whole numbers SPSS will add a decimal place with two zeros after it—this can be disconcerting initially!). It is easy enough to change the number of decimal places for a given variable by simply replacing the 2 with a new value depending on the level of precision you require.

The *define variable type* dialog box also allows you to specify a different type of variable. For the most part you will use numeric values. However, the other variable type of use is a string variable. A string variable is simply a line of text and could represent comments about a certain subject, or other information that you don't wish to analyze as a grouping variable (such as the subject's name). If you select the string variable option, SPSS lets you specify the width of the string variable (which by default is 8 characters) so that you can insert longer strings of text if necessary.



Define Variable Type:

● Numeric
○ Comma
○ Dot
○ Scientific notation
○ Date
○ Dollar
○ Custom currency
○ String

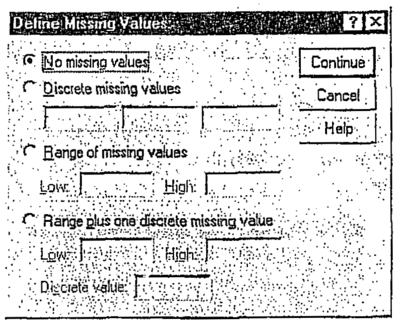
Width: 8

Decimal Places: 2

Figure 1.10: Defining the type of variable being used

1.2.3.3. Missing Values

Although as researchers we strive to collect complete sets of data, it is often the case that we have missing data. Missing data can occur for a variety of reasons: in long questionnaires participants accidentally miss out questions; in experimental procedures mechanical faults can lead to a datum not being recorded; and in research on delicate topics (e.g. sexual behaviour) subjects may exert their right not to answer a question. However, just because we have missed out on some data for a subject doesn't mean that we have to ignore the data we do have (although it sometimes creates statistical difficulties). However, we do need to tell the computer that a value is missing for a particular subject. The principle behind missing values is quite similar to that of coding variables in that we choose a numeric value to represent the missing data point. This value simply tells the computer that there is no recorded value for a participant for a certain variable. The computer then ignores that cell of the data editor (it does not use the value you select in the analysis). You need to be careful that the chosen code doesn't correspond with any naturally occurring data value. For example, if we tell the computer to regard the value 9 as a missing value and several subjects genuinely scored 9, then the computer will treat their data as missing when, in reality, it is not.



Define Missing Values:

● No missing values

○ Discrete missing values

○ Range of missing values

Low: High:

○ Range plus one discrete missing value

Low: High:

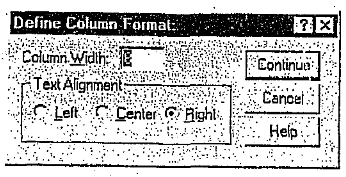
Discrete value:

Figure 1.11: Defining missing values

To specify missing values you simply click on Missing Values... in the *define variable* dialog box to activate the *define missing values* dialog box (see Figure 1.11). By default SPSS assumes that no missing values exist but if you do have data with missing values you can choose to define them in one of three ways. The first is to select discrete values (by clicking on the circle next to where it says *Discrete missing values*) which are single values that represent missing data. SPSS allows you to specify up to three discrete values to represent missing data. The reason why you might choose to have several numbers to represent missing values is that you can assign a different meaning to each discrete value. For example, you could have the number 8 representing a response of 'not applicable', a code of 9 representing a 'don't know' response, and a code of 99 meaning that the subject failed to give any response. As far as the computer is concerned it will ignore any data cell containing these values; however, using different codes may be a useful way to remind you of why a particular score is missing. Usually, one discrete value is enough and in an experiment in which attitudes are measured on a 100-point scale (so scores vary from 1 to 100) you might choose 999 to represent missing values because this value cannot occur in the data that have been collected. The second option is to select a range of values to represent missing data and this is useful in situations in which it is necessary to exclude data falling between two points. So, we could exclude all scores between 5 and 10. The final option is to have a range of values and one discrete value.

1.2.3.4. Changing the Column Format

The final option available to us when we define a variable is to adjust the formatting of the column within the data editor. Click on Column Format... in the *define variable* dialog box and the dialog box in Figure 1.12 will appear. The default option is to have a column that is 8 characters wide with all numbers and text aligned to the right-hand side of the column. Both of these defaults can be changed: the column width by simply deleting the value of 8 and replacing it with a value suited to your needs, and the alignment by clicking on one of the deactivated circles (next to either *Left* or *Center*). It is very useful to adjust the column width when you have a coding variable with value labels that exceed 8 characters in length.



Column Width: 8

Text Alignment

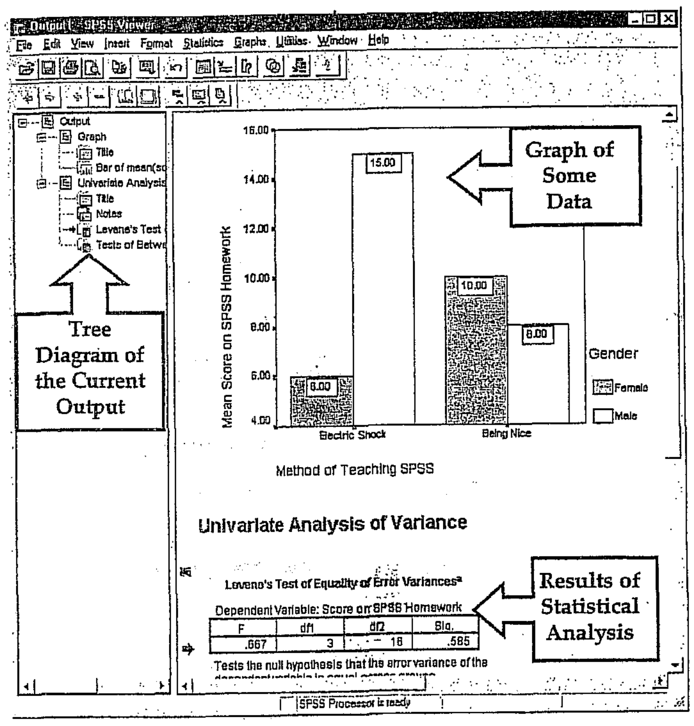
○ Left ○ Center ● Right

Figure 1.12: Defining the format of the column

1.2.4. The Output Viewer

Alongside the main SPSS window, there is a second window known as the output viewer (or *output navigator* in versions 7.0 and 7.5). In earlier versions of SPSS this is simply called the output window and its function is, in essence, the same. However, whereas the output window of old displayed only statistical results (in a very bland font I might add), the new, improved and generally amazing output viewer will happily display graphs, tables and statistical results and all in a much nicer font. Rumour has it that future versions of SPSS will even include a tea-making facility in the output viewer (I live in hope!).

Figure 1.13 shows the basic layout of the output viewer. On the right-hand side there is a large space in which the output is displayed. SPSS displays both graphs and the results of statistical analyses in this part of the viewer. It is also possible to edit graphs and to do this you simply double-click on the graph you wish to edit (this creates a new window in which the graph can be edited). On the left-hand side of the output viewer there is a tree diagram illustrating the structure of the output. This tree diagram is useful when you have conducted several analyses because it provides an easy way of accessing specific parts of the output. The tree structure is fairly self-explanatory in that every time you conduct a procedure (such as drawing a graph or running a statistical procedure), SPSS lists this procedure as a main heading.



Univariate Analysis of Variance

Levene's Test of Equality of Error Variances[a]

Dependent Variable: Score on SPSS Homework

F	df1	df2	Sig.
.667	3	16	.585

Tests the null hypothesis that the error variance of the

Figure 1.13 The output viewer

In Figure 1.13 I conducted a graphing procedure and then conducted a univariate analysis of variance (ANOVA) and so these names appear as main headings. For each procedure there are a series of sub-procedures, and these are listed as branches under the main headings. For example, in the ANOVA procedure there are a number of sections to the output such as a Levene's test (which tests the assumption of homogeneity of variance) and the between-group effects (i.e. the *F*-test of whether the means are significantly different). You can skip to any one of these sub-components of the ANOVA output by clicking on the appropriate branch of the tree diagram. So, if you wanted to skip straight to the between-group effects you should move the on-screen arrow to the left-hand portion of the window and click where it says *Tests of Between-Subjects Effects*. This action will highlight this part of the output in the main part of the viewer. You can also use this tree diagram to select parts of the output (which is useful for printing). For example, if you decided that you wanted to print out a graph but you didn't want to print the whole output, you can click on the word *Graph* in the tree structure and that graph will become highlighted in the output. It is then possible through the print menu to select to print only the selected part of the output. In this context it is worth noting that if you click on a main heading (such as *Univariate Analysis of Variance*) then SPSS will highlight not only that main heading but all of the sub-components as well. This is extremely useful when you want to print the results of a single statistical procedure.

There are a number of icons in the output viewer window that help you to do things quickly without using the drop-down menus. Some of these icons are the same as those described for the data editor window so I will concentrate mainly on the icons that are unique to the viewer window.

As with the data editor window, this icon activates the print menu. However, when this icon is pressed in the viewer window it activates a menu for printing the output. When the print menu is activated you are given the default option of printing the whole output, or you can choose to select an option for printing the output currently visible on the screen, or most useful is an option to print a selection of the output. To choose this last option you must have already selected part of the output (see above).

This icon returns you to the data editor in a flash!

This icon takes you to the last output in the viewer (so, it returns you to the last procedure you conducted).

This icon *promotes* the currently active part of the tree structure to a higher branch of the tree. For example, in Figure 1.13 the *Tests of Between-Subjects Effects* are a sub-component under the heading of *Univariate Analysis of Variance*. If we wanted to promote this part of the output to a higher level (i.e. to make it a main heading) then this is done using this icon.

This icon is the opposite of the above in that it *demotes* parts of the tree structure. For example, in Figure 1.13 if we didn't want the *Univariate Analysis of Variance* to be a unique section we could select this heading and demote it so that it becomes part of the previous heading (the *Graph* heading). This button is useful for combining parts of the output relating to a specific research question.

This icon collapses parts of the tree structure, which simply means that it hides the sub-components under a particular heading. For example, in Figure 1.13 if we selected the heading *Univariate Analysis of Variance* and pressed this icon, all of the sub-headings would disappear. The sections that disappear from the tree structure don't disappear from the output itself; the tree structure is merely condensed. This can be useful when you have been conducting lots of analyses and the tree diagram is becoming very complex.

This icon expands any collapsed sections. By default all of the main headings are displayed in the tree diagram in their expanded form. If, however, you have opted to collapse part of the tree diagram (using the icon above) then you can use this icon to undo your dirty work.

This icon and the following one allow you to show and hide parts of the output itself. So, you can select part of the output in the tree diagram and click on this icon and that part of the output will disappear. It isn't erased, but it is hidden from view. So, this icon is similar to the collapse icon listed above except that it affects the output rather than the tree structure. This is useful for hiding less relevant parts of the output.

This icon undoes the previous one, so if you have hidden a selected part of the output from view and you click on this icon, that part of the output will reappear. By default, all parts of the output are shown and so this icon is not active: it will become active only once you have hidden part of the output.

Although this icon looks rather like a paint roller, it unfortunately does not paint the house for you. What it does do is to insert a new heading into the tree diagram. For example, if you had several statistical tests that related to one of many research questions you could insert a main heading and then demote the headings of the relevant analyses so that they all fall under this new heading.

Assuming you had done the above, you can use this icon to provide your new heading with a title. The title you type in will actually appear in your output. So, you might have a heading like 'Research Question number 1' which tells you that the analyses under this heading relate to your first research question.

This final icon is used to place a text box in the output window. You can type anything into this box. In the context of the previous two icons, you might use a text box to explain what your first research question is (e.g. 'My first research question is whether or not boredom has set in by the end of the first chapter of my book. The following analyses test the hypothesis that boredom levels will be significantly higher at the end of the first chapter than at the beginning').

1.2.5. Saving Files

Although most of you should be familiar with how to save files in Windows it is a vital thing to know and so I will briefly describe what to do. To save files simply use the 💾 icon (or use the menus: **File⇒Save** or **File⇒Save As...**). If the file is a new file, then clicking this icon will activate the *Save As ...* dialog box (see Figure 1.14). If you are in the data editor when you select *Save As ...* then SPSS will save the data file you are currently working on, but if you are in the viewer window then it will save the current output.

There are a number of features of the dialog box in Figure 1.14. First, you need to select a location at which to store the file. Typically, there are two types of locations where you can save data: the hard drive (or drives) and the floppy drive (and with the advent of rewritable CD-ROM drives, zip drives, jaz drives and the like you may have many other choices of location on your particular computer). The first thing to do is select either the floppy drive, by double clicking on 🖴, or the hard drive, by double clicking on 🖴. Once you have chosen a main location the dialog box will display all of the available folders on that particular device (you may not have any folders on your floppy disk in which case you can create a folder by clicking on 📁). Once you have selected a folder in which to save your file, you need to give your file a name. If you click in the space next to where it says *File name*, a cursor will appear and you can type a name of up to ten letters. By default, the file will be saved in an SPSS format, so if it is a data file it will have the file extension *.sav*, and if it is a viewer document it will have the file extension *.spo*. However, you can save data in different formats such as

Microsoft Excel files and tab-delimited text. To do this just click on ▾ where it says *Save as type* and a list of possible file formats will be displayed. Click on the file type you require. Once a file has previously been saved, it can be saved again (updated) by clicking 💾. This icon appears in both the data editor and the viewer, and the file saved depends on the window that is currently active. The file will be saved in the location at which it is currently stored.

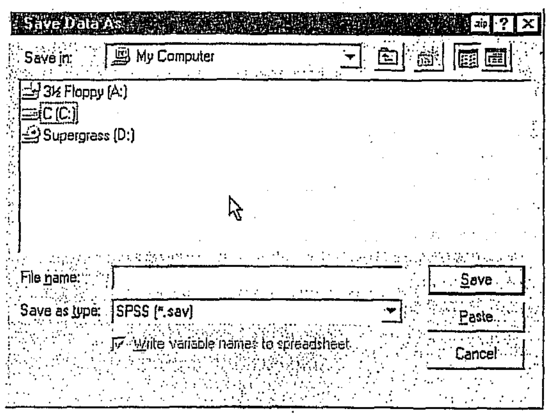


Figure 1.14: The *save data as* dialog box

1.2.6. Retrieving a File

Throughout this book you will work with data files that have been provided on a floppy disk. It is, therefore, important that you know how to load these data files into SPSS. The procedure is very simple. To open a file, simply use the 📂 icon (or use the menus: **File⇒Open**) to activate the dialog box in Figure 1.15. First, you need to find the location at which the file is stored. If you are loading a file from the floppy disk then access the floppy drive by clicking on ▾ where it says *Look in* and a list of possible location drives will be displayed. Once the floppy drive has been accessed you should see a list of files and folders that can be opened. As with saving a file, if you are currently in the data editor then SPSS will display only SPSS data files to be opened (if you are in the viewer window then only output files will be displayed). You can open a folder by double-clicking on the folder icon. Once you have tracked down the required file you can open it either by selecting it with the mouse and then clicking on Open, or by double-clicking on the icon next to the file you want (e.g. double-clicking on ▦). The data/output will then appear in the appropriate window. If you are in the data editor and you want to open a viewer file, then click on ▾ where it says *Files of type* and a list of alternative file formats will be displayed. Click on the appropriate file type (viewer document (**.spo*), Excel file (**.xls*), text file (**.dat*, **.txt*)) and any files of that type will be displayed for you to open.

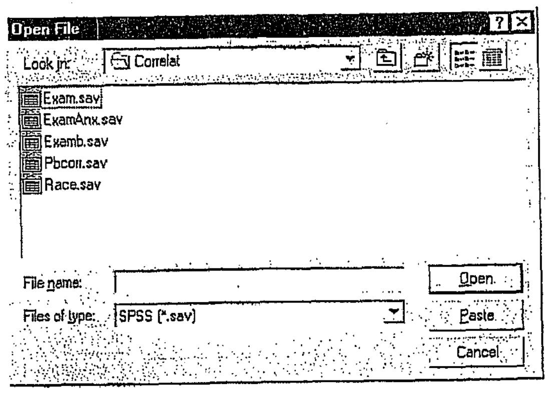


Figure 1.15: Dialog box to open a file

2. lekce

ROZLOŽENÍ KATEGORIZOVANÝCH DAT: ZÁKLADY UNIVARIAČNÍ ANALÝZY (TŘÍDĚNÍ I. STUPNĚ - Modul ANALYZE: procedura Frequencies).

ŘEHÁK, J., ŘEHÁKOVÁ, B.1986. *Analýza kategorizovaných dat v sociologii.* Academia: Praha.

Kapitola 4 Rozložení četností

4.1 Výskyt jevu, četnost, procento

Statistická analýza kategorizovaných dat je založena na studiu výskytů jednotlivých jevů a vztahů mezi nimi. *Jev* ve statistice chápeme obecně jako souhrn určitých projevů, vlastností, vztahů, podmínek, který je empiricky identifikovatelný, o němž můžeme vždy jednoznačně prohlásit, že buď nastal, nebo nenastal. Takový jednoznačný výsledek bývá však v praxi zatížen chybou a my se můžeme dopouštět dvou omylů s různými pravděpodobnostmi: jev nastal. ale prohlásíme, že nenastal, nebo jev nenastal, ale prohlásíme, že nastal. Proto je empirická identifikovatelnost jevu odstupňována podle pravděpodobnosti správného určení. Statistický jev je vždy vztažen k určitému komplexu podmínek, za nichž má odlišení „nastal-nenastal" smysl. Statistická analýza dat vychází z obou uvedených aspektů a jejich rozbor je jedním ze základních východisek postupů i konečné interpretace výsledků.

Jako příklad může sloužit často sledovaný jev **a** ≡ „spokojenost v práci". Místo něho však zjišťujeme jev **b** ≡ „respondent prohlásil, že je v práci spokojen". Vztah mezi oběma je sociologicky i metodologicky velmi složitý. Zatímco výskyty druhého jevu, **b**, zjišťujeme u respondentů vybraného souboru přesně (až na technické chyby v záznamu a přenosu dat), údaje o prvním jevu (**a**) jsou *jím* reprezentovány jen do jistého (většinou neurčeného) stupně spolehlivosti a platnosti. Respondenti nemusí vyjádřit skutečnou spokojenost, ať už záměrně či neúmyslně, pod vlivem nevhodně voleného sběru dat nebo v důsledku nějaké okamžité situace na pracovišti apod. Negací druhého jevu je jev opačný, „respondent neprohlásil, že je spokojen", který však může znamenat: respondent není spokojen, nechce svou spokojenost vyjádřit, nebo to dokonce vůbec prohlásit nemohl, protože nepracuje. Základní podmínkou sledování výskytu uvedeného jevu je tedy pracovní aktivita respondenta, která je např. u výzkumů pracovních kolektivů automaticky splněna, jindy však musí být zjišťována.

V praxi sociologicko-statistické analýzy je určení kontextu, v němž *má smysl* o jevu mluvit, totožné s určením souboru, k němuž má význam vztáhnout výskytovost jevu. Takový soubor lze vymezit pomocí vhodně zvoleného doplňkového jevu ($\bar{a}$) k jevu zkoumanému tak, že sjednocení obou charakterizuje kontext.

V uvedeném příkladě je možno doplněk k jevu „respondent prohlásil, že je v práci spokojen" volit jako jev „respondent prohlásil, že je v práci nespokojen". Opačné oběma jevům je „spokojen — nespokojen", společné (a tím i definicí souboru, k němuž vztahujeme výskytovost) je „prohlásil", „v práci" a všechna obecně platná omezení v daném šetření. Logicky tedy z analýzy vynecháme ekonomicky neaktivní respondenty a ty, kteří o spokojenosti nevypověděli.

Kontext, ke kterému vztahujeme výskytovost jevu, může být pro různé cíle definován různými způsoby, v rozmanité šíři a podmíněnosti. Každá statistická analýza je podmíněná: zvolené omezující podmínky jsou základním kvalitativním východiskem pro interpretaci statistických výsledků. Prakticky je podmíněná analýza prováděna vhodnou redukcí souboru dat. Obzvlášť silně se nutnost určení významových souvislostí projeví u komparace souborů, která je možná jen při srovnatelném základu. V běžné praxi se setkáváme s celou řadou rušivých vlivů, s nimiž není vždy lehké se vyrovnat: mateřská dovolená a vojenská služba jako důvody absence v podniku pro ženy a muže a různé věkové skupiny; otázka po důvodech změny zaměstnání pro osoby, které nejsou zaměstnány, nebo u nichž ke změně nedošlo; zjišťování postojů a názorů u osob, které si je nevytvořily, či si je dokonce ani vytvořit nemohly; typy školního vzdělání pro různé věkové kategorie apod.

Statistické jevy, jejich identifikovatelnost i způsob identifikace a komplex podmínek, za kterých má smysl o nich mluvit, se určují nejen u každé analýzy, ale už při přípravě sběru dat, při jejich záznamu a přenosu, při tvorbě dotazníků a záznamových listů, instrukcí pro pozorování, tazatele apod. Interpretace výsledků se opírá nejen o číselné závěry, ale i o rozbor empirické situace, metodologie sběru i teorii *vztahů mezi podstatovými jevy, jež nás zajímají*, a zjišťovanými empirickými jevy, které jsou vlastním předmětem statistické analýzy.

Statistický jev se váže ke statistické jednotce, u níž nastává, k jejímu místu, času, kontextu. Při sběru dat zjišťujeme u každé statistické jednotky, zda u ní jev nastal, nenastal, či nastat nemůže. U n^* jednotek souboru tak máme empirický údaj: m = počet jednotek, u nichž jev nastal, $\bar{m}$ = počet jednotek, u nichž jev nenastal, m^* = počet jednotek, u nichž jev nemá smysl, m' = počet jednotek, u nichž chybí informace, nebo je zjevně chybná. Z analýzy vynecháme $M = m^* + m'$ jednotek (tzv. vynechávaná data) a pro daný jev pracujeme se souborem o velikosti $n = m + \bar{m} = n^* - m^* - m'$.

Rozdíl *mezi absolutním a poměrovým* ukazatelem výskytu je dán otázkami: „kolik?" a „jaká část? (jaký podíl?)". V sociologické analýze ve většině případů pracujeme s poměrovými údaji, které jsou charakterizovány *relativními četnostmi* jevů $f = \frac{m}{n}$, tj. podílem souboru, u něhož jev nastal. Doplňková relativní četnost opačného jevu je $g = 1 - f = \dfrac{\bar{m}}{n}$. Někdy určujeme také podíl vynechávaných dat pro daný jev: $v = \dfrac{m' + m^*}{n^*}$. V praxi většinou uvádíme stonásobky relativních četností, kterým říkáme *procenta*. Vlastnosti relativních četností jsou velmi jednoduché:

a) Relativní četnost můžeme určovat vždy, existuje-li neprázdný ($n > 0$) soubor jednotek, pro něž má jev smysl.

b) $f = 0$ právě když jev vůbec nenastal.

c) $f = 1$ právě když jev nastal u všech jednotek.

d) Čím vyšší je f, tím častější je jev.

Při analýze více jevů **a**, **b**, **c**, ... značíme obvykle četnosti n_a, n_b, n_c, ... resp. f_a, f_b, f_c, ...

Absolutní četnost m má někdy sama o sobě praktický význam (např. počet osob, které odešly z pracovního kolektivu, musí být nahrazen bez ohledu na velikost skupiny), většinou však nás zajímá výskytovost jako podíl počtu výskytů v souboru (onemocní-li pět osob v třicetičlenném kolektivu, je to méně závažné, než onemocní-li stejný počet osob v patnáctičlenném kolektivu).

Každý jev **a** určuje jednoznačně dichotomickou proměnnou $\boldsymbol{A} = (a, \bar{a})$ = („jev **a** nastal", „jev **a** nenastal"); proto analýzu výskytovosti jevu provádíme také pomocí metod dalších paragrafů. Rozložení proměnné $\boldsymbol{A}$ je $(f, 1-f)$ resp. $(f_a, f_{\bar{a}})$.

4.2 Rozložení četností

Kategorizovanou proměnnou můžeme statisticky chápat jako soubor jevů, pro který platí:

— každé dva jevy jsou neslučitelné (žádné dva nemohou nastat současně);
— soubor jevů je úplný (alespoň jeden z jevů musí nastat);
— každý z jevů má smysl (každý z jevů může nastat);
— každý z jevů je identifikovatelný (v určitém stupni spolehlivosti);
— při identifikaci určujeme jednoznačně, který z jevů nastal.

Každá kategorie pak odpovídá jednomu z jevů; určení toho z jevů, který u statistické jednotky nastal, je totožné s určením kategorie, do které ji zařadíme. Proto kategorie znaku $\boldsymbol{A} = \{a_1, a_2, ..., a_K\}$ považujeme za soubor možných jevů, které lze zjistit.

Statistická analýza vychází ze vztahů všech K četností $\{n_1, n_2, ..., n_K\}$ resp. $\{f_1, f_2, ..., f_K\}$, a navíc z typu znaku, tj. z relací, které platí mezi $\{a_k\}$ tak, jak byly určeny vnějším sociologickometodologickým kritériem. Jde-li o prostý seznam jevů, hovoříme o nominálním znaku, jsou-li jevy uspořádány, jde o ordinální znak, přiřazujeme-li jevům čísla, dostáváme kardinální kategorizovaný znak. Zvláštní roli hraje znak *dichotomický, jehož dvě hodnoty se vzájemně vylučují a k jehož* statistickému popisu postačuje údaj o jedné kategorii, tj. f_1 nebo f_2 (druhý údaj plyne automaticky, $f_2 = 1 - f_1$).

Tabulka četností v třídění 1. stupně zahrnuje K nezávislých parametrů. Buď je to rozložení $\{n_k\}_K$, z něhož plyne výběrový rozsah $n = \Sigma n_k$, nebo $\{n, f_k\}_K$, kde jedna z relativních četností je odvoditelná z ostatních ($\Sigma f_k = 1$). V praxi analýzy je

vhodné využít grafická zobrazení rozložení četností, která mají celou řadu tvarů. Nejvhodnější je *histogram* (*sloupkový graf*) a pro nominální znaky také *kruhový graf*. Existuje celá řada dalších vhodných i méně vhodných ilustrativních metod, které lze vidět v publikacích statistické služby, v odborných článcích a knihách.

Příklad 4.1. *Důvody změny zaměstnání.* Ve výzkumu „Životní dráhy mládeže" byla položena otázka: „Změnil jste zaměstnání? Jestliže ano, jaký jste k tomu měl důvod?" Při záznamu odpovědí byly kódovány statistické jevy, které odpovídaly předem určeným kategoriím znaku, „*důvody změn*", a doplňkové jevy: „absence změny", „chybějící informace". Výsledky třídění 1. stupně jsou uvedeny v tab. 4.1a.

Tabulka 4.1. *Změna zaměstnání*
a) *Rozložení četností pro změnu zaměstnání a její důvody* (soubor mládeže ČSSR, 18—29 let)

Kód	Kategorie	Absolutní četnost	Relativní četnost	Procento
1	*neměnil zaměstnání*	1 628	0.8555	86
2	rodinné důvody	74	0.0389	4
3	finanční důvody	57	0.0300	3
4	zlepšení podmínek resp. výhodnější dojíždění	48	0.0252	3
5	nová práce lépe odpovídá zájmům a schopnostem	22	0.0116	1
6	lepší možnost růstu a postupu	8	0.0042	0
7	zdravotní důvody	17	0.0089	1
8	reorganizace	6	0.0032	0
9	ostatní	12	0.0063	1
0	chybí informace	31	0.0163	2
	Celkem	1 903	1.0001	101

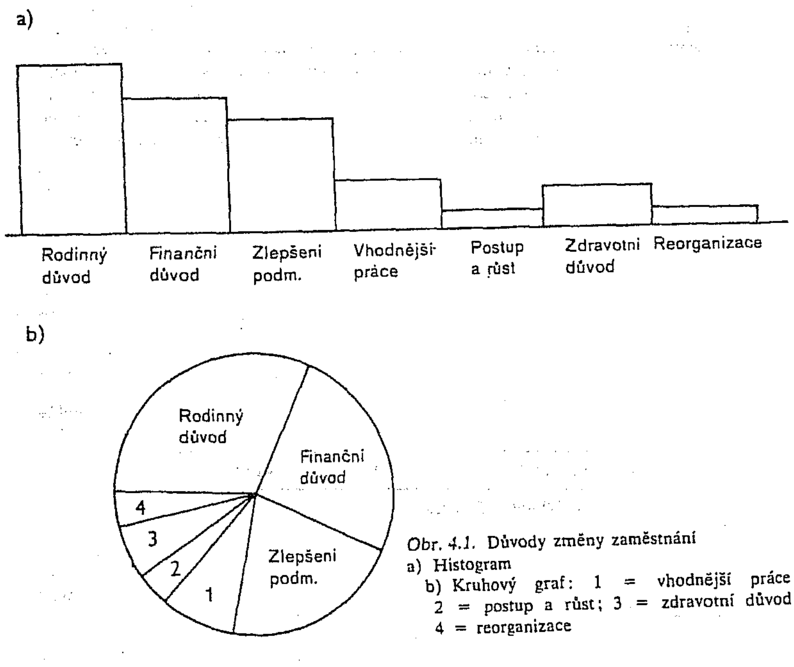
(Zdroj: V. Dubský, Životní dráhy mládeže, výzkumný soubor, ÚFS ČSAV, Praha 1978).

b) *Důvody změny zaměstnání* (Výzkum „Životní dráhy mládeže", soubor mládeže 15—29 let, $n = 232$).

Důvod	Rodinný důvod	Finanční důvod	Zlepšení podmínek	Vhodnější práce	Postup a růst	Zdravotní důvody	Reorganizace	Celkem
Procentní zastoupení	32%	25%	21%	9%	3%	7%	3%	100%

Proměnná „důvody změny zaměstnání" obsahuje však jen kategorie, které mají význam za podmínky, že respondent změnil zaměstnání. Proto z analýzy vynecháme kategorii 1 a 0 (= kód pro chybějící informaci). Nakonec vynecháme i málo obsazenou kategorii „ostatní důvody", která nemá interpretační význam (z důvodů obsahové heterogenity i nízkého procenta zastoupení). Po redukci dostaneme tab. 4.1b, která charakterizuje výskytovost důvodů změny na redukovaném souboru, a která je vhodným východiskem pro analýzu dat.

Rozložení z tab. 4.1b můžeme zobrazit histogramem nebo kruhovým grafem (viz. obr. 4.1).



Obr. 4.1. Důvody změny zaměstnání
a) Histogram
b) Kruhový graf: 1 = vhodnější práce; 2 = postup a růst; 3 = zdravotní důvod; 4 = reorganizace

Pro publikaci tabulek rozložení četností platí obvyklé zásady:

1. Každá tabulka je plně informativní a vypovídá sama o sobě. Obsahuje název nebo přesnou charakteristiku proměnné, charakteristiku souboru, místa, času, kontextu, případně i metodu, která je využita, a výsledky statistické analýzy.

2. Řádky a sloupce jsou jasně označeny slovním popisem (především jde o význam kategorií proměnné, význam charakteristik a čísel v tabulce), pouze obecně přijaté a dobře definované statistické symboly mohou být výjimkou.

3. Hlavní informace se umisťuje do záhlaví tabulky, doplňková informace do poznámek k tabulce (nikoliv pod čáru).

4. V poznámkách pod tabulkou (případně v záhlaví) je uveden zdroj dat, pokud nejde o data, která patří autorům, o data určená jinde (např. v rejstříku použitých dat) nebo společná pro celou publikaci.

5. Absolutní četnosti se uvádějí pouze tehdy, mají-li vlastní informativní hodnotu. Relativní čísla jsou většinou vyjádřena v procentech, a to zaokrouhleně

na celá čísla (méně často na jedno desetinné místo), vždy k nim uvádíme velikost souboru n.

6. Zaokrouhlování při dělení $\frac{n_i}{n}$ vede k tomu, že součet procent nemusí být přesně 100, ale může dávat 99, 101, či 100,1, 99,9 apod. Dříve se procenta v jednotlivých kategoriích upravovala tak, aby součet dával 100%, v současné době se od takových úprav upouští.

7. V poznámkách u tabulky (nebo i přímo v tabulce) zpravidla uvádíme procento vynechávaných hodnot.

Obdobná pravidla platí pro přípravu grafů: plná informativnost, vhodné měřítko, které zajišťuje přehlednost, slovní popis, případná slovní informace přímo v grafu nesmí rušit vjem, uvedení zdroje.

V tabulce rozložení četností pro *nominální* znak můžeme kategorie řadit sestupně podle četností jejich obsazení. Tím získáváme větší přehled a rychlejší informaci. Někdy uvádíme jen ty kategorie, které hrají v rozložení výraznou a interpretovatelnou roli. Takovou formu volíme, především jde-li o tzv. dlouhé znaky (velké K), a u znaků s předem neomezeným počtem hodnot. Typickými příklady proměnných, které většinou tabelujeme tímto způsobem, jsou: respondentův *nejoblíbenější* zpěvák (sportovec, kniha, film, opera), příčina pracovní neschopnosti, záměr trávení dovolené. Uvedená forma se však nehodí pro ordinální a kardinální znaky, neboť by porušila vztahy mezi kategoriemi.

4.3 Kumulativní četnosti (distribuční funkce)

U ordinálních a kardinálních znaků jsou kategorie seřazeny podle vnějšího kritéria určeného obsahem. Z tohoto jednoznačného řazení vychází řada analytických metod založených na kumulativních četnostech, vyjadřujících postupné přibývání výskytů podél stupnice uvažované proměnné. Používáme *absolutní* i *relativní kumulativní četnosti*

$$M_k = \sum_{j=1}^{k} n_j = \text{počet jednotek v kategoriích } 1, 2, ..., k,$$

$$F_k = \frac{M_k}{n} = \sum_{j=1}^{k} f_j = \text{podíl jednotek v kateg. } 1, 2, ..., k, \tag{4.1}$$

$$P_k = \sum_{j=1}^{k} p_j = \text{podíl jednotek v kateg. } 1, 2, ..., k \text{ v základním souboru.}$$

Pro popis vzorců v dalších částech zavedeme úmluvu

$$M_0 = F_0 = P_0 = 0. \tag{4.2}$$

Poznamenejme, že $M_K = n$, $F_K = P_K = 1$. Souboru relativních čísel $\{F_k\}_K$ resp. $\{P_k\}_K$ říkáme *distribuční funkce*. Smysl a využití kumulativních četností ilustruje příklad 4.2.

Příklad 4.2. Příchody do zoo.

Při sociologickém šetření struktury návštěvníků Pražské zoo a délky jejich pobytu bylo zjišťováno rozložení příchodů (metodou načítání příchozích u vchodu). Četnosti získané během jednoho výzkumného dne uvádí tab. 4.2.

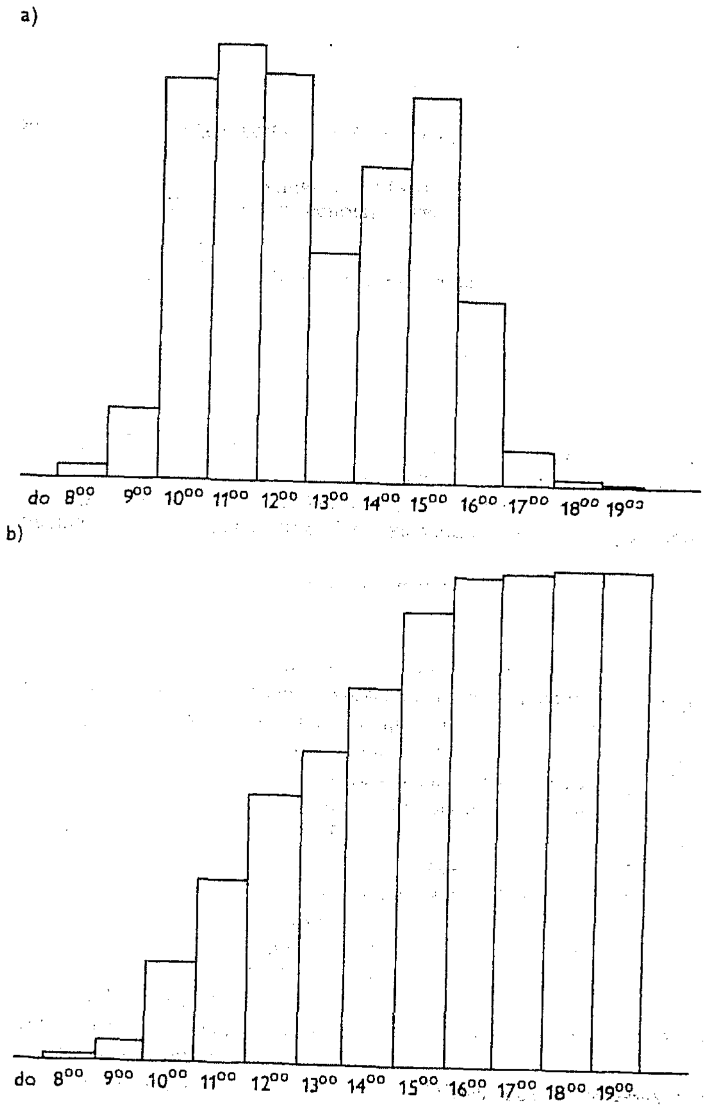
Tabulka 4.2. *Příchody do ZOO Praha v sobotu 12. 8. 1978*
(charakteristika dne: skoro zataženo, chladno)

	Příchody v hodinách												
	do 08.00	do 09.00	do 10.00	do 11.00	do 12.00	do 13.00	do 14.00	do 15.00	do 16.00	do 17.00	do 18.00	do 19.00	Celkem
Počet příchozích	20	219	956	1 034	971	547	759	924	438	75	11	0	5 954
Procento	0.3	3.7	16.1	17.4	16.3	9.2	12.7	15.5	7.3	1.3	0.2	0.0	100.0
Kumulativní četnost	20	239	1 195	2 229	3 200	3 747	4 506	5 430	5 868	5 943	5 954	5 954	—
Kumulativní procento	0.3	4.0	20.1	37.5	53.8	63.0	75.7	91.2	98.5	99.8	100.0	100.0	—

Absolutní četnosti jsou důležité pro zhodnocení náporu na pokladnu, pro služby uvnitř zahrady, pro požadavky na městskou dopravu. Relativní četnosti dobře ukazují rozložení náporu během dne a umožňují porovnání podobných údajů z jiných dnů. Kumulativní četnosti skýtají okamžitou informaci o tom, kolik lidí již do zoo přišlo, ale také podíl, kolik jich do určité hodiny přišlo, a tudíž jaká část jich ještě přijde. Znak „hodina příchodu" může být chápán nominálně (charakteristika určité části dne), ordinálně (průběžný posun během dne) i kardinálně (kvantifikovat *můžeme např. časovým odstupem* od otevírací nebo zavírací hodiny, od vrcholného zatížení restaurace apod.). Kumulativní četnosti umožňují také rychlý výpočet četností výskytů v určitém intervalu složeném ze sousedních kategorií: relativní četnost kategorií

$$(i, i+1, \ldots, j) = f_i + f_{i+1} + \ldots + f_j = F_j - F_{i-1}. \tag{4.3}$$

Tak např. mezi 10. a 14. hod. přišlo 75.7% − 20.1% = 55.6% návštěvníků. Četnosti lze graficky zobrazit pomocí obr. 4.2.



Obr. 4.2. Přehled příchodů do Zoo Praha, 12. 8. 1978
a) Četnosti příchodů v hodinových intervalech
b) Kumulativní četnosti příchodů

4.5 Zvláštní případ tabelací: vícenásobné výběrové otázky

Tabelace vícenásobných výběrových otázek není ve striktním slova smyslu tříděním 1. stupně. Vzhledem k častému výskytu v sociologických šetřeních se o ní však zmíníme. Vícenásobné výběrové otázky jsou instrukce typu: „Z přiloženého seznamu vyberte dvě položky, které považujete za nejdůležitější", „Jmenujte

Tabulka 4.3. *Názor na důležitost cílů v zaměstnání*
(Pokyn: „Vyberte dva z předložených cílů, které považujete za nejdůležitější", $n = 1903$)

Cíl	Počet voleb	Procento z počtu voleb	Procento z počtu respondentů
Řídit lidi, být vedoucím	164	5	9
Materiální zajištění	949	26	50
Společenská úcta, vážnost, prestiž	198	5	10
Možnost přinést maximální užitek lidem, společnosti	552	15	29
Tvůrčí činnost, možnost vytvářet nové	313	9	16
Každodenní svědomité plnění svých povinností	427	12	22
Radost z vykonané práce	828	23	44
Možnost rozšiřovat obzor	200	6	11
Celkem	3 631	101%	191%

(Zdroj: V. Dubský, Životní dráhy mládeže, výzkumný soubor, ÚFS ČSAV, Praha 1978)

nejvýše tři oblíbené zpěváky", „Uveďte tři nejpodstatnější příčiny jevu". Přitom instrukce neobsahuje pokyn k seřazení položek. Analýza těchto dat je složitá, neboť jevy jsou specifickým způsobem závislé. Třídění se provádí tak, že zjišťujeme

$$m_j = \text{počet voleb, které dostala položka „}j\text{"}, \tag{4.8}$$

a odhad četnosti pro každou z J položek (J je počet buď předložených, nebo jmenovaných možností):

$$r_j = \frac{m_j}{n}. \tag{4.9}$$

Jiným způsobem je tabelace jednorozměrné tabulky relativních četností vzhledem k počtu realizovaných voleb

$$g_j = \frac{m_j}{M}, \quad M = \sum_{j=1}^{J} m_j, \tag{4.10}$$

tj. podílu voleb kategorie „j" na všech realizovaných volbách.

Můžeme též uvést index R:

$$R = \frac{M}{Ln} = \frac{\sum_{j=1}^{J} r_j}{L}, \tag{4.11}$$

který vyjadřuje, do jaké míry respondenti využili povolených L voleb.

Častou chybou při zpracování odpovědí na vícenásobné výběrové otázky je to, že děláme třídění 1. stupně pomocných a jen formálně zavedených znaků, které vzniknou tak, že např. 3 možné volby kódujeme: 1. znak = první zatržená hodnota v seznamu, 2. znak = druhá zatržená hodnota v seznamu, 3. znak = třetí zatržená hodnota v seznamu. Rozložení těchto pomocných znaků nemá smysl a žádný interpretační význam. Např. lze snadno ověřit, že kód první položky se nemůže vůbec vyskytnout u 2. a 3. znaku, kód druhé položky se nemůže vyskytnout u 3. znaku, kód třetí položky se může vyskytnout u 1. znaku jen tehdy, využil-li respondent pouze jednu volbu. Hodnoty $\{m_j\}_J$ vznikají součtem rozložení uvedených pomocných znaků.

Příklad 4.3. Cíle v povolání.

Mladým lidem ve věku *18—29 let byl dán* tazatelem pokyn: „Ve svém povolání se lidé snaží dosáhnout nejrůznějších cílů. Vyberte dva z nich, které jsou podle Vašeho názoru nejdůležitější." Osm důvodů bylo předloženo na kartě, uvedené volby byly zakroužkovány v záznamovém listě. Kódování bylo provedeno pomocí dvou pomocných znaků **A** = kód první zakroužkované kategorie, **B** = kód druhé zakroužkované kategorie. Tabulka 4.3 vznikla jako součet absolutních *četností* znaku **A** a **B** (dílčí tabulky nemají smysl).

(Může překvapit nízké procento kategorie, „materiální zajištění", neboť při samostatném dotazu bychom očekávali téměř stoprocentní odpověď „*ano*, je to důležité".) Součet 191% ukazuje, že využití

NORUŠIS, Marija J. 1998. Guide to Data Analysis. Upper Saddle River, Prentice Hall.

Counting Responses 3

How can you summarize the various responses people give to a question?

- What is a frequency table, and what can you learn from it?
- How can you tell from a frequency table whether there have been errors in coding or entering data?
- What are percentages and cumulative percentages?
- What are pie charts and bar charts, and when do you use them?
- When do you use a histogram?
- What are the mode and the median?
- What do percentiles tell you?

Whenever you ask a number of people to answer the same questions, or when you measure the same characteristics for several people or objects, you want to know how frequently the possible responses occur. This can be as simple as just counting up the number of yes or no responses to a question. Or it can be considerably more complicated if, for example, you've asked people to report their annual income to the nearest penny. In this case, simply counting the number of times each unique income occurs may not be a useful summary of the data. In this chapter, you'll use the Frequencies procedure to summarize and display values for a single variable. You'll also learn to select appropriate statistics and charts for different types of data.

▶ The data analyzed in this chapter are in the *gss.sav* data file. For instructions on how to obtain the Frequencies output shown in the chapter, see "How to Obtain a Frequency Table" on p. 48.

Describing Variables

To see what's actually involved in examining and summarizing data, you'll use the nine variables from the General Social Survey described in Table 3.1. (You will use data from only 1500 respondents, since the SPSS student system is restricted in the number of cases in a data file.)

What's the General Social Survey? The General Social Survey is administered yearly by the National Opinion Research Center to a sample of about 1500 persons 18 years of age and older. The sample represents the population of non-institutionalized adults living in the United States. (College dormitories are excluded from the survey!) Questions on many different topics—from how often you pray to where you were living at age 16—are included. Data from the General Social Survey are distributed at a nominal cost and are widely used by researchers and students (Davis & Smith, 1993). ■■■

All of these variables are defined as numeric in SPSS, but in most cases the numbers are just codes for non-numeric information. Value labels for each variable specify what the codes really mean.

In the SPSS Data Editor, to display (or hide) value labels, from the menus choose:

View
Value Labels

Table 3.1 Variables from the General Social Survey

Variable Name	Description
age	Age of respondent in years
sex	1=Male, 2=Female
educ	Years of education
income91	Total family income in 1993 (classified into one of 21 income categories)
wrkstat	Work status (1=Full-time work, 2=Part-time work, 3=Temporarily not working, 4=Unemployed (laid off), 5=Retired, 6=In school, 7=Keeping house, 8=Other)
richwork	"Would you continue or stop working if you became rich?" (1=Continue, 2=Stop)
satjob	Job satisfaction (1=Very satisfied, 2=Moderately satisfied, 3=A little dissatisfied, 4=Very dissatisfied)
life	"Do you find life exciting, pretty routine, or dull?" (1=Dull, 2=Routine, 3=Exciting)
impjob	"How important to your life is having a fulfilling job?" (1=One of the most important, 2=Very important, 3=Somewhat important, 4=Not too important, 5=Not at all important)

Start by looking at the variable *impjob*, which tells you how important a fulfilling job is to the respondent. Since there are only five possible responses, you can easily count how many people gave each of them.

A Simple Frequency Table

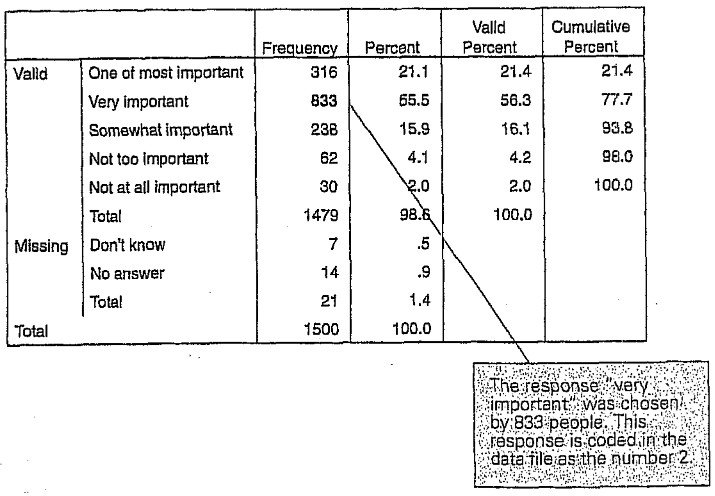
In Figure 3.1, you see the frequency table for the job importance variable.

To obtain this frequency table, from the menus choose:

Statistics
Summarize ▸
Frequencies...

In the Frequencies dialog box, select the variables impjob, as shown in Figure 3.11.

Figure 3.1 Frequency table of job importance



		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	One of most important	316	21.1	21.4	21.4
	Very important	833	55.5	56.3	77.7
	Somewhat important	238	15.9	16.1	93.8
	Not too important	62	4.1	4.2	98.0
	Not at all important	30	2.0	2.0	100.0
	Total	1479	98.6	100.0	
Missing	Don't know	7	.5		
	No answer	14	.9		
	Total	21	1.4		
Total		1500	100.0		

The response "very important" was chosen by 833 people. This response is coded in the data file as the number 2.

From a frequency table, you can tell how frequently people gave each response. The first row is for the response *one of the most important* (coded in the data with the value 1). The second row is for the response *very important* (coded in the data with the number 2). To determine how many people gave each response, look at the column labeled *Frequency*. For example, you find that 316 people find a fulfilling job to be *one of the most important* things to them, and 238 find it to be *somewhat important.* Only 30 people find having a fulfilling job *not at all important.* In the row labeled *Total*, you see that 1479 people selected one of the five possible valid responses.

The second part of the table tells you how many people did not select one of the five choices. There are two rows in the frequency table for the responses *don't know* and *no answer. Don't know* is used for people unwilling to commit themselves to a response. *No answer* is used when the response is illegible, lost, or not recorded by the interviewer. When the data file was defined, both *don't know* and *no answer* were identified as missing-value codes. That is, you don't have a valid answer for people whose responses are coded as *don't know* or *no answer*. In the *Frequency* column, you see that the response *don't know* was selected by 7 people and that the response was not available for 14 people. A total of 21 failed to select a valid response; that is, their response was identified as missing.

In the last row of the frequency table, you see that a total of 1500 people participated in the survey. Of these, 21 failed to select one of the five available responses; that is, their response was identified as *missing*. The other 1479 provided a valid response.

Why do you use different codes for don't know *and* no answer? It's important to pinpoint why data values are missing. A response of *don't know* tells you that a person probably doesn't have strong feelings about the topic. It's unlikely that they find a job to be very important. A response of *no answer* doesn't tell you anything about a person's opinion of the importance of a job. The number of *no answer* responses tells you whether the survey was carefully conducted. You'll see later that if there are many cases with missing values, you may have serious problems in drawing conclusions from your data. ■■■

In the frequency table, value labels, which are descriptions of the codes assigned when you define a variable, are used to identify rows. If you don't assign these descriptions, the actual codes are shown. If your codes are not inherently meaningful, you should assign value labels to them so that the output is easier to understand. Assigning a value label once is much easier than repeatedly having to look up the meanings of codes.

Only responses actually selected by the participants are included in the frequency table. If no one selected the response *not at all important*, it would not be included in the table. Similarly, if you accidentally enter a code that does not correspond to a valid response—say a code of 0, 6, or 7 for the job importance variable—you will find it as a row in the frequency table. That's why frequency tables are useful for detecting mistakes in the data file. If you find wrong codes in your data values, you must correct the data file before proceeding.

Percentages

A frequency count alone is not a very good summary of the data. For example, if you want to compare your results to those of another survey, it won't do you much good to know simply that 762 people in that survey chose the response *very important*. From the count alone, you can't tell if the other survey's results are similar to yours. To compare the two surveys, you must convert the observed counts to percentages.

From a percentage, you can tell what proportion of people in the survey gave each of the responses. Unlike counts, you can compare percentages across surveys with different numbers of cases. You compute a percentage by dividing the number of cases that gave a particular response by the total number of cases. Then you multiply the result by 100.

To change the number of decimal places shown in the output, double-click the pivot table to activate it, select the cell or column of interest, then choose:

Format
Cell Properties...

and change the Decimals specification.

In Figure 3.1, you find percentages in the column labeled *Percent*. Note that the 316 people who gave the response *one of the most important* are 21.1% of the 1500 people in your survey. Similarly, the 238 people who gave the response *somewhat important* are 15.9% of your sample. The 7 people who *don't know* are 0.5% of the total sample. (The actual percentage is 0.47%, but by default only one decimal place is shown.) The sum of the percentages over all the possible responses, including *don't know* and *no answer*, is 100%.

Percentages Based on Valid Responses

To get the numbers in the column labeled *Percent*, you divide the observed frequency by the total number of cases in the sample and multiply by 100. Cases with codes identified as *missing* are included in the denominator. That can be a problem. For example, the General Social Survey does not ask all questions of all people. The question "Would you continue or stop working if you became rich?" was asked of only two-thirds of people who were working or temporarily unemployed. Figure 3.2 shows the responses of people to this question.

To obtain this table, in the Frequencies dialog box select the variable richwork, as shown in Figure 3.11.

Figure 3.2 Frequency table of continue working

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Continue working	448	29.9	69.8	69.8
	Stop working	194	12.9	30.2	100.0
	Total	642	42.8	100.0	
Missing	Not applicable	842	56.1		
	Don't know	11	.7		
	No answer	5	.3		
	Total	858	57.2		
Total		1500	100.0		

The percentage of people giving the response *continue working* is 29.9. What does that mean? Does it mean that about 30% of people in the survey would continue working if they became rich? No. It means that about 30% of the people in the sample, regardless of whether they were asked the question or volunteered an answer, gave the response *continue working*. Of the 1500 people in the survey, 56.1% weren't even asked the question (recorded in the table as *Not applicable)*. An additional 1% were asked and either gave the response *don't know* or their response was lost (*no answer*). All of these missing people are included in the denominator of the *Percent* calculation.

If you want to know what percentage of people who gave an acceptable answer selected *continue working*, look at the *Valid Percent* column. Almost 70% of people who answered the question claim that they would continue working if they struck it rich. (It's up to you whether you believe that percentage!) That's quite different from 30%. To calculate the entries in the *Valid Percent* column, you must exclude all people who gave an answer identified as *missing*. Valid percentages sum to 100 over all possible answers that are not missing. In this example, there are only two valid answers: *continue working* and *stop working*. Of the people who gave one of these answers, 69.8% selected the first and 30.2% selected the second. These two percentages sum to 100.

Problems with Missing Data

Removing people who aren't asked a question from the calculation of percentages is not troublesome. They don't make interpretation of the results difficult. However, if a lot of people who are asked the question refuse to answer, that can be a problem. In Figure 3.2, you see that only 11 people gave an answer of *don't know*. They represent fewer than 2% of the 653 people who were actually asked the question. So, you don't have to worry much about their impact on any conclusions you draw.

In contrast, however, consider the following situation. You conduct an employee satisfaction survey among 100 employees and find that 55 of them rate themselves as satisfied, 4 rate themselves as unsatisfied, and the remaining 41 decline to answer your question. That means that 55% of the polled employees consider themselves satisfied. However, if you exclude those who refused to answer from the denominator, 93% of the employees who answered the question consider themselves satisfied.

Which is the correct conclusion? Unfortunately, you don't know. It's possible that you have a company full of satisfied employees, many of whom don't like to answer questions. It's also possible that almost half of your employees are unhappy but are wary of voicing their dissatisfaction. When your data have many missing values because of people refusing to answer questions, it may be difficult, if not impossible, to draw correct conclusions. When you report percentages based on cases with nonmissing values, you should also report the percentage of cases that refused to give an answer.

Cumulative Percentages

There's one more percentage of interest in the frequency table. It's called the cumulative percentage. For each row of the frequency table, the **cumulative percentage** tells you the percentage of people who gave that response and any response that precedes it in the frequency table. It is the sum of the valid percentages for that row and all rows before it. Since there are only two possible valid answers for the continue working variable, the cumulative percentages in Figure 3.2 are of little interest. Instead, consider Figure 3.1 again. The cumulative percentage for *somewhat important* is 93.8. This means that over 93% of the people who answered the question said that a fulfilling job was at least somewhat important to their lives. Only 6.2% of the people rated the importance of a fulfilling job as less than *somewhat important*. Cumulative percentages are most useful when there is an underlying order to the codes assigned to a variable.

Sorting Frequency Tables

Unless you specify otherwise, SPSS produces a frequency table in which the order of the rows corresponds to the values of the codes you assign to the responses. The first row is for the smallest number found in the data values, and the last is for the largest. Codes that have been declared missing are at the end of the table. For example, if you had assigned the code 1 to *stop working*, it would have appeared first in the frequency table in Figure 3.2.

When you have several possible responses and the codes are not arranged in a meaningful order, you may want to rearrange the frequency table so that it's easier to use. You can determine the order of the rows in the table based on the frequency of values in the data. For example, Figure 3.3 shows a frequency table for the work status variable when the table is sorted in descending order of frequencies. Look at the column labeled *Frequency*. The frequencies go from largest to smallest.

To obtain this output, select Format in the Frequencies dialog box. Then select Descending counts, as shown in Figure 3.12.

Figure 3.3 Frequency table sorted by counts

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Working fulltime	747	49.8	49.8	49.8
	Retired	231	15.4	15.4	65.2
	Keeping house	200	13.3	13.3	78.5
	Working parttime	161	10.7	10.7	89.3
	Unempl, laid off	51	3.4	3.4	92.7
	School	42	2.8	2.8	95.5
	Other	36	2.4	2.4	97.9
	Temp not working	32	2.1	2.1	100.0
	Total	1500	100.0	100.0	

Table is sorted by the counts in the Frequency column

Sorting a frequency table will usually change the values in the *Cumulative Percent* column, since the cumulative percentages depend on the order of the rows in the table. When the work status table is sorted by decreasing frequency, the cumulative percentage for *retired* is the percentage of people retired or working full time. In the default frequency table, however, in which the rows are sorted by the values of the codes,

NORUŠIS, Marija J. 1998. *Guide to Data Analysis*. Upper Saddle River, Prentice Hall.

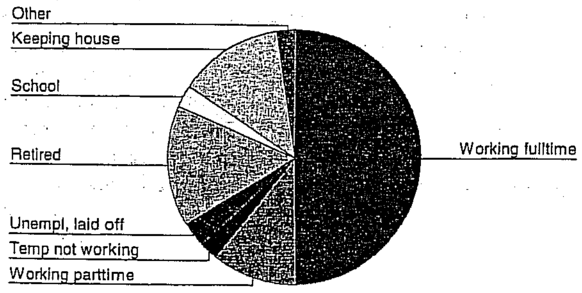
the cumulative percentage for *retired* is the sum of the valid percentages for codes 1 through 5.

Pie Charts

The information in a frequency table is easier to see if you turn it into a visual display, such as a bar chart or a pie chart. In Figure 3.4, you see a pie chart of the frequency table in Figure 3.3. There is a "slice" for each row of the frequency table. From the pie chart, you can easily see that almost half of your sample is *working full time*. It's also easy to see that the number of people who are *retired, keeping house,* and *working part time* are roughly equal. If you have many small slices in a pie chart, you can combine them into an *other* category. For example, Figure 3.5 is the pie chart for the same frequency table, except that all slices that have fewer than 5% of the cases (*in school, temporarily not working, unemployed,* and *other*) are combined into a single slice.

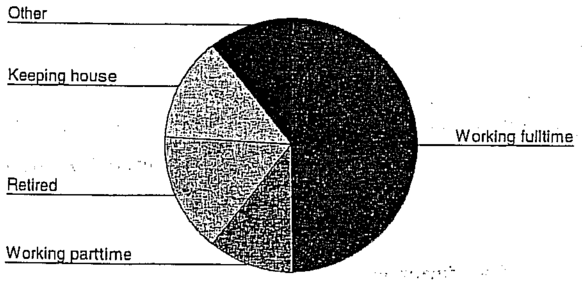
To obtain a pie chart, select Pie charts in the Frequencies Charts dialog box, as shown in Figure 3.14.

Figure 3.4 Pie chart of work status



You can collapse categories in a pie chart after it has been created. See "Modifying Chart Options" on p. 518 in Appendix A.

Figure 3.5 Work status with categories collapsed

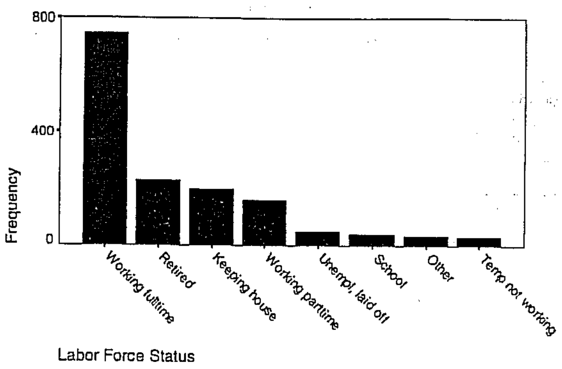


would expect, the tallest bar is for the *working full time* category. It's about three times as tall as the next largest bar, which represents *retired.*

To obtain this output, select Bar charts in the Frequencies Charts dialog box, as shown in Figure 3.14.

You can also obtain bar charts using the Graphs menu, as discussed in Appendix A.

Figure 3.6 Bar chart of work status



3. lekce

ROZLOŽENÍ SPOJITÝCH DAT: ZÁKLADY UNIVARIAČNÍ ANALÝZY (TŘÍDĚNÍ I. STUPNĚ - Modul ANALYZE: procedury Frequencies, Descriptives, Explore).

Summarizing the Age Variable

Although you can produce frequency tables for any kind of data, a frequency table becomes less useful as the number of possible responses increases. For example, you can construct a frequency table for the variable *age*, which tells you the ages of the people in your survey, but you will have as many rows in the frequency table as there are different ages in the data file. In Figure 3.7 you see that there is a row for every age from 18 to 89.

What's this? Nobody in the General Social Survey sample was 90 years or older? Actually, this is just a quirk in the way ages are coded in the General Social Survey. For obscure historical reasons, the General Social Survey assigns an age of 89 to everyone with an age of 89 or older. The code 99 indicates that the age is not known. Because so few people are that old, this quirk has very little effect on analyses that use the age variable. When you design a study, record the actual age—or better yet, the birth date, since it's harder to fudge. (To remain 30 forever, you have to remember to change your birth year annually!)

Figure 3.7 Frequency table of age

The Pivot Table Editor is used to hide the Percent column.

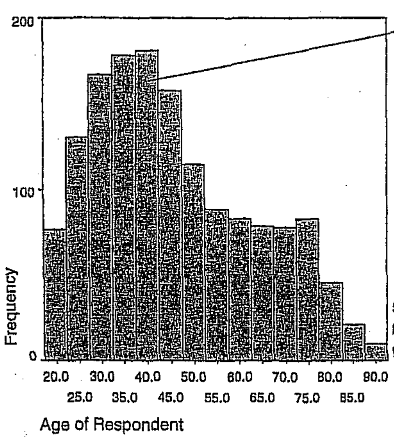
		Frequency	Valid Percent	Cumulative Percent
Valid	18	5	.3	.3
	19	17	1.1	1.5
	20	18	1.2	2.7
	21	22	1.5	4.1
	22	15	1.0	5.2
	23	28	1.9	7.0
	24	23	1.5	8.6
	25	30	2.0	10.6
	26	27	1.8	12.4
	27	22	1.5	13.8
	28	42	2.8	16.7
	29	30	2.0	18.7
	30	36	2.4	21.1
	31	31	2.1	23.1
	32	28	1.9	25.0
	33	33	2.2	27.2
	34	26	1.7	28.9
	35	41	2.7	31.6
	36	42	2.8	34.4
	37	37	2.5	36.9
	38	41	2.7	39.7
	39	38	2.5	42.2
	40	36	2.4	44.6
	41	36	2.4	47.0
	42	30	2.0	49.0
	43	39	2.6	51.6
	44	28	1.9	53.5
	45	30	2.0	55.5
	46	29	1.9	57.5
	47	32	2.1	59.6
	48	20	1.3	60.9
	49	27	1.8	62.7
	50	21	1.4	64.1
	51	26	1.7	65.9
	52	21	1.4	67.3
	53	18	1.2	68.5
	54	19	1.3	69.8
	55	22	1.5	71.2
Valid	56	12	.8	72.0
	57	18	1.2	73.2
	58	25	1.7	74.9
	59	14	.9	75.9
	60	16	1.1	76.9
	61	11	.7	77.7
	62	17	1.1	78.8
	63	19	1.3	80.1
	64	13	.9	80.9
	65	17	1.1	82.1
	66	19	1.3	83.3
	67	11	.7	84.1
	68	16	1.1	85.2
	69	19	1.3	86.4
	70	9	.6	87.0
	71	15	1.0	88.0
	72	19	1.3	89.3
	73	20	1.3	90.6
	74	18	1.2	91.8
	75	17	1.1	93.0
	76	13	.9	93.8
	77	15	1.0	94.8
	78	14	.9	95.8
	79	7	.5	96.3
	80	6	.4	96.7
	81	9	.6	97.3
	82	10	.7	97.9
	83	3	.2	98.1
	84	3	.2	98.3
	85	4	.3	98.6
	86	5	.3	98.9
	87	6	.4	99.3
	88	3	.2	99.5
	89	7	.5	100.0
	Total	1495	100.0	
Missing	NA	5		
	Total	5		
Total		1500		

Histograms

You won't find pie charts and bar charts of the age variable to be useful either. There will be as many slices and bars as there are distinct ages. The arrangement of the values in the charts can be troublesome as well. Both bar charts and pie charts arrange bars and slices in ascending order of the values. However, if a particular age doesn't occur, an empty space is not left for it. That means that in a bar chart, the bar for 46 years may be right next to the bar for 50 years. You won't see a gap to remind you that ages 47 through 49 don't occur in your data.

A better display for a variable like *age*, for which it makes sense to group adjacent values, is a histogram. A **histogram** looks like a bar chart, except that each bar represents a range of values. For example, a single bar may represent all people in their twenties. In a histogram, the bars are plotted on a numerical scale that is determined by the observed range of your data.

Figure 3.8 Histogram of age



181 people are between 37.5 and 42.5 years of age

To obtain this output, select Charts in the Frequencies dialog box. Then select Histograms, as shown in Figure 3.14.

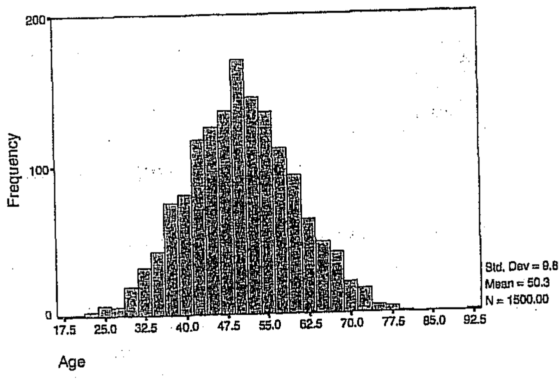
You can also obtain histograms using the Graphs menu, as discussed in Appendix A.

You see a histogram for the age variable in Figure 3.8. Age values are on the horizontal axis, and frequencies are on the vertical axis. The first bar represents cases with ages between 17.5 and 22.5. The middle value in this interval, the **midpoint**, is 20, which becomes the label used for the bar. From the histogram, you see that about 80 cases fall into this interval. Similarly, the second bar represents cases with ages between 22.5 and 27.5. This bar represents 130 cases.

A histogram tells you about the distribution of the data values. That is, it tells you how likely various values are. From it, you can see whether the cases cluster around a central value. You can also see whether large and small values are equally likely and whether there are values far removed from the rest. This is important not only to understand the data you've collected, but also for choosing appropriate statistical techniques for analyzing them. In Figure 3.8, you see that the age distribution has a peak corresponding to the interval 37.5 to 42.5. Additionally, you can see that the distribution of ages is not symmetric but has a "tail" extending to the older ages. That's because the General Social Survey interviews only respondents who are 18 or older.

What's a symmetric distribution? A distribution is symmetric if a vertical line going through its center divides it into two halves that are mirror images of each other. Figure 3.9 shows what a symmetric distribution of a hypothetical age variable might look like. Note that small and large values of age are equally likely.

Figure 3.9 Symmetric distribution



Mode and Median

You can use a variety of statistics to further summarize the information in a frequency table. In Chapter 4, you'll learn about a large number of such summary statistics. In the remainder of this chapter, you'll focus on summary measures that are easily obtained from the frequency table.

The mode is defined as the most frequently occurring value in your data. From the frequency table in Figure 3.7, you see that two ages (28 and 36) are tied for the mode. There are 42 people with each of these ages. Although these ages occur most frequently, they represent a small percentage of the total cases. Less than 3% of the total has an age of 28. Knowing the mode tells you very little about the data.

What are the modes for the job importance and What if you were rich variables? In Figure 3.1, you see that *very important* is the most frequently occurring response to the job importance question. That makes it the mode. Similarly, for the if rich variable shown in Figure 3.2, *Continue working* is the most common response, so it's the mode.

Scales on which variables are measured are discussed in Chapter 4.

If you can meaningfully order your data values from smallest to largest, you can compute additional summary measures. These measures are better than the mode, since they make use of the additional information about the order of the data values. For example, the **median** is the value that is greater than half the data values and less than the other half.

To display the median and mode along with your frequency table, in the Frequencies dialog box, select Statistics. Then click Median and Mode (see Figure 3.13).

You calculate the median by finding the middle value when values for all cases are ordered from smallest to largest. If you have an odd number of cases, the median is just the middle value. If you have an even number of cases, the median is the value midway between the two middle ones. For example, the median of the five values 12, 34, 57, 92, and 100 is 57. For the six numbers 13, 20, 40, 60, 89, and 123, the median is 50, because 50 is the value midway between 40 and 60, the two middle numbers. (Add the two middle values and divide by two.)

You can calculate the median very easily from a frequency table. Find the first value for which the cumulative percentage exceeds or is equal to 50%. For the age variable in Figure 3.7, the median is 43. That means that half of the people in your sample are less than 43 years of age, and half are older. That's much more useful information than knowing that 28 and 36 years are tied for the mode.

What's the median for the work status variable? Since there is no meaningful order of the codes assigned to the work status variable, it doesn't make sense to talk about median work status. You should use the median only when the data values can be ranked from smallest to largest.

Percentiles

When you calculate the median, you find the number that splits the sample into two equal parts. Half of the cases have values smaller than the median, and the other half have values larger than the median. You can compute values that split the sample in other ways. For example, you can find the value below which 25% of the data values fall. Such values are called percentiles, since they tell you the percentage of cases with values below and above them. Twenty-five percent of the cases have values smaller than the 25th percentile, and 75% of the cases have values larger than the 25th percentile. The median is the 50th percentile, since 50% of the cases have values less than the median, and 50% have values greater than the median.

To obtain this output, select Statistics in the Frequencies dialog box. Then select Quartiles, as shown in Figure 3.13.

Figure 3.10 Quartiles for age

	N		Percentiles		
	Valid	Missing	25	50	75
AGE	1495	5	32.00	43.00	59.00

See "Percentiles" on p. 95 in Chapter 6 for further discussion of percentiles.

You can compute percentiles from a frequency table by finding the first value with a cumulative percentage larger than or equal to the percentile you're interested in. You can see the 25th, 50th, and 75th percentiles for the age variable in Figure 3.10. From these, you know that 25% of the cases are 32 or younger, 50% are 43 or younger, and 75% are 59 or younger. Together, the 25th, 50th, and 75th percentiles are known as **quartiles**, since they split the sample into four groups with roughly equal numbers of cases. That is, 25% of the cases are 32 years old or younger, 25% are between 32 and 43, 25% are between 43 and 59, and 25% are 59 or older.

Summary

How can you summarize the various responses people give to a question?

- A frequency table tells you how many people (cases) selected each of the responses to a question. It contains the number and percentage of the people who gave each response, as well as the number of people for whom responses are not available.
- If you find codes in the frequency table that weren't used in your coding scheme, you know that an error in data coding or data entry has occurred.
- A count can be transformed into a percentage by dividing it by the total number of responses and multiplying by 100.
- A cumulative percentage is the percentage of cases with values less than or equal to a particular value.
- Pie charts and bar charts are graphical displays of counts.
- A histogram is a graphical display of counts for ranges of data values.
- The mode is the data value that occurs most frequently.
- The median is the middle value when data values are arranged from smallest to largest.
- Percentiles are values below which and above which a certain percentage of case values fall.

Computing Descriptive Statistics 4

How can you summarize the values of a variable?

- What are scales of measurement, and why are they important?
- How does the arithmetic mean differ from the mode and the median?
- When is the median a better measure of central tendency than the mean?
- What does the variance tell you? The coefficient of variation?
- What are standardized scores, and why are they useful?

In the previous chapter, you used frequency tables, bar charts, pie charts, histograms, and percentiles to examine the distribution of values for a variable. These are essential techniques for getting acquainted with the data. Often, however, you want to summarize the information even further by computing summary statistics that describe the "typical" values, or the **central tendency**, as well as how the data spread out around this value, or the **variability**. In this chapter, you'll learn how to use the Frequencies and Descriptives procedures to compute the most commonly used summary statistics for central tendency and variability.

▶ This chapter continues to use the *gss.sav* data file. For instructions on how to obtain the Descriptives output discussed in the chapter, see "How to Obtain Univariate Descriptive Statistics" on p. 69.

What's a statistic? Often when you collect data, you want to draw conclusions about a broader base of people or objects than are actually included in your study. For example, based on the responses of people included in the General Social Survey, you want to draw conclusions about the population of adults in the United States. The people you observe are called the **sample**. The people you want to draw conclusions about are called the **population**. A **statistic** is some characteristic of the sample. For example, the median age of people in the General Social Survey is a statistic. The term **parameter** is used to describe characteristics of the population. If you had the ages of all adults in the United States, the median age would be called a parameter value. Most of the time, population values, or parameters, are not known. You must estimate them based on statistics calculated from samples. ■■■

Summarizing Data

Consider again the data described in the previous chapter. Suppose you want to summarize the data values further. You want to know the typical age for participants in the survey, or their typical status in the workplace, or typical satisfaction with their job. A unique answer to these questions doesn't exist, since there are many different ways to define "typical." For example, you might define it as the value that occurs most often in the data (the mode), or as the middle value when the data are sorted from smallest to largest (the median), or as the sum of the data values divided by the number of cases (the arithmetic mean). To choose among the various measures of central tendency and variability you must consider the characteristics of your data as well as the properties of the measures. Although the mode may be a plausible statistic to report for status in the labor force, it may be a poor selection for a variable like age.

Scales of Measurement

One of the characteristics of your data that you must always consider is the scale on which they are measured. Scales are often classified as nominal, ordinal, interval, and ratio, based on a typology proposed by Stevens (1946). A nominal scale is used only for identification. Data measured on a **nominal scale** cannot be meaningfully ranked from smallest to largest. For example, status in the work force is measured on a nominal scale, since the codes assigned to the categories, although numeric, don't really mean anything. There is no order to *retired, in school, keeping house,* and *other*. Place of birth, hair color, and favorite statistician are all examples of variables measured on a nominal scale.

Variables whose values indicate only order or ranking are said to be measured on an **ordinal scale**. Job satisfaction and job importance are examples of variables measured on an ordinal scale. There are limitations on what you can say about data values measured on an ordinal scale. You can't say that someone who has a job satisfaction rating of 1 (*very satisfied*) is twice as satisfied as someone with a rating of 2 (*moderately satisfied*). All you can conclude is that one person claims to be more satisfied than the other. You can't tell how much more. The variable *income91* described in Chapter 3 is also measured on an ordinal scale. That's because the income is grouped into 21 unequal categories. You can't tell exactly how much more one person earned than another.

In the Define Variable dialog box, select Scale as the Measurement alternative for variables measured on a ratio or an interval scale.

If you record people's actual annual incomes, you are measuring income on what is called a **ratio scale**. You can tell how much larger or smaller one value is compared with another. The distances between values are meaningful. For example, the distance between incomes of $20,000 and $30,000 is the same as the distance between incomes of $70,000 and $80,000. You can also legitimately compute ratios of two values. An income of $50,000 is twice as much as an income of $25,000. Age and years of education are both examples of variables measured on a ratio scale.

An **interval scale** is just like a ratio scale except that it doesn't have an absolute zero. You can't compute ratios between two values measured on an interval scale. The standard example of a variable measured on an interval scale is temperature. You can't say that a 40°F day is twice as warm as a 20°F day. Few variables are measured on an interval scale, and the distinction between interval and ratio scales is seldom, if ever, important in statistical analyses.

Although it is important to consider the scale on which a variable is measured, statisticians argue that Stevens' typology is too strict to apply to real world data (Velleman & Wilkinson, 1993). For example, an identification number assigned to subjects as they enter a study might appear to be measured on a nominal scale. However, if the numbers are assigned sequentially from the first subject to enter the study to the last, the identification number is useful for seeing whether there is a relationship between some outcome of the study and the order of entry of the subjects. If the outcome is a variable like how long it takes a subject to master a particular task, it's certainly possible that instructions have improved during the course of a study and later participants fare better than earlier ones.

It's an oversimplification to conclude that the measurement scale dictates the statistical analyses you can perform. The questions that you want to be answered should direct the analyses. However, you should always make sure that your analysis is sensible. Using the computer, it's easy to calculate meaningless numbers, such as percentiles for place of birth or the median car color. In subsequent discussion, we'll occasionally refer to the scale of measurement of your data when describing various statistical techniques. These are not meant to be absolute rules but useful guidelines for performing analyses.

Mode, Median, and Arithmetic Average

The mode, median, and arithmetic average are the most commonly reported measures of central tendency. In Chapter 3, you saw how to compute the mode and median. You calculate the **mode** by finding the most frequently occurring value. The mode, since it does not require that the values of a variable have any meaning, is usually used for variables measured on a nominal scale. The mode is seldom reported alone. It's a useful statistic to report together with a frequency table or bar chart. You can easily find fault with the mode as a measure of what is typical. Even accompanied by the percentage of cases in the modal category or categories, it tells you very little.

If you are summarizing a variable whose values can be ranked from smallest to largest, the median is a more useful measure of central tendency. You calculate the **median** by sorting the values for all cases and then selecting the middle value. A problem with the median as a summary measure is that it ignores much of the available information. For example, the median for the five values 28, 29, 30, 31, and 32 is 30. For the five values 28, 29, 30, 98, and 190, it is also 30. The actual amounts by which the values fall above and below the median are ignored. The high values in the second example have no effect on the median.

The most commonly used measure of central tendency is the **arithmetic mean**, also known as the average. (For a sample, it's denoted as $\bar{X}$.) The mean uses the actual values of all of the cases. To compute the mean, add up the values of all the cases and then divide by the number of cases. For example, the arithmetic mean of the five values 28, 29, 30, 98, and 190 is

$$\text{Mean} = \frac{28 + 29 + 30 + 98 + 190}{5} = 75 \qquad \textbf{Equation 4.1}$$

Don't calculate the mean if the codes assigned to the values of a variable are arbitrary. For example, average car manufacturer and average religion don't make sense, since the codes are not meaningful.

Can I use the mean for variables that have only two values? Many variables, such as responses to yes/no or agree/disagree questions, have two values. If a variable has only two values, coded as 0 or 1, the arithmetic mean tells you the proportion of cases coded 1. For example, if 5 out of 10 people answered yes to a question and the coding scheme used is 0=no, 1=yes, the arithmetic mean is 0.50. You know that 50% of the sample answered yes. ■ ■ ■

Comparing Mean and Median

Figure 4.1 contains descriptive statistics from the Frequencies procedure for the age and education variables.

You can obtain these statistics using the Frequencies procedure, as discussed in Chapter 3. In the Frequencies Statistics dialog box (see Figure 3.13), select Mean, Median, and Mode.

Figure 4.1 Mean, median, and mode for age and education

Statistics

	N		Mean	Median	Mode
	Valid	Missing			
AGE Age of Respondent	1495	5	46.23	43.00	28[1]
EDUC Highest Year of School Completed	1496	4	13.04	12.00	12

[1] Multiple modes exist. The smallest value is shown

You see that the average age of the participants of the General Social Survey is 46.23 years. The median is somewhat lower, 43 years. The average number of years of school completed is 13.04, and the median is 12. For both of these variables, the arithmetic mean is somewhat greater than the median. The reason is that both of these variables have a "tail" toward larger values. Remember the histogram for age from Chapter 3. Since the General Social Survey is restricted to adults at least 18 years of age, young ages do not occur in the data. There is no such restriction for older ages. The older ages drive up the mean, which is based on all data values. They have no effect on the median, since it depends only on the values of the middle cases. If the distribution of data values is exactly symmetric, the mean and median are equal. If the distribution has a long tail (that is, the distribution is skewed), the mean is larger than the median if the tail extends toward larger values, and smaller than the median if the tail extends toward smaller values. In this example, the differences between the mean and the median are not very large. This is not always true.

Consider the following example. You ask five employees of a company how much money they earned in the past year. You get the following replies: $45,000, $50,000, $60,000, $70,000, and $1,000,000. The average salary received by these five people is $245,000. The median is $60,000. The arithmetic mean doesn't really represent the data well. The CEO salary makes the employees appear much better compensated than they really are. The median better represents the employees' salaries.

Measures of central tendency that are less affected by extreme values are discussed in Chapter 6.

Whenever you have data values that are much smaller or larger than the others, the mean may not be a good measure of central tendency. It is unduly influenced by extreme values (called **outliers**). In such a situation, you should report the median and mention that some of the cases had extremely small or large values.

Measures of Variability

Measures of central tendency don't tell you anything about how much the data values differ from each other. For example, the mean and median are both 50 for these two sets of ages: 50, 50, 50, 50, 50 and 10, 20, 50, 80, 90. However, the distribution of ages differs markedly between the two sets. **Measures of variability** attempt to quantify the spread of observations. We'll discuss the most common measures of variability in this chapter. Chapter 6 contains discussion of additional measures.

To obtain this output, from the menus choose:

Statistics
Summarize ▸
Descriptives...

Select the variables age and educ, as shown in Figure 4.5. In the Descriptives Options dialog box, select the variance, as shown in Figure 4.6.

Figure 4.2 Descriptive statistics for age and education

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
AGE Age of Respondent	1495	18	89	46.23	17.42	303.386
EDUC Highest Year of School Completed	1496	0	20	13.04	3.07	9.450
Valid N (listwise)	1491					

Range

The **range** is the simplest measure of variability. It's the difference between the largest and the smallest data values. Since the values for a nominal variable can't be meaningfully ordered from largest to smallest, it doesn't make sense to compute the range for a nominal variable such as status in the work force. In Figure 4.2, you see that for the variable *age*, the smallest value (labeled *Minimum*) is 18. The largest value (labeled *Maximum*) is 89. The range is 71 years. A large value for the range tells you that the largest and smallest values differ substantially. It doesn't tell you anything about the variability of the values between the smallest and the largest.

You can use the Explore procedure, described in Chapter 6, to calculate the range and the interquartile range.

A better measure of variability is the **interquartile range**. It is the distance between the 75th and 25th percentile values. The interquartile range, unlike the ordinary range, is not easily affected by extreme values. In Chapter 3, you calculated the 25th percentile for the age variable as 32 years, the 75th percentile as 59. The interquartile range is therefore 27, the difference between the two.

Variance and Standard Deviation

The most commonly used measure of variability is the **variance**. It is based on the squared distances between the values of the individual cases and the mean. To calculate the squared distance between a value and the mean, just subtract the mean from the value and then square the difference. (One reason you must use the squared distance instead of the distance is that the sum of distances around the mean is always 0.) To get the variance, sum up the squared distances from the mean for all cases and divide the sum by the number of cases minus 1.

The formula for computing the variance of a sample (denoted s^2) is

You can obtain the variance by selecting Options in the Descriptives dialog box. See Figure 4.6.

$$\text{Variance} = \frac{\text{sum of squared distances from the mean for all cases}}{(\text{number of cases} - 1)}$$

Equation 4.2

For example, to calculate the variance of the numbers 28, 29, 30, 98, and 190, first find the mean. It is 75. The sample variance is then

$$s^2 = \frac{(28-75)^2 + (29-75)^2 + (30-75)^2 + (98-75)^2 + (190-75)^2}{4}$$
$$= 5{,}026$$

Equation 4.3

If the variance is 0, all of the cases have the same value. The larger the variance, the more the values are spread out. In Figure 4.2, the variance for the age variable is 303.39 square years; for the education variable it is 9.45 square years. To obtain a measure in the same units as the original data, you can take the square root of the variance and obtain what's known as the **standard deviation**. Again in Figure 4.2, the standard deviation (labeled *Std Dev*) for the age variable is 17.42 years; for the education variable, it is 3.07 years.

Why divide by the number of cases minus 1 when calculating the sample variance, rather than by the number of cases? You want to know how much the data values vary around the population mean, but you don't know the value of the population mean. You have to use the sample mean in its place. This makes the sample values have less variability than they would if you used the population mean. Dividing by the number of cases minus 1 compensates for this. ■ ■ ■

The Coefficient of Variation

The magnitude of the standard deviation depends on the units used to measure a particular variable. For example, the standard deviation for age measured in days is larger than the standard deviation of the same ages measured in years. (In fact, the standard deviation for age in days is 365.25 times the standard deviation for age in years.) Similarly, a variable like salary will usually have a larger standard deviation than a variable like height.

The **coefficient of variation** expresses the standard deviation as a percentage of the mean value. This allows you to compare the variability of different variables. To compute the coefficient of variation, just divide the standard deviation by the mean and multiply by 100. (Take the absolute value of the mean if it is negative.)

$$\text{coefficient of variation} = \frac{\text{standard deviation}}{|\text{mean}|} \times 100$$

Equation 4.4

The coefficient of variation equals 100% if the standard deviation equals the mean. The coefficient of variation for the age variable is 37.68%. For the education variable, the coefficient of variation is 23.54%. Compared to their means, age varies more than education.

13

Standard Scores

The mean often serves as a convenient reference point to which individual observations are compared. Whenever you receive an examination back, the first question you ask is, How does my performance compare with the rest of the class? An initially dismal-looking score of 65% may turn stellar if that's the highest grade. Similarly, a usually respectable score of 80 loses its appeal if it places you in the bottom quarter of the class. If the instructor just tells you the mean score for the class, you can only tell if your score is less than, equal to, or greater than the mean. You can't say how far it is from the average unless you also know the standard deviation.

For example, if the average score is 70 and the standard deviation is 5, a score of 80 is quite a bit better than the rest. It is two standard deviations above the mean. If the standard deviation is 15, the same score is not very remarkable. It is less than one standard deviation above the mean. You can determine the position of a case in the distribution of observed values by calculating what's known as a standard score, or *z* score.

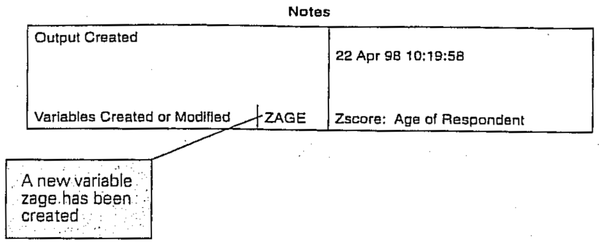
To calculate the standard score, first find the difference between the case's value and the mean and then divide this difference by the standard deviation.

$$\text{standard score} = \frac{\text{value} - \text{mean}}{\text{standard deviation}} \qquad \textbf{Equation 4.5}$$

A standard score tells you how many standard deviation units a case is above or below the mean. If a case's standard score is 0, the value for that case is equal to the mean. If the standard score is 1, the value for the case is one standard deviation above the mean. If the standard score is –1, the value for the case is one standard deviation below the mean. (For many types of distributions, including the normal distribution discussed in Chapter 10, most of the observed values fall within plus or minus two standard deviations of the mean.) The mean of the standard scores for a variable is always 0, and their standard deviation is 1.

You can use the Descriptives procedure in SPSS to obtain standard scores for your cases and to save them as a new variable. Figure 4.3 shows the notes from the Descriptives procedure that indicate that a new variable, the standard score for age, has been created. In addition, a new variable, *zage*, has been saved in the Data Editor, containing the standard scores for age (see Figure 4.4).

Figure 4.3 Descriptive statistics in the Viewer



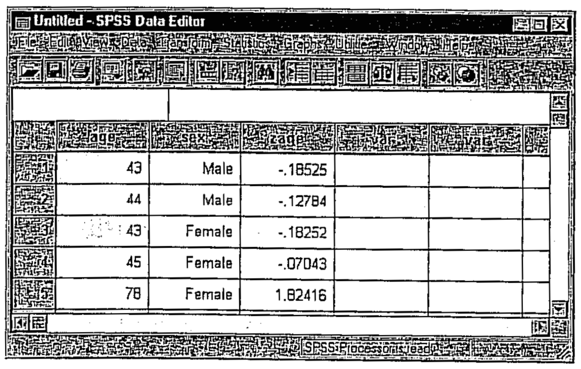
Notes

Output Created		22 Apr 98 10:19:58
Variables Created or Modified	ZAGE	Zscore: Age of Respondent

A new variable zage has been created

Figure 4.4 Data Editor with standard scores saved as a new variable

To save standardized scores, select Save standardized values as variables in the Descriptives dialog box, as shown in Figure 4.5.



Untitled - SPSS Data Editor
File Edit View Data Transform Statistics Graphs Utilities Window Help
SPSS Processor is ready

	age	sex	zage	var	var
1	43	Male	-.18525		
2	44	Male	-.12784		
3	43	Female	-.18252		
4	45	Female	-.07043		
5	78	Female	1.82416		

You see that the first case has an age of 43. From the standard score, you know that the case has an age less than average, but not very much. The age for the case is less than a quarter of a standard deviation below the mean. The fifth case has an observed age of 78, which is almost two standard deviations above the mean.

Standard scores allow you to compare relative values of several different variables for a case. For example, if a person has a standard score of 2 for income, and a standard score of –1 for education, you know that the person has a larger income than most and somewhat fewer years of education. You couldn't meaningfully compare the original values, since the variables all have different units of measurement, different means, and different standard deviations.

Summary

How can you summarize the values of a variable?

- Scales of measurement tell you about the properties of the values of a variable.
- The arithmetic mean is calculated by summing the values of a variable and dividing by the number of cases. Unlike the median and mode, the arithmetic mean uses all of the values of a variable.
- The median is a better measure of central tendency than the mean when there are data values that are far removed from the rest.
- The variance is a measure of the spread of data values around the mean. The coefficient of variation tells you the percentage the standard deviation is of the mean.
- A standardized score tells you how many standard deviation units above or below the mean an observation is.

Looking at Distributions 6

What additional displays are useful for summarizing the distribution of a variable for several groups?

- What is a stem-and-leaf plot?
- How does a stem-and-leaf plot differ from a histogram?
- What is a boxplot?
- What can you tell from the length of a box?
- How is the median represented in a boxplot?

Since most statistical analyses of data involve comparisons of groups, SPSS contains many procedures that help you to examine the distribution of values for individual groups of cases. In Chapter 5, you used the Means procedure to calculate descriptive statistics for education when the cases were subdivided on the basis of job satisfaction and gender. To examine each of the groups in more detail, you need to use the Explore procedure, which contains additional descriptive statistics as well as plots. That's what this chapter is about. (The statistics and displays described in this chapter are also useful for looking at the distribution of values for the entire sample.)

▶ This chapter continues to use the *gssft.sav* data file. For instructions on how to obtain the Explore output shown in the chapter, see "How to Explore Distributions" on p. 102.

Age and Job Satisfaction

In Chapter 5, you looked at the relationship between education and job satisfaction for full-time workers. You found that the average years of education did not differ much among people in the different categories of job satisfaction. Now you'll consider the relationship between age and job satisfaction for the same group of cases. You'll be able to examine the groups in considerably more detail. Consider first the descriptive statistics for age among workers in the different satisfaction categories.

To obtain these descriptive statistics, from the menus choose:

Statistics
Summarize ▶
Explore...

Select the variables age and satjob, as shown in Figure 6.9.

Figure 6.1 Case Processing Summary

		Cases					
		Valid		Missing		Total	
	Job Satisfaction	N	Percent	N	Percent	N	Percent
Age of Respondent	Very satisfied	325	99.4%	2	.6%	327	100.0%
	Mod satisfied	319	99.7%	1	.3%	320	100.0%
	A little dissatisfied	74	100.0%	0	.0%	74	100.0%
	Very dissatisfied	26	100.0%	0	.0%	26	100.0%

From Figure 6.1, you see that there are 325 cases in the *very satisfied* group for whom age is available. The number of missing cases is 2. That means that 2 *very satisfied* cases do not have a valid value for the age variable. Since these cases represent only 0.6% of all the cases in the group, you don't have to worry about the effect of missing age values on your analysis. Age is available for almost all of the cases in the other groups as well. Note that the number of cases varies considerably among the four groups. Over 300 people classify themselves as *very satisfied* and a similar number as *moderately satisfied*. However, 74 people are *a little dissatisfied* and only 26 *very dissatisfied*. You'll have to be careful about what you say about the last two groups since they are based on small numbers of cases.

The Pivot Table Editor was used to hide some statistics and to rearrange the default table

Figure 6.2 Average age and job satisfaction

	Age of Respondent			
	Job Satisfaction			
	Very satisfied	Mod satisfied	A little dissatisfied	Very dissatisfied
Mean	41.50	39.49	40.26	38.58
5% Trimmed Mean	41.05	39.11	39.83	38.19
Median	40.00	39.00	38.00	36.50
Std. Deviation	11.54	10.89	10.72	9.91
Minimum	19	20	23	22
Maximum	82	75	72	63
Range	63	55	49	41
Interquartile Range	15.50	16.00	14.25	17.00

The means of the ages, shown in Figure 6.2, range from a high of 41.5 years in the *very satisfied* group to 38.58 in the *very dissatisfied* group. The median ages are slightly less in all of the groups because the age distributions have tails toward larger values. As you've learned in Chapter 4, one of the shortcomings of the arithmetic mean is that very large or very small values in the data can change its value substantially. The trimmed mean avoids this problem. A **trimmed mean** is calculated just like the usual arithmetic mean, except that a designated percentage of the cases with the largest and smallest values are excluded. This makes the trimmed mean less sensitive to outlying values. The 5% trimmed mean excludes the 5% largest and the 5% smallest values. It's based on the 90% of cases in the middle. The trimmed mean provides an alternative to the median when you have some data values that are far removed from the rest.

In Figure 6.2 you see that the 5% trimmed mean doesn't differ much from the usual mean. That's not surprising since the largest age in all groups (*Maximum*) is 82 and the smallest age (*Minimum*) is 19. The person with an age of 82 is in a group with 325 cases. You'd need a surviving Roman warrior to have a real effect on the mean.

Again from Figure 6.2, you see that the standard deviation ranges from 11.54 years in the *very satisfied* group to 9.91 in the *very dissatisfied* group. The range is largest in the very satisfied group since it contains the 82 year old. The range is based on the largest and smallest values so a single outlying value can have a large effect on the range. Unlike the ordinary range, the interquartile range is not easily affected by extreme values, since the bottom 25% and the top 25% of the data values are excluded from its computation. It's the difference between the 75th and the 25th percentile values. In Figure 6.2, you see that the interquartile ranges, are fairly similar in all of the groups.

Identifying Extreme Values

Since many statistics are affected by data values that are much smaller or larger than most, it's always important to examine your data to see if extreme values are present. You can obtain from the Explore procedure a list of the cases with the five largest and the five smallest values in each group. Always check any suspicious values to make sure they are not the result of an error in data recording or entry. If you find a mistake, it's easy to change the values for a case using the Data Editor. If the extreme values are correct, make sure to select summary measures that are not unduly affected by these outliers.

To obtain extreme values, select Statistics in the Explore dialog box. Then select Outliers, as shown in Figure 6.10.

Figure 6.3 Outliers from the very satisfied group

	Job Satisfaction			Case Number	Value
Age of Respondent	Very satisfied	Highest	1	344	82
			2	223	78
			3	263	77
			4	401	77
			5	208	73
		Lowest	1	173	19
			2	364	20
			3	714	20
			4	320	21
			5	665	21

At 82 years, case 344 is the oldest respondent in the group

Figure 6.3 shows for the *very satisfied* group the cases with the five largest and smallest ages. The column labeled *Case Number* contains the sequence number in the file for each case. That makes it easier for you to track the suspicious values. (SPSS can also show a name or any other identifier to label the cases.) You see that case 344 is the oldest, at age 82. Case 173 is the youngest, at age 19. Neither of these values is usual. Since only five cases with the smallest and largest values are listed, it's possible that not all cases with those values are listed. For example, there may be more than one 73-year-old or more than two 21-year-olds. If that's true, a note is printed beneath the table. Just because a value is included in the extreme value table doesn't mean it really is an outlier. It's only one of the largest or smallest values; you must decide if it is really unusual.

Percentiles

Using Explore, you can also obtain percentiles for each of the groups. Figure 6.4 shows the percentiles for age for the satisfaction subgroups.

To obtain percentiles, select Statistics in the Explore dialog box. Then select Percentiles. (See Figure 6.10.)

Figure 6.4 Age percentiles

			Percentiles						
		Job Satisfaction	5	10	25	50	75	90	95
Weighted Average (Definition 1)	Age of Respondent	Very satisfied	24.00	27.00	33.50	40.00	49.00	57.40	61.00
		Mod satisfied	24.00	26.00	31.00	39.00	47.00	55.00	60.00
		A little dissatisfied	25.00	27.00	32.75	38.00	47.00	55.50	60.25
		Very dissatisfied	24.10	28.00	30.00	36.50	47.00	51.60	59.50
Tukey's Hinges	Age of Respondent	Very satisfied			34.00	40.00	49.00		
		Mod satisfied			31.00	39.00	47.00		
		A little dissatisfied			33.00	38.00	47.00		
		Very dissatisfied			30.00	36.50	47.00		

10% of cases are 27 or younger in the very satisfied group

10% are 57.4 or older in the very satisfied group

Two sets of percentiles are shown: The first set is obtained using a method called *Weighted Average*. Other ways can be used, but most of the time they give pretty much the same results. You see that for the *very satisfied* group 10% of the cases are 27 years of age or younger (the 10th percentile) and 10% are 57.4 or older (the 90th percentile). The percentiles are comparable for the four satisfaction groups.

The second part of the percentile table shows *Tukey's Hinges*, which are quartiles (values that divide the sorted cases into four equal groups) calculated using a slightly different method than the weighted average percentiles.

How can you get different numbers for the same percentiles? Percentiles don't have a single, unique definition. For example, consider the eight numbers 25 26 27 27 27 27 30 31. What's the 25th percentile? Any number between 26 and 27 is a plausible value. One definition of percentiles gives the answer 26.5, since that's the average of 26 and 27, the interval within which the percentile falls. Another definition results in the answer 26, since that's the first value for which the cumulative percentage is equal to or greater than 25%.

For small data sets, especially when several cases have the same values, different percentiles may have the same value. For the previous example, it's possible for percentiles greater than the 25th and less than the 75th to have the value 27. For small data sets, percentile values can vary a lot for samples from the same population, so you shouldn't place too much confidence in their exact values. (You also shouldn't worry about where the "equal" goes. That is, whether 25% of the cases have values less than the 25th percentile, or whether 25% of the cases have values less than or *equal* to the 25th percentile. Statistical software packages implement arbitrary rules about where the "equal" goes.)

Plots

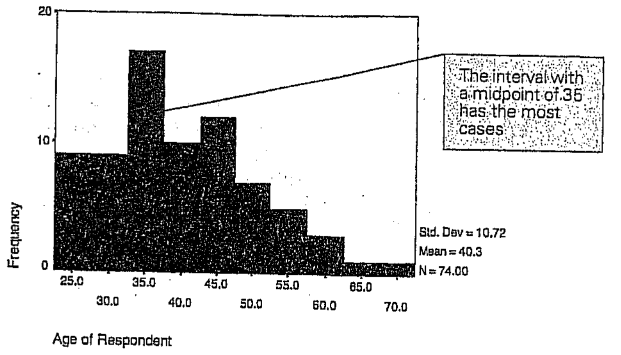
One of the easiest ways to see the distributions of your variables is literally with a picture. The Explore procedure provides several plots that let you evaluate the shape of a distribution. From these plots, you see how often different values of a variable occur in your data. As you will see in Part 3, your choice of the statistical analysis for a particular problem depends on the assumptions you are willing to make about the distributions of the variables of interest. That's why it's important to examine them.

Histograms and Stem-and-Leaf Plots

The Explore procedure can produce separate histograms for groups of cases. The histograms are identical to those produced by the Frequencies procedure, as described in Chapter 3. Figure 6.5 shows a histogram of age for the people *a little dissatisfied* with their jobs.

Figure 6.5 Age histogram for a little dissatisfied

To obtain a histogram, select Plots in the Explore dialog box. Then select Histogram, as shown in Figure 6.11.



The interval with a midpoint of 35 has the most cases

Note the main peak, centered at 35 years of age, with a smaller peak at 45. From the histogram you can only tell the number of cases in each of the intervals: you don't know the actual values of the cases. For example, for the interval centered around 50, all of the cases could be 50 years old, or they could be any combination of 48-, 49-, 50-, 51- and 52-year-olds. From the histogram, you see that the distribution of age values in the groups is not symmetric. There is a tail toward larger values. You know that's because only adults are included in the General Social Survey.

What kinds of things should I look for in a histogram? You already know that you should look for cases with values very different from the rest. In fact, if there are such cases, they can cause most of your data values to bunch in one or two bars of the histogram, since the horizontal axis of the histogram is selected so that all data values can be shown. You should see also whether the distribution is symmetric, since many of the statistical procedures described in Part 3 require that the distribution be more or less symmetric.

You should also look for separate clumps of data values. For example, if young men and mature women made up most of the *a little dissatisfied* group, you would see a bunch of cases with values in the 20's, perhaps, and another bunch of cases with values in the 60's. There wouldn't be many cases in between. That's an important finding, since then you know that a mean age of 40-something for the *a little dissatisfied group* is meaningless. It doesn't represent the data well. In this situation, you'd want to analyze the data for men and for women separately.

A stem-and-leaf plot is a display very much like a histogram. However, more information about the actual data values is preserved. Consider Figure 6.6, which is a stem-and-leaf plot for age in the group *a little dissatisfied* with their jobs. It looks like a histogram, because the length of each line corresponds to the number of cases in the interval. However, the cases are represented with different symbols. Each observed value is divided into two components—the leading digit or digits, called the stem, and a trailing digit, called the leaf. For example, the value 23 has a stem of 2 and a leaf of 3.

In a stem-and-leaf plot, each row corresponds to a stem and each case is represented by its leaf. More than one row can have the same stem. For example, in Figure 6.6 each stem is subdivided into two rows.

Figure 6.6 Stem-and-leaf plot of age for a little dissatisfied

To obtain a stem-and-leaf plot, select Plots in the Explore dialog box. Then select Stem-and-leaf, as shown in Figure 6.11.

```
Age of Respondent Stem-and-Leaf Plot for
SATJOB= A little dissatisfied

 Frequency    Stem &  Leaf
     2.00        2 .  33
    13.00        2 .  5556777899999
     7.00        3 .  0123334
    18.00        3 .  555566666777788899
     7.00        4 .  0012234
    13.00        4 .  5556666677888
     5.00        5 .  02223
     5.00        5 .  55679
     3.00        6 .  013
     1.00 Extremes    (>=72)

 Stem width:  10
 Each leaf:        1 case(s)
```

Multiply stem by stem width and add leaf values to get actual data values (60, 61, and 63)

Look at the row with the stem of 6 in Figure 6.6. The three leaves are 0, 1, and 3. What does this mean? In order to translate the stem-and-leaf values into actual numbers, you must look at the stem width given below the plot. In this case it's 10. You multiply each stem value by 10 and then add it to the leaf to get the actual value. The resulting age values are 60, 61, and 63. If the stem width were 100, you would multiply each stem by 100 and each leaf by 10 before adding them together. The values for the indicated row would be 600, 610, and 630.

If there are few values of the stem (for example, if most cases are in one or two decades of age), each stem can be subdivided into more than two rows. Consider, for example, Figure 6.7, which is a stem-and-leaf plot for years of education for all *very dissatisfied* people, regardless of their status in the work force.

Why all of a sudden are we looking at all people instead of full-time workers? The data values determine the type of stem-and-leaf plot that Explore makes. To illustrate this particular version of the plot, we had to look for a set of data that would generate it. Including all cases in the *very dissatisfied* group worked.

Figure 6.7 Stem-and-leaf plot of education for very dissatisfied

```
Highest Year of School Completed Stem-and-Leaf Plot for
SATJOB= Very dissatisfied

 Frequency    Stem &  Leaf
      .00        0 .
     4.00        0 .  8999
     7.00        1 .  0001111
    15.00        1 .  222222222222233
    13.00        1 .  4444444555555
     9.00        1 .  666666666
     5.00        1 .  88889

 Stem width:   10
 Each leaf:        1 case(s)
```

Four cases have 11 years of education

One case has 19 years of education

In Figure 6.7, the stem value 1 is subdivided into five rows—each representing two leaf values. The first row is for leaves of 0 and 1, the second row is for leaves of 2 and 3, the third for leaves of 4 and 5, the fourth for 6 and 7, and the last for 8 and 9. You see that the *very dissatisfied* group is made up of people of various educational levels. Having a college degree is no guarantee of job satisfaction.

How would you make a stem-and-leaf plot of a variable like income? For a variable like income, which has many digits, it's unwieldy and unnecessary to represent each case by the last digit. (Think of how many stems you would have!) Instead, you can look at income to the nearest thousand. For example, you can take a number like 25,323 and divide it into a stem of 2 and a leaf of 5. In this case, the stem is the ten thousands, and the leaf is the thousands. You no longer retain the entire value for the case, but that's not of concern, since income differences in the hundreds seldom matter very much. The Explore procedure always displays the stem width under the plot.

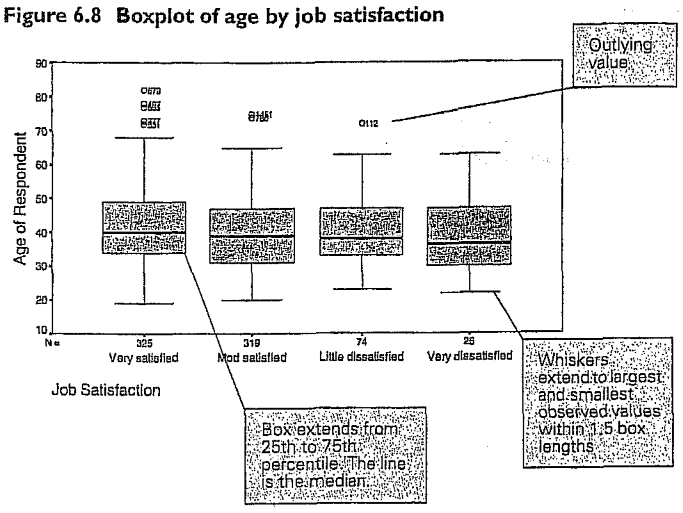
Boxplots

Another display that helps you visualize the distribution of a variable is the boxplot. It simultaneously displays the median, the interquartile range, and the smallest and largest values for a group. A boxplot is more compact than a histogram but doesn't show as much detail. For example, you can't tell if your distribution has a single peak or if there are intervals that have no cases.

You can use the Explore procedure to produce a display that contains boxplots for all the groups of interest. Consider Figure 6.8, which is an annotated boxplot of the age of the respondent for the four categories of job satisfaction.

To obtain this boxplot, select Plots in the Explore dialog box. Then select Factor levels together. (See Figure 6.11.)

Figure 6.8 Boxplot of age by job satisfaction



The lower boundary of the box represents the 25th percentile. The upper boundary represents the 75th percentile. (The percentile values known as Tukey's hinges are used to construct the box.) The vertical length of the box represents the interquartile range. Fifty percent of all cases have values within the box. The line inside the box represents the median. Note that the only meaningful scale in the boxplot is the vertical scale. All values are plotted on this scale. The width of a box doesn't represent anything.

In a boxplot, there are two categories of cases with outlying values. Cases with values between 1.5 and 3 box lengths from the upper or lower edge of the box are called **outliers** and are designated with (O). Cases with values more than 3 box-lengths from the upper or lower edge of the box are called **extreme values**. There aren't any such cases here, but if there were, they would be designated with asterisks (*). Lines are drawn from the edges of the box to the largest and smallest values that are outside the box but within 1.5 box lengths. (These lines are sometimes called **whiskers**, and the plot is sometimes called a **box-and-whiskers plot**.)

What can you tell about your data from a boxplot? From the median, you can get an idea of the typical value (the central tendency). From the length of the box, you can see how much the values vary (the spread or variability). If the line representing the median is not in the center of the box, you can tell that the distribution of your data values is not symmetric. If the median is closer to the bottom of the box than to the top, there is a tail toward larger values (this is also called **positive skewness**). If the line is closer to the top of the box, there is a tail toward small values (negative skewness). The length of the tail is shown by the length of the whiskers and the outlying and extreme points.

The Explore procedure also has more specialized charts and statistics for examining groups. These are discussed in Chapter 14 and Chapter 21.

In Figure 6.8, you see that the *very satisfied* group has the highest median age, though the differences among the groups are small. (Chapter 13 tests the hypothesis that the average age of people who are *very satisfied* is the same as the average age of those who are less than *very satisfied*.) The *very satisfied* and *moderately satisfied* groups have some large outliers. They are identified by case number in the plot. These are the satisfied old-timers who continue to work. If you specify a case label, the extreme and outlying points will be identified with this label.

What should I do if I find outliers and extremes on my boxplots? Use the case numbers to track down the data points and make sure the values are correct. If these points are the results of data entry or coding errors, correct them.

Interpretation Questions

Examine the printouts for the descriptive statistics.

1. Using Output 4.1, look for any obvious errors or problems in the data. What will you look for?

2. Name the nominal variables in Output 4.2. Can you interpret the mean and standard deviation? Explain.

3. Using Output 4.2: a) How many participants are there all together? b) How many have complete data (nothing missing)? c)What percentage took algebra 1 in high school? d) What is the range of father's education scores? Does this agree with the codebook?

Outputs and Interpretations

```
GET
  FILE='A:\hsbdata.sav'.
EXECUTE .
```

These are the instructions or syntax that you produced to retrieve the hsbdata file from the disk.

Output 4.1: Descriptives With Errors

Syntax from Problem 1. If you save it, you can use it later to modify and rerun the analysis. See Appendix C.

Syntax for the mean, standard deviation, minimum, and maximum for all variables

```
DESCRIPTIVES
  VARIABLES=alg1 alg2 calc ethnic faed gend geo grades id maed mathach mathgr
  mosaic q01 q02 q03 q04 q05 q06 q07 q08 q09 q10 q11 q12 q13 trig visual
  /STATISTICS=MEAN STDDEV MIN MAX .
```

Interpretation of Output 4.1

The Output provides the number of subjects (*N*), the lowest and highest score, mean or average, and standard deviation for each variable. At the beginning of your data analysis, check to make sure that all means seem reasonable (given the information in your codebook), and check to see that the minimum and maximum are within the appropriate range for each variable. For example, note in the codebook that *alg1* has to be **0 = not taken or 1 = taken** so the minimum should be 0 and maximum 1. If not, you have an error to correct before proceeding. Did you find the two errors in the data? You will correct them in **Problem 2**. Note from the bottom of Output 4.1 that the valid number (*N*) of observations/subjects (listwise) is 67 rather than 75, the number of participants in the data file. This is because the listwise *N* only includes the persons with *no* missing data on any variable. Notice that several variables (e.g., ethnicity) each have a few participants missing.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
algebra 1 in h.s.	75	0	1	.79	.41
algebra 2 in h.s.	75	0	1	.47	.50
calculus in h.s.	75	0	1	.11	.31
ethnicity	73	1	4	1.77	1.02
father's education	73	2	10	4.73	2.83
gender	75	1	2	1.55	.50
geometry in h.s.	75	0	1	.48	.50
grades in h.s.	75	2	8	5.68	1.57
identification	75	1	75	38.00	21.79
mother's education	75	2	10	4.11	2.24
math achievement	75	-1.67	23.67	12.5645	6.6703
math grades	75	0	1	.41	.50
mosaic, pattern test	75	-4.0	56.0	27.413	9.574
question 1	74	1	10	3.08	1.21
question 2	75	1	40	4.00	4.31
question 3	74	1	4	2.82	.90
question 4	74	1	4	2.16	.92
question 5	75	1	4	1.61	.97
question 6	75	1	4	2.43	.98
question 7	75	1	4	2.76	1.05
question 8	75	1	4	1.95	.91
question 9	74	1	4	3.32	.76
question 10	75	1	4	1.41	.74
question 11	75	1	4	1.36	.75
question 12	75	1	4	3.00	.82
question 13	75	1	4	2.67	.79
trigonometry in h.s.	75	0	1	.27	.45
visualization score	75	-.25	14.75	5.2433	3.9120
Valid N (listwise)	67				

Remember, 0=not taken 1=taken

Sometimes you get misleading information. There is no average ethnicity.

This N is different because some of the data is missing.

These two are errors and need correcting.

N only includes the persons with no missing data on any variable.

It's good to check this against the codebook. Why do you think the first number is negative?

When variables are 1 and 0, the mean indicates the percent who had 1, i.e., 27% took trig.

algebra 1 in h.s.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	not taken	16	21.3	21.3	21.3
	taken	59	78.7	78.7	100.0
	Total	75	100.0	100.0	
Total		75	100.0		

algebra 2 in h.s.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	not taken	40	53.3	53.3	53.3
	taken	35	46.7	46.7	100.0
	Total	75	100.0	100.0	
Total		75	100.0		

geometry in h.s.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	not taken	39	52.0	52.0	52.0
	taken	36	48.0	48.0	100.0
	Total	75	100.0	100.0	
Total		75	100.0		

trigonometry in h.s.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	not taken	55	73.3	73.3	73.3
	taken	20	26.7	26.7	100.0
	Total	75	100.0	100.0	
Total		75	100.0		

calculus in h.s.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	not taken	67	89.3	89.3	89.3
	taken	8	10.7	10.7	100.0
	Total	75	100.0	100.0	
Total		75	100.0		

Output 5.3: Frequencies, Statistics, and Histograms

Syntax for the frequency distribution, descriptive statistics, and histograms

```
FREQUENCIES
  VARIABLES=mosaic visual grades mathach
  /PERCENTILES= 33 67
  /STATISTICS=STDDEV VARIANCE RANGE MEAN MEDIAN MODE SKEWNESS SESKW KURTOSIS
  SEKURT
  /HISTOGRAM  NORMAL.
```

Interpretation of Output 5.3

The output file provides all the requested statistics for the four variables as a group. Then the four frequency distributions and four histograms with the normal curve superimposed over them are given individually so you can visualize whether the frequency distribution (histogram) looks normal. However, visual inspection can be deceiving because distributions need only be approximately normal. In the statistics tables, note columns for the skewness and kurtosis of the four variables. Divide each of the statistics by its standard error. If the result is not more than 5.5 (which is approximately the .01 level) that skewness or kurtosis is *not* significantly different from normal. Note that, using this measure, none of the four variables is markedly skewed, but the distribution of the *mosaic* scores is too peaked; i.e., it has a positive kurtosis almost six times its standard error. You can see this visually in the histogram.

Notice also the 33rd and 67th percentile columns in the statistic tables. You could use these percentiles if you wanted to divide your participants into three approximately equal size groups such as low, medium, and high. You can see from Output 5.3 that the 33rd and 67th percentiles for *mosaic* are 24.04 and 29.50. Thus, the low mosaic group would have scores from lowest to 24.04, the medium group from 24.04 to 29.50; and the high achievement group from 29.50 to highest. This could be done using the **Recode** command described in Assignment C.

The average, middle, and most frequent score.

Common measure of the variability of the scores.

The standard deviation squared.

Statistics

	N		Mean	Median	Mode	Std. Deviation	Variance	Range
	Valid	Missing						
visualization score	75	0	5.2433	4.7500	1.00[a]	3.9120	15.3040	15.00
mosaic, pattern test	75	0	27.413	27.000	25.0[a]	9.574	91.658	60.0
grades in h.s.	75	0	5.68	6.00	7	1.57	2.46	6
math achievement	75	0	12.5645	13.0000	14.33	6.6703	44.4930	25.33

a. Multiple modes exist. The smallest value is shown.

The highest minus the lowest score.

Statistics

	Skewness		Kurtosis		Percentiles	
					33.0000	67.0000
	Statistic	Std. Error	Statistic	Std. Error	Statistic	Statistic
visualization score	.536	.277	-.398	.548	3.5000	6.4600
mosaic, pattern test	.529	.277	3.106	.548	24.040	29.500
grades in h.s.	-.332	.277	-.763	.548	5.00	7.00
math achievement	.044	.277	-.940	.548	9.0000	15.5870

Determines if the curve is nonsymmetrical.

Tells the shape of a curve (flat or peaked).

Is the statistic 2.5 times greater than the standard error?

Divided students into three groups: Low, middle, and high thirds.

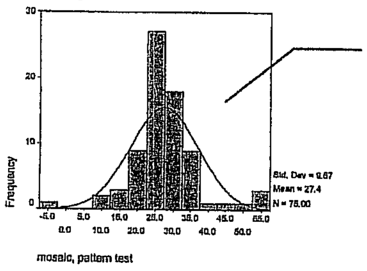
MORGAN George A. and Griego V. ORLANDO. 1998. *Easy Use and Interpretation of SPSS for Windows: Answering Research Questions With Statistics*. Mahwah: Lawrence Erlbaum Associates.

mosaic, pattern test

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	-4.0	1	1.3	1.3	1.3
	-4.0	1	1.3	1.3	2.7
	11.0	1	1.3	1.3	4.0
	13.0	1	1.3	1.3	5.3
	13.5	1	1.3	1.3	6.7
	16.0	1	1.3	1.3	8.0
	17.5	1	1.3	1.3	9.3
	18.0	1	1.3	1.3	10.7
	20.0	1	1.3	1.3	12.0
	20.5	2	2.7	2.7	14.7
	22.0	4	5.3	5.3	20.0
	22.5	2	2.7	2.7	22.7
	23.0	4	5.3	5.3	28.0
	23.5	2	2.7	2.7	30.7
	24.0	2	2.7	2.7	33.3
	24.5	3	4.0	4.0	37.3
	25.0	5	6.7	6.7	44.0
	26.0	3	4.0	4.0	48.0
	26.5	1	1.3	1.3	49.3
	27.0	5	6.7	6.7	56.0
	27.5	1	1.3	1.3	57.3
	28.0	3	4.0	4.0	61.3
	28.5	1	1.3	1.3	62.7
	29.0	2	2.7	2.7	65.3
	29.5	2	2.7	2.7	68.0
	30.0	2	2.7	2.7	70.7
	30.5	2	2.7	2.7	73.3
	31.0	4	5.3	5.3	78.7
	32.0	1	1.3	1.3	80.0
	32.5	1	1.3	1.3	81.3
	33.0	3	4.0	4.0	85.3
	34.0	1	1.3	1.3	86.7
	35.0	1	1.3	1.3	88.0
	35.5	1	1.3	1.3	89.3
	36.0	1	1.3	1.3	90.7
	37.0	1	1.3	1.3	92.0
	41.0	1	1.3	1.3	93.3
	44.0	1	1.3	1.3	94.7
	51.5	1	1.3	1.3	96.0
	53.0	1	1.3	1.3	97.3
	56.0	2	2.7	2.7	100.0
	Total	75	100.0	100.0	
Total		75	100.0		

Do you know what this means?

Histogram

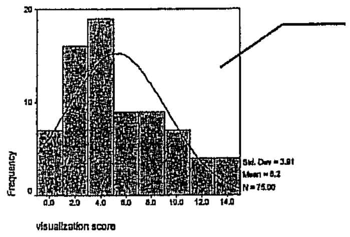


This is not normally distributed. The kurtosis is high.

visualization score

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	-.25	7	9.3	9.3	9.3
	1.00	10	13.3	13.3	22.7
	2.25	5	6.7	6.7	29.3
	2.50	1	1.3	1.3	30.7
	3.50	7	9.3	9.3	40.0
	3.75	2	2.7	2.7	42.7
	4.75	10	13.3	13.3	56.0
	5.00	2	2.7	2.7	58.7
	6.00	6	8.0	8.0	66.7
	6.50	1	1.3	1.3	68.0
	7.25	5	6.7	6.7	74.7
	8.50	2	2.7	2.7	77.3
	8.75	2	2.7	2.7	80.0
	9.50	1	1.3	1.3	81.3
	9.75	6	8.0	8.0	89.3
	11.00	4	5.3	5.3	94.7
	13.50	2	2.7	2.7	97.3
	14.75	2	2.7	2.7	100.0
	Total	75	100.0	100.0	
Total		75	100.0		

Histogram

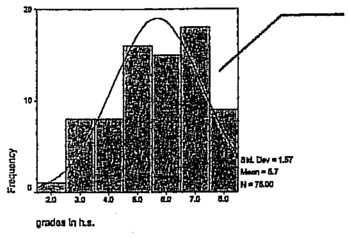


This is approximately normally distributed.

grades in h.s.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	mostly D	1	1.3	1.3	1.3
	half CD	8	10.7	10.7	12.0
	mostly C	8	10.7	10.7	22.7
	half BC	16	21.3	21.3	44.0
	mostly B	15	20.0	20.0	64.0
	half AB	18	24.0	24.0	88.0
	mostly A	9	12.0	12.0	100.0
	Total	75	100.0	100.0	
Total		75	100.0		

Histogram



This is approximately normally distributed.

BLACK, Thomas, R. 1999. *Doing Quantitative Research in the Social Sciences. An Integrated Approach to Research Design, Measurement and Statistics*. London: Sage.

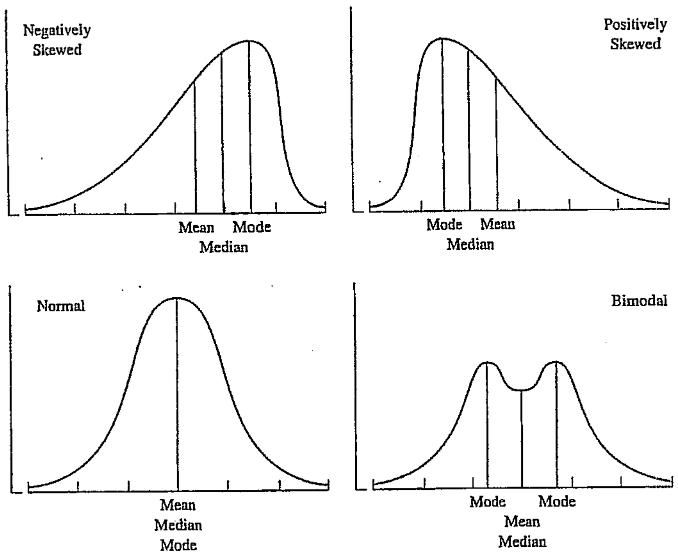


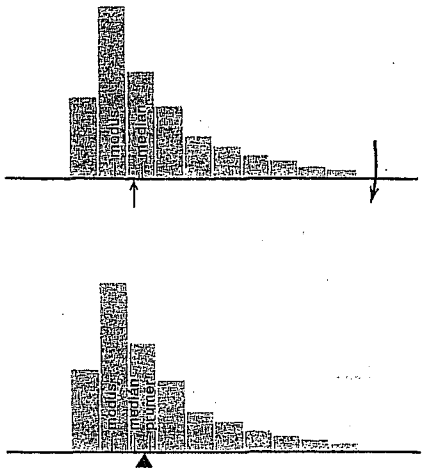
FIGURE 12.20
Mean, median and mode for a variety of distributions

WONNACOT Thomas H a Ronald J. WONNACOT. 1993. *Statistika pro obchod a hospodářství*. Praha: Victoria Publishing.

Co když je ale rozdělení nesymetrické? Na obr. 2-4 a) vidíme, že rozdělení dlouze klesá vpravo. Bude v tomto případě např. medián totožný s modem? Vzhledem k velkému množství pozorování v pravé části je zřejmé, že pro rozdělení výběru na dvě poloviny je třeba, aby medián ležel více vpravo od vrcholu rozdělení, tedy i vpravo od modu.

A kde se bude nacházet průměr? Někde blízko mediánu? V obr. 2-2 a) vidíme, co se stane, když se pokusíme vyvážit rozdělení v mediánu. Na každé straně je stejný počet pozorování, ale pozorování napravo dosahují dále, a celé rozdělení se tedy naklání doprava. K nalezení skutečného těžiště je třeba jít dále doprava, jak vidíme na obr. 2-4 b). Průměr se tedy nachází napravo od mediánu.

Jaké jsou tedy závěry z nesymetrického rozdělení funkce? **Vzhledem k modu leží medián ve směru delší části rozdělení a průměr ještě dále v tomto směru.**



OBRÁZEK 2-4
Modus, medián a průměr v rozdělení klesajícím vpravo.
a) Medián je vpravo od modu.
b) Těžiště (průměr) je vpravo od mediánu. (Rozdělení nebude v rovnováze, umístíme-li těžiště do mediánu, neboť pozorování vpravo převáží.

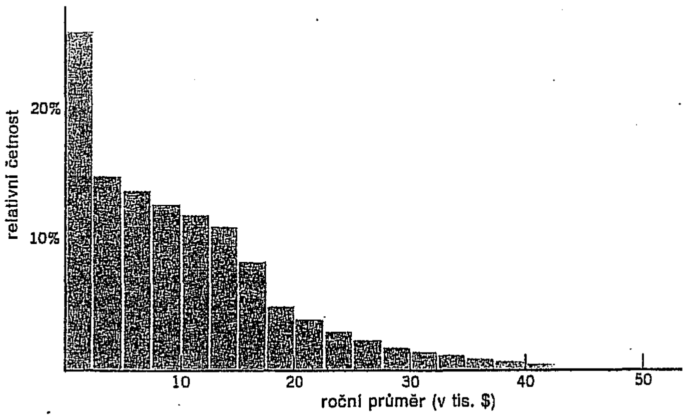
F—KTERÁ CHARAKTERISTIKA POLOHY JE NEJVHODNĚJŠÍ — MODUS, MEDIÁN NEBO PRŮMĚR?

Pro některé účely je vhodná jedna charakteristika, pro jiné jiná. Pro ilustraci se podívejme na rozdělení příjmů 78 miliónů Američanů v roce 1975 znázorněné na obr. 2-5.

Modus se nachází v blízkosti 0, a ukazuje nám pouze, že nejvíce lidí je prakticky bez příjmů – nezaměstnaní a důchodci. Největší příjmy jsou zastoupeny v mnohem menším počtu v rozmezí 2 až 40 tisíc – tento jev však není modem vůbec postižen. Budeme-li používat modus, nebudeme moci srovnat příjmy v roce 1975 s mnohem nižšími příjmy v roce 1875! V tomto případě nám je modus k ničemu.

Medián se nachází v asi 8 tisících dolarech a tato hodnota je mnohem reprezentativnější – 50 % nad a 50 % pod. Možná je to nejlepší hodnota „typického“ amerického příjmu. Navíc je **resistentní**, tj. nereaguje na extrémní hodnoty jediného pozorování. Například, zvýšíme-li nejvyšší příjem desetkrát, medián se nezmění.

Konečně průměr je okolo 10 tisíc dolarů. Tato hodnota byla získána rovnocenným zahrnutím všech dolarů — dolarů žebráka i milionáře. To má své výhody i nevýhody — je to nejužitečnější měřítko pro berní úřad, neboť vyjadřuje celkový příjem (78 miliónů lidí × 10 tisíc dolarů = 780 miliard dolarů); přesto není tak dobrým měřítkem pro typický příjem jako medián, neboť se může značně změnit vychýlením jen jediného pozorování (jediným velmi vysokým nebo velmi nízkým příjmem), tj. není rezistentní jako medián.



OBRÁZEK 2-5
Příjmy amerických mužů, 1975. (Stat. Abst. of U.S., 1980, str. 462)

4. lekce
UMĚLÉ PROMĚNNÉ (modul TRANSFORM: procedury Recode, Compute, Count, Rank Cases) .

B Transforming and Selecting Data

SPSS includes a powerful set of facilities for transforming data values and selecting which cases should be analyzed. This appendix covers two general types of data manipulation: data transformation and case selection.

Data transformation procedures change the actual values of your variables or create new variables. For example, you can create a new variable that contains the natural log of an existing variable. **Case selection** procedures do not change data values but restrict the number of cases used in the analysis. For example, you can restrict your analysis to people who are married or who are holding full-time jobs.

SPSS also provides a number of advanced data manipulation utilities that are not used in this book and are not discussed here. These utilities are described, however, in the online Help system.

Data Transformations

Often you need to make modifications to your data before you can perform your analysis. For small changes, such as the urbanization of Bhutan in Chapter 8, it is easy to enter the corrected value into the Data Editor. But suppose you want to take the natural log of several variables, each with 1500 cases, as you do for the analysis in Chapter 22? SPSS provides data transformation facilities to handle such tasks easily and accurately.

Data transformations affect the values of existing variables or create new variables. Transformations affect only the working data file; the changes do not become permanent unless you save the working data file to your disk.

Transformations at a Glance

This appendix describes the following transformations, available using the SPSS Data Editor's Transform menu:

Compute. Compute calculates data values according to a precise expression. With this option, you can do anything from set a variable to 0 for all cases to calculate an elaborate expression involving the values of other variables. You can assign the computed values to a new variable, or you can assign them to an existing variable (replacing the current values). You can also request that the computation be carried out selectively based on a conditional expression.

Recode. Recode assigns discrete values to a variable, based solely on the present values of the variable being recoded. You can assign the recoded values to the variable being recoded, or you can assign them to a new variable. You can also request that the computation be carried out selectively based on a conditional expression.

Automatic Recode. Automatic recode assigns successive integer codes—1, 2, 3, and so on—to a new variable, based on the existing codes of another variable. This saves you the effort of specifying how the recoding should be carried out.

The following options are also available on the Transform menu but are not discussed in this book. These transformations are described in the online Help system.

Random Number Seed. Lets you reproduce the pseudo-random numbers generated by SPSS for sampling and certain functions in the transformation language.

Count. Creates a new variable that counts for each case the number of times certain specified values occur in other variables. You can count, for example, the number of times that values of 1 or 2 occur in a group of existing variables.

Rank Cases. Creates rank scores, which show each case's rank among all the cases in the file according to the values of a particular variable.

Create Time Series. Creates new time series, containing functions such as the differences between successive cases, in a time series data file.

Replace Missing Values. Supplies nonmissing values to replace missing values, according to any of several functions that might provide plausible values.

Run Pending Transformations. Forces SPSS to execute transformations that are pending as a result of the Transform & Merge Options setting. (See "Delaying Processing of Transformations" below.)

Saving Changes

Bear in mind when transforming your data that you are only changing the working data file.

- ▶ To make the changes permanent, save the working data file to your hard disk.
- ▶ To discard the changes, exit SPSS (or open a new data file) without saving the working data file.

Delaying Processing of Transformations

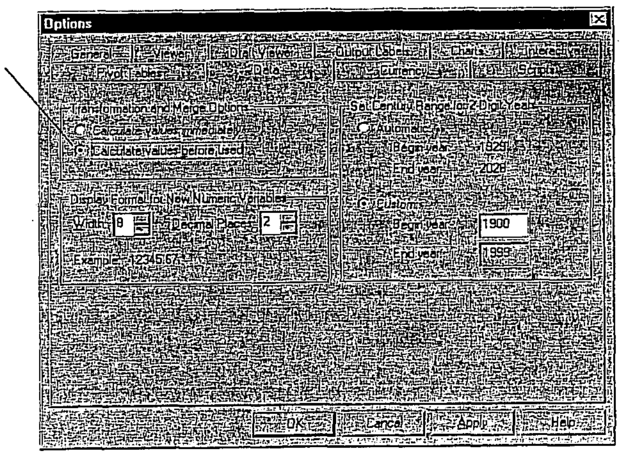
SPSS normally executes transformation commands as soon as you request them. However, since transformations can take several minutes to execute for a very large data file, there are times when you want to enter a dozen or more transformation commands one after another and then let the computer process them all at once.

- ▶ To prevent SPSS from processing transformations immediately, from the menus choose:

  Edit
    Options...

- ▶ In the tabbed SPSS Options dialog box, click the Data tab. This displays the SPSS Options Data tab, as shown in Figure B.1

Figure B.1 SPSS Options dialog box



Transformation and Merge Options
○ Calculate values immediately
● Calculate values before used

Display Format for New Numeric Variables
Width: 8 Decimal Places: 2
Example: 12345.67

Set Century Range for 2-Digit Years
○ Automatic
Begin year: 1929
End year: 2028
● Custom
Begin year: 1900
End year: 1999

Select Calculate values before used

- ▶ Set Transformation & Merge Options to Calculate values before used and click OK.

With this setting, SPSS does not execute Compute and Recode transformations until it needs the data. In the meantime, the status bar displays the message Transformations pending and the results of transformations are not yet visible.

- ▶ To execute pending transformations, run a procedure that requires SPSS to use the data or choose Run Pending Transformations from the Transform menu.

When transformations are pending, the Data Editor will not allow you to make certain changes to your working data file.

Recoding Values

Recoding is done with a series of specifications, of the form, "If the old value is this, assign a new value of that." A case's existing value is checked against each of these specifications until one of them matches. Then the new value is assigned, and SPSS moves on to process the next case.

NORUŠIS, Marija J. 1998. *Guide to Data Analysis*. Upper Saddle River, Prentice Hall.

There are two Recode commands: Recode into Same Variables and Recode into Different Variables. The former changes the values of variables based solely on their existing values, while the latter creates new variables with values that depend only on the existing values of single variables.

- A case is never changed by more than one of a group of recode specifications.
- If a case doesn't match any of the recode specifications, its value remains unchanged (if recoding into same variable) or becomes system-missing (if recoding into a new variable).

Example: Recoding Age into Age Categories

This example recodes the variable *age* (age in integer years) into a new variable that contains age in one of three categories: 14 through 29, 30 through 49, and 50 or older. (If *age* is not an integer you must modify the recode statement so that ages between 29 and 30, and between 49 and 50, are assigned to the proper groups.)

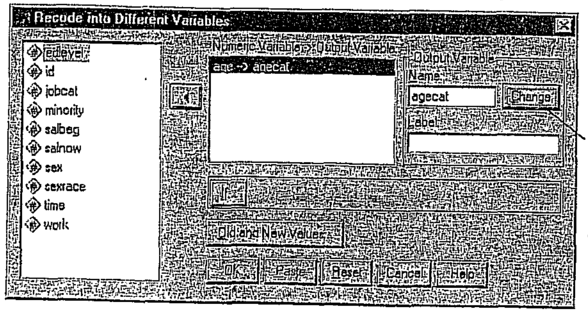
▶ Open the *salary.sav* data file.

▶ From the menus choose:

Transform
 Recode ▶
  Into Different Variables...

This opens the Recode into Different Variables dialog box, as shown in Figure B.2.

Figure B.2 Recode into Different Variables dialog box



Type agecat and click Change

▶ Move *age* into the Input Variable -> Output Variable list. The name of the list changes to reflect that a numeric variable has been selected, as shown in Figure B.2.

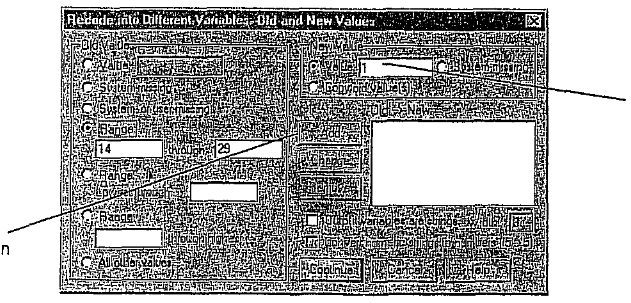
▶ In the Output Variable box, type *agecat* for the output variable and click Change.

This adds agecat to the Numeric Variable -> Output list. A new variable *agecat* will be created, which contains the recoded values of *age*.

▶ Click Old and New Values.

This opens the Old and New Values dialog box, as shown in Figure B.3.

Figure B.3 Old and New Values dialog box



Enter new value

Click Add to add specification to list

▶ In the Old Value group, select the first Range alternative.

▶ Type 14 in the first range box and 29 in the second range box.

▶ Type 1 in the New Value box.

▶ Click Add.

The specification 14 thru 29 -> 1 is added to the Old -> New list. All ages between 14 and 29 will be coded 1 in the new *agecat* variable.

▶ Click again on Range. Type 30 in the first range box and 49 in the second range box.

▶ Type 2 in the New Value box and click Add.

▶ Click Range: through highest and type 50 in the box.

NORUŠIS, Marija J. 1998. *Guide to Data Analysis*. Upper Saddle River, Prentice Hall.

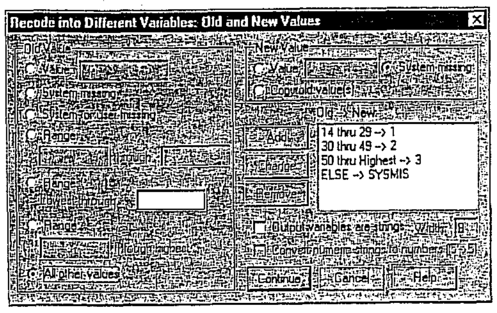
▶ Type 3 in the New Value box and click Add.

That should take care of all age groups in this file. But what if someone is coded with an age less than 14? Since the file contains data about adults who work for a bank, that would surely be a coding mistake, but it could happen. It's best to be safe.

▶ In the Old Values box, click All other values.

▶ In the New Value box, click System-missing and click Add one more time.

Figure B.4 Completed Old and New Values dialog box



```
14 thru 29 -> 1
30 thru 49 -> 2
50 thru Highest -> 3
ELSE -> SYSMIS
```

The Old and New Values dialog box should now look like Figure B.4. If it doesn't—if one of your specifications is incorrect—click the incorrect specification in the Old -> New list, make the needed correction, and click Change.

▶ Click Continue to return to the Recode into Different Variables dialog box. Then click OK to execute the transformation.

You have now changed the working data file; however, you don't want to make these changes a permanent part of the *salary.sav* data file.

▶ To avoid saving changes to the *salary.sav* data file, exit SPSS *without* saving changes or clear the Data Editor by selecting New from its File menu.

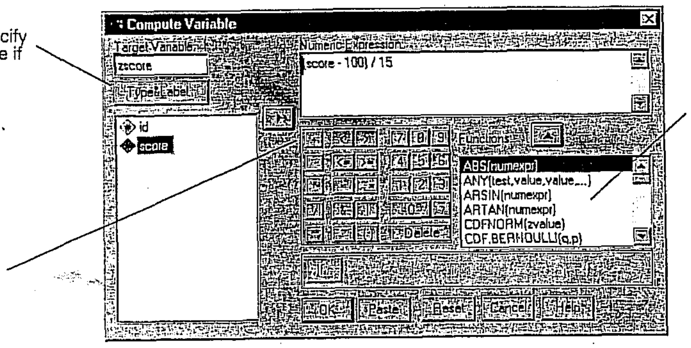
Computing Variables

The Compute Variable dialog box assigns the result of a single expression to a "target variable" for each case. The target variable can be a new variable or an existing variable (in which case the existing values will be overwritten). For example, you can compute standard scores for a variable, as described in the first example below. A great number of functions are available, so expressions can be quite complex.

▶ To open the Compute Variable dialog box, as shown in Figure B.5, from the Data Editor menus choose:

Transform
 Compute...

Figure B.5 Compute Variable dialog box



Click to specify variable type if other than numeric

Function list

Calculator pad

Unlike a spreadsheet, SPSS does not remember the formula used to compute data values or automatically update them. (In the example mentioned above, if you go back and change the values for the variable *score*, the *zscore* values will not be automatically recalculated to reflect the change.)

The Calculator Pad

The calculator pad allows you to paste operators and functions into your formula. You don't have to use the calculator pad: you can click anywhere in the Numeric Expression box and start typing. Often that's the simplest and quickest way to build an expression. The visual controls in the calculator pad are there to remind you of the possibilities and to reduce the likelihood that you won't remember how to spell one of the many functions available in SPSS.

To use the calculator pad, just click on buttons to paste symbols and operators at the insertion point. Use the mouse to move the insertion point.

To paste a function, select it in the scrolling list and click the ▲ button. You must then fill in the arguments, which are the values that the function operates on.

See "Example: Cumulative Distribution Function" on p. 535 for an example of using functions in an expression.

A few basic calculator pad operators are described in Table B.1. The Help system contains a more detailed description of the calculator pad, with definitions of all the functions.

Table B.1 Calculator pad operators

*	Multiply
/	Divide
**	Raise to power
+	Add
–	Subtract

Example: Computing Z Scores

Suppose you have a sample of IQ scores, and you wish to calculate standard scores (*z* scores) for the sample. Assuming that in the population IQ scores have a mean of 100 and a standard deviation of 15 (as was long assumed to be true), the formula is

zscore = (*score* – 100)/15

To compute standard scores for a variable according to this formula:

▶ Open the *iq.sav* data file.

This file contains IQ scores for a hypothetical group of students.

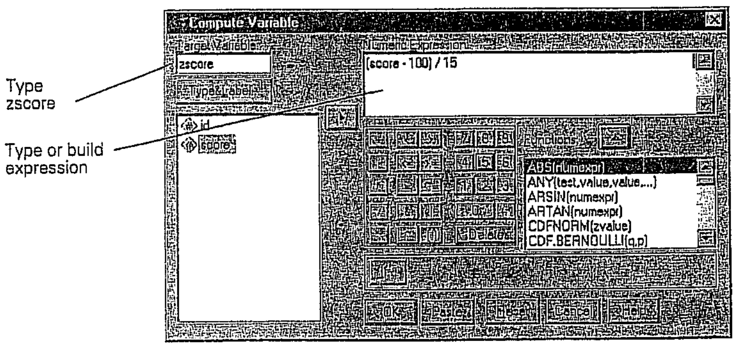
▶ Activate the Data Editor window.

▶ From the menus choose:

Transform
  Compute...

This opens the Compute Variable dialog box, as shown in Figure B.6.

Figure B.6 Compute Variable dialog box



▶ Click in the Target Variable box and type zscore.

Since zscore is a numeric variable, you don't have to click the Type & Label button.

▶ Click in the Numeric Expression box.

You can simply type the expression (score–100)/15 directly in the Numeric Expression box or build it using the calculator pad, as follows:

▶ Select score in the variable list and click ▶.

The variable name *score* is pasted into the expression at the insertion point.

▶ Enter –100.

▶ Select the entire expression score –100 and click the ( ) button.

The expression now reads (score –100).

▶ Enter /15.

The expression now reads (score –100)/15.

▶ Click OK.

SPSS computes *z* scores for all cases in the working data file.

Example: Cumulative Distribution Function

You can calculate the proportion of the population with *z* scores greater in absolute value than each of the *z* scores in your sample, as discussed in question 10 in the exercises for Chapter 10. Assuming the variable *zscore* contains the *z* scores for your sample, the formula is

twotailp = 2*(1 – cdfnorm(abs(zscore)))

▶ If you want to attempt this example, you can substitute any variable that contains *z* scores for the *zscore* variable named in the formula above. (You can use the Descriptives procedure to save *z* scores for any variable, as described in Chapter 4.)

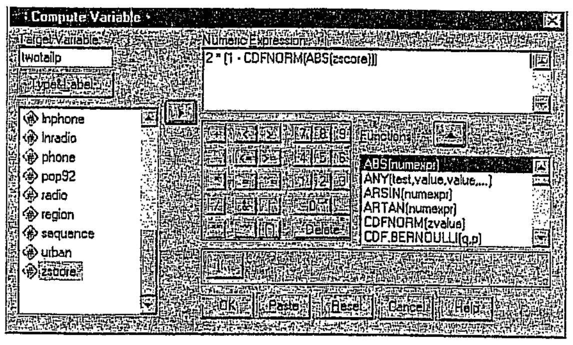
▶ From the menus choose:

Transform
  Compute...

This opens the Compute Variable dialog box.

▶ Type twotailp in the Target Variable box.

Figure B.7 Compute Variable dialog box



You can simply type the expression 2*(1-CDFNORM(ABS(zscore))), as shown in Figure B.7, or build the expression as follows:

▶ Enter 2*(1–).

▶ With the cursor inside the right parenthesis, select CDFNORM(zvalue) in the Functions list and click ▲.

The CDFNORM function is pasted into the formula at the insertion point. The expression now reads 2 * (1– CDFNORM(?)), with the question mark selected. You must replace the question mark with an argument for the CDFNORM function.

▶ Select ABS(numexpr) in the Functions list and click ▲.

The expression now reads 2 * (1– CDFNORM (ABS(?))). Once again, the question mark is selected; you must now supply an argument for the ABS function.

▶ Select *zscore* in the variable list and click ▶.

The variable *zscore* is now pasted in as the argument for the ABS function. The expression is now complete.

▶ Click OK.

SPSS computes the proportions for all cases in the working data file.

Automatic Recoding

SPSS's Recode facility is quite useful but requires you to enter detailed specifications. The Automatic Recode facility needs no specifications. It simply converts all the codes of a current variable into new codes—1, 2, 3, and so on—for a new variable.

Automatic Recode is particularly useful as a way of converting a string variable into a numeric variable.

Example: Creating Numeric Country Codes

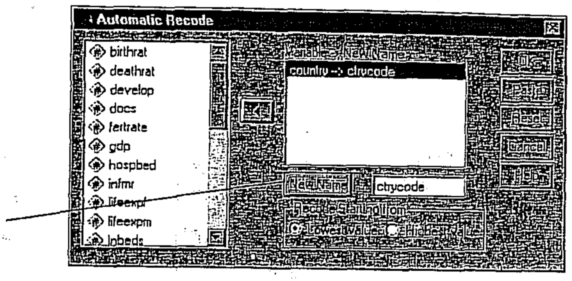
In the *country.sav* data file, the string variable *country* contains the name of each country. Suppose you want to create numeric country codes. You can do this as follows:

▶ From the Data Editor menus choose:

Transform
  Automatic Recode...

This opens the Automatic Recode dialog box, as shown in Figure B.8.

Figure B.8 Automatic Recode dialog box



Type name for new variable and click New Name

▶ Select *country* in the variable list and move it into the Variable -> New Name list.

▶ Type ctrycode in the New Name box and click New Name.

▶ Click OK.

SPSS creates the new variable *ctrycode*, which contains a unique numeric code for each country. The codes are assigned in sequence; the first country will have a code of 1, the second 2, and so on. If there were several cases for the same country, they would all be assigned the same code value.

Since the original variable *country* does not have value labels, the actual values of *country* (Afghanistan, Albania, Algeria, and so on) are used as value labels for the new variable *ctrycode*.

Conditional Transformations

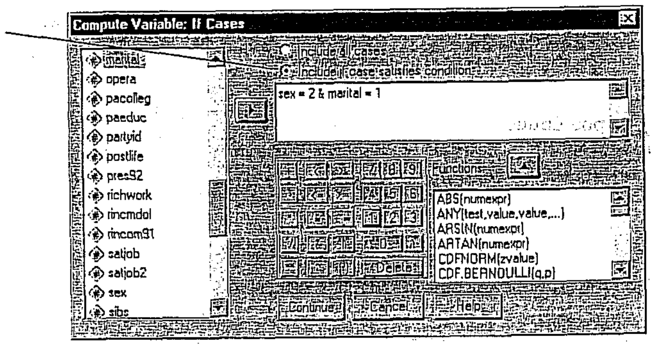
If you want to transform the values of only some cases, depending on their data values, you need a **conditional transformation**, one that is carried out only if a logical condition is true. For example, you might want to transform *only* cases for people who are full-time workers.

The Compute Variable dialog box, both Recode dialog boxes, and the Count dialog box (not shown) allow you to specify such a logical condition.

▶ To specify a logical condition for a transformation, click If in the Compute Variable, Recode, or Count dialog box.

This opens a dialog box where you can specify a logical condition. For example, the Compute Variable If Cases dialog box is shown in Figure B.9.

Figure B.9 Compute Variable If Cases dialog box



Select to specify a logical condition

See "The Calculator Pad" on p. 532.

This dialog box contains the familiar Calculator Pad. Here you use it to build a logical condition, one that is either true or false for a case, depending on the case's data values. Table B.2 describes some operators that are particularly useful in building logical conditions.

Table B.2 Operators useful in logical expressions

<	Less than
>	Greater than
<=	Less than or equal to
>=	Greater than or equal to
=	Equal to
~=	Not equal to
&	And
|	Or
~	Not

The logical expression sex = 2 & marital = 1, for example, is true only for those cases in which *both* conditions are met: the variable *sex* must equal 2 *and* the variable *marital* must equal 1. The logical expression sex = 2 | marital = 1, by contrast, is true if either of the conditions is met.

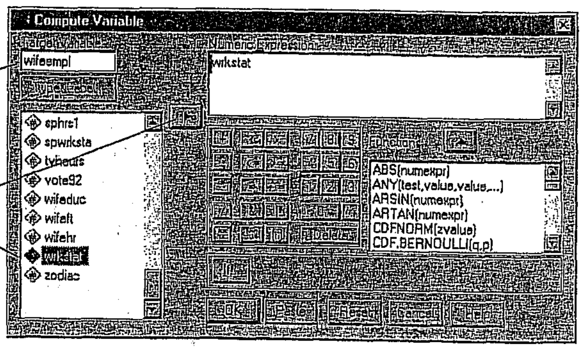
Example: Wife's Employment Status

The General Social Survey contains employment status questions for the respondent and for the respondent's spouse. The respondent could be either husband or wife, depending on who was interviewed. This means that for each household, the wife's work status could be coded in either the variable *wrkstat* (if the wife was interviewed) *or* in the spouse's work status variable *spwrksta* (if the husband was interviewed). To create a variable containing, for all married couples, the wife's employment status, you might proceed as follows:

▶ To open the Compute Variable dialog box (see Figure B.10), from the menus choose:

Transform
  Compute...

Figure B.10 Compute Variable dialog box



Type wifeempl

Paste wrkstat into the expression

▶ In the Compute Variable dialog box, type wifeempl into the Target Variable box.

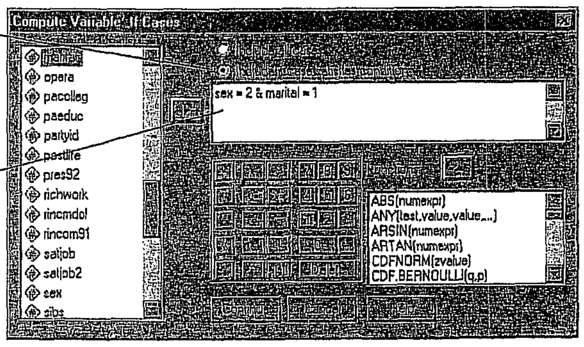
▶ Select *wrkstat* in the variable list and press ▶ to move it into the Numeric Expression box.

The new variable *wifeempl* will have the same value as the variable *wrkstat*. However, you must specify that this expression will only be evaluated for cases where the respondent is a married woman.

▶ Click If.

This opens the Compute Variable If Cases dialog box, as shown in Figure B.11.

Figure B.11 Compute Variable If Cases dialog box



Select Include if case satisfies condition

Enter expression

▶ Select Include if case satisfies condition.

▶ Using either the calculator pad or the keyboard, enter the condition sex = 2 & marital = 1.

This condition specifies that the new value should be computed only for cases for whom the value of the variable *sex* equals 2 (the code for female) *and* for whom the value of *marital* equals 1 (married). For cases that do not meet this condition, the new variable *wifeempl* will be equal to the system-missing value.

▶ Click Continue to return to the Compute Variables dialog box. Then click OK.

This creates a new variable *wifeempl*, which is equal to *wrkstat* for married women. For cases where the respondent is not married, or is a man, the value of *wifeempl* is not defined (system-missing).

At this point, you're halfway there. But what about respondents who are married men? In that case the wife's employment status would be coded in the variable *spwrksta*, which contains the work status of the respondent's spouse.

▶ Open the Compute Variable dialog box again.

▶ Delete *wrkstat* from the Numeric Expression box.

NORUŠIS, Marija J. 1998. *Guide to Data Analysis*. Upper Saddle River, Prentice Hall.

▶ Select the variable *spwrksta* and paste it into the Numeric Expression box.

▶ Click If.

The logical expression still reads sex = 2 & marital = 1.

▶ Delete the 2 and type 1 in its place.

The expression now reads sex = 1 & marital = 1.

▶ Click Continue and then click OK.

This sets *wifeempl* equal to *spwrksta* for married men. To summarize, the first transformation creates a new variable *wifeempl*, which is equal to *wrkstat* for married women and not defined for others. The second conditional transformation sets *wifeempl* equal to *spwrksta* for married men. The end result is a variable equal to wife's employment status for all married couples. For unmarried respondents, neither transformation is executed and *wifeempl* is never changed. Since it's a new variable, it is assigned the system-missing value for the unmarried respondents.

Bryman Alan and Duncan Cramer. 1997. *Quantitative Data Analysis with SPSS for Windows*. London, Routledge.

RELATIONAL OPERATORS

A relational operator like = compares the value on its left (for example, **gender**) with that on its right (for example, **1**). There are six such operators, which are represented by the following symbols:

=	equal to
~=	not equal to
<	less than
<=	less than or equal to
>	greater than
>=	greater than or equal to

The question of which is the most appropriate operator to use in selecting cases will depend on the selection criteria. To select cases under 40 years of age, we could use less than (<):

age < 40

It would also, of course, have been possible to use less than or equal to (<=) 39 in this instance since we are dealing with whole numbers:

age <= 39

To select non-whites, we could use not equal to (~=) 1 since whites are coded 1:

ethnicgp ~= 1

COMBINING LOGICAL RELATIONS

We can combine logical expressions with the logical operators & (**and**) and | (**or**). For example, we can select white men under 40 with the following conditional expression:

ethnicgp = 1 & gender = 1 & age < 40

To select people of only West Indian and African origin, we would have to use the | (**or**) logical operator:

ethnicgp = 3 | ethnicgp = 4

Note that it is necessary to repeat the full logical relation. It is *not* permissible to abbreviate this command as:

ethnicgp = 3 | 4

An alternative way of doing the same thing is to use the **any** logical function where any case with a value of either **3** or **4** for the variable **ethnicgp** is selected:

any (ethnicgp,3,4)

The variable and the values to be selected are placed in parentheses.

To select people between the ages of 30 and 40 inclusively, we can use the expression:

age >= 30 & age <= 40

Here, we have to use the **&** (**and**) logical operator. If we used | (**or**), we would in effect be selecting the whole sample since everybody is either above 30 or below 40 years of age.

Another way of selecting people aged 30 to 40 inclusively is to use the **range** logical function where any case with a value in the range of 30 to 40 for the variable **age** is selected:

range (age,30,40)

APPENDIX C

Working with SPSS Syntax (Log) Files

Mei-Huei Tsay

It is possible to modify syntax files to run slightly different statistics and/or complex, customized statistics. Sometimes output files are too large to save on a disk or take up too much space on your hard drive. Therefore, it is a good idea to understand how to use the syntax or log files that contain SPSS commands. You can use the SPSS logs from your output file to run these commands.

It is possible to open a syntax window and type in commands, but sometimes it is easier to build your syntax file by using one of the following methods:

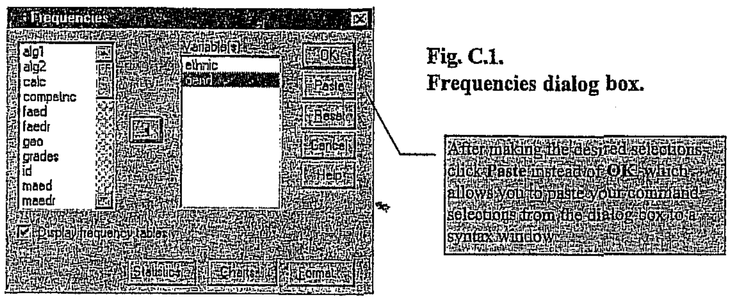
- Paste syntax commands from dialog boxes.
- Copy syntax from the output log.
- Copy syntax from the journal file.

Creating Syntax Commands From Dialog Boxes: Using Paste Instead of OK

The easiest way to generate a syntax command file is to make selections in dialogue boxes and paste the syntax of the selections into a syntax window. By pasting the syntax in the syntax window, you can generate a job file which allows you to repeat the analysis, edit it, save the syntax in a syntax file, and copy/cut it into an output log.

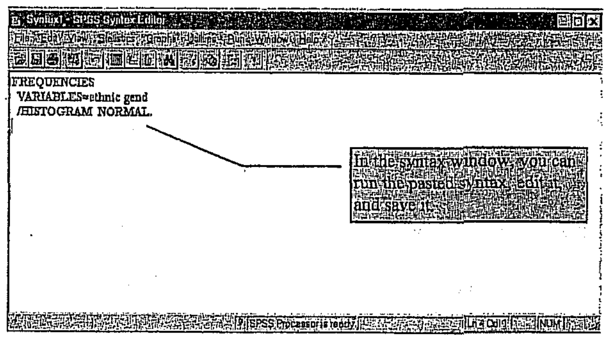
To paste syntax commands from a dialog box:

- Retrieve the data file.
- Open the dialog box and make desired selections. For example: **Statistics => Summarize => Frequencies.**
- After making all the desired selections, click **Paste** instead of **OK** (see Fig. C.1). The syntax command is pasted to the syntax window. If you don't have an open syntax window, SPSS will open a new syntax window and paste the syntax there (see Fig. C.2).



After making the desired selections, click **Paste** instead of **OK** which allows you to paste your command selections from the dialog box to a syntax window.

Fig. C.1. Frequencies dialog box.


Syntax1 - SPSS Syntax Editor
File Edit View Statistics Graphs Utilities Run Window Help
SPSS Processor is ready
Ln 4 Col 1
NUM

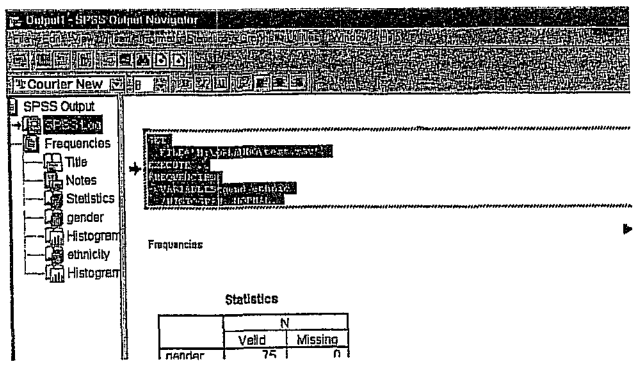
```
FREQUENCIES
  VARIABLES=ethnic gend
  /HISTOGRAM NORMAL.
```

In the syntax window, you can run the pasted syntax, edit it, and save it.

Fig. C.2. The syntax window.

- To run/use a syntax file, when there is only one syntax file in the window, just simply click on **Run** in the menu bar.
- If there are many syntax files in the window, *highlight* the desired syntax first, then click **Run** then **Selection.** The output will show on the output window.
- If you need to repeat all the analyses in the syntax window, you can click on **Run** and then **All.**

- To do this, from the menus click on **Edit => Options => Navigator => Display commands in the log.** Each command you did will then be recorded in the SPSS log as in Fig. C.3.



Statistics

	N	
	Valid	Missing
gender	75	0

Fig. C.3. Syntax commands in the SPSS log.

- To copy the syntax from the Output Navigator, first *double click* on the syntax file table to activate it (which allows you to edit the table), then **highlight** the desired syntax file (see Fig. C.3); from **Edit,** click on **Copy.**
- Second, open a previously saved syntax file or create a new one.
- To create a new syntax file, from the menus choose **File => New => Syntax;** then in the syntax window, choose **Edit** then **Paste.** You can run or change the pasted syntax as we did above.

Using Syntax From the Journal File

This is a more complicated way of doing things, but it is worth mentioning here.

By default, SPSS records all commands executed during a session in a journal file named *spss.jnl* (set with **Options** on the **Edit** menu). You can edit the journal file and save it as a syntax file that you can use to repeat the previous analysis.

To open the journal file,

- from the menus choose **File => Open;** under the **Files of Type,**
- choose **All files** (*.*);
- then choose *spss.jnl* from the file name box or enter **.jnl* in the **File Name box,**
- then click **Open** (see Fig. C.4).

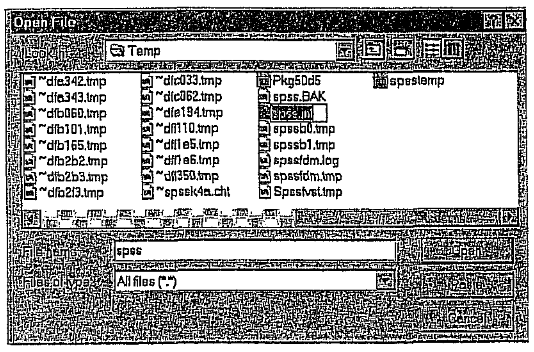
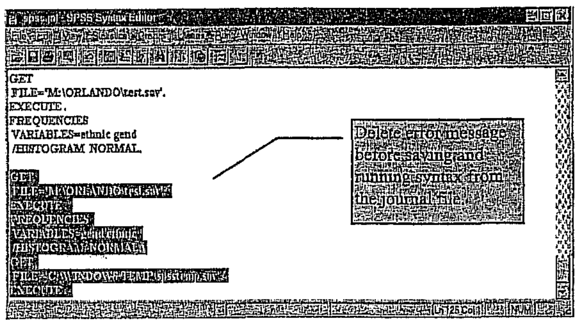


Fig. C.4. Open spss.jnl.

The journal file is a text file that can be edited like any other text file. But notice, because error and warning messages are also recorded in the journal file along with your commands, you must delete any of these messages that appear before saving or running the syntax file (see Fig. C.5).


spss.jnl - SPSS Syntax Editor
File Edit View Statistics Graphs Utilities Run Window Help
Ln 25 Col 1
NUM

```
GET
 FILE='M:\ORLANDO\tert.sav'.
EXECUTE .
FREQUENCIES
 VARIABLES=ethnic gend
 /HISTOGRAM NORMAL.
```

Delete error message before saving and running syntax from the journal file.

Fig. C.5. Editing the journal file.

To run or edit the journal file, see above.

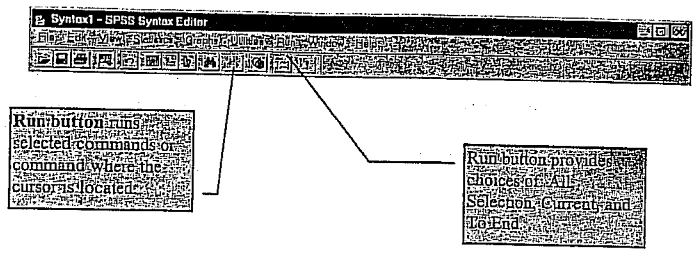
Running Syntax Commands

You can run single commands, selected groups of commands, or all commands in a syntax window. The following options are available on the **Run** menu (see Fig. C.6):

- **All.** Runs all commands in the syntax window.
- **Selection.** Runs the currently selected commands. This includes any commands partially highlighted.
- **Current.** Runs the command where the cursor is currently located.

MORGAN George A. and Griego V. ORLANDO. 1998. *Easy Use and Interpretation of SPSS for Windows: Answering Research Questions With Statistics*. Mahwah: Lawrence Erlbaum Associates.

- **To End**. Runs all commands from the current cursor location to the end of the command syntax file.


Syntax1 - SPSS Syntax Editor
File Edit View Statistics Graphs Utilities Run Window Help

Run button runs selected commands or command where the cursor is located.

Run button provides choices of All, Selection, Current, and To End.

Fig. C.6. Syntax editor toolbar.

5. lekce

NORMÁLNÍ ROZLOŽENÍ A ZÁKLADY TESTOVÁNÍ HYPOTÉZ. STATISTICKÁ INFERENCE ANEB ZOBECŇOVÁNÍ VÝBĚROVÝCH VÝSLEDKŮ NA ZÁKLADNÍ SOUBOR.

3.3 Normální rozdělení

Jsou-li všichni havrani černí, není třeba vytvářet výběrové soubory a uvažovat o tom, kolik by mohlo být havranů šedých nebo světle modrých. Je-li pravděpodobnost, že havrani jsou černí, $p = 1$, jde o určitost.

Něco jiného je, zkoumám-li např. váhu sta, tisíce nebo ještě většího množství havranů. I když se přitom vyloučí mláďata, objeví se uvnitř zkoumaného souboru výkyvy. Podobně se navzájem odlišuje tělesná výška dospělých mužů, kvocient inteligence školních dětí, počet slov na plně potištěných stránkách

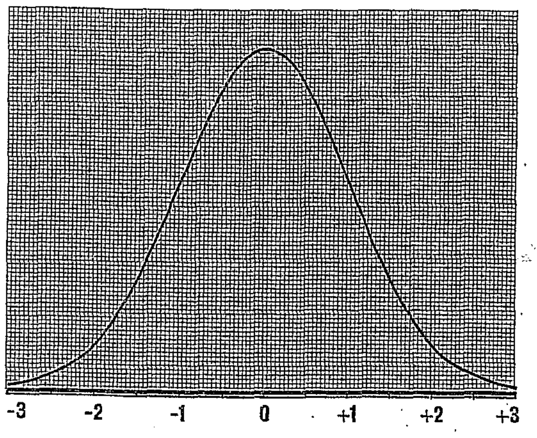
knihy, přesná délka sériově vyráběných věšáků na zeď, životnost elektrických žárovek, ba dokonce navzájem nezávislá měření jedné a téže vzdálenosti. Jistě jste si vzpomněli i na rozdělení příjmů našich 25 abiturientů a obyvatel Zbohatlíkova, avšak právě tato rozdělení musíme z našeho pojednání prozatím vyloučit. Naproti tomu se sem velmi dobře hodí úvahy, které jsme rozvíjeli v souvislosti s binomickým rozdělením a zákonem velkých čísel. Výsledky dvaceti hodů mincí stejně jako téměř každého druhu výběrových šetření vykazují totiž také charakteristické znaky tzv. normálního rozdělení, jímž se nyní budeme zabývat.

Může být sporné, zda označení „*normální rozdělení*" je zvoleno šťastně (kritika někdy uvádí, že slovo „normální" naznačuje jakési souhlasné hodnocení), nicméně ve všech světových jazycích se přesto mluví o „*normálním rozdělení*", „*Normalverteilung*", „*normal distribution*", „*distribution normale*" — tzn. že statistická terminologie je bez tohoto slova nemyslitelná —, aniž by statistika proto napadlo, že jiná rozdělení než normální jsou „abnormální".

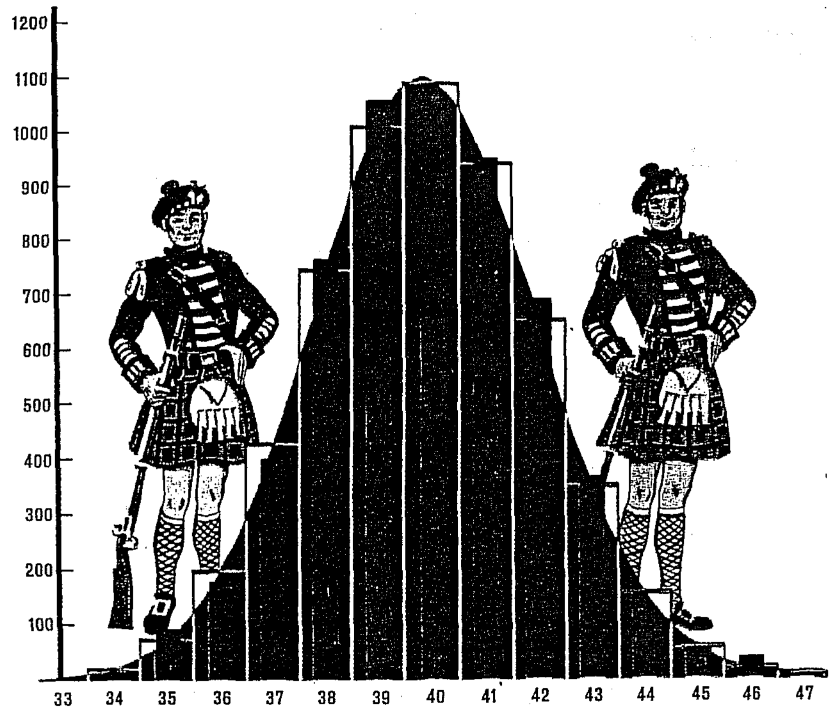
Nebude také ani požadovat, aby se daná čísla a měřené hodnoty daly ve všech případech úzkostlivě uvést do souladu s ideálním obrazem normálních rozdělení. *Normální rozdělení* je jako každé jiné statistické rozdělení *především myšlenkovým modelem a početní pomůckou*, tedy nikoliv exaktním přírodním zákonem, který by musel být naplněn s malichernou přesností.

Jestliže se statistika vždy, kdykoli vystupuje jako induktivní statistika, snaží usuzovat o celku na základě vzorků a dílčích pozorování, a proto může poskytnout jen více nebo méně pravděpodobné odhady, bylo by dvojnásob absurdní, kdyby se předstírala přesnost, která se především ve skutečnosti v takovém stupni nikdy nevyskytuje a která je dále vyloučena samým nahodilým charakterem každého výběrového souboru.

Teorie a praxe však již prokázaly správnost domněnky, že normální rozdělení platí pro téměř všechny výběry a pro velmi mnohá rozdělení podchytitelných souborů.



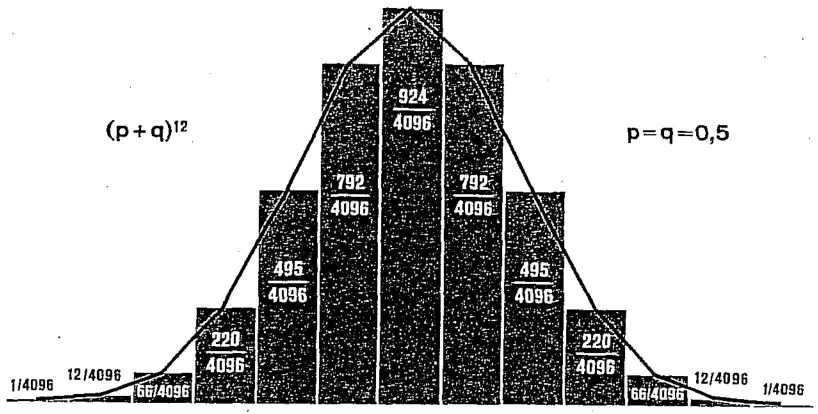
Normální křivka nebo lépe jedna z normálních křivek vzhledem k tomu, že by ji bylo možno nakreslit strmější nebo mnohem plošší. Podstatný je jen vztah výše křivky v bodech, které vymezují směrodatnou odchylku. Body ležící na křivce ve vzdálenosti + nebo — 2 σ směrodatné odchylky se nacházejí v $^1/_8$ výšky, kterou má křivka normálního rozdělení ve vrcholu (nad průměrem 0).



Obvod hrudi skotských vojáků podle Quételetovy statistiky. Shoda pozorovaných hodnot s normálním rozdělením je až zarážející ($\mu = 39{,}8$ palce).

Normální rozdělení má především velmi příjemnou vlastnost, která lehce vysvětluje jeho oblibu: ať už jde o jakékoliv objekty, úkazy, měření nebo sčítání, *jsou vždy jednoznačně určeny střední hodnotou a rozptylem*. Pokud jde o „střední hodnotu", má se zpravidla, a nikoliv neprávem, na mysli aritmetický průměr $\bar{x}$ nebo μ, a to platí i v případě normálního rozdělení. Avšak stejnou hodnotu jako aritmetický průměr mají i modus a medián (nejčetnější hodnota a prostřední hodnota). Krátká úvaha hned prokáže, že normální rozdělení s největší četností „uprostřed" musí být symetrické. Jak taková normální křivka, která je grafickým znázorněním normálního rozdělení, ve skutečnosti vypadá, ukazuje obrázek na str. 74.

Základem normálního rozdělení je zkušenost, že bezpočetné znaky a hodnoty jsou rozloženy tak, že jeden výsledek měření nebo sčítání je „nejčetnější" a „na obě strany od něho" jsou výsledky ponenáhlu stále méně četné, až konečně vykazují jen ojedinělou extrémní hodnotu. Jeden z prvních příkladů takového normálního rozdělení podal Quételet na základě měření obvodu prsou 5738 skotských vojáků. Nejčetnější hodnota činila (zaokrouhleno na celé coule) 40 coulů, skoro stejnou četnost vykazovalo 39 coulů, 41 a 38 coulů se vyskytovalo již vzácněji, 42 a 37 byly ještě vzácnější a konečně zjištěných 33, resp. 48 coulů představovalo jen ojedinělé extrémní hodnoty.



Binomické rozdělení $(p + q)^{12}$ dovoluje již jasně poznat podobu normální křivky.

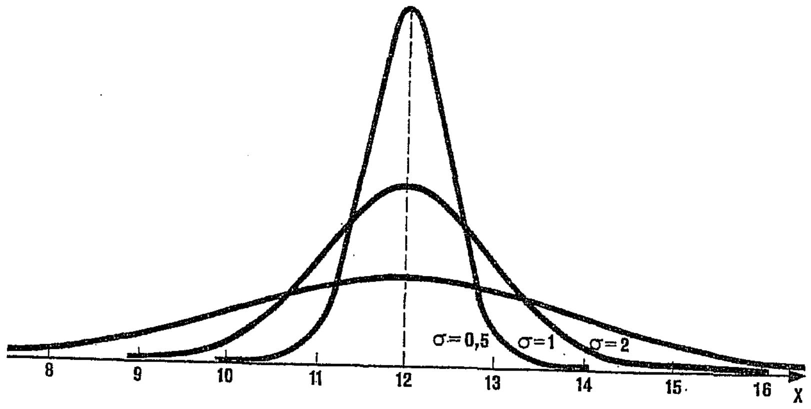
Podobné uspořádání vykazují i výsledky dalších četných měření, např. váha cigaret vyráběných cigaretovým automatem: v nejčetnějších případech vážily 1,18 až 1,20 g nebo 1,20 až 1,22 g, jen málokteré byly lehčí než 1,08 g nebo těžší než 1,32 g.

Podle rozsahu rozptylu i podle měřítka zvoleného pro grafické znázornění vzniká v idealizované formě plošší nebo strmější křivka.

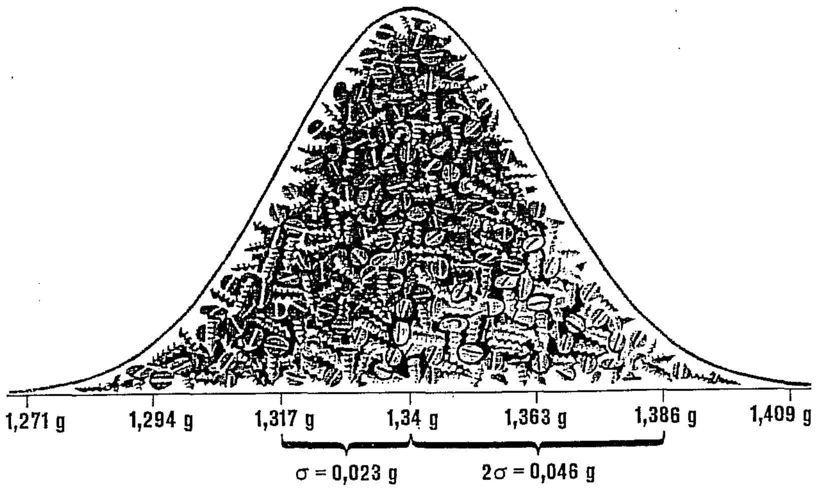
Mluví se o „*normální křivce*", v němčině o „Glockenkurve" (*zvonovité křivce*) se zřetelem na středně vysokou křivku, která se také uvádí jako model pro normovanou normální křivku, ve Francii o „*courbe en chapeau de gendarme*" (křivku policejního klobouku) se zřetelem na plošší normální křivku. Jako vědecká označení se také používají názvy „*Gaussova křivka rozdělení chyb*" a „*de Moivrova stochastika*" s odvoláním na oba praotce zvonovité křivky.

Ať už se zvolí jakékoliv měřítko a ať je rozptyl k průměru v jakémkoliv poměru, normální křivka má vždy některé charakteristické znaky, z nichž uvedeme ty, které jsou pro praxi nejdůležitější. Jestliže pozorujeme plochu ležící pod „zvonem" jako soubor, *leží na obě strany od maxima* (střední hodnota, nejčetnější hodnota) *vždy přesně stejné části této plochy*, a to v úseku mezi $+\sigma$ a $-\sigma$ leží 68,26 %, tj. nepatrně více než $^2/_3$ celkové plochy, v úseku mezi $+2\sigma$ a -2σ skoro přesně 95 % a mezi $+3\sigma$ a -3σ již 99,7 % plochy.

Hodnoty za třemi směrodatnými odchylkami se proto berou v úvahu již jen velmi zřídka, ačkoliv normální křivka se teoreticky rozkládá od —∞ po +∞, tzn. v jisté míře od nekonečna do nekonečna. Tímto *postupem do nekonečna a plynulým průběhem* se normální křivka liší od binomického rozdělení. Přesto však mezi těmito oběma rozděleními je tak těsná souvislost, že *normální rozdělení* je možno v téměř všech prakticky důležitých případech pokládat za dostatečně přesné *vyjádření binomického rozdělení*, čímž si lze ušetřit svízelné *početní operace, které jsme* alespoň v náznaku poznali při našich úvahách o binomickém rozdělení. V historii normální křivky nelze zamlčet její původ z teorie hazardní hry. U její kolébky stál stařešina počtu pravděpodobnosti Abraham de Moivre.



Tři normální rozdělení kolem střední hodnoty $\mu = 12$. Má-li rozdělení silný rozptyl ($\sigma = 2$), je křivka plochá a rozložená; je-li rozptyl malý ($\sigma = 0{,}5$), je strmá a vysoká. Střední křivka vykazuje proporce „normovaného normálního rozdělení".



Průměrná váha určitého druhu šroubů je 1,34 g, směrodatná odchylka 0,023 g; protože jde o normální rozdělení, budou se velmi vzácně vyskytovat šrouby, které váží méně než 1,294 nebo více než 1,386 g.

3.32 Normované normální rozdělení

I když jsou normální křivky pravidelné, symetrické a stejnorodé, získávají velký praktický význam teprve dalším procesem *standardizace (normování)*. K tomu, abychom porozuměli procesu standardizace, musíme uvést ještě několik příkladů nestandardizovaného normálního rozdělení. Charakteristické chování křivky je vždy stejné: bod obratu křivky leží vždy ve vzdálenosti $+\sigma$ a $-\sigma$, tečna v bodu obratu vždy protíná souřadnici ve vzdálenosti $+2\sigma$ a -2σ, teoreticky je vždy křivka rozložena na obě strany donekonečna, avšak již při $+3\sigma$ a -3σ se prakticky dotýká osy úseček (souřadnice x).

Nanášené měrné hodnoty jsou však nestejné. Váha šroubů se může někdy odchylovat o směrodatnou odchylku 0,023 g od průměru 1,34 g; jindy mohou psychometrická šetření skupiny studentů vykazovat rozptyl 36 ($\sigma = 6$) kolem kvocientu inteligence 112; tělesná váha, výsledky sklizně, stejně tak jako libovolný počet sčítaných a měřených výsledků mohou být rozděleny přibližně normálně kolem střední hodnoty se směrodatnou odchylkou, která může jednou činit půldruhého dne, jindy 0,85 kg, 3,2 cm, 0,8 ohmů, 35 marek nebo 3,5 mm.

Mají-li být všechna tato rozdílná měření a sčítání opravdu účelně zobrazena normálním rozdělením, je žádoucí, aby byl k dispozici *standardizovaný soubor nástrojů*, který umožňují odpověď na velmi rozličné otázky: „Kolik % šroubů je mimo toleranční meze ± 0,06 g?" — „Odchyluje se výrazně rozptyl inteligenčního kvocientu určité skupiny od rozptylu přibližně dvojnásobně velké skupiny, s níž byl proveden stejný test?" — „Vytvoříme-li výběrový soubor 50 žárovek, bude v něm s pravděpodobností větší než 68 % nejméně jeden zmetek?" — „Jak velká je pravděpodobnost, že v ruletě padne v průběhu nejbližšího tisíce sázek číslo 13 přesně třináctkrát?"

Převedení všech těchto rozmanitých měření, otázek a odpovědí na jediné schéma je možné jen tehdy, když je normální rozdělení *jednoznačně určeno* směrodatnou odchylkou a průměrem a když struktura normálního rozdělení *se nemění*, ať už jde o centimetry, hektolitry, ohmy anebo čísla v ruletě. Použijme pro objasnění daného problému např. šroubů: nejdříve máme průměrnou hodnotu 1,34 g a směrodatnou odchylku 0,023 g. První krok: vynecháme gramy a zůstává $\mu = 1{,}34$ a $\sigma = 0{,}023$. Během druhého kroku provedeme další abstrakci: průměr $\mu = 1{,}34$ je pro standardizované normální rozdělení právě průměrem a průměr standardizované normální křivky se zásadně rovná nule. To není pouhá libovůle, nýbrž účelná konvence, protože od kterého jiného čísla než od nuly je možno tak lehce zjistit zrcadlově stejné odchylky nahoru a dolů? Odchylce „nahoru" (třeba +2) odpovídá stejná odchylka „dolů" (—2). Z těchto důvodů také tabulky standardizovaného normálního rozdělení, které jsou nepostradatelnou pomůckou při statistické práci, nerozlišují mezi kladnými a zápornými odchylkami. Přesto však je možno z nich vyčíst pravděpodobnost kteréhokoliv jevu a četnosti.

Jak to vypadá v praxi? Tak např. chceme zjistit, kolik šroubů se odchyluje v tom nebo onom směru o více než 0,06 g od průměrné váhy. Známe: $\sigma = 0{,}023$. Víme také, že rozdělení ve standardizovaném normálním rozdělení závisí již jen na směrodatné odchylce, protože jsme průměr posunuli na nulu. Zcela jednoduše proto porovnáme hledané 0,06 g se směrodatnou odchylkou a bez velké námahy vypočítáme, že 0,06 je rovno 2,6 směrodatné odchylky standardizovaného normálního rozdělení.

Danou skutečnost lze označit také jinak. Třeba takto: šrouby nemají být těžší než 1,40 g a lehčí než 1,28 g. Jaká je pravděpodobnost, že odchylky budou přesahovat uvedené váhy? Pak bychom měli podle vzorce pro standardizaci normálního rozdělení uvést hledané hodnoty do spojitosti se střední hodnotou:

$$\text{normovaná hodnota} = \frac{\text{hodnota minus aritmetický průměr}}{\text{směrodatná odchylka}}$$

Jestliže tuto „normovanou hodnotu" označíme z, může být vyjádření ještě kratší:

$$\boxed{z = \frac{x - \mu}{\sigma}.}$$

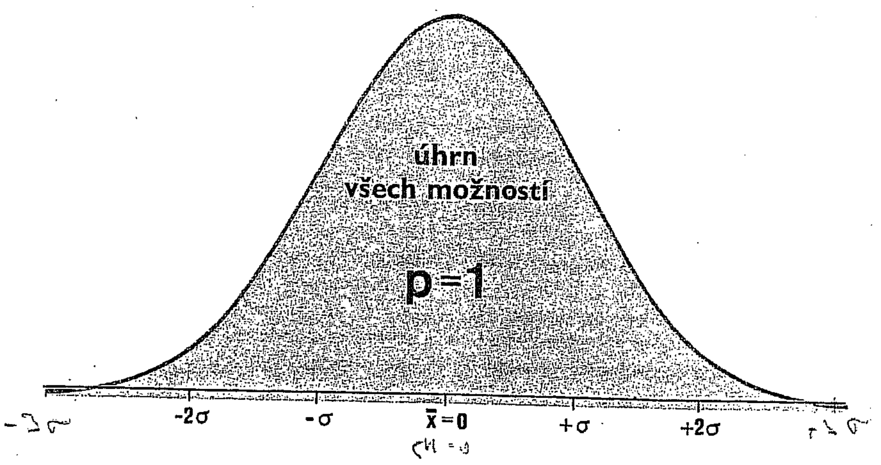
Dosadíme-li naše čísla, dostaneme:

$$\frac{1{,}40 - 1{,}34}{0{,}023} = 2{,}6, \text{ a stejně}$$

$$\frac{1{,}28 - 1{,}34}{0{,}023} = -2{,}6.$$

Záporné znaménko označuje odchylku pod střední hodnotu.

Nyní tedy víme, že $z = 2{,}6$. Ale co vlastně je z (nebo ať už tuto hodnotu nazveme jakkoliv, protože se vyskytuje pod různými názvy)? Chceme-li to tak vyjádřit, je to standardizovaná směrodatná odchylka. Stejně jako jsme zprvu přeměnili reálnou střední hodnotu $\mu = 1{,}34$ g na $\mu = 0$, tak i nyní jsme přeměnili reálnou směrodatnou odchylku $\sigma = 0{,}023$ na standardizovanou (normalizovanou) odchylkovou veličinu $z = 2{,}6$.



Standardizace normálního rozdělení: průměrná hodnota je vždy 0, odchylky se už neudávají v gramech, litrech nebo absolutních četnostech, nýbrž pouze v blíže nespecifikovaných směrodatných odchylkách.

SWOBODA Helmut. 1977. *Moderní statistika*. Praha, Svoboda.

NORUŠIS, Marija J. 1998. *Guide to Data Analysis*. Upper Saddle River, Prentice Hall.

10 The Normal Distribution

What is the normal distribution, and why is it important for data analysis?

- What does a normal distribution look like?
- What is a standard normal distribution?
- What is the Central Limit Theorem, and why is it important?

In Chapter 9, you learned how to evaluate a claim about the mean of a variable that has two possible values. Using the binomial test, you calculated the probabilities of getting various sample results when the probability of a success was assumed to be known. In this chapter, you'll learn how to test claims about the mean of a variable that has more than two values. You'll also learn about the normal distribution and the important role it plays in statistics.

▶ This chapter examines data on serum cholesterol levels from the *electric.sav* data file. In addition, some figures use simulated data sets included in the file *simul.sav*. The histograms and output shown can be obtained using the SPSS Graphs menu (see Appendix A) and the Descriptives procedure (see Chapter 4).

The Normal Distribution

You may have noticed that the shapes of the two stem-and-leaf plots in Chapter 9 are similar. They look like bells (on their sides). The same data are displayed as histograms in Figure 10.1 and Figure 10.2, where a bell-shaped distribution with the same mean and variance as the data is superimposed. You can see that most of the values are bunched in the center. The farther you move from the center, in either direction, the fewer the number of observations. The distributions are also more or less symmetric. That is, if you divide the distribution into two pieces at the peak, the two halves of the distribution are very similar in shape, but mirror images of each other. (The theoretical bell distribution is perfectly symmetric.)

You can obtain histograms using the Graphs menu, as described in Appendix A.

In the Histograms dialog box, select the variables cured10 and cured40.

Figure 10.1 Simulated experiments: sample size 10

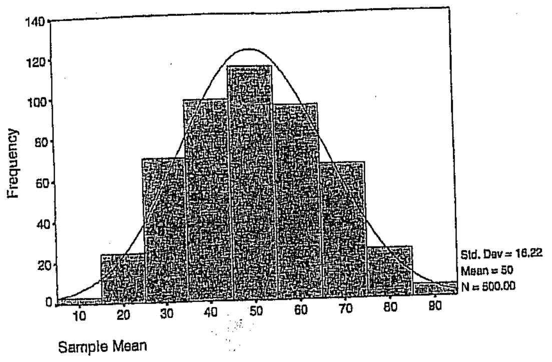
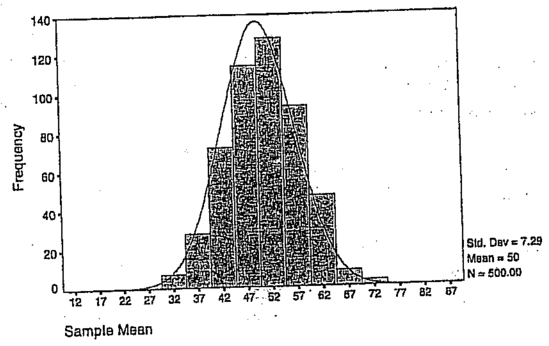


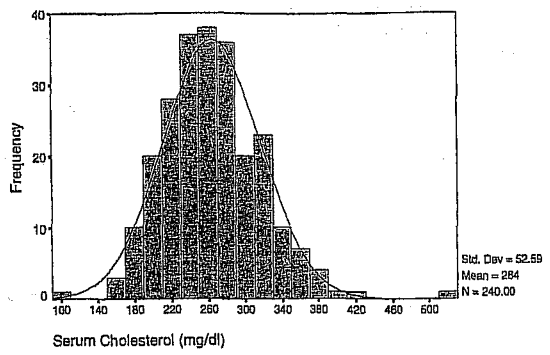
Figure 10.2 Simulated experiments: sample size 40



Many variables—such as blood pressure, weight, and scores on standardized tests—turn out to have distributions that are bell-shaped. For example, look at Figure 10.3, which is a histogram of cholesterol levels for a sample of 239 men enrolled in the Western Electric study (Paul et al., 1963). Note that the shape of the distribution is very similar to that in Figure 10.2. That's a pretty remarkable coincidence, since Figure 10.2 is a plot of many sample means from a distribution that has only two values (1=cured, 0=not cured), while Figure 10.3 is a plot of actual cholesterol values.

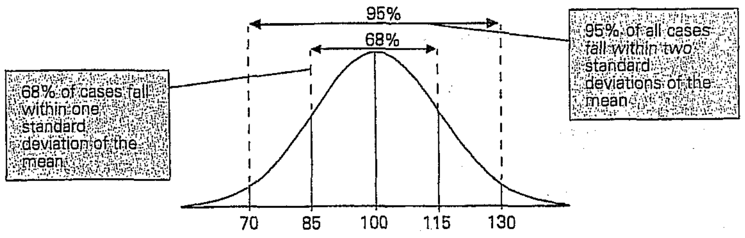
Figure 10.3 Histogram of cholesterol values

To obtain this histogram, open the electric.savdata file and select chol58 in the Histograms dialog box.



The bell distribution that is superimposed on Figure 10.1, Figure 10.2, and Figure 10.3 is called the **normal distribution**. A mathematical equation specifies exactly the distribution of values for a variable that has a normal distribution. Consider Figure 10.4, which is a picture of a normal distribution that has a mean of 100 and a standard deviation of 15. The center of the distribution is at the mean. The mean of a normal distribution has the same value as the most frequently occurring value (the mode), and as the median, the value that splits the distribution into two equal parts.

Figure 10.4 A normal distribution



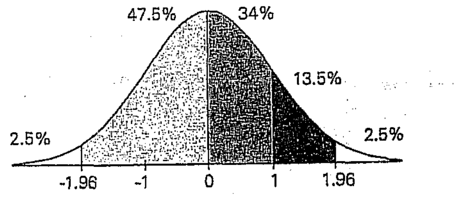
68% of cases fall within one standard deviation of the mean

95% of all cases fall within two standard deviations of the mean

If a variable has exactly a normal distribution, you can calculate the percentage of cases falling within any interval. All you have to know are the mean and the standard deviation. Suppose that scores on IQ tests are normally distributed, with a mean of 100 and a standard deviation of 15, as was once thought to be true. In a normal distribution, 68% of all values fall within one standard deviation of the mean, so you would expect 68% of the population to have IQ scores between 85 (one standard deviation below the mean) and 115 (one standard deviation above the mean). Similarly, 95% of the values in a normal distribution fall within two standard deviations of the mean, so you would expect 95% of the population to have IQ scores between 70 and 130.

Since a normal distribution can have any mean and standard deviation, the location of a case within the distribution is usually given by the number of standard deviations it is above or below the mean. (Recall from Chapter 4 that this is called a **standard score**, or **z score**.) A normal distribution in which all values are given as standard scores is called a **standard normal distribution**. A standard normal distribution has a mean of 0, and a standard deviation of 1. For example, a person with an IQ of 100 would have a standard score of 0, since 100 is the mean of the distribution. Similarly a person with an IQ of 115 would have a standard score of +1, since the score is one standard deviation (15 points) above the mean, while a person with an IQ of 70 would have a standard score of −2, since the score is two standard deviation units (30 points) below the mean.

Figure 10.5 The standard normal distribution



Some of the areas in a standard normal distribution are shown in Figure 10.5. Since the distribution is symmetric, half of the values are greater than 0, and half are less. Also, the area to the *right* of any given positive score is the same as the area to the left of the same negative score. For example, 16% of cases have standardized scores greater than +1, and 16% of cases have standardized scores less than −1. Appendix D gives areas of

the normal distribution for various standard scores. The exercises show you how to use SPSS to calculate areas in a normal distribution.

If you're more than two standard deviations from the mean on some characteristic, does that mean you're abnormal? Not necessarily. For example, pediatricians often evaluate a child's size by finding percentile values. They may tell the parents that their child is at the 2.5th percentile, or 97.5th percentile for height. (For a normal distribution, these percentiles correspond to standardized scores of −2 and +2.) The small or large percentile values don't necessarily indicate that something is wrong. Even if you took a group of healthy children and looked at their height distribution, some of them would be more than two standard deviations from the mean. Somebody has to fall into the tails of the normal distribution. This also leads to a convincing argument against grading on the curve. Even in a brilliant, hard-working class, some students will receive scores more than 2 standard deviations below the mean. Does that make their performance unacceptable? Not necessarily. ■■■

Samples from a Normal Distribution

If you look again at Figure 10.3, you'll see that the normal distribution that is superimposed on the cholesterol data doesn't fit the data values exactly. The observed data are not perfectly normal. Instead, the distribution of the data values can be described as approximately normal. That's not surprising. Even if you assume that cholesterol values have a perfect normal distribution in the population, you wouldn't expect a sample from this distribution to be exactly normal. You know that a sample is not a perfect picture of the population. You expect that samples from a normal population would appear to be more or less bell shaped, but it would be unrealistic to expect that every sample is exactly normal. In fact, even the population distribution of most variables is not exactly normal. Instead, it's usually the case that the normal distribution is a good approximation. Slight departures from the normal distribution have little effect on statistical analyses that assume that the distribution of data values is normal.

Means from a Normal Population

Since we've established that the normal distribution is a reasonable representation of the distribution of data values for many variables, we can use this information in testing statistical hypotheses about such variables. For example, suppose you want to test whether highly paid CEO's have average cholesterol levels which are different from the population as a whole. In 1991, *Forbes* sent out a survey to the 200 most highly compensated CEO's requesting their cholesterol levels. The 21 CEO's who responded had an average cholesterol of 193 mg/dL. Assume that, in the population, cholesterol levels are normally distributed with a mean of 205 and a standard deviation of 35. Based on this information, how would you determine if the CEO's differ from the rest of us not only in their net worth but in average cholesterol as well?

To answer this question, you need to know whether 193 is an unlikely sample value for the mean, when the true population value is 205. To arrive at this information, you'll follow the same procedure as you did in Chapter 9. However, instead of taking samples from a population in which only two values can occur, you'll take repeated samples from a normal population.

Figure 10.6 Distribution of 500 sample means

To obtain this histogram, open the simul.sav file and select the variable normal21 in the Histograms dialog box.

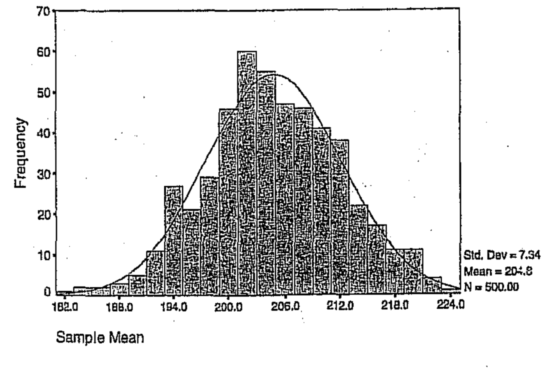


Figure 10.6 shows the distribution of 500 sample means from a normal distribution with a mean of 205 and a standard deviation of 35. Each mean is based on 21 cases. As you can see, the distribution of sample means is also approximately normal. That's always the case when you

NORUŠIS, Marija J. 1998. *Guide to Data Analysis*. Upper Saddle River, Prentice Hall.

calculate sample means for data from a normal population. The mean of the sample means is very close to 205, the population value. In fact, for the theoretical sampling distribution of the means, the value is exactly 205. (Remember, the theoretical distribution of sample means is mathematically derived and tells you precisely what the distribution of the sample means is for all possible samples of a particular size.) In Figure 10.6, the standard deviation of the means, also known as the standard error of the mean, is 7.34.

Standard Error of the Mean

You saw in Chapter 9 that the standard error of the mean tells you how much sample means from the same population vary. It depends on two things: how large a sample you take (that is, the number of cases used to compute the mean) and how much variability there is in the population. Means based on large numbers of cases vary less than means based on small numbers of cases. Means calculated from populations with little variability vary less than means calculated from populations with large variability.

If you know the population standard deviation (or variance) and the number of cases in the sample, you can calculate the standard error of the mean by dividing the standard deviation by the square root of the number of cases. In this example, the population standard deviation is 35 and the number of cases is 21, so the standard error of the mean is:

$$\frac{35}{\sqrt{21}} = 7.64 \qquad \textbf{Equation 10.1}$$

Note that the value we calculated based on the 500 samples with 21 hypothetical CEO's in each sample was not exactly 7.64, but very close. What we obtained was an *estimate* of the true value. That's because we did not take all possible samples from the population, but restricted our attention to 500.

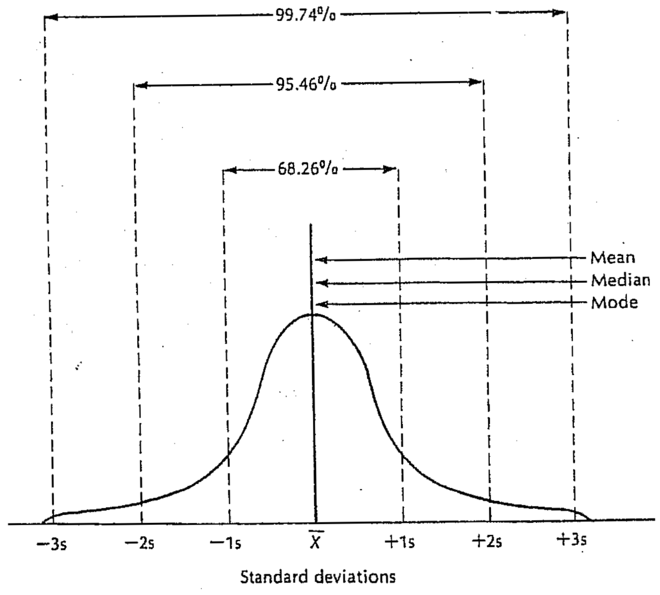
Summary

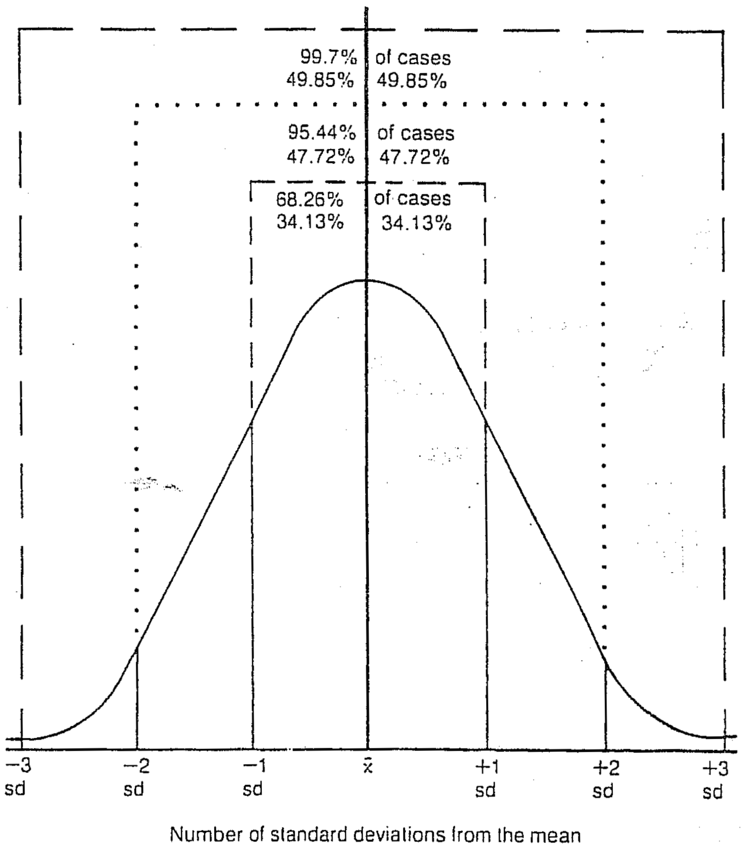
What is the normal distribution, and why is it important for data analysis?

- A normal distribution is bell shaped. It is a symmetric distribution in which the mean, median, and mode all coincide. In the population, many variables, such as height and weight, have distributions that are approximately normal.
- Although normal distributions can have different means and variances, the proportional distribution of the cases about the mean is always the same.
- A standard normal distribution has a mean of 0 and a standard deviation of 1.
- The Central Limit Theorem states that for samples of a sufficiently large size, the distribution of sample means is approximately normal. (That's why the normal distribution is so important for data analysis.)

WILLIAMSON, J. B., David A. KARP, John R. DALPHIN, Paul S. GRAY. 1982. *The Research Craft. An Introduction to Social Research Methods*. Boston: Little, Brown and Company.

FIGURE 16.3 Percentages of Observations within Various Standard Deviation Units of the Mean for the Normal Curve





2.5. Testing whether a Distribution is Normal

2.5.1. *Running the Analysis*

It is all very well to look at histograms, but they tell us little about whether a distribution is close enough to normality to be useful. Looking at histograms is subjective and open to abuse (I can imagine researchers sitting looking at a completely distorted distribution and saying 'yep, well Bob, that looks normal to me', and Bob replying 'yep, sure does'). What is needed is an objective test to decide whether or not a distribution is normal. Fortunately, such tests exist: the Kolmogorov-Smirnov and Shapiro-Wilk tests. These tests compare the set of scores in the sample to a normally distributed set of scores with the same mean and standard deviation. If the test is non-significant ($p > 0.05$) it tells us that the distribution of the sample is not significantly different from a normal distribution (i.e. it is probably normal). If, however, the test is significant ($p < 0.05$) then the distribution in question is significantly *different from a normal distribution* (i.e. *it is non-normal*). These tests are great: in one easy procedure they tell us whether our scores are normally distributed (nice!).

The Kolmogorov-Smirnov (K-S from now on) test can be accessed through the *explore* command (**Analyze⇒Descriptive Statistics⇒Explore...**).[2] Figure 2.6 shows the dialog boxes for the *explore* command. First, enter any variables of interest in the box labelled *Dependent List* by highlighting them on the left-hand side and transferring them by clicking on ▸. For this example, just select the exam scores and numeracy scores. It is also possible to select a factor (or grouping variable) by which to split the output (so, if you select **uni** and transfer it to the box labelled *Factor List*, SPSS will produce exploratory analysis for each group—a bit like the *split file* command). If you click on Statistics... a dialog box appears, but the default option is fine (it will produce means, standard deviations and so on). The more interesting option for our purposes is accessed by clicking on Plots... . In this dialog box select the option ☑ Normality plots with tests, and this will produce both the K-S test and normal Q-Q plots for all of the variables selected. By default, SPSS will produce boxplots (split according to group if a factor has been specified) and stem and leaf diagrams as well. Click on Continue to return to the main dialog box and then click OK to run the analysis.

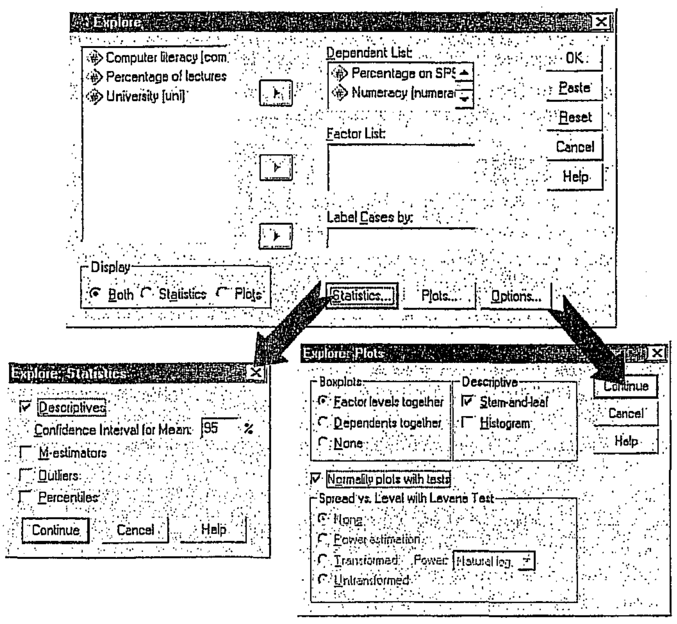


Figure 2.6: Dialog boxes for the *explore* command

2.5.2. *Output*

The first table produced by SPSS contains descriptive statistics (mean etc.) and should have the same values as the tables obtained using the frequencies procedure. The important table is that of the Kolmogorov-Smirnov test. This table includes the test statistic itself, the degrees of freedom (which should equal the sample size) and the significance value of this test. Remember that a significant value (*Sig.* less than 0.05) indicates a deviation from normality. For both numeracy and SPSS exam, the K-S test is highly significant, indicating that both distributions are not normal. This result is likely to reflect the bimodal distribution found for exam scores, and the positively skewed distribution observed in the numeracy scores. However, these tests confirm that these deviations were *significant*. This finding is important because the histograms tell us only that our sample distributions deviate *from normal*; they do not tell us whether this deviation is large enough to be important.

[2] This menu path would be **Statistics⇒Summarize⇒Explore...** in version 8.0 and earlier.

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
Percentage on SPSS exam	.102	100	.012
Numeracy	.153	100	.000

a. Lilliefors Significance Correction

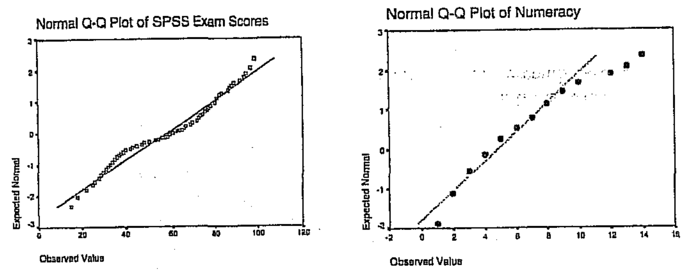


Figure 2.7: Normal Q-Q plots of numeracy and SPSS exam scores

SPSS also produces a normal Q-Q plot for any variables specified (see Figure 2.7). The normal Q-Q chart plots the values you would expect to get if the distribution were normal (expected values) against the values actually seen in the data set (observed values). The expected values are a straight diagonal line, whereas the observed values are plotted as individual points. If the data are normally distributed, then the observed values (the dots on the chart) should fall exactly along the straight line (meaning that the observed values are the same as you would expect to get from a normally distributed data set). Any deviation of the dots from the line represents a deviation from normality. So, if the Q-Q plot looks like a straight line with a wiggly snake wrapped around it then you have some deviation from normality! In both of the variables analysed we already know that the data are not normal, and these plots confirm this observation because the dots deviate substantially from the line. It is noteworthy that the deviation is greater for the numeracy scores, and this is consistent with the higher significance value of this variable on the Kolmogorov-Smirnov test. A deviation from normality such as this tells us that we cannot use a parametric test, because the assumption of normality is not tenable. In these circumstances we can sometimes turn to non-parametric tests as a means of testing the hypothesis of interest. In the next section we shall look at some of the non-parametric procedures available on SPSS.

BLACK, Thomas, R. 1999. *Doing Quantitative Research in the Social Sciences. An Integrated Approach to Research Design, Measurement and Statistics*. London: Sage.

The Normal Distribution of Probabilities and Z-Scores

We will now progress to the situation where discrete or continuous data have been collected and we are confident that we are dealing with a normally distributed trait. Even though our sample data may not provide a perfectly normal curve, they are close enough to assure us we can proceed, and we have some independent evidence that the trait is intrinsically normally distributed. The next question is: what can the data reveal?

It is possible to glean a certain amount of information when provided with the mean and standard deviation for a distribution. Such information will assume a normal distribution for the population and therefore will use its intrinsic shape as the basis for discussing probabilities of occurrence of events. For example, IQ tests are actually designed to have a mean of 100 and a standard deviation of 15. Referring to Figure 12.21, one would expect that about 68% of all persons taking an IQ test will have an IQ of between 85 and 115. One way of indicating an individual's performance is to state his or her position on the horizontal axis in terms of the percentage of examinees performing below this position, the *percentile group*. In other words, if John did better than 67% of the other people taking an exam, then John was in the 67th percentile group. If you have an IQ score of 115, one standard deviation above the mean, then your score is better than 84% of all persons taking that examination (50% below the mean plus 34% up to the first standard deviation). This also means that visually, 84% of the area under the curve is to the left, as shown in Figure 13.5.

It is possible to identify where in a distribution an individual score lies when the mean and standard deviation are known. It is relatively easy to convert a raw score into a number of standard deviations, called a *z-score*, which can be found in a table to see exactly in what percentile group that score falls:

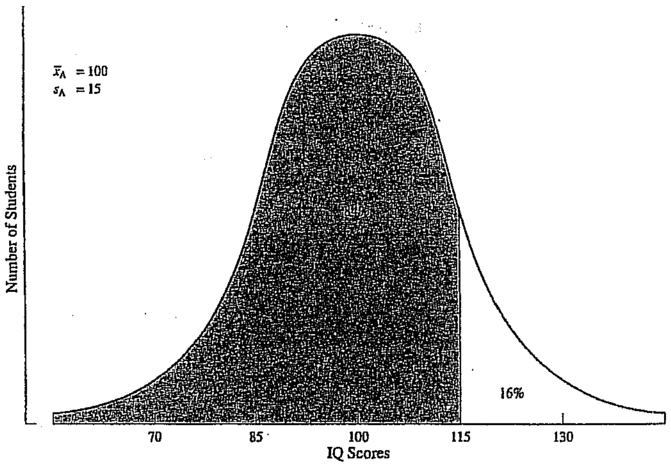


FIGURE 13.5
The 84th percentile group for IQ scores

$$z\text{-score} \equiv \frac{\text{raw score} - \text{mean}}{\text{standard deviation}}$$

Expressed in mathematical symbols,

$$z_i = \frac{x_i - \bar{x}}{S} \qquad (13.1)$$

where x_i is the individual score for person i, $\bar{x}$ is the mean of the distribution of scores and S is the standard deviation of the distribution from equation (12.4). For example, an IQ score of 92 would be:

$$z = \frac{92 - 100}{15} = \frac{-8}{15} = -0.53$$

or 0.53 standard deviations *below* the mean. Looking this up in Table B.1 in the Appendix B, reveals that the score corresponds to a percentage score of 20.19% below the mean. Subtracting this from the 50% total below the mean results in this score being in the 29.81 percentile. In other words, this person scored higher than 29.81% of the persons taking this test and 70.19% did better than this person. This simply tells how an individual with this score performed with respect to all the others. What decisions are made based upon such results is the domain of the researchers or other persons using these data. Now try Activity 13.4, where you are asked to find equivalent z-scores for raw scores.

> Activity 13.4
>
> Find the percentile group for IQ scores of 110, 98 and 120, using Table B.1 in Appendix B.
> *Answers (rounded values)*: $z = 0.67$ *thus* 75%, $z = -0.13$ *thus* 45%, $z = 1.33$ *thus* 91%

The IQ score distribution is based upon population data, whereas in many situations one would be finding z-scores based upon an estimate of the population mean and standard deviation provided by sample data. The assumption is that the distribution will not be greatly different if the sample is truly representative. As noted earlier, in most situations, population data will simply not be available anyway.

LOETHER Herman J. and Donald G. McTAVISH. 1988. *Descriptive and Inferencial Statistics. An Introduction.* Boston, Allyn and Bacon.

15

The Logic of Hypothesis Testing

Suicide is obviously an individual act with psychological overtones. Emile Durkheim, the famous French sociologist, maintained, however, that the regularity and predictability of suicide rates over time could not be explained by psychological variables. He was convinced that suicide rates were explainable, rather, in terms of "social facts." He said,

> If, instead of seeing in them only separate occurrences, unrelated and to be separately studied, the suicides committed in a given society during a given period of time are taken as a whole, it appears that this total is not simply a sum of independent units, a collective total, but is itself a new fact *sui generis*, with its own unity, individuality, and consequently its own nature—a nature, furthermore, dominantly social. (Durkheim, tr. 1951:46.)

With the collective nature of the suicidal act in mind, Durkheim attributed the regularity of its rates for various populations to the social fact of group solidarity, or the lack thereof. He theorized that people lacking support from group solidarity are most vulnerable to suicide. His theory did not seek to explain the dynamics of individual suicides; it sought, rather, to explain suicide *rates* in terms of the differential vulnerability of various cohorts of people.

Although Durkheim's work was completed more than 90 years ago, it is considered a sociological classic because it illustrates the proper relationship between theory and data. In spite of soft spots in his methods, his general theoretical notions about suicide as a sociological phenomenon are still relatively sound.*

* Douglas (1967) critiques Durkheim's work and reviews further studies of suicide.

From his theory Durkheim derived some specific hypotheses about the relationship between social solidarity and suicide rates. Even though he never defined social solidarity in a rigorous manner, he did indicate various indices of it that could be observed and measured. For instance, he felt that social solidarity varied from one religious persuasion to another. He reasoned that social solidarity was highest among Jews, next among Catholics, and lowest among Protestants. He attributed these differences to differences in the degree to which the lives of individual members were dominated by their religions. Because Protestantism allowed for more free inquiry and placed more responsibility on the shoulders of the individual, it fostered less social solidarity than the other two. Accordingly, Durkheim hypothesized that suicide rates would be highest for Protestants and lower for Catholics and Jews. To test his hypothesis, he examined suicide statistics for various European countries. In general, he found that his hypothesis was supported by the data. For example, in the states of Germany suicide rates varied in direct proportion to the number of Protestants and in inverse proportion to the number of Catholics (Durkheim, tr. 1951:153). Moreover, those European countries that were predominantly Catholic (*e.g.*, Portugal, Spain, and Italy) had low suicide rates as compared with high rates for predominantly Protestant countries, and the rates for mixed Catholic–Protestant countries were intermediate (Durkheim, tr. 1951:152). When the rates of Protestant, Catholic, and Jewish groups were compared, Protestant rates were consistently higher than those of the other two, and Jewish rates were generally lower than those for Catholics (Durkheim, tr. 1951:155).*

Another index of social solidarity examined was marital status. Durkheim reasoned that the married enjoyed more social solidarity than the unmarried or widowed; thus, suicide rates would be lower for them. Again, the data tended to support his hypothesis. Rather consistently, for each age category, married persons had lower rates than unmarried or widowed persons (Durkheim, tr. 1951: 176–177). Furthermore, suicide rates for married persons with children were consistently lower than those for married persons without children (Durkheim, tr. 1951:186 ff.).

Durkheim went on to examine several other variables that he took to be indices of social solidarity, in each case comparing suicide rates from several different sources. The data generally supported his theory.†

Durkheim's basic approach was to develop theory to account for the regularity and predictability of suicide rates in Europe in the latter part of the 19th century. His major explanatory concept was social solidarity. He *selected* a number of measurable variables to be indices of social solidarity, *deduced* a number of specific hypotheses relating these indices to suicide rates, *gathered* all available data bearing on his hypotheses, and *examined* them to see whether they lent

* It should be noted that Durkheim's statistics relating suicide rates and religion were for countries or areas of countries and not for individuals. If predominantly Protestant countries have high suicide rates, it does not necessarily follow that those who are committing suicide are Protestants (*see also* Robinson, 1950.)

† Although the statistical techniques known to us were not available to Durkheim, on a raw level he duplicated the reasoning underlying modern statistics. Sir Francis Galton invented correlation earlier (Galton, 1886), but the technique was not generally known nor understood at the time that Durkheim conducted his study.

LOETHER Herman J. and Donald G. McTAVISH. 1988. Descriptive and Inferencial Statistics. An Introduction. Boston, Allyn and Bacon.

support to his predictions. He *concluded* that the data, on the whole, supported his hypotheses and theory. Therefore, he offered his theory as a fruitful explanation of suicide rates.

The point of this exercise in social science was to develop a theoretical explanation for a social phenomenon useful for predicting the same class of phenomena in the future. Durkheim did not *prove* his theory any more than any scientific theory is ever proven. He did demonstrate, however, that his findings were useful for predicting suicide rates. This is the way science uses theory. A theory is never proven; it is demonstrated to be useful or not useful, and is used or revised. If a competing theory is developed that predicts better, it is substituted for the previous theory and is used until it is replaced, in turn, by another theory that is an even better predictor.

The process of developing theoretical explanations of social behavior is what sociology is all about. Science is, after all, a continual interplay between theory and data. The examination of data in a systematic manner gives rise to theory; theory gives direction to the collection of new data; and new data either give additional support to the theory or contribute to its refutation. The game of science is concerned with the selection of the most useful theory from a number of competing ones. The *application* of scientific knowledge by practitioners and policymakers involves the practical use of the most fruitful theory currently in vogue.

When Durkheim developed his theory he had in mind an explanation that would transcend particular populations. He did not, however, use sampling techniques nor did he worry whether his data were representative of either general or special populations. The data he used were descriptive of specific geographic areas at specific points in time. He did examine sets of data descriptive of a number of *different* special populations. In effect, he replicated his hypothesis tests. A very important principle of science is that hypothesis tests should be repeated independently a number of times to gauge their soundness. The impressive aspect of Durkheim's work was the consistency with which these separate sets of data upheld his hypotheses.

15.1 STATISTICS AND HYPOTHESIS TESTING

When concepts that appear in sociological theories are not defined in measurable form, the hypotheses linking the concepts are not directly testable. The usual procedure in such cases is to specify measurable indices of the concepts, frame hypotheses relating the indices, and test these "**working hypotheses**" against empirical data.

Figure 15.1 illustrates the relationship between a general theoretical hypothesis and a working hypothesis using Durkheim's study of suicide as an example. Robert E. Clark conducted a study designed to test the general hypothesis that incidence of mental disorders varies with occupational status (Clark, 1949). As indices of mental disorders Clark used diagnostic categories assigned to patients in mental hospitals in the Chicago area. He looked separately at rates for alcoholic psychoses, senile psychoses, paresis, manic-depressive psychoses, and schizophrenia among patients whose occupations were known. He used two indices of occupational status: a measure of the prestige of the occupation on the North-Hatt scale,* and the median income for the occupation in the Chicago area at the time of the study.

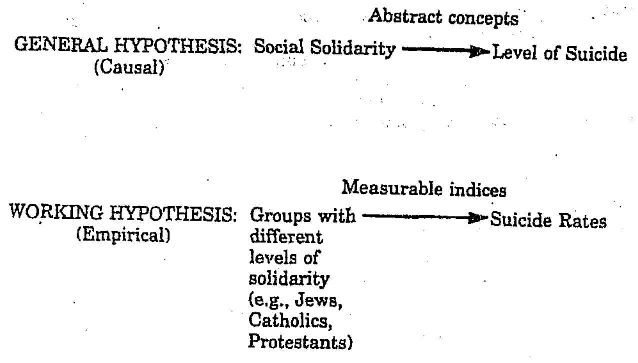

Abstract concepts
GENERAL HYPOTHESIS: Social Solidarity → Level of Suicide
(Causal)
Measurable indices
WORKING HYPOTHESIS: Groups with different levels of solidarity (e.g., Jews, Catholics, Protestants) → Suicide Rates
(Empirical)

FIGURE 15.1 RELATIONSHIP BETWEEN A GENERAL THEORETICAL HYPOTHESIS AND A WORKING HYPOTHESIS: DURKHEIM'S THEORY OF SUICIDE

From the one general hypothesis Clark framed several working hypotheses, relating indices of mental disorders to indices of occupational status. When he tested his working hypotheses against the data, he found that his general hypothesis had to be qualified. Alcoholic psychoses, senile psychoses, paresis, and schizophrenia were inversely related to occupational status (the higher the occupational status, the lower the incidence of the disorder), but manic-depressive psychoses were unrelated (Clark, 1949:440). Although Clark's substantive findings are interesting in themselves, we are concerned primarily with the procedures used to test the general hypothesis. Clark selected indices of his theoretical concepts and recast his general hypothesis in terms of these indices, thus deriving working hypotheses; then he subjected his working hypotheses to empirical tests and, on the basis of these tests, drew conclusions about the general hypothesis.

When population data are available, the indices serve as parameters; and decisions about hypotheses can be made simply by examining the parameters (no test of significance is needed). When Durkheim compared suicide rates in Catholic Bavaria with those in Protestant Prussia he merely had to take note of the fact that the Prussian rates were higher. Since these rates were parameters for his populations, the differences between them were actual differences (assuming that the data were error free). Therefore, descriptive statistics allow for direct tests of hypotheses without further complications.

Unfortunately, population data are not usually available. We are faced with the necessity of examining sample data and making generalizations about the population. The procedure for testing hypotheses with sample data is as follows:

* For a description of the North-Hatt Scale see Reiss (1961).

LOETHER Herman J. and Donald G. McTAVISH. 1988. Descriptive and Inferencial Statistics. An Introduction. Boston, Allyn and Bacon.

1. *Specify indices of the concepts included in the general hypothesis.*
2. *Derive working hypotheses that link the indices.*
3. *Draw a sample of data from the population to which the hypotheses apply.*
4. *Use statistical techniques to analyze the sample data.*
5. On the basis of that analysis, *decide whether the data support the working hypotheses.*
6. *Decide whether the general hypothesis fruitfully describes the population.*

When hypotheses are tested using sample data, the complication introduced is that parameters must be estimated, since they cannot be examined directly because they are not available to the researcher. For example, if Durkheim's comparisons of suicide rates in Catholic Bavaria and Protestant Prussia had been based on sample data rather than population data, he would have been faced with the necessity of estimating suicide rates from his sample data, and then deciding whether the estimated parameters actually differed.

We found in Chapter 14 that estimating parameters involves the use of probability theory applied to sampling distributions. As you will see later, there is a close relationship between interval estimates of parameters and statistical tests of hypotheses. The difference between them is a difference in orientation rather than kind.

> BOX 15.1 SAMPLING DISTRIBUTION
>
> If the concept of the sampling distribution is not yet quite clear to you, perhaps you should go back and review Chapter 13, particularly the units under Section 13.2. An understanding of the concept of sampling distribution is essential to the discussion that follows.

15.1a The Statistical Hypothesis

When we use sample data to test hypotheses, it is necessary to introduce a third type of hypothesis—the **statistical hypothesis**. *We start with a general hypothesis, which we translate into working hypotheses, and from our working hypotheses we derive statistical hypotheses.* Statistical hypotheses make statements about population parameters, but are tested by examination of statistics computed from sample data. As a result of the outcomes of tests of statistical hypotheses, we decide what conclusions about the working hypotheses are warranted and, in turn, these decisions help us to make decisions about the general hypothesis.

Again, we will rely heavily on sampling distributions to help us make decisions. The questions we will ask, however, will be somewhat different from those raised in estimating parameters. We will examine the following questions:

Is it reasonable to conclude that the statistic we have computed from the sample is an estimate of a specific, given parameter?

Are the statistics we have computed reasonably seen as separate sample estimates of a common parameter or do they estimate different, distinct parameters?

In each case primary concern will be with making decisions about hypotheses that refer to parameters or to relationships between parameters. In classical hypothesis testing we must choose between two competing hypotheses.

LOETHER Herman J. and Donald G. McTAVISH. 1988. Descriptive and Inferencial Statistics. An Introduction. Boston, Allyn and Bacon.

15.2 TESTING STATISTICAL HYPOTHESES

Thus far our discussion of hypothesis testing has been fairly abstract. It might be helpful at this point to take a concrete example, run through the process involved in testing a hypothesis, then analyze the procedure involved. In the process of presenting the example we will introduce the concepts that play an integral part in the testing of statistical hypotheses.

One characteristic that particularly distinguishes the developing countries of the world from the others is the rate at which infants and young children die. Infant and childhood diseases that have long since ceased to be serious killers in the industrialized countries still take a terrible toll of babies and small children in developing countries. According to World Health Organization estimates, 3,450,000 children die every year of diseases that are preventable through vaccination (United Nations Children's Fund, 1985). Measles alone is estimated to kill 2 million annually. Is the distinction between the developing and the developed countries merely that vaccinations are more common in the latter than in the former? According to McKeown (1976), malnutrition is an important factor in the whole equation. People who are malnourished are more vulnerable to infection, and thus fall victim to diseases that, in other circumstances, are much less virulent (1976:35).

One might assume that malnutrition would not be a common characteristic of countries that are largely rural because the inhabitants could subsist on foodstuffs they grew themselves. According to Bogue (1969:46), however, predominantly rural settlement patterns are characteristic of developing countries and it is in the developing countries that death rates are high.

To understand better this whole process whereby underdevelopment, malnutrition, infectious disease, and high infant death rates are linked, an informative preliminary step would be to establish the nature of the relationship between rural-urban settlement and the level of nutrition existing nationally. The Food and Agriculture Organization of the United Nations (U.N.) has collected international data on daily per capita calorie supply as a percentage of the requirement necessary for satisfactory nutrition (United Nations Children's Fund, 1985:134–35). These data may be used as a measure of the level of nutrition available for each country covered.

Consistent with what has been said thus far about underdevelopment, nutrition, disease, and death rates, **our working hypothesis will be that countries that are predominantly rural will be less likely to provide the required per capita daily calorie intake than will the countries of the world in general.** In order to test this hypothesis we used the U.N. data to divide the

countries of the world into two categories: those that meet or exceed the percentage requirement of daily calorie intake and those that do not. Using this distinction, we found that 65% of the countries of the world met or exceeded the daily percentage requirement.

Furthermore, we singled out the countries of the world in which more than 50% of their populations were rural and drew a simple random sample of 30 of those countries. When these countries were examined to determine whether their per capita daily calorie intake met or exceeded the percentage requirement it was found that 9 of the countries did and 21 did not. The research question that we wish to answer is whether this distribution for the 30 predominantly rural countries provides evidence in support of our working hypothesis.

LOETHER Herman J. and Donald G. McTAVISH. 1988. Descriptive and Inferencial Statistics. An Introduction. Boston, Allyn and Bacon.

15.2.1 The Null Hypothesis

Framing a statistical hypothesis in a positive manner, we would come up with some such hypothesis as the following: **Countries that are more than 50% rural are significantly less likely to meet or exceed the daily per capita calorie intake requirement than are the countries of the world in general.** The kind of sample data we would consider as evidence of significance would have to be stated explicitly so that we could test the hypothesis.

Since statistical inference is based upon probability theory, tests of statistical hypotheses are probabilistic rather than absolute. It is not possible to prove or disprove statistical hypotheses in an absolute sense. The best that can be achieved is an estimate of their truth or falsity.

It so happens that the rejection of a statistical hypothesis is much more clear-cut than is its acceptance. Therefore, the usual procedure is to frame a statistical hypothesis contrary to that which we are hoping to prove. Such a hypothesis is known as a **null hypothesis.** If sample data warrant rejection of the null hypothesis, that is regarded as evidence for its alternatives—those hypotheses our theory predicted and those we proposed as explanations in the first place. The null hypothesis gets its name from the fact that it is the hypothesis to be nullified by statistical test.

The advantage of using the null hypothesis is that it serves as a basis for selecting a specific sampling distribution, that is, the sampling distribution that would be found if the null hypothesis were, in fact, true. This sampling distribution is then used to determine whether sample data warrant rejection of the null hypothesis in favor of some set of alternatives to it.

Since we wish to seek evidence in support of the contention that the rural countries of the world are less likely to meet daily calorie intake requirements than are countries in general, we wish to show that significantly fewer than 0.65 of the rural countries meet or exceed those requirements. We use the 0.65 because that is the proportion of *all* of the countries of the world that meet or exceed the requirements. It is, therefore, the parameter of interest to us in generating the sampling distribution necessary to test our null hypothesis. The null hypothesis would be formulated as follows: **The proportion of rural countries that meet or exceed the daily per capita calorie intake requirement will not differ significantly from 0.65.** If this null hypothesis were true, we would have a sampling distribution of proportions with an expected value of 0.65, which is the value for countries in which the daily per capita calorie intake met or exceeded the requirement.

The procedure we follow in testing the null hypothesis is: (a) assume that it is true, (b) generate a sampling distribution from the null hypothesis, (c) draw a random sample, (d) collect the relevant sample data, (e) compute the relevant statistic, and (f) decide whether it is reasonable to assume that the statistic came from the given sampling distribution. If the probability that the statistic came from the given sampling distribution is as small as or smaller than some predetermined level, we reject the null hypothesis in favor of its alternatives. If the probability is not as small as or smaller than the predetermined level, we fail to reject the null hypothesis.

Notice that we *fail to reject* the null hypothesis rather than accept it. If we accepted the null hypothesis, we would be saying, in effect, that it is true. However, if we fail to find reason to reject the null hypothesis, it does not necessarily follow that it is true. We are saying, rather, that the data we collected did not provide us with sufficient basis for concluding that the null hypothesis is false. Perhaps, our data collection was merely inadequate.

By the same reasoning, if we *do* reject the null hypothesis, it does not follow that we accept its alternatives. All we imply by rejecting the null hypothesis is that some set of alternatives to it is more probable than the null hypothesis itself.

Note, also, that both the null hypothesis and its alternatives apply not to sample data but to population data—not to statistics but to parameters. We test the null hypothesis with sample data and generalize from statistics to parameters.

NULL HYPOTHESES

The rather convoluted thinking of null hypotheses is necessary if we are going to set the scene for testing hypotheses using statistical tools. As noted in Chapter 1, theories survive and gain support as a result of not being disproved, rather than being proven conclusively. For sound theories, this does not imply a ticking bomb waiting to explode in the form of some researcher in the future proving it wrong. What it does suggest is that researchers are usually trying out components of a theory in different situations or with different groups; they are looking for the *limits* of applicability or refinements in detail. Hypotheses, as described above, express anticipated outcomes as predicted by a given theory or the expected consequence of an application of principles to a situation, stated in more specific terms than those of a general research question.

When it comes to testing hypotheses, all that statistics can tell us is whether the outcomes we ultimately see could have happened due to some causal relationship *or* simply by chance alone. In other words, the effect has to be big enough, whether it is the difference in average scores on some performance task for two groups, or the size of a correlation coefficient. The null hypothesis simply states that 'no significant difference' is expected between what we obtain and what would happen by chance alone. If the difference observed is greater than some minimum, then it is considered significant and whatever has happened (probably) did not occur by chance alone. It is still up to the researcher to prove through sound design and data collection that nothing could have caused the observed effect other than what is described in the hypothesis.

So the next stage in refining our statement of hypotheses would be to try to express them as null hypotheses related to the data that will be collected. As a consequence of a given study, several types of null hypothesis could be generated – for example, describing differences in scores or frequencies of events between the sample and the population (normative), *or* between two groups or among three or more groups – i.e., they actually belong to the same population, not to separate populations (experimental, quasi-experimental or ex post facto). The statements simply anticipate that any difference(s) will be too small to be attributable to anything but chance.

Alternatively, if one were carrying out a correlational study, the null hypothesis of 'no significant correlation' anticipates correlations that will be so small that they could have happened by chance alone. To illustrate this, the hypotheses of Table 2.2 above are provided in Table 2.3 with corresponding possible null hypotheses.

The process of specifying a null hypothesis is one that focuses the attention on what will happen next, stating the implications of the proposed relationship among variables in terms that can be resolved by statistical instruments (see Figure 2.14). At this stage, it is sometimes possible to identify potential difficulties in carrying out the research. For example, where are we going to find the sample of twins implied by the fourth proposal in Table 2.3? Some of the more interesting questions generate very difficult scenarios for resolving them, compelling researchers to rethink the hypotheses resulting from a question. Obviously, it is better to consider such issues early in the research process before too much is invested in an impossible task.

TABLE 2.3 Hypotheses from Table 2.2 and potential corresponding null hypotheses

Hypotheses	Null hypotheses
A random sample of assembly-line workers in factories in Birmingham will be found to suffer a greater frequency of sleep interruptions, and a longer amount of time awake after going to bed, than the population as a whole.	(Both of the hypotheses assume that population data exist.) There will be no significant difference between the mean number of times per night that assembly-line workers in Birmingham awaken and the mean for the population of employed adults as a whole, or between the mean number of minutes that these workers are awake per night and that for the *population of employed* adults.
One of three counselling approaches, A, B or C, will produce a greater reduction in frequency of return to drinking among alcoholics.	There will be no significant difference frequencies of 'dry' and return drinkers across three equivalent sets of alcoholics participating in the three counselling approaches, A, B, C.
It is expected that there will be a negative correlation between social class and drug use, and a negative correlation between educational achievement and drug use for a representative selection of 18–24-year-olds.	There will be no significant correlation between social class and frequency of drug use, or between educational achievement and frequency of drug use for a random selection of 18–24-year-olds (i.e., any correlation will not differ from that which could be expected by chance alone).
For a sample of identical twin boys who are the sons of alcoholic fathers *and* fostered or adopted from infancy separately from each other, one to a family with at least one alcoholic parent, one group will show a greater tendency towards alcoholism than the other.	There will be no significant difference in frequency of alcoholism between groups of separated twins, all sons of alcoholics, when one twin goes to a family with at least one alcoholic parent and the other goes to a family with no alcoholic parents.
In a given hospital, patients on 24-hour prescriptions will be expected to feel more rested if they are awakened for medicines at times that follow REM rather than just at equal time intervals.	There will be no significant difference in the perception of feeling rested, as measured by the Bloggs Restedness Scale completed by patients, between two groups: those whose medication was administered at regular time intervals and those whose medication was administered at times close to times prescribed but following a period of REM.

BLACK, Thomas, R. 1999. Doing Quantitative Research in the Social Sciences. An Integrated Approach to Research Design, Measurement and Statistics. London: Sage.

Testing the null hypothesis

For normally distributed traits, those that produce sample means out in either of the tails of a distribution of sampling means are highly unlikely. Social science researchers commonly accept that events which occur less frequently than 5% of the time are unlikely to have occurred by chance alone and consequently are considered statistically significant. To apply this to a normal distribution would mean that the 5% must be divided between the top and the bottom tails of the distribution, with 2.5% for each (there are occasions when all 5% would occur in one tail, but that is the exception, to be discussed later). Consulting Table B.1 in Appendix B, the top 2.5% is from 47.5% onward, or (interpolating) 1.96 standard deviations (SEMs) or more from the mean. The two ranges of sample means that would be considered *statistically significant*, and result in the rejection of the null hypothesis since they probably did not occur as part of the natural chance variation in the means, are shown shaded in Figure 13.8.

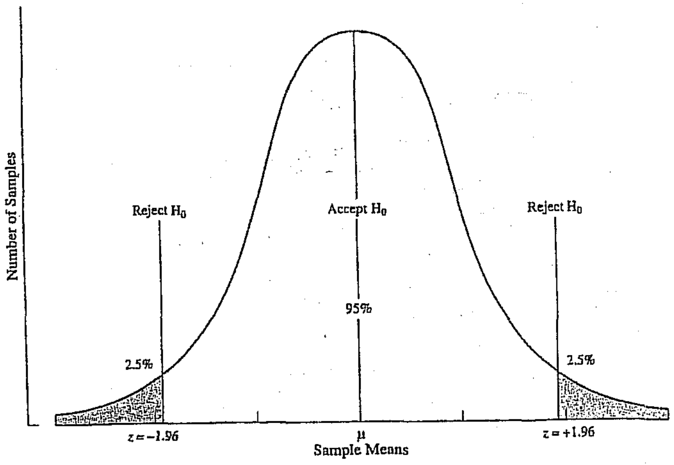


FIGURE 13.8 Normal distribution of sample means with 5% significance levels, where μ is the population mean.

Thus for the situation above involving the mean IQ of the sample of 11-year-olds, the null hypothesis and the statement of expected outcomes need an addition:

> . . . and, is the probability that the difference between the sample mean and the population mean would occur naturally more or less than 5% (the chosen level of significance that will be used as the test criteria)?

The cut-off point of 1.96 standard deviations (SEMs) would correspond to $1.96 \times 2.5 = 4.9$ points above or below the mean. Thus a sample mean IQ of less than 95.1 or greater than 104.9 would be considered significant and the sample not representative of the population. Therefore, in the example, the group with a mean IQ of 106 would be considered statistically significant and the group not typical, and it is unlikely that they are a representative sample of the whole population, for IQ.

Some researchers present results that are supported by an even lower level of probability, usually designated by the Greek letter α, to support their argument, such as 1% ($\alpha = 0.01$), 0.5% ($\alpha = 0.005$), or even 0.1% ($\alpha = 0.001$). Two problems arise with such a practice. First, for the test to be legitimate, one school of thought says the level of significance should be set *before* the test (or even the study) is conducted. Remember that the hypothesis is a statement of expectation, one that should include what will be expected in terms of statistical outcome. It is not fair to write the rules after the game has begun. Second, there is a feeling that a lower significance level than 5% ($p < 0.05$), such as 1% ($p < 0.01$), provides greater support for the results. In other words, if the probability of the relationship existing is only 1 in 100, that must be a stronger statement than if it were only 1 in 20. This supposition will be challenged in Chapter 14 when the concept of the power of a statistical test is introduced.

Poznámky k analýze sociologických dat

JAN ŘEHÁK
Ústav pro filosofii a sociologii ČSAV, Praha

Poznámky shrnuté v této stati nepředstavují úplnou inventarizaci úloh a problémů, s kterými se v běžné praxi analýzy sociologických dat setkáváme; týkají se několika vybraných aspektů statistické analýzy a některých aspektů, které se statistickou analýzou a statistickou inferencí úzce souvisí.

Jde hlavně o to, ukázat na omezení aplikačních možností statistických metod v sociologické praxi. Tato omezení vyplývají jednak z principů metod samých, jednak z kvality dat, s nimiž pracujeme. Statistika je vědní disciplína, která má svou vlastní teorii a metodologii a z těch vyplývají její procedury a přístupy. Teorii samu lze studovat pouze ze speciálních učebnic a se značnými matematickými znalostmi. Určitý pohled na přístupy ke statistické teorii a alespoň letmé a velmi stručné seznámeni se s principy může však snad přispět k fundovanějšímu aplikačnímu postřehu.

Statistika stoji v postavení metody analýzy dat; kvalifikovaná aplikace nutně vyžaduje znalosti teorie statistiky na jedné straně a teoretické znalosti aplikační vědy na straně druhé a proto je taková aplikace většinou obtížným problémem. Kromě toho jsou tu stálé vtíravé otázky o kvalitě dat, s nimiž pracujeme. Proto závěry z dat se vytvářejí součinností inference metodologické, sociologické a statistické. Už tato jednota ukazuje, že analýza dat není vůbec jen rutinní záležitostí, ale součástí tvořivé inference vědecké.

1. Typy statistických závěrů

Statistické závěry činíme o *základním souboru, populaci,* tj. o definované množině statistických jednotek (statistických objektů), která je předmětem našeho zájmu. Základní soubor může být dvojího typu:

a) jednak je to přesně vymezený *konečný* soubor jednotek, který je jednoznačně určitelný například tím, že pořídíme jeho seznam (teoreticky je možno vytvořit seznam všech jednotek konečného souboru vždycky, prakticky to ale nebývá možné u větších souborů);

b) jednak je to soubor určený svými vlastnostmi, ale neomezený na daný konečný počet jednotek; je to soubor, jehož velikost je neznáma a někdy ani známa být nemůže — takový základní soubor můžeme nazvat *otevřeným,* respektive hypotetickým nebo *neurčitým.*

Příkladem souboru a) může být soubor všech školních dětí v určitém obvodu Prahy, příkladem b) může být soubor žáků experimentální školy v Bratislavě a všech žáků, kteří se v podobných podmínkách kdy ocitali, ocitají nebo budou ocitat v budoucnu; soubor je tu definován teoreticky a není možno jej určit jednoznačně seznamem, protože nevíme kolik dětí a které děti se budou učit za stejných podmínek v budoucnu.

Statistické údaje však zřídka získáváme od celého základního souboru. Většinou jsou data k dispozici pouze od jeho vybrané části: každou takovou část základního souboru nazveme *výběrovým souborem, respektive výběrem.* Tak výběrovým souborem v případě a) je například soubor dětí ze tří vybraných pražských tříd; v případě b) je výběrem ona bratislavská experimentální třída.

Data, která získáváme, mohou být klasifikována také z hlediska jejich proměnlivosti u statistické jednotky. Některé znaky jsou *konstantní,* nemění se podle nálady, okamžitého stavu respondenta, nepodléhají chybám při zjišťování, nemění se v čase. Jiné znaky (a těch je většina) jsou *proměnlivé.* Můžeme je považovat za náhodné proměnné, neboť jejich zjištění je podmíněno spoustou nejrůznějších faktorů — subjektivní pocity, chyba při počítání, nepozornost, špatná interpretace otázky či výběr nežádoucího obsahového prvku otázky. Tato data měříme s chybou. Zatímco pohlaví je neproměnné, věk v průběhu

krátkého časového období se nemění, zaměstnání se nemění apod. Otázka „Jak se cítíte spokojen v zaměstnání?" bude zodpovězena v závislosti na mnoha okolnostech a odpóvěď se může měnit den ze dne či někdy z hodiny na hodinu, obzvláště tam, kde odpovědi jsou sémanticky neurčité: velmi spokojen, spokojen atd.

Při typologii situací statistických závěrů můžeme vzít v úvahu obě hlediska současně a naznačit je v tabulce 1.

Tabulka 1.

	Data mají konstantní charakter	Data jsou realizacemi náhodných proměnných — měření s chybou
Data se získávají z celého základního souboru (ZS)	1. není problém zobecnění, úlohy se redukují na vhodné vytváření přehledných měr vlastností souborů	2. eliminujeme chyby měření a náhodnost při získávání dat
Data jsou získávána pouze z výběrového souboru (VS)	3. úlohou je tu zobecnit data z VS na ZS a vyjádřit přesnost zobecnění	4. současně 1. a 3.: tato situace je nejobtížnější a většinou se redukuje na 2. nebo 3.

Otevřený základní soubor je věcí modelu a přijaté konvence. Výběr z takového základního souboru (například experimentální třídu) můžeme považovat za základní soubor s tím, že nás nezajímá jen současný stav, ale proces, resp. výsledek procesu, který vede k současnému stavu. Dostáváme se tím z pole 3 do pole 2. Podle toho, co bylo řečeno o základních souborech, musíme rozlišovat v poli 3 dvě situace — zobecnění na konečný (3a) a na neurčitý (3b) základní soubor. V každém případě pak musíme k úloze přistupovat jinak, tj. tvořit jiné statistické, resp. pravděpodobnostní modely.

Většinou se při analýze dat dostáváme do pole 2 resp. do pole 4 a po redukci do pole 2. To je proto, že na složité otázky sociologického výzkumu odpovídá respondent většinou s možností chyby a subjektivního hodnocení obsahu kategorií odpovědi. Nebude třeba snad argumentovat pro to, že poznání možného zdroje nepřesných odpovědí a informací je velice důležité pro konečnou interpretaci. Některé pomocné „klasifikace" znaků uvádím v dalším paragrafu. (Slovo klasifikace je v uvozovkách, protože jde jen o pomocná třídění, která jsou spíše výhodná než nutná.)

2. Některé pohledy na znaky z hlediska chyb zjišťování

Většina znaků tedy podléhá různým chybám při zjišťování jejich hodnot pro jednotlivce nebo jiné statistické jednotky. Pomíjím zde problémy validity a reliability jako takové a uvádím typy znaků z hlediska chyb, které lze a priori předpokládat a tím i vhodnou metodou měření (či zjišťování) potlačovat na nejmenší míru. Přitom nemám na mysli chyby náhodné, tj. náhodné odchylky od skutečné hodnoty znaku; ty jsou relativně málo nebezpečné, neboť jsou identifikovatelné. (Příklad takové chyby je zjišťování skóre I.Q. testu: při každém zjišťování I.Q. skóre se dá předpokládat, že respondent bude mít poněkud jiný výsledek, přičemž nelze identifikovat žádný faktor, který tuto variabilitu způsobuje.)

Kritéria typů:

a) *Formální vztahy hodnot znaku:* zde rozeznáváme znaky kardinální, ordinální a nominální. Informace čerpaná z těchto vztahů mezi hodnotami znaku je podstatná pro vytváření nejrůznějších populačních měr. Obzvlášť obtížné je zpracování ordinálních znaků, tj. znaků, jejichž hodnoty jsou uspořádané kategorie (obvyklé „kvantifikace" pořadovým číslem odpovědi v uspořádání nejsou považovány vždy za vyhovující); to plyne jednak z faktu, že není známa vzdálenost kategorií na předpokládané škále a dále, že odpovědi bývají často sémanticky neurčité a subjektivně interpretovatelné respondentem.

b) *Přímé zjišťování* × *konstrukce znaků* (složené znaky, škály). Je třeba vyjádřit

vytvářet tak, že jej rozprostřeme symetricky kolem bodového odhadu $\bar{p}$:

$$\bar{p}-d \qquad \bar{p} \qquad \bar{p}+d$$

Šířka d bude zřejmě závislá na přesnosti měření a na spolehlivosti, s jakou chceme náš *konfidenční interval* (interval spolehlivosti) konstruovat. Předpokládejme, že spolehlivost má být 95%; pak

$$d = 1{,}96\,.\,s_{\bar{p}}\,,$$

kde $S_{\bar{p}}$ je výběrová chyba. (Kdybychom chtěli spolehlivost např. 99 %, pak by se koeficient změnil z 1,96 na 2,58 apod.)

Výběrová chyba se vyjádří vzorcem:

$$S_{\bar{p}} = \sqrt{\left(1-\frac{n}{N}\right)\frac{\bar{p}\,(100-\bar{p})}{n-1}}$$

kde N je velikost populace (N = 1000)
n je velikost výběru (n = 200)
$\bar{p}$ je procento ve výběru (postupně 62,5; 78,0; 56,3).

Pro naše tři výzkumníky dostaneme postupně výpočtem (a po zaokrouhlení) tabulku 2.

Tabulka 2

$\bar{p}$	62,5%	78,0%	56,3%
$1-\frac{n}{N}$	0,8	0,8	0,8
$\frac{\bar{p}(100-\bar{p})}{n-1}$	11,78	8,62	12,93
$s_{\bar{p}}^2$	9,42	6,90	10,34
$s_{\bar{p}}^2$	3,07	2,63	3,22
$1{,}96\ s_{\bar{p}}$	6,02	5,15	6,30
interval spolehlivosti	(56,48;68,52)	(72,85;83,15)	(50,00;62,60)

Každý výzkumník dostává zcela odlišné výsledky, pouze první (aniž o tom ví) pokrývá skutečnou hodnotu. Všichni však doufají, že jejich výsledky jsou správné a mají za sebou spolehlivost 95%.

Jak interpretovat slovo *spolehlivost?* Jestliže vybereme náhodně jeden ze souborů o 200 statich, máme 95% šanci, že neuděláme tímto statistickým postupem chybu; v 95 % všech možných výběrů o velikosti 200 vede tento postup k pokrytí skutečného neznámého parametru. A to je též klíč k volbě tohoto čísla. Ve společenských vědách se většinou rozhodujeme pro 95 %, někdy pro 99 %. Podle problému však ve statistických úlohách může být hladina spolehlivosti volena až na 99,9 %, což je případ některých medicínských úloh nebo situací, kdy se na základě měření rozhodujeme pro velké investice.

Máme přirozeně zájem na tom, aby interval spolehlivosti byl co nejkratší. Platí ovšem, že čím větší spolehlivost požadujeme, tím širší interval dostaneme a naopak. Je tu tedy konflikt mezi oběma hledisky. Jednou krajností je bodový odhad, který má spolehlivost 0 % a na druhé straně je tu 100% spolehlivost, té však odpovídá celá škála možných hodnot, takže takový výsledek je prakticky bezcenný.

Výhodou postupu je to, že je v něm už přímo zabudována přesnost — čím užší interval při dané spolehlivosti, tím méně entropičnosti v rozhodování a tím přesnější informace. Nevýhodou však je právě to, že jde o interval, s nímž nemůžeme snadno pracovat při vytváření dalších měr a při srovnávání různých výsledků. Kromě toho je tu stále základní fakt statistických závěrů — jsou to závěry za neurčitosti. Pouze doufáme, že nám je náhoda při měření příznivá. V dlouhé řadě konstrukcí intervalů spolehlivosti očekáváme 5 % případů nepokrytí. Je-li zvolená spolehlivost dostatečně vysoká, musí každý výzkumník rozhodnout sám.

6. Testování hypotéz

Nejčastější úlohy formulované ve společenskovědním výzkumu jsou úlohy testování hypotéz. To jsou případy, kdy nás nezajímá odhadování hodnoty parametru přímo, ale rozhodnutí, zda lze přijmout ten či jiný výrok o parametrech nebo o rozložení veličin vůbec.

Existuje celá řada typů hypotéz. Tak např. výzkumník uvedený ad 4. si mohl postavit hypotézu, že skutečné procento p = 60 %. Test hypotézy pak mohl být ekvivalentní s konstrukcí intervalů spolehlivosti; jestliže interval spolehlivosti pokryje hodnotu 60 %, pak není důvodů hypotézu odmítnout, je-li pak 60 % vně intervalu, považujeme hypotézu za nesprávnou (málo pravděpodobnou). I zde je vidět omezení statistických aplikací. Předpokládejme, že naše tři osoby, každá nezávisle, vyslovily tři různé hypotézy: p = 60 %, p = 65 %, p = 55 %. První a třetí osoba přijímají své hypotézy, které jsou ale odlišné pro tentýž soubor, tj. pro tutéž realitu; druhá osoba odmítá hypotézu (ač hypotetická hodnota je tatáž jako skutečná v souboru).

Další příklady hypotéz:

a) průměrný výsledek bodovaného testu z matematiky je větší u studentů sociologie než u studentů psychologie;

b) existuje průkazná závislost mezi znakem „informovanost o daném objektu" a znakem „postoj k objektu";

c) na základě minulých měření předpokládáme, že se v průměru zlepší počasí a budou létat letadlové linky mezi Prahou a Bratislavou;

d) na základě zjištěných symptomů přijímáme diagnózu D.

Mluvíme tu o statistických hypotézách, tj. o hypotézách týkajících se pravděpodobnostních vlastností náhodných veličin, které měříme a zjišťujeme. Mluvíme-li o statistice, předpokládáme, že překlad z jazyka sociologie do jazyka statistiky, tj. statistická operacionalizace, je už provedena. Konečný výsledek rozhodnutí o veličinách musí být opět převeden zpět do jazyka sociologie.

Mnohdy při testování hypotéz nás výsledky překvapují svou neočekávatelností a zdají se zcela nemožnými — odmítáme hypotézy, které by měly být podle teoretického východiska jednoznačně potvrzeny. Interpretace takovýchto výsledků musí být krajně zdrženlivá a opatrná. Je pochopitelně nutné takový výsledek vysvětlit a na bázi teorie se s ním vyrovnat. Nesmíme však zapomínat, že to může být způsobeno nejrůznějšími faktory a ukvapené závěry o revizi východisek nelze přijmout bez prověření všech kroků ve výzkumu. Uveďme některé možné příčiny:

1. chyby v operacionalizaci teoretických pojmů;
2. špatná konstrukce modelů — tj. špatný překlad ze sociologie do matematiky;
3. chybný plán a realizace sběru dat;
4. špatná volba znaků vypovídajících o objektech; špatná volba *referenčního systému znaků;*
5. nevhodná formulace otázek;
6. sémantická neurčitost otázek — respondent je chápe jinak než výzkumník;
7. špatný zpětný překlad statistického rozhodnutí do pojmového rámce výchozí teorie;
8. je k dispozici příliš málo dat, aby mohly být prokázány formulované hypotézy;
9. data nejsou analyzována podle modelu, který byl aplikován;
10. mohl nastat případ, že právě u této konkrétní hypotézy se projevilo pravděpodobnostní riziko, se kterým pracujeme ve statistice vždy.

Není možné tedy zamítnout škrtem pera východiska, aniž bychom nesledovali celou dlouhou řadu vlivů (z nichž některé zde byly jmenovány), které mohou ovlivňovat celkové závěry a celý proces vědecké inference — nejen její statistickou část.

Statistické testování hypotéz, jako všechny úlohy řešené matematikou, vycházejí z modelu, který je abstrakcí skutečnosti. Z toho plynou další omezení aplikace metody — vždy je zachycen jen některý z aspektů reality.

Shrňme jen stručně některé z možných přístupů k testování hypotéz.

A) Formulujeme hypotézu H a za předpokladu, že platí, odvodíme teoretické pravděpodobnostní chování sledovaných veličin a jevů. Jestliže data, která získáme jsou za této hypotézy málo pravděpodobná, pak hypotézu zamítáme. Tento přístup nazveme přístupem *R. A. Fishera.* Bylo by možno jej charakterizovat jako zeslabenou analogii logického postupu

$$(H \Rightarrow D) \Leftrightarrow (\text{non}\,D \Rightarrow \text{non}\,H)$$

(z hypotézy (H) plyne chování dat (D); nechovají-li se data podle předpokladu, hypotéza neplatí).

Uvažujeme vždy riziko α (odpovídající riziku nepokrytí skutečné hodnoty při intervalovém odhadu, např. 5% resp. v jednotkách pravděpodobnosti 0,05), které je stanoveno jako horní pravděpodobnostní hranice proti hypotéze, jestliže je správná. Nastane-li málo pravděpodobný jev, pak buď hypotéza neplatí, nebo nastal zázrak (R. A. Fisher).

ŘEHÁK, Jan. 1971. „Poznámky k analýze sociologických dat." *Sociologický časopis*, 27 (4), str. 425-433.

B) Přístup *Neyman-Pearsonův* lze charakterizovat dvojicí hypotéz, které jsou postaveny proti sobě. Základní nulová hypotéza (H_0) je testována proti nějaké jiné možnosti — alterativní hypotéze (H_1). Předpokládáme teoreticky dva možné stavy skutečnosti — zde tedy vstupuje teorie ještě o krok dále do statistické procedury — vymezuje možnou alternativu. To umožňuje lepší výběr mezi rozhodovacími pravidly. Je tu možné formulovat hledisko optimálnosti pro takový výběr.

Rozhodovací situace může být naznačena tabulkou

Tabulka 3

		Rozhodnutí pro	
		H_0	H_1
Ve skutečnosti platí	H_0	správné rozhodnutí	chyba 1. druhu
	H_1	chyba 2. druhu	správné rozhodnutí

Rozhodovací pravidla jsou volena tak, aby pravděpodobnost chyby 1. druhu nepřevýšila dané číslo (např. 0,05; 0,01; 0,001; značíme ji obvykle α) a přitom aby pravděpodobnost chyby 2. druhu byla co nejmenší (značíme obvykle β). Rozhodnutí je tu závislé na schopnosti přesně formulovat oba modely odpovídající oběma hypotézám, na počtu pozorování a (jako vždy) na náhodě.

C) *Bayesovský přístup* je založen na Bayesově větě (publikované v roce 1763 Thomasem Bayesem). Tento přístup formuluje hypotézy $H_1 \dots H_k$ a předpokládá, že jsou a priori (před sběrem dat) známy pravděpodobnosti těchto hypotéz; tyto apriorní pravděpodobnosti mívají nejrůznější interpretace a celý tento přístup má mnoho různých škol podle názoru na apriorní pravděpodobnosti. Můžeme snad souhrnně říci, že apriorní pravděpodobnosti odrážejí stupeň našich znalostí a zkušeností o možných jevech, ať už je vyjadřujeme empiricky a na základě dřívějších zkoumání, nebo jako subjektivní názor; mohou reprezentovat také subjektivní důvěru v platnost hypotézy.

Pomocí Bayesovy formule pak opravujeme apriorní pravděpodobnosti na základě evidence ze sebraných dat. Takto získané aposteriorní pravděpodobnosti jsou základem pro statistické rozhodování o hypotézách. Hypotézu, která má aposteriori dostatečně vysokou pravděpodobnost, můžeme přijmout za správnou. Bayesovská analýza dat je technicky značně náročná. Závěry závisí na apriorních pravděpodobnostech a na modelu, který byl zvolen.

D) V případě, že nejsme s to, či nechceme, nebo se neodvažujeme konstruovat matematický model, který umožňuje inferenci o parametrech, můžeme volit *neparametrické techniky*, v nichž jsou předpoklady daleko volnější a nezavádí se parametry, kromě velice obecných (jako je posunutí počátku, změna měřítka apod.). Výhodou těchto technik je, že se obejdou bez modelu; jedna technika zahrnuje daleko více aplikačních situací. Jsou většinou velice jednoduché, snadno pochopitelné a jejich heuristické odvození je jasné. Přesto že jsou „do šíře“ aplikabilnější, nelze doporučit jejich použití tam, kde lze sestrojit model. Do modelu totiž vkládáme už apriorní informaci a ta nám umožňuje analyzovat data efektivněji; pro stejnou přesnost závěru je třeba méně pozorování. Chyby druhého druhu jsou menší. Přínos informace skrze model umožňuje snížit entropičnost rozhodovací situace. Také formální podoba neparametrických technik tento fakt potvrzuje. Jsou to většinou metody založené na pořadí nebo na konstatování výskytu nějakých jevů. Kardinální znaky pak pro použití takových testů ordinalizujeme, což je pochopitelně jistou ztrátou informace. Formální stránka, která vede k využití pořadí pozorování však předurčuje neparametrickým technikám velice široké použití ve společenských vědách — mnohdy totiž nejsme schopni získávat kardinální znaky a naše měření jsou pouze pořadového charakteru. (Např. nejsme schopni měřit schopnost žáků, ale jsme schopni je podle schopnosti uspořádat.) Formulace úlohy je obdobná jako u Neyman-Pearsonova přístupu; zavádíme též nulovou a alternativní hypotézu a chyby 1. a 2. druhu. Rozdíl spočívá v tom, že zde nezavádíme parametrické modely.

7. Typy statistických hypotéz

Pro lepší orientaci v různých úlohách testování hypotéz je možné rozčlenit je do

tří nejčastěji se vyskytujících typů. Toto členění je pomocné a má úlohu pouze orientační (z hlediska formálního i terminologického mu lze leccos vytknout). Velká většina hypotéz je v současné době založena na Neyman-Pearsonově přístupu a neparametrických technikách, které jsou jeho rozšířením.

Ve výzkumné praxi se vyskytují nejčastěji tři typy statistických hypotéz:

1. *Hypotézy o stavu populace — testy dobré shody*

H_0 je hypotéza, která říká, že pro danou populaci platí určitý model, resp. určitý výrok o neznámých parametrech. Termín testy dobré shody je tu chápán šířeji než ve statistické literatuře. Jde obecně o tvrzení, že populace je v nějakém daném hypotetickém stavu. Jako příklady mohou sloužit:

a) veličina má normální rozložení;

b) víme, či předpokládáme, že veličina má normální rozložení a H_0 specifikuje, že má průměr 38,5;

c) průměrné skóre I.Q. testu v dané konečné populaci je $\overline{X} = 70$ bodů (ze 100 možných).

V případě c) např. můžeme formulovat řadu alterativních hypotéz:

$H_1 : \overline{X} \neq 70$ $\qquad$ $H_4 : \overline{X} > 70$

$H_2 : \overline{X} = 80$ $\qquad$ $H_5 : \overline{X} < 70$

$H_3 : \overline{X} = 62$ $\qquad$ $H_6 : \overline{X} < 50$

a mnoho dalších; výběr z nich záleží na situaci a teorii, kterou vkládáme do procesu inference.

Nulová hypotéza může být specifikována na rozložení četností, průměr, procento, rozptyl, medián, symetrii apod.

2. *Hypotézy o srovnání dvou nebo více populací — testy homogenity*

V těchto testech se promítá do statistiky srovnávací metoda. Těmito testy se srovnávají dvě nebo více populací, resp. některé rysy populací. Tak např. je možno porovnávat rozložení četností, aritmetické průměry apod.

Nulová hypotéza bývá formulována obvykle jako hypotéza homogenity, tj. hypotéza, že soubory jsou z daného hlediska statisticky podobné (je možno je smíchat a výsledné rozložení, či zkoumaná vlastnost je stejná pro celý soubor právě tak, jako pro jeho původní části). Alternativní hypotézy specifikují rozdílnost: například rozdílnost průměrů nebo jejich uspořádání atd.

3. *Testy o struktuře vztahu mezi proměnnými — testy závislosti*

V tomto případě (který by bylo možno obecně zahrnout do první kategorie) je předmětem rozhodovacího procesu pouze jedna populace, ale více proměnných, jejichž vztah nás zajímá.

Nulová hypotéza je většinou formulována jako nezávislost znaků. Alternativní hypotéza pak je specifikována kauzálním modelem, který nás zajímá, tj. strukturou vztahů, které chceme prokázat statisticky.

Je přirozené, že existuje řada dalších typů hypotéz a testů pro ně a že toto dělení je jen zcela hrubé a pomocné. Přesto může být snad užitečné při formulaci statistických úloh a při následné interpretaci.

8. Závěr

Tyto poznámky neměly být receptem jak analyzovat data, ale měly čtenáře seznámit s tím, jaké úlohy před námi ve statistice stojí, jaké požadavky můžeme formulovat a hlavně naznačit, s jakými omezeními při použití musíme počítat, jaká omezení máme. Jsou do jisté míry reakcí na současný stav aplikace statistických metod v sociologickém výzkumu . Musím tu varovat před přespřílišným a hlavně nekvalifikovaným používáním testů a jiných procedur matematické statistiky. Nekvalifikovaná aplikace udělá obvykle více škod než užitku. Domnívám se, že aplikace matematické statistiky a matematiky vůbec má velký význam pro analýzu dat i pro modelové účely a výstavbu dílčích teorií, že se bez nich společenská věda, která chce kvalifikovaně hodnotit jednotlivé měřené aspekty skutečnosti, neobejde; soudím však, že je třeba postupovat opatrně a odborně. Chtěl bych varovat před magií čísel a před absolutizováním četnosti jako důkazového materiálu. Doufám, že jsem uvedl dost různých okolností, které takovýto přístup zpochybňují.

SWOBODA Helmut. 1977. *Moderní statistika*. Praha, Svoboda.

5.5 Spolehlivost výběrových souborů — intervaly spolehlivosti

Pravděpodobnost jistoty, spolehlivost, vypovídací schopnost, přesnost, interval spolehlivosti — je to vše totožné? Pokusíme se ještě zřetelněji ukázat souvislosti a rozdíly a především se nad těmito pojmy poněkud zamyslíme.

Co je například *spolehlivost*? Zpravidla jsme tím mysleli onu spolehlivost, která říká, že výpověď je z 90 nebo 95 % správná. To je pravděpodobnost jistoty, která vymezuje také interval spolehlivosti a kterou se budeme ještě zabývat podrobněji.

„Spolehlivost“ se však může vztahovat také na všeobecnou kvalitu statistické výpovědi. V tomto smyslu např. systematická chyba narušuje spolehlivost.

Další otázkou je, zda vytváření výběrových souborů je vůbec spolehlivou vědeckou metodou. Kromě toho experiment musí být zásadně opakovatelný. Specifikem statistických experimentů však je, že nelze téměř nikdy docílit přesně stejných výsledků, nýbrž jen velmi podobných. Zkoumání výběrových souborů má být opakovatelné, jeho výsledky však mohou být pouze navzájem porovnatelné, nikoli totožné.

Další požadavek, který se klade na všechny vědecké pokusy, je *požadavek významnosti, vypovídací schopnosti a důležitosti:* musí se skutečně dokázat to, co má být dokázáno, nikoliv jen něco podobného.

Vraťme se však zpět k 95% *pravděpodobnosti* jistoty, ke stupni jistoty. Co si pod tím máme představit? Můžeme opět pomyslit na rozdělení četnosti pod normální křivkou a předpokládat soubor všech možných výběrových souborů v rozsahu *n* jako normálně rozdělený. Pak 95% pravděpodobnost znamená, že výběrový soubor, který jsme právě náhodně vytvořili, je právě v toleranci 95 %, které dávají střední hodnotu $\bar{x}$, ležící maximálně 2σ od skutečného průměru μ.

Zda postačí 95 % jistoty nebo nikoliv, nelze říci všeobecně. O tom je třeba rozhodnout v každém jednotlivém případě samostatně. Jestliže se nemá vydat moc peněz a postačí-li přibližný přehled (jako např. u mnohých otázek průzkumu trhu), je postačující 90% nebo ještě nižší jistota. Jde-li o zvlášť významné, velmi závažné rozhodnutí, bude snaha dosáhnout 99% nebo ještě vyšší pravděpodobnosti. Za postačující se však zpravidla pokládá 95% pravděpodobnost.

Jestliže chceme dosáhnout vyšší spolehlivosti, je nutno zkoumat větší výběrové soubory. Není-li to z nějakých důvodů možné, pak také nelze dosáhnout žádaného stupně jistoty — a dost! Nepomohou žádné početní triky či okliky a ani jiné komplikované kejkle neposkytnou východisko. Rozsah výběrového souboru a spolehlivost výsledku výběrového souboru jsou také v nejužší logické a matematické souvislosti s úkazem, který jsme pozorovali při prvním výkladu binomického rozdělení a Pascalova trojúhelníku: v protikladu k naivní představě o „zákonu velkých čísel“ neexistuje sice přesné ustálení výkyvů na očekávané hodnotě, přesto však při rostoucím rozsahu výběrového souboru se budou proporce výběrového souboru relativně (ne absolutně!) méně odchylovat od skutečné proporce souboru základního.

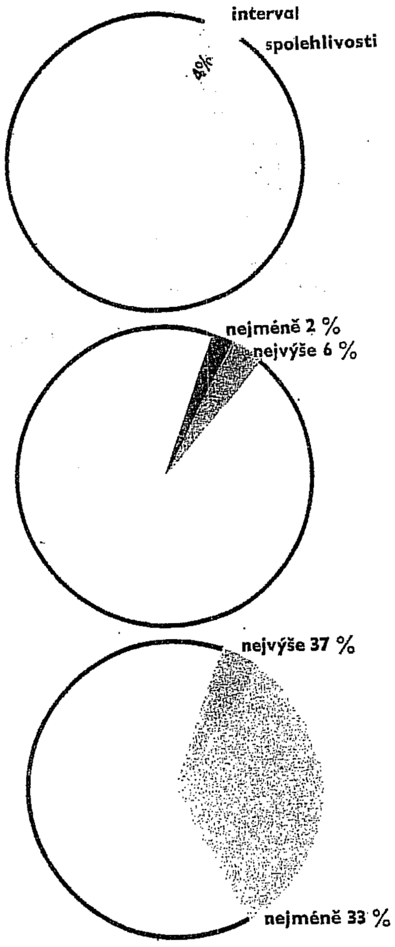
V ruletě to například znamená: je zcela dobře myslitelné, že ve „výběrovém souboru“ dvaceti po sobě jdoucích her je 16 červených a jen čtyři černé, to znamená odchylka o 6 od očekávané hodnoty 10. Naproti tomu je zcela vyloučeno, aby z 2000 her bylo 1600 červených a jen 400 černých. I poměr 1100 ku 900 by byl velmi nepravděpodobný. S rostoucím rozsahem výběrového souboru bude tedy stále přesněji zaměřován skutečný průměr základního souboru (který v případě rulety známe). Při rozsahu výběrového souboru $n =$ $= 20$ se mohu velice zklamat (např. jako u neuvěřitelného výsledku 16 červených, 4 černé), při $n = 200$ se však mohu spolehnout, že neskončím příliš daleko od pravdy. To je „zákon velkých čísel“ v teorii a praxi výběrových souborů.

Pravděpodobnost výběrového souboru mohu také definovat ze záporného hlediska a obdržím „*riziko chyb*“; 95% pravděpodobnost znamená 5% riziko chyb. Máme-li pro „pozitivní" a „negativní" dva různé výrazy, stane se situace ještě zmatenější, když se začne mluvit o „hladině významnosti". Není to zvlášť šťastně zvolený výraz, ale přesto je velmi rozšířen. „*Významnost*" je prakticky totéž co pravděpodobnost: výsledek je „statisticky významný na úrovni 5 %", protože by jen čirou náhodou nenastal v 95 % případů. Celá rozsáhlá oblast testování hypotéz (viz kap. 6), která nás bude ještě velmi zaměstnávat, se zakládá na takovýchto úvahách a výpočtech: je nebo není dosažený výsledek slučitelný s tou nebo onou hypotézou?

„Významnost" není tedy nic jiného než dohoda mezi těmi, kdo statistické metody používají. A zcela stejně jako u pravděpodobnosti lze různě posuzovat úroveň významnosti v závislosti na kladení otázek. *Zpravidla se všeobecně pokládá úroveň 5 %* (nazývaná také často úroveň 95 %) *za „významné", úroveň 1 % (99 %) za „vysoce významné"*.

Pravděpodobností jistoty je také zároveň určen *interval spolehlivosti*. Tento interval se měří jako směrodatná odchylka normovaného normálního rozdělení nebo také i v absolutních hodnotách. Měli jsme již příklady pro oba způsoby vyjádření: Úspory mají rozptyl $s = 60$ kolem průměru $\bar{x} = 480$, interval spolehlivosti (95 %) sahá od 360 do 600 DM, totiž od $(480 - 2\sigma)$ až do $(480 + 2\sigma)$. Mám 95 % jistoty, že náhodně vybraný rolník má více než 360 a méně než 600 DM úspor.

Poznámka: Pozor na záměnu *pravděpodobnost jevu a pravděpodobnost statistického rozboru*. Pravděpodobnost četnosti nějakého znaku zjištěná ve vzorku (proporce výběrového souboru p) je nezávislá na pravděpodobnosti jistoty tohoto vzorku. Mohu vytvořit tři vzorky o rozsahu $n_1 = 20$, $n_2 = 100$ a $n_3 = 500$, které (náhodou) všechny mají četnost výběrového souboru $p = 0{,}7$ (tedy 70 %) znaku. Avšak kvalita, vypovídací schopnost, pravděpodobnost jistoty malého výběrového souboru budou podstatně nižší než u většího souboru.



Interval spolehlivosti ve výši 4 % má různou váhu u malých a velkých podílů. Jestliže podíl vzorku činí např. 4 %, sahá interval spolehlivosti od 2 do 6 %, tedy až do trojnásobku. Činí-li podíl vzorku naproti tomu 35 %, je rozpětí „mezi 33 a 37 %" dostatečně informativní a přesné; 37 % je totiž jen asi o desetinu větší než 33 %.

STATISTICAL INFERENCE

Now that we have established a background in probability and have seen what a normal distribution can reveal, what can a statistical test tell a researcher? It *cannot* prove that a change in one variable caused a change in another, but it can tell whether the difference in mean scores observed between those experiencing one treatment and the usual population that did not could have occurred as a random event. If the test says that it is unlikely that the difference occurred by chance alone, it is still up to the researcher to prove that the one variable was the only possible cause. Statistical tests are like the 'idiot lights' on the dashboard of

your car: they only tell you that *something* has happened, but not exactly why. For example, if the oil light comes on, we assume something is not right. It could mean the engine is low on oil, the engine bearings have worn out, the oil pump has perished, the signal-sending device on the engine is broken, or a wire has shorted out to the light. The motorist obviously checks the oil level first, but if that is adequate, then it is time to call the mechanic, who will try to find the reason for the light being on. In the social sciences, the researcher should plan a study such that when the light comes on (the statistics indicate that something probably happened), then there is only one predicted, defensible link or potential cause. As seen in Chapters 1–7, designing a study to resolve such issues is not trivial.

The term *inferential statistics* refers to the process of using data collected from samples to make inferences about a larger population or populations. The research process introduces complications since:

- most research involves samples (which are *probably* representative);
- the traits usually result in distributions of scores, thus the group characteristics or tendencies are best described as measures of central tendency, such as means;
- the natural variability (with hopefully little error due to low reliability of the instrument) is best indicated by standard deviations.

Using this information, there is a desire to compare groups to determine relationships that will ultimately extend back to the original population(s). Thus any comparisons will require the nature of the distribution to be considered as well as the central tendency of the group. All of this depends heavily upon probability, and it is never possible to speak about relationships with absolute certainty, a fact that causes a distinct amount of mental anguish for most people who feel that events should have some degree of certainty.

There is a need to state the expected outcomes of inferential statistical research in terms of the *null hypothesis*: that there will *not* be any statistically significant difference. In other words, it is expected that any differences or changes or relationships found will be attributable to chance alone. Even if the null hypothesis is rejected, it only means that the difference or occurrence witnessed *probably* did not occur by chance alone. This probability level traditionally has been set at a critical level of 5%, which basically means that if a statistical test says that the probability of this event occurring by chance alone is less than 5%, then it probably did *not* occur as a random event. At this level, there is something probably influencing the event(s), or at least the event(s) has/have occurred as the result of some external influence other than natural random fluctuation. Exactly what this influence is, is not made clear by the statistical test. As noted before, it is still up to the researcher to justify that what he or she did or the variables identified, were the only possible source of influence.

This section will bring together earlier concepts from probability and combine them with research questions and hypotheses, and apply them to the cases where the variables are normally distributed. Before the actual choice of statistical tests can be considered, it is necessary to take a brief mathematical look at what underlies statistical inference and significance. This will be done graphically as much as possible, since most decisions are made on the basis of where the means of sets of data are in a normal distribution. It will provide a

basis for later chapters that will continue the review of inferential statistics by considering a variety of specific tests which can be used as part of experimental, quasi-experimental and ex post facto designs to decide the acceptability of stated hypotheses.

Linking probability to statistical inference

Just as individual scores for a trait vary around a mean to form a normal distribution, the means of samples themselves will vary if a number of representative samples are taken from a population. Thus, if the frequencies of these means are plotted on a graph it is not surprising that we find yet another normal distribution. This *distribution of sampling means* will be quite useful in making inferences about the population. This was introduced earlier in Chapter 5 with reference to sampling error. Figure 13.6 shows all three types distributions for IQ scores: (a) an exemplar population distribution with parameters provided; (b) a single-sample distribution with its statistics; and (c) a distribution of sampling means.

The IQ score is used here simply because it is one distribution for which the population parameters are known, since the tests are designed to produce a mean of 100 and a standard deviation of 15. As will be seen later, this is the exception, since we rarely know what the population mean is. The situation where the population mean is known is used here primarily because it is the simplest and easiest to use to illustrate the principles behind statistical significance. Once this foundation is laid, all the others are basically variations on this ideal.

Remember that when the term *population* is used, it refers to a group sharing a limited set of common characteristics. In social sciences, these are often not

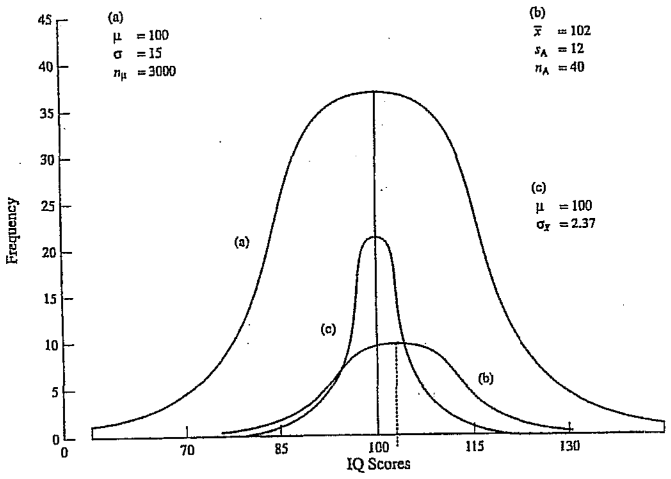


FIGURE 13.6
(a) The population distribution of IQ scores for all 3000 11-year-olds in a local education authority (LEA); (b) a single exemplar sample distribution of IQ scores of a random selection of 40 11-year-olds in the LEA; and (c) the distribution of sample means for a number of such random samples of 40 students

obvious to the casual observer and require some form of detailed observation, measurement or questioning of the subjects. So initially, the issue is whether or not a sample as a group is similar enough to the population for the trait or characteristic in question to be considered representative. A statistical test should be able to resolve what is enough.

The first thing to notice in Figure 13.6 is that the standard deviation (and width of the bell-shaped curve) for the distribution of sample means is relatively small compared to the standard deviations for the population and any single sample. Thus it is very unlikely that a truly representative sample will have a mean very different from that of the population. This fact is used in the most basic of inferential statistical tests, deciding whether a sample is to be considered part of a defined population, or part of some other population. To distinguish this standard deviation from that of a sample of the population, the standard deviation of the distribution of sampling means is used, which is known as the *standard error of the mean* (SEM). This will be designated by $\sigma_{\bar{x}}$ if it is calculated from the population parameter and is found by

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{13.2}$$

where σ is the population standard deviation (equation (12.2)), and n is the sample size. Obviously, the standard error of the mean depends on the sample size: for a very large sample size the standard error of the mean, and consequently the width of the curve for the sampling distribution, will be very small.

It is illustrative to consider an example: in order to carry out a study, a researcher selects a sample of 40 students from the LEA population of 11-year-olds described in Figure 13.6. They are given an IQ test: the group mean is found to be 106. Is this group typical? Let us first state this question as a null hypothesis:

H_0: There is no significant difference between the IQ of the sample group and that of the population.

In everyday English, we would say that we expect that the sample *is* representative of the population for this trait. Here the sample mean will be used to resolve the issue. To make the decision, it is necessary to zoom in on distribution (c) in Figure 13.6, the sample means, shown enlarged in Figure 13.7. The question now becomes one that is stated in terms of probabilities:

What is the probability that a sample with a mean of 106 would be randomly chosen from the population?

Recall that the area under the distribution for a range of scores represents the percentage of people having scores within that range (see Figure 12.21). In this situation, we are considering a distribution of sample means. Using Table B.1 in Appendix B, the number of standard deviations of sampling means (SEMs) can be used to determine what percentage of sample means one would expect below this group's. Here, a sample mean of 106 is 2.40 standard deviations (SEMs) above the population mean, as marked on Figure 13.7. From Table B.1, this

BLACK, Thomas, R. 1999. *Doing Quantitative Research in the Social Sciences. An Integrated Approach to Research Design, Measurement and Statistics*. London: Sage.

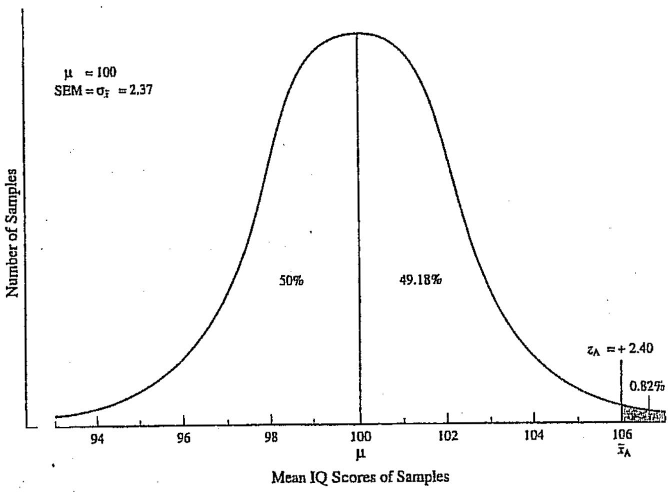


FIGURE 13.7 Distribution of sampling means (each sample size = 40), showing the position of a single sample mean, $\bar{x}_A$

tells us that 49.18% of the sample means would be expected to be between this score and the population mean. Add to this the 50% below the population mean and we find that 99.18% of the sample means should be below this, as shown in Figure 13.7. To put it another way, the probability of this event or any one beyond it occurring as a random event is 100% − 99.18% = 0.82%, or 0.82 of a chance in 100 or 8.2 chances in 1000. Thus this sample mean does seem to be a highly unlikely outcome for a random sample, but what is *unlikely enough* for researchers?

6. lekce

SROVNÁVÁNÍ SKUPIN NA ZÁKLADĚ STŘEDNÍCH HODNOT JEJICH KARDINÁLNÍCH CHARAKTERISTIK (modul Analyze: procedura means). HYPOTÉZA O SHODĚ DVOU PRŮMĚRŮ PRO NEZÁVISLÁ A PÁROVANÁ DATA: T-TESTY A EKVIVALENTNÍ NONPARAMETRICKÉ TESTY (modul ANALYZE: procedury Compare means: one-sample t-test; independent-samples t-test, paired-samples t-test).

Comparing Groups

How can you determine if the values of the summary statistics for a variable differ for subgroups of cases?

- What are subgroups of cases?
- What can you learn from calculating summary statistics separately for subgroups of cases?
- How can you graph means for subgroups of cases?

In Chapter 3 and Chapter 4, you used the Frequencies and Descriptives procedures to calculate summary statistics for all of the cases in your study. Often, however, you are interested in comparing summary statistics for different groups of cases. For example, you want to compare hours studied per week for college freshmen, sophomores, juniors, and seniors. Or you want to find the average income for people living in different geographical areas. There's no easy way with the Frequencies or Descriptives procedure to produce such information. In this chapter, you'll use the Means procedure to calculate simple summary statistics for subgroups of cases. You'll see if you can find a relationship between the average years of education and job satisfaction. You'll also see whether the relationship appears to be similar for men and women. (The Explore procedure described in Chapter 6 lets you examine the values of a variable for subgroups of cases in much greater detail.)

▶ This chapter uses the *gssft.sav* data file, which contains some of the variables in the *gss.sav* file, but for full-time workers only. (See "Case Selection" on p. 541 in Appendix B if you want to know how this smaller file was created.) For instructions on how to obtain the SPSS output discussed in this chapter, see "How to Obtain Subgroup Means" on p. 83.

Education and Job Satisfaction

In Figure 5.1, you see the average years of education and the standard deviation for people in each of four job satisfaction categories. To make comparisons easier to interpret, only people employed full time are included in the analysis.

Figure 5.1 Pivoted Means output for education and job satisfaction

To obtain this output, from the menus choose:

Statistics
Compare Means ▶
Means...

Select the variables educ and satjob, as shown in Figure 5.6.

The Pivot Table Editor is used to put statistics into the columns.

Highest Year of School Completed

Job Satisfaction	Mean	N	Std. Deviation
Very satisfied	14.25	327	2.79
Mod satisfied	13.80	320	2.60
A little dissatisfied	13.92	74	2.75
Very dissatisfied	14.65	26	2.37
Total	14.04	747	2.70

Statistics for individual subgroups

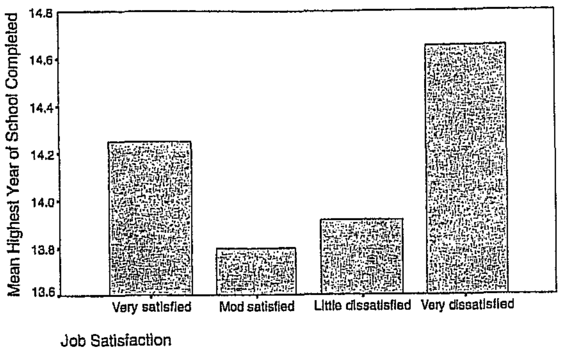
Statistics for all cases analyzed

Looking at the last row of Figure 5.1, you see that the 747 people who are employed full time have 14.04 years of education on average, with a standard deviation of 2.70 years. These 747 people are assigned to one of four subgroups, based on their job satisfaction. The first subgroup, *very satisfied* employees, are somewhat more educated than the group as a whole, while the fourth subgroup, *very dissatisfied* respondents, are the most educated of all. People in the two subgroups in the middle—*moderately satisfied* and *a little dissatisfied* employees—have somewhat less education than the group as a whole.

Plotting Mean and Standard Deviation

A plot of the mean years of education for the four subgroups is shown in Figure 5.2. There is a bar for each of the subgroups. The height of the bar depends on the average years of education. You can easily see that the *very dissatisfied* people have the largest mean years of education. The mean years of education for the middle two satisfaction groups appear to be similar. Note, however, that the scale for the axis on which mean education is plotted doesn't start at 0. That makes even small differences in years of education look large on the plot.

Figure 5.2 Bar chart of education by job satisfaction



You can obtain bar charts using the Graphs menu, as described in Appendix A.

In the Define Simple Bar Chart Summaries for Groups of Cases dialog box, select Other summary function and select the variable educ. Select satjob for Category axis.

Layers: Defining Subgroups by More than One Variable

In Figure 5.1, all of the people who are employed full time are subdivided into four groups based only on their answer to the job satisfaction question. If you want to see whether the relationship between education and job satisfaction is similar for males and females, you must subdivide each of the rows of Figure 5.1 further. Figure 5.3 shows summary statistics for full-time workers, subdivided first by job satisfaction and then by gender.

Figure 5.3 Pivoted means output for job satisfaction and gender subgroups

To obtain this output, select the variable sex for layer 2 in the Means dialog box, as shown in Figure 5.7.

The Pivot Table Editor is used to move the statistics to the columns.

Highest Year of School Completed

Job Satisfaction	Respondent's Sex	Mean	N	Std. Deviation
Very satisfied	Male	14.20	179	3.00
	Female	14.31	148	2.54
	Total	14.25	327	2.79
Mod satisfied	Male	13.71	173	2.68
	Female	13.90	147	2.50
	Total	13.80	320	2.60
A little dissatisfied	Male	14.02	41	2.95
	Female	13.79	33	2.53
	Total	13.92	74	2.75
Very dissatisfied	Male	15.27	15	2.40
	Female	13.82	11	2.14
	Total	14.65	26	2.37
Total	Male	14.01	408	2.85
	Female	14.06	339	2.51
	Total	14.04	747	2.70

The total line in each category is the same as Figure 5.1.

Within each category, separate statistics are shown for men and women.

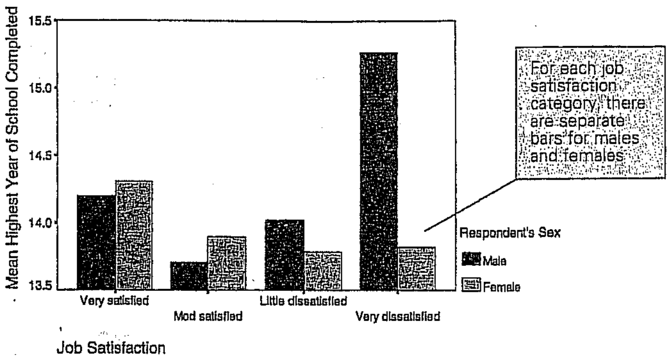
You see that 327 people rate themselves as *very satisfied* with their job. They have an average of 14.25 years of education. Of the 327 *very satisfied* people, 179 are men and 148 are women. The men have an average of 14.20 years of education, while the women have 14.31 years of education. That's not much of a difference. Looking at the *moderately satisfied* males, you see that their average years of education is about half a year less than that of the *very satisfied* males.

One of the more interesting observations gleaned from Figure 5.3 is that the *very satisfied* women have the highest average years of education of all the women. Women in the remaining three satisfaction categories have very similar average years of education. In contrast, the *very dissatisfied* males have the highest average years of education, 15.27. (However, there are few cases in the *very dissatisfied* group, so your conclusions are necessarily tentative.) The *moderately satisfied* males have the smallest average years of education.

Figure 5.4 is a bar chart that displays the results of Figure 5.3. There are four sets of bars corresponding to the job satisfaction categories. Each set of bars has separate bars for males and for females. The conclusions we reached based on the summary table are easier to see from this display. By looking at a corresponding pair of bars, you can see if the average years of education are similar for men and women within each category of job satisfaction. (Boxplots, which are a better way of comparing summary statistics for groups of cases, are described in Chapter 6.)

You can obtain this clustered bar chart using the Graphs menu, as described on p. 500. Select the variables educ, satjob, and sex in the Define Clustered Bar Summaries for Groups of Cases dialog box.

Figure 5.4 Bar chart of education by job satisfaction and sex



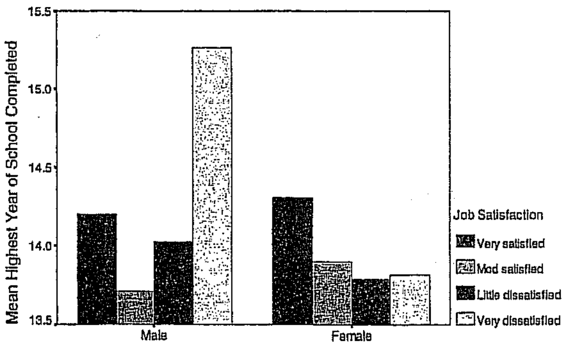
For each job satisfaction category, there are separate bars for males and females

You can also group all of the bars for men and all of the bars for women together, as shown in Figure 5.5.

You can obtain this chart by modifying Figure 5.4, as described in "Bar Charts" on p. 520 in Appendix A. Activate the chart into a chart editor window and from the menus choose:

Series
Transpose Data

Figure 5.5 Bar chart of education by sex and job satisfaction



Now you have two subgroups: men and women. Within each subgroup, you see the four categories of job satisfaction. This plot makes it easy to see that the relationship between job satisfaction and education is not the same for men and women.

Means and standard deviations for groups of cases can be displayed with error bar charts. See "Error Bar Charts" on p. 523 in Appendix A.

What problems are associated with calculating statistics for subgroups of cases? As the number of subgroups you want to compare increases, the sample size in each of the subgroups diminishes. When your means are based on a small number of cases, they are not very reliable. That is, the subgroup means can change substantially if you select another sample from the same population. You'll learn more about the variability of sample means in Part 3.

Summary

How can you determine if the values of the summary statistics for a variable differ for subgroups of cases?

- Subgroups are formed when cases are subdivided into groups based on the values of one or more variables.
- By calculating summary statistics separately for subgroups of cases, you can see if there is a relationship between the summary statistics and the subgroups.
- You can make bar charts of the means of a variable for different subgroups.

Testing a Hypothesis about Two Independent Means

13

How can you test the null hypothesis that two population means are equal, based on the results observed in two independent samples?

- Why can't you use a one-sample *t* test?
- What assumptions are needed for the two independent-samples *t* test?
- Can you prove the null hypothesis is true?
- What is power, and why is it important?

You know how to test whether a single sample of data comes from a population with a known mean. You've tested whether the average cholesterol level for CEO's is the same as the average for the general population, whether college graduates work a 40-hour week on average, and whether the average change in β-endorphin values is 0 during a half-marathon run. In Chapter 12, although you had pairs of observations, you analyzed the differences between the two values and tested the hypothesis that these differences come from a population with a mean of 0.

In this chapter, you'll learn how to test whether two population means are equal based on the results observed in two independent samples—one from each of the populations of interest. You'll use a statistical technique called the two independent-samples t test. You can use the two independent-samples *t* test to see if, in the population, men and women have the same scores on a test of physical dexterity or if two treatments for high cholesterol result in the same mean cholesterol levels.

▶ This chapter uses the *gssft.sav* data file, which includes only cases for people holding full-time jobs. For instructions on how to obtain the independent-samples *t* test output shown in this chapter, see "How to Obtain an Independent-Samples T Test" on p. 250.

Looking at Age Differences

In Part 2 of this book, you examined the relationship between job satisfaction, age, and education for full-time employees. You saw that the average values of age and education vary among the different job satisfaction groups. That isn't surprising, since you know that even if the average ages and educational levels in the population are the same for all job satisfaction groups, the sample means will not be equal. Different samples from the same population result in different sample means and standard deviations. To determine if any of the observed sample differences among groups might be real, that is, not simply the result of the usual variability of sample means from a single population, you need to determine if the observed sample means would be unusual when the population means are equal.

Let's consider what happens if you form two independent groups of people—those who are very satisfied with their jobs and those who are not. You want to determine whether the population values for average age and average education are the same for the two groups. First we'll look at age.

What do you mean by independent groups? Samples from different groups are called **independent** if there is no relationship between the people or objects in the different groups. For example, if you select a random sample of males and a random sample of females from a population, the two samples are independent. That's because selecting a person for one group in no way influences the selection of a person for another group. The two groups in a paired design are not independent, since either the same people or closely matched people are in both groups.

Since you have means from two independent groups, you can't use the one-sample *t* test to test the null hypothesis that two population means are equal. That's because you now have to cope with the variability of two sample means: the mean for *very satisfied* people and the mean for the *not very satisfied* people. When you test whether a single sample comes from a population with a known mean, you only have to worry about how much individual means from the same population vary. The population value to which you compare your sample mean is a fixed, known number. It doesn't vary from sample to sample. You assumed that the value of 205 mg/dL for the cholesterol of the general population is an established norm based on large-scale studies. Similarly, the value of 40 hours for a work week is a commonly held belief.

The two independent-samples t test is basically a modification of the one-sample t test that incorporates information about the variability of the two independent-sample means. The standard error of the mean difference is no longer estimated from the variance and number of cases in a single group. Instead, it is estimated from the variances and sample sizes of the two independent groups.

Descriptive Statistics

Look at Figure 13.1, which shows descriptive statistics for the age variable, when full-time workers are classified into one of two distinct groups—the *very satisfied* and the *not very satisfied.*

You can obtain these statistics using the Means procedure, which is described in Chapter 5. Select the variables age and satjob2 in the Means dialog box.

Figure 13.1 Descriptive statistics for age by job satisfaction category

Age of Respondent

Job Satisfaction	Mean	Std. Deviation	N
Very satisfied	41.50	11.54	325
Not very satisfied	39.57	10.79	419
Total	40.41	11.16	744

You see that the average age of the *very satisfied* group is 41.5 years, while the average age of the *not very satisfied* group is 39.6. The standard deviation of the *very satisfied* group, 11.5 years, is slightly larger than the standard deviation of the *not very satisfied* group, 10.8 years. In the General Social Survey sample, the *very satisfied* people are on average 1.9 years older than those less content with their jobs. Based on these sample results, what can you reasonably conclude about the population of American adults who are employed full time? Can you conclude that there is a difference in average ages between the two groups?

Distribution of Differences

To answer this question, you have to determine if your observed age difference would be unusual if the two populations have the same average age. In the previous chapters, you answered similar questions by looking at the distribution of all possible means from a population. Now you'll look at the distribution of all possible *differences* between sample means from two independent groups.

Fortunately, the Central Limit Theorem works for differences of sample means as well as for individual means. So if your data are samples from approximately normal populations, or your sample size is large enough so that the Central Limit Theorem holds, the distribution of differences between two sample means is also normal. It's always a good idea to obtain stem-and-leaf plots or histograms for each of the two groups. From these, you can tell what the distribution of values looks like.

Standard Error of the Mean Difference

If two samples come from populations with the same mean, the mean of the distribution of differences is 0. However, that's not enough information to determine if the observed sample results are unusual. You also need to know how much the sample differences vary. The standard deviation of the difference between two sample means, the **standard error of the mean difference**, tells you that. When you have two independent groups, you must estimate the standard error of the mean difference from the standard deviations and the sample sizes in each of the two groups.

How do I estimate the standard error of the difference? The formula is

$$S_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{S_1^{\,2}}{n_1}+\frac{S_2^{\,2}}{n_2}}$$

where $S_1^{\,2}$ is the variance for the first sample and $S_2^{\,2}$ is the variance for the second sample. The sample sizes for the two samples are n_1 and n_2. If you look carefully at the formula, you'll see that the standard error of the mean difference depends on the standard errors of the two sample means. You square the standard error of the mean for each of the two groups. Next you sum them, and then take the square root. ■■■

Computing the T Statistic

Once you've estimated the standard error of the mean difference, you can compute the t statistic the same way as in the previous chapters. You divide the observed mean difference by the standard error of the difference. This tells you how many standard error units from the population mean of 0 your observed difference falls. That is,

$$t = \frac{(\bar{X}_1-\bar{X}_2)-0}{S_{\bar{X}_1-\bar{X}_2}}$$ **Equation 13.1**

If your observed difference is unlikely when the null hypothesis is true, you can reject the null hypothesis.

How is this different from the one-sample t *test?* The idea is exactly the same. What differs is that you now have two independent-sample means, not one. So you estimate the standard error of the mean difference based on two sample variances and two sample sizes. ■■■

Output from the Two Independent-Samples T Test

Look at Figure 13.2, which shows the results from SPSS of testing the null hypothesis that in the population the average age of *very satisfied* full-time workers and *not very satisfied* full-time workers is the same.

To obtain this output, from the menus choose:

Statistics
Compare Means ▸
Independent Samples T Test...

Select the variables age and satjob2, as shown in Figure 13.8.

Figure 13.2 Independent-samples t test of age by job satisfaction

			Age of Respondent	
			Equal variances assumed	Equal variances not assumed
Levene's Test for Equality of Variances	F		.377	
	Sig.		.540	
t-test for Equality of Means	t		2.347	2.327
	df		742	672.439
	Sig. (2-tailed)		.019	.020
	Mean Difference		1.93	1.93
	Std. Error Difference		.82	.83
	95% Confidence Interval of the Mean	Lower	.32	.30
		Upper	3.54	3.56

There is only a 1.9% chance of observing a mean difference at least this large if the null hypothesis is true

The mean age for the two samples differs by 1.93 years

In the output, there are two slightly different versions of the t test. One makes the assumption that the variances in the two populations are equal; the other does not. This assumption affects how the standard error of the mean difference is calculated. You'll learn more about this distinction later in this chapter.

Consider the column labeled *Equal variances assumed.* You see that for the observed difference of 1.93 years, the t statistic is 2.35. (To calculate the t statistic, divide the observed difference of 1.93 by 0.82, the standard error of the difference estimate when the two population variances are assumed to be equal.) The degrees of freedom for the t statistic are 742, the sum of the sample sizes in the two groups minus 2.

The observed two-tailed significance level is 0.019. This tells you that only 1.9% of the time would you expect to see a sample difference of 1.93 years or larger, when the two population means are equal. Since 1.9% is less than 5%, you reject the null hypothesis that the two groups of workers come from populations with the same average age. Your observed results are unusual if the null hypothesis is true.

Confidence Intervals for the Mean Difference

Take another look at Figure 13.2. The 95% confidence interval for the true difference is from 0.32 years to 3.54 years. This tells you it's likely that the true mean difference is anywhere from a third of a year to slightly more than three and one-half years. Since your observed significance level for the test that the two population means are equal was less than 5%, you know that the 95% confidence interval will not contain the value of 0. (Remember, only likely values are included in a confidence interval. Since you found 0 to be an unlikely value, it won't be included in the confidence interval.)

To calculate a 99% confidence interval, specify 99 in the T Test Options dialog box (see Figure 13.10).

If I compute a 99% confidence interval for the true mean difference, will it also not include 0? The 99% confidence interval for the mean difference extends from −0.194 to 4.053. This interval does include the value of 0. That's because your observed significance level is greater than 1%. If your criterion for unusual is 1 in a 100 or less, you cannot reject the null hypothesis based on the t test or on the corresponding 99% confidence interval for the mean difference. ■■■

Another Way of Looking at It

You found a small, but statistically significant, age difference between people who are *very satisfied* with their jobs and those who aren't. Since the observed sample difference is less than two years, it's tempting to dismiss this finding as not particularly interesting. However, there are many different ways you can look at the relationships between the two variables. Sometimes uninteresting information can become more interesting when looked at in another way.

You can obtain crosstabulations using the Crosstabs procedure, as discussed in Chapter 7.

Figure 13.3 Crosstabulation of job satisfaction and age

		Job Satisfaction		
		Very satisfied	Not very satisfied	Total
18-29	Count	46	91	137
	% within age in four categories	**33.6%**	66.4%	100.0%
30-39	Count	112	131	243
	% within age in four categories	46.1%	53.9%	100.0%
40-49	Count	89	121	210
	% within age in four categories	42.4%	57.6%	100.0%
50+	Count	78	76	154
	% within age in four categories	**50.6%**	49.4%	100.0%
Total	Count	325	419	744
	% within age in four categories	43.7%	56.3%	100.0%

33.6% of people less than 30 are very satisfied with their jobs

50.6% of people age 50 and over are very satisfied

Look at Figure 13.3, which is a crosstabulation of the two categories of job satisfaction and four categories of age. From the row percentages, you see that only 33.6% of people less than 30 are *very satisfied* with their jobs, while over 50% of people age 50 and over are *very satisfied*. Overall, 43.7% of full-time workers claim to be *very satisfied* with their jobs. The crosstabulation provides findings that are more interesting and easier to interpret. (In Chapter 16, you'll learn how to test hypotheses that the two variables in a crosstabulation are independent.)

Testing the Equality of Variances

You saw that there are two different *t* values in Figure 13.2. That's because there are two different ways to estimate the standard error of the difference. One of them assumes that the variances are equal in the two populations from which you are taking samples, the other one does not.

In Figure 13.1, you see that the observed standard deviations in the two samples are fairly similar. You can test the null hypothesis that the two samples come from populations with the same variances using the Levene test, which is shown in Figure 13.4. If the observed significance level for the Levene test is small, you can reject the null hypothesis that the two population variances are equal.

For this example, you can't reject the equal variances hypothesis, since the observed significance level for the Levene test is 0.54. That means you can use the results labeled *equal variances* in Figure 13.4.

To obtain this output, select the variables age and satjob2 in the Independent-Samples T Test dialog box, as shown in Figure 13.8.

Figure 13.4 Levene test for equality of variances

			Age of Respondent	
			Equal variances assumed	Equal variances not assumed
Levene's Test for Equality of Variances	F		.377	
	Sig.		.540	
t-test for Equality of Means	t		2.347	2.327
	df		742	672.439
	Sig. (2-tailed)		.019	.020
	Mean Difference		1.93	1.93
	Std. Error Difference		.82	.83
	95% Confidence Interval of the Mean	Lower	.32	.30
		Upper	3.54	3.56

You don't reject the hypothesis that the two population variances are equal based on the Levene test

If the Levene test leads you to reject the null hypothesis that the two population variances are equal, or if you are unsure, you should use the results from the column labeled *Equal variances not assumed* in Figure 13.4. Notice that the estimate of the standard error of the difference is not the same in the two columns. This affects the *t* value and confidence interval. When you use the estimate of the standard error of the difference that does not assume that the two variances are equal, the degrees of freedom for the *t* statistic are no longer the sum of the two sample sizes minus two. They are calculated based on both the sample sizes and the standard deviations in each of the groups. In this example, both *t* tests give very similar results, but that's not always the case.

Why do you get different numbers for the standard error of the mean difference depending on the assumptions you make about the population variances? If you assume that the two population variances are equal, you can compute what's called a pooled estimate of the variance. The idea is similar to that of averaging the variances in the two groups, taking into account the sample size. The formula for the pooled variance is

$$S^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1-1)+(n_2-1)}$$

It is this pooled value that is substituted for both S_1^2 and S_2^2 in the equation on p. 236. If you do not assume that the two population variances are equal, the individual sample variances are used in the equation on p. 236.

Comparing Education

From the previous analysis, you concluded that there appears to be a difference in average ages between those who are *very satisfied* with their jobs and those who are not. Younger people tend to be less satisfied with their jobs than older people. Now consider education. As always, your first step should be to look at the data values in the two groups. Look at the distributions, and try to see if there's anything unusual going on. For small sample sizes, see if the distribution of data values is approximately normal.

Figure 13.5 contains the descriptive statistics for the two groups. You see that the very satisfied people are somewhat better educated (or at least went to school longer) than those who are *not very satisfied*. The difference is slightly more than a third of a year. Figure 13.6 shows the distribution of education graphically. From the boxplot, you see that the variability of the two groups is similar. Since the median for the *very satisfied* group is in the middle of the box, the distribution of values for the group is more or less symmetric. One case stands out in the plot. That's the person who claims no formal education. For the *not very satisfied* group, the median is close to the bottom edge, indicating that there is a tail toward higher educational levels. If your sample is small, and the departures from normality are severe, you may want to substitute one of the nonparametric tests described in Chapter 17 for the independent-samples *t* test. In this case, the sample sizes are large, so the independent-samples *t* test should work just fine.

You can obtain these statistics using the Means procedure, as discussed in Chapter 5.

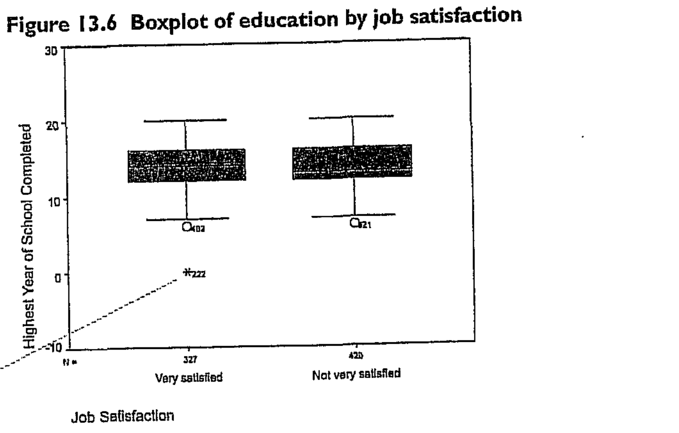
Figure 13.5 Descriptive statistics for education by job satisfaction

Highest Year of School Completed

Job Satisfaction	Mean	N	Std. Deviation
Very satisfied	14.25	327	2.79
Not very satisfied	13.87	420	2.62
Total	14.04	747	2.70

You can obtain this boxplot using the Explore procedure, as discussed in Chapter 6.

Figure 13.6 Boxplot of education by job satisfaction



Respondent with no formal education

Based on the Levene test in Figure 13.7, there is no reason to doubt that the population variances are equal, so you can use the *t* value in the column labeled *Equal variances assumed* to test the null hypothesis that in the population, the average years of education are the same for those who are *very satisfied* with their jobs and those who are not. The two-tailed significance level is 0.057, so you don't reject the null hypothesis. As expected, the 95% confidence interval for the mean difference includes the

NORUŠIS, Marija J. 1998. Guide to Data Analysis. Upper Saddle River, Prentice Hall.

value of 0. (The lower bound of the 95% confidence interval is given in scientific notation. The lower limit is –0.011 years.)

What kind of number has an E in it? When SPSS displays a very small or very large number, it uses scientific notation. The number that follows the letter E tells you how many places the decimal must be moved. If the number following E is negative, move the decimal to the left. If the number following E is positive, move the decimal point to the right. For example, –1.1E–02 is –0.011; –1.1E02 is –110.

If you don't like the format SPSS uses to display a number, activate the pivot table, select the cell, and from the menus choose:

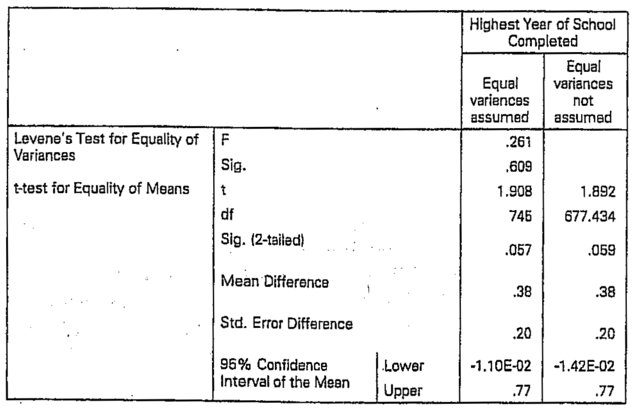
Format
Cell properties...

Select the format you prefer in the Value tab of the Cell properties dialog box.

To obtain this output, select the variables educ and satjob2 in the Independent-Samples T Test dialog box. (See Figure 13.8.) Then activate the pivot table and from the menus choose:

Pivot
Transpose Rows and Columns

Figure 13.7 Independent-samples t test of education by job satisfaction



			Highest Year of School Completed	
			Equal variances assumed	Equal variances not assumed
Levene's Test for Equality of Variances	F		.261	
	Sig.		.609	
t-test for Equality of Means	t		1.908	1.892
	df		746	677.434
	Sig. (2-tailed)		.057	.059
	Mean Difference		.38	.38
	Std. Error Difference		.20	.20
	95% Confidence Interval of the Mean	Lower	-1.10E-02	-1.42E-02
		Upper	.77	.77

7. lekce

JAK TESTOVAT NULOVOU HYPOTÉZU O SHODĚ NĚKOLIKA POPULAČNÍCH PRŮMĚRŮ.

NORUŠIS, Marija J. 1998. *Guide to Data Analysis*. Upper Saddle River, Prentice Hall.

14 One-Way Analysis of Variance

How can you test the null hypothesis that several population means are equal?

- What is analysis of variance?
- What assumptions about the data are needed to use analysis-of-variance techniques?
- How is the *F* ratio computed, and what does it tell you?
- Why do you need multiple comparison procedures?

You've already learned how to test hypotheses about two population means using the paired-samples *t* test and the independent-samples *t* test. Often, however, you want to compare more than two population means. For example, if you are studying four methods for teaching mathematics, you want to compare average test scores for all four groups. Or, if you are testing seven different treatments for lowering cholesterol, you may want to compare the average final cholesterol levels for all seven methods. In this chapter, you'll learn how to test the null hypothesis that several independent population means are equal. The technique you'll use is called **analysis of variance**, usually abbreviated as ANOVA.

▶ This chapter uses the *gssft.sav* data file, which includes only people holding full-time jobs. For instructions on how to obtain the One-Way ANOVA output shown in the chapter, see "How to Obtain a One-Way Analysis of Variance" on p. 274.

Hours in a Work Week

In Chapter 11, you looked at the average number of hours worked in a week by college graduates. Based on the results from the General Social Survey, you rejected the null hypothesis that the average work week is 40 hours. You found the 95% confidence interval for the population value for the average number of hours worked to be from 46.16 hours to 49.30 hours. So it's not inconceivable that the average college graduate works almost an extra 8-hour day each week.

An obvious question that arises is whether it's just college graduates who suffer from the expansion of the work week, or is everyone, regardless of educational background, working more? Using the General Social Survey, you can look at the average number of hours worked by full-time employees of various educational backgrounds.

Describing the Data

You see in Figure 14.1 that the average work week for all full-time employees is 46.29 hours. (It's the entry in the column labeled *Total.*) The average work week ranges from a low of 43.69 hours for people without a high school diploma to a high of 50.27 hours for people with graduate degrees.

To obtain this output, from the menus choose:

Statistics
Compare Means ▸
One-Way ANOVA...

Select the variables hrs1 and degree, as shown in Figure 14.5.

In the One-Way ANOVA Options dialog box, select Descriptive, as shown in Figure 14.7.

Figure 14.1 Descriptive statistics for hours worked

		HRS1 Number of Hours Worked Last Week					
		DEGREE RS Highest Degree					
		Less than HS	High school	Junior college	Bachelor	Graduate	Total
N		52	387	54	162	86	741
Mean		43.69	45.77	45.87	46.38	50.27	46.29
Std. Deviation		8.72	10.58	11.66	12.89	11.44	11.27
Std. Error		1.21	.54	1.59	1.01	1.23	.41
95% Confidence Interval for Mean	Lower Bound	41.26	44.72	42.69	44.38	47.82	45.48
	Upper Bound	46.12	46.83	49.05	48.38	52.72	47.10
Minimum		20	8	25	5	34	5
Maximum		70	80	80	89	89	89

The average work week ranges from 43.69 hours to 50.27 hours

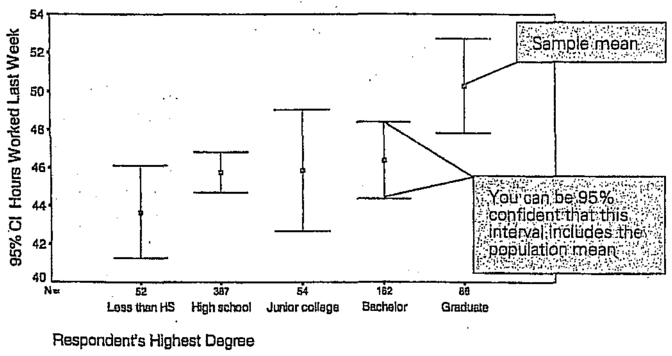
NORUŠIS, Marija J. 1998. *Guide to Data Analysis*. Upper Saddle River, Prentice Hall.

In the row labeled *Std. Deviation,* you see that the smallest variability in hours worked is for people with less than a high school diploma, while the largest is for people with bachelor's degrees. The next column, labeled *Standard Error,* tells you how much the sample means vary in repeated samples from the same population. For each group, it's the standard deviation divided by the square root of the sample size. The smallest standard error is for high school graduates, since they are the largest group.

Confidence Intervals for the Group Means

In the last two columns of Figure 14.1, you see for each group the 95% confidence interval for the population value of the average hours worked per week. You are 95% confident that the true work week for those with less than a high school diploma is between 41.26 and 46.12 hours. For those with a graduate degree, you are 95% confident that the true average work week is between 47.82 and 52.72 hours.

Figure 14.2 Plot of sample means and 95% confidence intervals



Sample mean

You can be 95% confident that this interval includes the population mean

You can obtain this error bar chart using the Graphs menu, as described in "Error Bar Charts" on p. 523 in Appendix A.

In the Define Simple Error Bar Summaries for Groups of Cases dialog box, select the variables hrs1 and degree.

Plots of the means and confidence intervals are shown in Figure 14.2. You see that the 95% confidence interval for high school graduates is the narrowest. That's because there are so many of them in the sample. Many of the confidence intervals in Figure 14.2 overlap. That tells you that some of the values that are plausible for the true work week in one group are also plausible for the true work week in the others. The exception is the confidence interval for those with graduate degrees. It doesn't overlap the confidence interval for those with less than a high school education, nor the interval for those with a high school education.

Can you tell from the plot if the 40-hour work week is a reasonable guess for the true hours worked per week? Sure. Remember, if a value doesn't fall in the 95% confidence interval for the mean, you can reject the hypothesis that it's a plausible population value. You see in Figure 14.2 that the value 40 is not included in any of the confidence intervals. That means you can reject the hypothesis that it's a reasonable value for any of the groups. It appears that the 40-hour work week may be a thing of the past, regardless of your education level.

Testing the Null Hypothesis

The descriptive statistics and plots suggest that there are differences in the average work week among the five education groups. Now you need to figure out whether the observed differences in the samples may be attributed to just the natural variability among sample means or whether there's reason to believe that some of the five groups have different values in the population for average hours worked.

The null hypothesis says that the population means for all five groups are the same. That is, there is no difference in the average hours worked for people in the five education categories. The alternative hypothesis is that there is a difference. The alternative hypothesis doesn't say which groups differ from one another. It just says that the groups means are not all the same in the population; at least one of the groups differs from the others.

The statistical technique you'll use to test the null hypothesis is called analysis of variance (abbreviated ANOVA). It's called analysis of variance because it examines the variability of the sample values. You look at how much the observations within each group vary as well as how much the group means vary. Based on these two estimates of variability, you can draw conclusions about the population means. If the sample means vary more than you expect based on the variability of the observations in the groups, you can conclude that the population means are not all equal.

SPSS contains several different procedures that perform analysis of variance. In this chapter, you'll use the One-Way ANOVA procedure. It's called one-way analysis of variance because cases are assigned to different groups based on their values for one variable. In this example, you form the groups based on the values of the *degree* variable. The variable used to form groups is called a factor. In Chapter 15, you'll learn how to test hypotheses when cases are classified into groups based on their values for two factors.

Assumptions Needed for Analysis of Variance

Analysis of variance requires the following assumptions:

- Independent random samples have been taken from each population.
- The populations are normal.
- The population variances are all equal.

The Kruskal-Wallis test, described in Chapter 17, requires more limited assumptions about the data.

Independence. The independence assumption means that there is no relationship between the observations in the different groups and between the observations in the same group. For example, if you administer four different treatments to each individual, you cannot use the one-way analysis-of-variance procedure to analyze the data. Observations from the same individual appear in each of the groups, so they are not independent. (In this situation, you must use an extension of the paired-samples *t* test. It's called repeated measures analysis of variance, a topic not covered in this book.) Observations within a group are also not independent if conditions are changing with time. For example, if you are explaining a task to subjects and your instructions get better with time, early subjects may not perform as well as later subjects. In this situation, the response of the subject depends on the point in time he or she entered into the study. Consecutive subjects will be similar to each other.

Normality. The normality assumption in analysis of variance can be checked by making histograms or normal probability plots for each of the groups. In practice, the analysis of variance is not heavily dependent on the normality assumption. As long as the data are not extremely non-normal, you do not have to worry. (If your sample sizes in the groups are small, you should be aware of the impact of unusual observations, which can have a big effect on the mean and standard deviation. You can rerun the analysis without the unusual point to make sure that you reach the same conclusions.)

Equality of Variance. The equality of variance assumption can be checked by examining the spread of the observations in the box plot. You can also compute the Levene test for equality of variance, which is available in the Explore and One-Way procedures. In practice, if the number of cases in each of the groups is similar, the equality of variance assumption is not too important.

What should I do if I suspect that my data violate the necessary assumptions? Well, it depends on which assumption is being violated. For example, if you're worried about the normality or equal-variance assumptions, sometimes you can transform your data so that the distribution of values is more normal or the variances in the groups are more similar. Taking logarithms or square roots of the data values is often helpful. If this fails, you can use a statistical test that makes fewer assumptions about the data. In particular, you may want to use the Kruskal-Wallis test described in Chapter 17.

The situation is considerably more complicated if you're worried about whether the groups are somehow biased. That is, you're concerned that one or more of your samples differs in some important way from the population of interest. For example, if you want to compare four medical treatments, and the participating physicians have assigned the sickest patients to a particular group, you've got a real problem. You may not be able to draw any correct conclusions from your data. That's why it's very important when comparing several treatments or conditions, to make sure that the subjects are randomly assigned to the different groups. *Randomly* doesn't mean *haphazardly*. It means that you must have a well organized system for random assignment of cases.

Analyzing the Variability

In analysis of variance, the observed variability in the sample is divided (partitioned, in statistical lingo) into two parts: variability of the observations *within* a group about the group mean, and variability *between* the group means.

Why are we talking about variability? Aren't we testing hypotheses about means? Yes, we're testing hypotheses about population means; but as you've seen in previous chapters, your conclusions about population means are always based on looking at the variability of sample means. You have to determine if your sample mean is outside the usual range of variability of sample means from the population.

In analysis of variance, you'll look at how much your observed sample means vary. You'll compare this observed variability to the expected variability if the null hypothesis that all population means are the same is true. If the sample means vary more than you'd expect, you have reason to believe that this extra variability is because some of groups don't have the same population mean. (If you have two independent groups, you'll get the same results using ANOVA or the equal variance *t* test.)

Let's look a little more closely now at the two types of variability and how they are used to test the null hypothesis that the population values for average hours worked per week are the same for people in the five education categories. The game plan is as follows: You want to know whether your sample means vary more than you would expect if the null hypothesis is true. First, you'll see how much the observations in a group vary, and then you'll see how much the sample means vary. If the sample means vary more than you expect, you'll reject the null hypothesis.

Within-Groups Variability

The within-groups estimate of variability, as its name suggests, tells you how much the observations within a group vary. The sample variance of each group estimates within-groups variability. One of the assumptions of analysis of variance is that all groups come from populations with the same variance. That makes it possible for you to average the variances in each of the groups to come up with a single number, which is the within-groups variance. (You'll see later how this averaging is done. You can't just add up the sample variances and divide by the number of groups.)

You might wonder why you can't just put all of your observations together and compute the variance. The reason is that you don't know if all of the groups have the same population mean. If they don't, pooling all the values together will give you the wrong answer. For example, suppose that all people without a high school diploma work exactly 40 hours a week; all people with a high school diploma work exactly 43 hours a week; and all people with a college degree work exactly 45 hours. The variance in each of the groups is 0, since the values within a group don't vary at all. The correct estimate of the within-groups variance is also 0. If you compute the variance for all cases together, it wouldn't be close to 0. The observed variability would be the result of differences in the means of the three groups.

Between-Groups Variability

You have a sample mean for each of the groups in your study. If all of the groups have the same number of cases, you can find the standard deviation of the sample means. What would that tell you? If all the groups come from populations with the same mean and variance, the standard deviation of the sample means tells you how much sample means from the same population vary. The standard deviation of the sample means is an estimate of the standard error of the mean.

From the standard error of the mean, you can estimate the standard deviation of the observations. You do this by multiplying the standard error of the mean by the square root of the number of cases in a group.

Where did that come from? The standard error of the mean is the standard deviation of the observations divided by the square root of the sample size. So, using simple algebra, the standard deviation is the standard error of the mean multiplied by the square root of the sample size. Thus,

$$\text{standard error} = \frac{\text{standard deviation}}{\sqrt{\text{sample size}}}$$

and

$$\text{standard deviation} = \text{standard error} \times \sqrt{\text{sample size}}$$

If you square the estimate of the standard deviation, you have a quantity that's called the **between-groups estimate of variability.** It's called the between-groups estimate of variability because it's based on how much sample means vary *between* the groups.

Comparing the Two Estimates of Variability

You now have two estimates of how much the observations within a group vary: the within-groups estimate and the between-groups estimate. These two estimates differ in a very important way: the between-groups estimate of variance will be correct only if the null hypothesis is true. If the null hypothesis is false, the between-groups estimate of variance will be too large. The observed variability of the sample means will be the result of two factors: the variability of the observations within a group and the variability of the population means. The within-groups estimate of variability doesn't depend on the null hypothesis being true. It's always a good estimate.

Your decision about the null hypothesis will be based on comparing the between-groups and the within-groups estimates of variability. You'll see how much the number of hours worked varies for individuals in the same education group. This will give you the within-groups estimate of variability. Then you'll see how much the means of the five groups vary. Based on this, you'll calculate the between-groups estimate of variability. If the between-groups estimate is sufficiently larger than the within-groups estimate, you'll reject the null hypothesis that all of the means are equal in the population.

The Analysis-of-Variance Table

The estimates of variability that we've been talking about are usually displayed in what's called an **analysis-of-variance table**. Figure 14.3 is the analysis-of-variance table for the test of the null hypothesis that the population value for average hours worked per week is the same for people in five categories of education. By looking at this table, you'll be able to tell whether you have enough evidence to reject the null hypothesis.

tain this output,
the variables
nd degree in the
Vay ANOVA
box, as shown
ure 14.5.

Figure 14.3 Analysis-of-variance table

		Sum of Squares	df	Mean Square	F	Sig.
Number of Hours Worked Last Week	Between Groups	1825.917	4	456.479	3.646	.006
	Within Groups	92148.280	736	125.201		
	Total	93974.197	740			

Ratio of mean squares

Probability of obtaining F ratio at least this large when null hypothesis is true

The two estimates of variability are shown in the column labeled *Mean Square*. Their ratio is in the column labeled *F*. If the null hypothesis is true, you expect the ratio of the between-groups mean square to the within-groups mean square to be close to 1, since they are both estimates of the population variance. Large values for the *F* ratio indicate that the sample means vary more than you would expect if the null hypothesis were true.

You can tell if your observed *F* ratio of 3.65 is large enough for you to reject the null hypothesis by looking at the observed significance level, which is labeled *Sig*. You see that the probability of obtaining an *F* ratio of 3.65 or larger when the null hypothesis is true is 0.006. Only 6 times in 1000, when the null hypothesis is true, would you expect to see a ratio this large or larger. So you can reject the null hypothesis. It's unlikely that the number of hours worked per week is the same for the five groups in the population.

Now that you know the punch line, let's see where all the numbers are coming from.

Estimating Within-Groups Variability

You need three steps to compute the within-groups estimate of variability:

1. First, you must compute what's called the **within-groups sum of squares**. Take all of the standard deviations in Figure 14.1 and square them to obtain variances. Then multiply each variance by one less than the number of cases in the group. Finally, add up the values for all of the groups. The within-groups sum of squares is:

$$(8.72^2 \times 51) + (10.58^2 \times 386) + (11.66^2 \times 53) + (12.89^2 \times 161) + (11.44^2 \times 85) = 92148.28 \qquad \textbf{Equation 14.1}$$

You see this number in the second row of Figure 14.3 in the column labeled *Sum of Squares*. (You have to use more decimal places for the standard deviation than shown above to get exactly the answer given.)

2. Next, you must compute the degrees of freedom. That's easy to do. For each group, you compute the number of cases minus 1, and then add up these numbers for all of the groups. In this example, the degrees of freedom are:

$$\text{degrees of freedom} = 51 + 386 + 53 + 161 + 85 = 736 \qquad \textbf{Equation 14.2}$$

This number is shown in the *Within Groups* row of Figure 14.3, in the column labeled *df* (for degrees of freedom).

3. Finally, divide the sum of squares by its degrees of freedom, to get what's called a **mean square**. This is the estimate of the average variability in the groups. It's really nothing more than an average of the variances in each of the groups, adjusted for the fact that the number of observations in the groups differs. Your estimate of the variance for the number of hours worked, based on the variability of the observations within each of the groups, is 125.20.

Estimating Between-Groups Variability

You also need three steps to calculate the between-groups estimate of variability.

1. First, you compute the **between-groups sum of squares**. Subtract the overall mean (the mean of all of the observations) from each group mean. Then square each difference, and multiply the square by the number of observations in its group. Finally, add up all the results. For this example, the between-groups sum of squares is:

$$\begin{aligned} &52 \times (43.69 - 46.29)^2 \\ &+ 387 \times (45.77 - 46.29)^2 \\ &+ 54 \times (45.87 - 46.29)^2 \\ &+ 162 \times (46.38 - 46.29)^2 \\ &+ 86 \times (50.27 - 46.29)^2 = 1825.92 \end{aligned} \qquad \textbf{Equation 14.3}$$

2. Next, you must compute the degrees of freedom. The degrees of freedom for the between-groups sum of squares is just the number of groups minus 1. In this example, there are five education groups, so the degrees of freedom for the between-groups sum of squares is 4.

3. Finally, calculate the between-groups mean square by dividing the between-groups sum of squares by its degrees of freedom. The between-groups mean square is 456.48.

Calculating the F Ratio

You now have the two estimates of the variability in the population: the within-groups mean square and the between-groups mean square. The *F* ratio is simply the ratio of these two estimates. Take the between-groups mean square and divide it by the within-groups mean square:

$$F = \frac{\text{between-groups mean square}}{\text{within-groups mean square}} = \frac{456.48}{125.20} = 3.65 \qquad \textbf{Equation 14.4}$$

(Remember, the within-groups mean square is based on how much the observations within each of the groups vary. The between-groups mean square is based on how much the group means vary among themselves.) If the null hypothesis that the average hours worked per week is the same for the five groups is true, the two numbers should be close to each other. If you divide one by the other, the result should be close to 1.

As you see, the ratio of the two estimates is not 1. Does that mean you automatically reject the null hypothesis? No. You know that your sample ratio will not be exactly 1, even if the null hypothesis is true. You need to figure out how often you would expect to see a sample value of 3.65 or greater when the null hypothesis is true. That is, you need to determine whether your sample results are unlikely if the null hypothesis is true.

The observed significance level is calculated by comparing your observed *F* ratio to values of the *F* distribution. The observed significance level depends on both the observed *F* ratio and the degrees of freedom for the two mean squares.

What's the F *distribution?* Like the normal and *t* distributions, the *F* distribution is defined mathematically. It's used when you want to test hypotheses about population variances. The Central Limit Theorem doesn't work for variances. Their distributions are not normal. The ratio of two sample variances from normal populations has an *F* distribution. The *F* distribution is indexed by two values for the degrees of freedom, one for the numerator and one for the denominator. The degrees of freedom depend on the number of observations used to calculate the two variances.

In Figure 14.3, you see that the observed significance level for this example is 0.006. Since the value is small, you can reject the null hypothesis that the average hours worked per week in the population is the same for the five groups. The observed sample results are not likely to occur when the null hypothesis is true.

Multiple Comparison Procedures

A statistically significant *F* ratio tells you only that it appears unlikely that all population means are equal. It doesn't tell you which groups are different from each other. You can reject the null hypothesis that all population means are equal in a variety of situations. For example, it may be that the average hours worked differs for all of the five groups. Or it may be that only one or two of the groups differ from the rest. Usually when you've rejected the null hypothesis, you want to pinpoint exactly where the differences are. To do this, you must use multiple comparison procedures.

NORUŠIS, Marija J. 1998. *Guide to Data Analysis*. Upper Saddle River, Prentice Hall.

Why do you need yet another statistical technique? Why can't you just compare all possible pairs of means using t *tests?* The reason for not using many *t* tests is that when you make many comparisons involving the same means, the probability increases that one or more comparisons will turn out to be statistically significant, even when all the population means are equal. This is known as the **multiple comparison problem.**

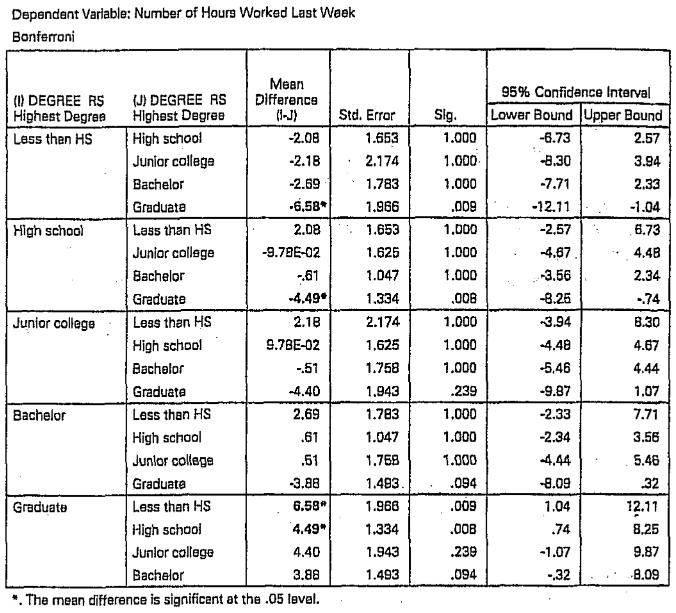
For example, if you have 5 groups and compare all pairs of means, you're making 10 comparisons. When the null hypothesis is true, the probability that at least 1 of the 10 observed significance levels is less than 0.05 is about 0.29. With 10 means (45 comparisons), the probability of finding at least one significant difference is about 0.63. The more comparisons you make, the more likely it is that you'll find 1 or more pairs to be statistically different, even if all population means really are equal.

Multiple comparison procedures protect you from calling differences *significant* when they really aren't. This is accomplished by adjusting the observed significance level for the number of comparisons that you are making, since each comparison provides another opportunity to reject the null hypothesis. The more comparisons you make, the larger the difference between pairs of means must be for a multiple comparison procedure to call it statistically significant. That's why you should look only at differences between pairs of means that you are interested in. When you use a multiple comparison procedure, you can be more confident that you are finding true differences. ■■■

Many multiple comparison procedures are available. They differ in how they adjust the observed significance level. One of the simplest is the **Bonferroni procedure.** It adjusts the observed significance level by multiplying it by the number of comparisons being made. For example, if you are making five comparisons, the observed significance level for each comparison must be less than 0.05/5, or 0.01, for the difference to be significant at the 0.05 level.

To obtain the Bonferroni test, select Post Hoc in the One-Way ANOVA dialog box. Then select Bonferroni, as shown in Figure 14.6.

Figure 14.4 Bonferroni multiple comparison test on hours worked



Dependent Variable: Number of Hours Worked Last Week

Bonferroni

(I) DEGREE RS Highest Degree	(J) DEGREE RS Highest Degree	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Less than HS	High school	-2.08	1.653	1.000	-6.73	2.57
	Junior college	-2.18	2.174	1.000	-8.30	3.94
	Bachelor	-2.69	1.783	1.000	-7.71	2.33
	Graduate	-6.58*	1.966	.009	-12.11	-1.04
High school	Less than HS	2.08	1.653	1.000	-2.57	6.73
	Junior college	-9.78E-02	1.625	1.000	-4.67	4.48
	Bachelor	-.61	1.047	1.000	-3.56	2.34
	Graduate	-4.49*	1.334	.008	-8.25	-.74
Junior college	Less than HS	2.18	2.174	1.000	-3.94	8.30
	High school	9.78E-02	1.625	1.000	-4.48	4.67
	Bachelor	-.51	1.758	1.000	-5.46	4.44
	Graduate	-4.40	1.943	.239	-9.87	1.07
Bachelor	Less than HS	2.69	1.783	1.000	-2.33	7.71
	High school	.61	1.047	1.000	-2.34	3.56
	Junior college	.51	1.758	1.000	-4.44	5.46
	Graduate	-3.88	1.493	.094	-8.09	.32
Graduate	Less than HS	6.58*	1.966	.009	1.04	12.11
	High school	4.49*	1.334	.008	.74	8.25
	Junior college	4.40	1.943	.239	-1.07	9.87
	Bachelor	3.88	1.493	.094	-.32	8.09

*. The mean difference is significant at the .05 level.

If you want to compare all five education groups to one another, you can form 10 unique pairs of groups. Statistics for all pairs of group comparisons using the Bonferroni multiple comparison procedure are shown in Figure 14.4. There are a lot of numbers, but they're not hard to understand. Each row corresponds to a comparison of two groups. The first row is for the comparison of the less than high school group to the high school group. The last row is for the comparison of the graduate group to the bachelor's degree group. The difference in hours worked between the two groups is shown in the column labeled *Mean Difference*. Pairs of means that are significantly different from each other are marked with an asterisk. You see that people with graduate degrees work significantly longer than people with less than a high school education and people with graduate degrees work significantly longer than people with just a high school education. No two other groups are significantly different from one another. The table shows all possible pairs of groups twice. There is a row for the comparison of bachelor to graduate and another row for the comparison of graduate to bachelor. These two rows are identical, except for the sign of the mean difference.

The column labeled *Std. Error* (of the difference) is calculated from the within-groups estimate of the standard deviation and the sample sizes in each of the two groups. The observed significance level for the test of the null hypothesis that the two groups come from populations with the same mean is shown in the column labeled *Sig.* Looking down the column of observed significance levels, you see that four of them are less than 0.05. (Note that the mean differences for these pairs are marked with an asterisk.) The 95% confidence interval for the mean difference gives you a range of values that you expect would include the true population difference between the two groups. For example, it's possible that the true difference between the hours worked by people with graduate degrees and people with less than a high school education is anywhere between 1 and 12 hours. Note that the confidence intervals for the pairs that are significantly different from one another do not include the value 0. The confidence intervals are also modified to take into account the fact that 10 pairs of means are being compared. They are wider than they would be if only one pair of means was being compared.

How come the graduate degree group isn't different from the junior college group too? Whether a difference between two groups is statistically significant depends on how big the difference is between the two groups and how many cases there are in each of the groups. (The same estimate of variance is used for all groups.) The average hours worked for junior college grads is very similar to the average hours worked for high school grads. However, there are only 54 people with junior college degrees. It's possible that you'll find in the pairwise table of differences that smaller differences between two groups may be significant, while larger differences between other groups are not. That's the result of differences in the sample sizes between the groups. ■■■

Summary

How can you test the null hypothesis that several population means are equal?

- Analysis of variance is a statistical technique that is used to test hypotheses about two or more population means.
- To use analysis of variance, your groups should be random samples from normal populations with the same variance.
- The *F* ratio is the ratio of two estimates of the population variance: the between-groups and the within-groups mean squares.
- The analysis-of-variance *F* test does not pinpoint which means are significantly different from each other. That's why multiple comparison procedures, which protect you against calling too many differences significant, are used to identify groups that appear to be different from each other.

8. lekce

ZÁKLADY BIVARIAČNÍ ANALÝZY: ROZLOŽENÍ DAT V KONTINGENČNÍ TABULCE - POVAHA VZTAHU MEZI HODNOTAMI PROMĚNNÝCH A POROVNÁVÁNÍ POZOROVANÝCH S OČEKÁVANÝMI ČETNOSTMI. MĚŘENÍ (SÍLY) ASOCIACE MEZI DVĚMA KATEGORIZOVANÝMI PROMĚNNÝMI: KOEFICIENTY ASOCIACE (modul ANALYZE: procedura Crosstabs).

LOETHER Herman J. and Donald G. McTAVISH. 1988. *Descriptive and Inferencial Statistics. An Introduction.* Boston, Allyn and Bacon.

6.3.1 Creating a Bivariate Frequency Distribution

To illustrate how a table is constructed, Table 6.4 shows a tolerance score and a gender score for each of 13 fictitious individuals. The bivariate distribution is set up as shown, with the categories of tolerance (here two categories, high and low) down the **stub** or side of the table, and the categories of gender across the **heading** at the top.

This table is a 2 by 2 table, or **fourfold table**, because it has two rows and two columns (or four cells in the **body** of the table where the rows and columns intersect). Tables can, of course, have any number of rows and column (in general we refer to an *r* by *c* table where *r* refers to the number of rows and *c* to the number of columns); *r* and *c* depend upon the number of categories that are distinguished for row and column variables.

The problem now is to count the number of cases that have various possible combinations of values on the two variables, and to enter these totals into the table to form a bivariate frequency distribution. Notice that 3 of the 13 cases in this sample are males who are "high" on tolerance, 3 are males who are "low" on tolerance, 5 are females who are "high" on tolerance, and 2 are females who are "low" on tolerance. These numbers are written in the boxes or cells in the body of the table corresponding to the appropriate row and column labels shown in Table 6.4. Sometimes it is helpful to create a tally within each cell as a workmanlike way to assure accuracy.

Each of the boxes in the table is called a **cell** and the frequency in a cell is called a **cell frequency**. Cell frequencies are sometimes symbolized by the small letter n_{ij}, where the first subscript (i) indicates the number of the row and the second subscript (j) indicates the number of the column, as follows:

	Column 1	*Column 2*	*Row Totals*
Row 1	n_{11}	n_{12}	$\sum_j n_{1j}$
Row 2	n_{21}	n_{22}	$\sum_j n_{2j}$
Column totals	$\sum_i n_{i1}$	$\sum_i n_{i2}$	N

They indicate the number of cases in the total sample that fall in a certain category of the row and column variables as indicated by the row and column labels.* The cell frequencies indicate the number of cases with two characteristics simultaneously. Row and column totals each add up to 13, the total number of cases there were. The cell frequencies constitute the conditional distributions, and the row and column totals reflect the marginals or univariate distribution of each variable.

* Some authors use a different set of symbols for row and column totals, where a dot is used in place of a subscript for totals. Thus $n_{.1}$ would be the sum of column 1 over all of the rows in the table (the row subscript is replaced by a dot) instead of $\sum_j n_{i1}$, and $n_{1.}$ would symbolize the total of row one, instead of $\sum_j n_{1j}$. One could use n rather than N or $\Sigma\Sigma n_{ij}$ to indicate the grand total number of cases.

6.3.2 Traditions of Table Layout

As mentioned above (Section 6.3) tables usually are set up so that the dependent variable is the one with categories listed down the *stub*, or left side of the table, and the independent variable is listed across the top in the *heading*. This convention, of course, is not always kept, but it does tend to aid the examination of conditional distributions in each column to have it set up this way. Table 6.3 illustrates proper labeling of a table. Notice that low categories of the independent variable, where there are low categories on that variable, are listed at the left and the high categories at the right. For the dependent variable the high categories are at the top of the table and the low categories are at the bottom. This is similar to the labeling of other graphs, although in the case of tables the convention is not as rigidly adhered to, and the investigator would do well to double check the table layout before proceeding to make any interpretation.

A table usually has a title that lists the dependent variable, whether the table contains frequencies or percentages (or some other measure), the independent variable(s), and the kind of case upon which the measurements were taken. Table 6.3 contains data on 7,714 individuals. If the table is a percentaged table, it is important to indicate the base upon which the percentage was computed in brackets, at the bottom by the column total percentages*, when this is done, cell frequencies may be omitted from a percentaged table. The source of data is indicated, typically, in a footnote to the table, and both the stub and heading are clearly marked with the variable and the name of each of the categories of each variable.

Table 6.4 is a frequency table. It has categories of the tolerance variable down the side (the stub), and categories of the variable, gender, across the top. In this case, tolerance played the role of dependent variable.

Notice that cell frequencies in columns are summed, and the sums are put at the bottom of the table. Rows are also summed, and the totals put at the right-hand side. These row and column totals are called **marginals**, or simply row totals and column totals, and they are merely the univariate distribution of each variable separately.

If the table shows percentages it is called a bivariate percentage distribution, and if frequencies are shown, it is called a bivariate frequency distribution.

6.3.3 Percentaged Tables

Probably the most often used type of table is the percentaged table. Its value lies in the way it helps one to make comparisons across the conditional distributions one wants to compare. The basic rule for computing percentages in a table is as follows:

Compute percentages in the direction of the independent variable.

This means that percentages should sum up to 100% for each category of the independent variable. For tables set up such as Table 6.3, the percentaging rule

* This is true if column totals are the bases of percentages. If rows sum to 100% then row total frequencies are given.

leads to computation with column totals as the base of the percentage: thus column percentages add up to 100% for each column. If the independent variable and the dependent variable were switched around, the percentages would have to be run in the other direction. There are three ways that a table can be percentaged, as shown in Table 6.5, using the hypothetical data from Table 6.4. Tables could be percentaged with *column totals* as the base of percentages, with *row totals* as the base of percentages, and with the *grand total* as the base of percentages. Since the dependent variable is down the stub of Table 6.4, the proper table to examine to see what differences there may be between categories of the gender variable would be to percentage with column totals (the number of males or the number of females) as the base of the percentages. One wants to contrast the distribution of the dependent variable between men and women, and the only way to do this is to take out the effect of different numbers of men and women by percentaging down (in the direction of the independent variable). This type of operation permits one to make comparisons in the *other direction*. *Comparisons are made in the opposite direction from the way percentages are run.*

TABLE 6.5 Illustration of Different Ways Percentages Can Be Computed on Tables

Original Frequency Distribution from Table 6.4

	Gender		
Tolerance Level	*Male*	*Female*	*Total*
High	3	5	8
Low	3	2	5
Total	6	7	13

A. Percentaging to Column Totals as the Base

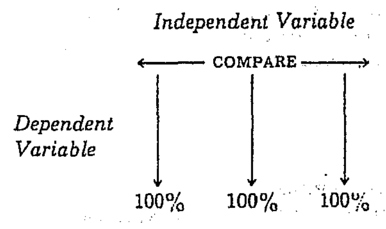
	Gender		
Tolerance Level	*Male*	*Female*	*Total*
High	50%	71%	62%
Low	50	29	38
Total	100%	100%	100%

B. Percentaging to Row Totals as the Base

	Gender		
Tolerance Level	*Male*	*Female*	*Total*
High	38%	62	100%
Low	60%	40	100%
Total	46%	54	100%

C. Percentaging to Overall Grand Total as the Base

	Gender		
Tolerance Level	*Male*	*Female*	*Total*
High	23%	39%	62%
Low	23%	15%	38
Total	46%	54	100%



Comparisons are made in a percentaged table by examining differences between percentages. In Table 6.5A, for example, the difference between percentage "high" on tolerance among men and women is 21% (71% − 50% = 21%). This value is called **epsilon**, the percentage difference in a table, and it is symbolized by the Greek letter ϵ. For tables larger than a 2 by 2 table, there are a number of percentage contrasts or epsilons that may be computed and used in interpretation. Epsilon will be discussed further later on in this chapter.

Sometimes an investigator will compute percentages, as in Table 6.5C, with the total number of cases (N) as *the base for all cell percentages*. Where this is done, we no longer can compare conditional distributions, but we can express the percentage of cases that have each of the different combinations of characteristics labeled by the rows and columns.

If it is not clear which variable is dependent or independent, or if we could think of the data in both ways, we might compute percentages to *both row and column totals* (as in Tables 6.5A and 6.5B) and *examine each table*. Table 6.5A would permit us to say that females are more likely to be higher on tolerance than are males. Table 6.5B would permit us to say that high-tolerance people are more likely to be female than are low-tolerance people—a subtle shift with worlds of import, as we shall soon see.

As shown in Table 6.5, percentaging *down* permits an examination of any influence gender may have on the distribution of tolerance; percentaging *across* shows the possible recruitment pattern into tolerance levels from each gender, and percentaging to the *grand total* permits us to examine the joint percentage distribution of tolerance levels and gender.

6.4 FOUR CHARACTERISTICS OF AN ASSOCIATION

Going back to a bivariate distribution such as that shown in Table 6.3, we can think of that distribution as a relationship between two variables. Suppose we want to know how the distribution of the dependent variable varies as we move from category to category of the other variable. The way two variables relate to each other is called an **association** between the variables. In Table 6.3, as city size increased, the percentage of individuals showing higher tolerance increased. The two variables were associated in that particular fashion.

We can speak of the association of any two variables and describe that association in terms of a percentaged table, as we have shown. There are other ways to summarize the association, however, and, in fact, there are four characteristics of an association that we will single out for summary, just as there are three characteristics of an univariate distribution that we summarized in terms of different index numbers (*i.e.*, central tendency, variation, and form). The four aspects of a bivariate association are:

1. Whether or not an association *exists*.
2. The *strength* of that association.
3. The *direction* of the association.
4. The *nature* of the association.

Each of these characteristics will be discussed in turn, and in the next chapter we will develop several alternative measures of them. In fact, we will create a single number that will be used to describe the first three features of an association listed above and in some cases a simple formula can be used as an efficient description of the last.

6.4.1 The Existence of an Association

An association is said to exist between two variables if the distribution of one variable differs in some respect between at least some of the categories of the other variable. This rather general statement can be pinned down in a number of ways, the first of which we have already discussed. If, after computing percentages in the appropriate direction in a table, there is *any* difference between percentage distributions, we would say that an association exists in these data. In the table, below, the distribution of education is slightly different for men compared with women. We know this by percentaging in the direction of the independent variable and comparing across.

Education	*Men*	*Women*	*Total*
High	40%	38%	38%
Low	60	62	62
Total	100%	100%	100%
	(43)	(56)	(99)

In the table below, however, there is *no* association between "toenail length" and "education," and this is shown by the fact that there is no difference in the percentage distribution of education (the dependent variable) regardless of the category of the independent variable within which we examine the dependent variable.

	Toenail Length		
Education	*Short*	*Long*	*Total*
High	33%	33%	33%
Low	67	67	67
Total	100%	100%	100%
	(521)	(1756)	(2277)

In the following table it is clear that there *is* an association between social class and the number of arrests, because the percentage distributions, comparing across the way percentages were run, are different.

	Social Class		
Number of Arrests	*Low*	*Medium*	*High*
None	16%	28%	45%
Few	18	18	35
Many	66	54	20
Total	100%	100%	100%
	(129)	(129)	(73)

Recall that there is a name for these comparisons: **epsilon** (ε), which is the percentage difference computed across the way percentages were run in a table. In a table where *all* of the epsilons are 0, there is *no* association. If any epsilon is non-0, there is an association in the data even though we may not choose to consider the very small differences important enough to talk about.

The second way to tell whether or not there is an association in a table is to compare the **actual observed table frequencies** with the frequencies we would expect if there were no association, or **expected frequencies**. If the match between actual data and our model of no association is perfect, then there is no association in the actual data between the two variables that were cross-tabulated in the table.

6.4.1a No-Association Models

A **model of no association** can be set up for a specific table as follows. Usually in setting up a model of the way frequencies in a table should look if there were no association, we assume that the marginal distribution of each variable is the way it is in the observed data table, and that the total number of cases is the same. The problem is to specify the pattern of cell frequencies in the body of the table in a way that shows no association. As an example, suppose the marginals for variables X and Y are as follows:

	(X)		
(Y)	Low	High	Total
High	a	b	57
Low	c	d	50
Total	34	73	107

The problem is to find a pattern of frequencies for cells a, b, c, and d such that they exhibit no association between X and Y. The reasoning goes like this. If there is no association in the table, then the ratio of "high" cell frequencies for variable Y as related to the corresponding column totals should be the same throughout the table, as it is in the overall distribution of Y itself, namely 57 to 107. In the table above. we would expect 57/107ths of the 34 cases in the "low" category of X to be in the "high" category of Y. Furthermore, we would expect the same ratio, 57/107ths of the 73 cases in the high column of X to be in the top row. This would mean that, relatively speaking, there is no difference between the proportion of cases in the top row for any column of the table.

$$\frac{57}{107}(34) = .533(34) = 18.1 \text{ cases } \textit{expected} \text{ in cell } a$$

$$\frac{57}{107}(73) = .533(73) = 38.9 \text{ cases } \textit{expected} \text{ in cell } b$$

Given that one of the above cell frequencies in a 2 by 2 table is computed, the other expected cell frequencies could be determined by subtraction. The resulting table of expected cell frequencies (expected if there were no association between the two variables X and Y for these 107 cases) is shown below.

"Expected" Cell Frequencies

	(X)		
(Y)	Low	High	Total
High	18.1	38.9	57.0
Low	15.9	34.1	50.0
Total	34.0	73.0	107.0

This is a hypothetical tabulation showing no association and thus fractional frequencies are acceptable.

Expected cell frequencies (f_e) can be computed for a given cell by multiplying the row total for that cell by the column total for that cell and dividing by N, which is the operation explained above.

$$f_{e_{ij}} = \frac{(n_{i.})(n_{.j})}{N} \tag{6.1}$$

where $f_{e_{ij}}$ refers here to the expected cell frequency for the cell in the ith row and jth column of the table; $n_{i.}$ is the total for the ith row and $n_{.j}$ is the total for the jth column; and N is the total number of cases. An expected cell frequency is computed (or found by subtraction) for each cell in the table.

Now the difference between the table of observed data and the model we could construct of how this table would look if there were no association can be compared. This comparison is made by subtracting an expected cell frequency, f_e, from the corresponding observed cell frequency, f_o. The difference is called **delta**, and in this text we will symbolize delta with the upper case Greek letter delta (Δ). For a given cell,

$$\Delta = f_o - f_e \tag{6.2}$$

A delta value can be computed for each cell in a table, regardless of the size of the table. If any of the deltas are *not* 0, then there is at least some association shown in the table. Whenever all deltas are 0, all epsilons will also be 0. Later we will discuss summary measures of association based on these ideas.

In summary, whether or not an association exists in a table of observed frequencies can be exactly determined in two ways that yield the same conclusions. One way is to compute percentages in one direction and compare across in the other direction, using epsilon. The other way is to create a table of expected cell frequencies and compare the observed and expected cell frequencies, cell by corresponding cell, using delta. If all of the epsilons that can be computed in a table, or if all of the delta values for a table, amount to 0, then there is no association between the two variables cross-tabulated in the bivariate distribution. This is called **statistical independence**. If, on the other hand, there is any epsilon or any delta that is not 0, then there is an association in the observed frequency table, however slight or large that association might be.

6.4.2 Degree (Strength) of Association

Where the differences between percentages (epsilons) are large, or where the deltas are large, we speak of a strong **degree of association** between the two variables; that is, the dependent variable is distributed quite differently within the different conditional distributions defined by the independent variable. This can be contrasted with a weak association where there is very little difference or where the epsilons and deltas are very small, approaching or equaling 0.

Often investigators use epsilon (or delta) as a crude first indicator of the strength of association. The problem with both delta and epsilon is that it is difficult to determine what a given-sized delta or epsilon means, other than that there is some association in the table. The reason for this is that both delta and epsilon values for any cell(s) can vary from 0 or near-0 up to a magnitude that is not, in general, fixed. They are not "normed" or standardized. Later, in this chapter and in the next, the problem of creating good standardized measures of the strength of association will be discussed and several alternative measures will be described. Suffice it to say here that some tables show a strong relationship between independent and dependent variables, and some show a weak association or no association at all.

6.4.3 Direction of Association

Where the dependent and independent variables in a table are at least ordinal variables, it makes sense to speak about the **direction** of an association that may exist. If the tendency in the table as shown by the percentage distribution is for the higher values of one variable to be associated with the higher values of the other variable (and the lower values of each variable also tend to go together). then the association is called a *positive* association. Height and weight tend to have a positive association, since taller persons tend to be heavier; in general. across the people in a general population.

On the other hand, if the higher values of one variable are associated with lower values of the other (and the lower values of the first with higher values of the second), the association is said to be *negative*. Sociologists generally expect that the higher the educational level of people, the lower their degree of normlessness will be—a negative association.

The association between city size and tolerance scores (see Table 6.3) is positive because the larger the city, in general, the higher the tolerance level becomes (*i.e.*, the higher the percentage of people who have high tolerance scores). The older a person's age, in general, the fewer the years left until retirement, a negative association.

6.4.4 The Nature of Association

Finally, the **nature** of an association is a feature of a bivariate distribution referring to the general *pattern* of the data in the table. This is often discovered by examining the pattern of percentages in a properly percentaged table. Often the pattern is irregular, and an investigator would cite many epsilons in describing where the various concentrations of cases are in the different categories of the independent variable. Sometimes there is a rather uniform progression in concentration of cases on the dependent variable as we move toward higher values of the independent variable. If, with an increase of one step in one variable, cases tend to move up (or down) a certain number of steps on the other variable we might call the nature of the association "linear." That is, the concentrations of cases on the dependent variable (the mode, for example) tend to fall along a straight line that could be drawn through the table.

The nature of association will be discussed at length in the next chapter. Simple linear associations have an intrinsic interest to investigators as one of the simplest natures of association, but some associations are curvilinear in nature, or of some more complex patterning. In most cases the nature of association will be determined from a percentage table or a scatter plot, but in some cases nature can be described in terms of an equation.

At this point we should pause to examine several tables and describe them in terms of these four features of an association. Table 6.9 presents a series of examples together with brief summary statements.

TABLE 6.9A Percentage Distribution of Total Money Income for Families by Ethnic Background of Householder, 1983

Income	White	Spanish	Black	Total
$35,000 and over	31.5	15.1	13.2	29.6
$25,000 to 34,999	20.3	14.3	14.0	19.5
$15,000 to 24,999	23.7	27.0	21.5	23.4
$ 7,500 to 14,999	16.0	23.6	23.7	16.8
Under $7,500	8.5	20.0	27.6	10.7
Total	100.0	100.0	100.0	100.0
N (millions)	(53.9)	(3.6)	(6.7)	(62.0)

Source: U.S. Bureau of the Census. (1985).

1. *Existence of Association.* This table is percentaged down in the direction of the independent variable, ethnicity, so that comparisons may be made across. The percentages across are different so that there is an association evident in the table. Compare any row, say the row for incomes of $35,000 or more; percentages range from a high of 31.5% down to 13.2%, all different from 29.6%, the total for that income category.

2. *Strength of Association.* Overall, if ethnicity makes any difference in the distribution of family income, we should find fairly substantial epsilon's. Here, the white-black epsilon for $35,000 and over incomes is 18.3. That for Spanish-black for the same income category is 1.9 and for white-Spanish is 16.4. These are all less than 100% but substantially larger than 0.

3. *Direction of Association.* Here, ethnicity is a nominal variable so that it is impossible to talk about direction of association.

4. *Nature of Association.* To see the pattern of association most clearly, it is helpful to underline the highest percentages in each row-wise comparison. Here, for the highest income category, 31.5% is clearly the largest percentage and is underlined. In the next row, 20.3% is largest and is underlined; 27.0% is underlined in the third row. We will underline both the 23.6% and 23.7% in the fourth row because they are essentially equivalent in magnitude. Finally, 27.6% is underlined in the bottom row. Notice that we are underlining the percentages from the body of the table, not the marginal distribution. The nature of association is the pattern of white's having higher percentages in the highest two income groups, compared with the other two ethnic groups; Spanish have higher percentages in the next group and are tied in percentages in the $7,500 to 14,999 income category. Finally, the black group has highest percentages in the lowest income category. The pattern of high and low percentages from comparisons across the way the percentages were run is the nature of the association of ethnicity and income for families in 1983.

TABLE 6.9B PERCENTAGE DISTRIBUTION OF FATHER'S OCCUPATION BY OCCUPATION OF 30–59-YEAR-OLD CHILD, SWEDEN, 1977

Occupation of Father	*Occupation of Child*				
	Farmer	Worker	Entrepreneur	Middle Class	Not Known
Middle Class	2	10	13	29	14
Entrepreneur	6	10	23	15	10
Worker	14	52	38	39	39
Farmer	77	24	21	14	21
Not Known	1	4	5	3	16
Total	100.	100.	100.	100.	100.
N	(241)	(2964)	(525)	(2557)	(166)

Source: Adapted from Sundstrom (1986:369).

1. *Existence of Association.* This table presents the results of a Swedish survey of occupations of adult (30–59-year-old) children and their fathers. The table is percentaged in the direction of the child's occupation to show the distribution of fathers' occupation. Comparing across, the percentages are different, thus there is an association in the table. Note that this way of percentaging the table permits one to make statements about background occupational experience of children (e.g., their father's occupation). Percentaging the table the other way would permit one to say something about the distribution of children's occupations for fathers in different occupations. Percentaging to the total would permit statements about the percentage of people who, for example, stayed in the same occupation that their father had. The way percentages are run permits quite different kinds of comparisons.

2. *Strength of Association.* In this table, the highest epsilon ought to be seen in comparing extreme categories of child's occupation for the highest (or lowest) category of father's occupation. Taking the middle class fathers, the farmer to middle-class epsilon is 27%. For fathers who are farmers, the same epsilon comparison is 63%. These are both very substantial epsilons. The association is quite strong.

3. *Direction of Association.* As an aid in finding the direction of an association between variables each of which is at least ordinal, a useful procedure is to make comparisons across the way percentages are run, underlining the highest percentage for each comparison. In this table we could make four comparisons (aside from the not known category that the author provides here). For the middle class row the highest percentage is "middle class." For the entrepreneur row, it is "entrepreneur"; it is "worker" for the worker row, and "farmer" for the farmer row. Notice that the highest percentage is for father having the same occupation as the child, or no social mobility between generations. One could draw a diagonal line through the underlined percentages in the table. In this case, the line would extend from the "farmer"–"farmer" or lower status occupational category to the highest category combination, "middle class"–"middle class." This indicates a positive association: the higher the occupational status of the father, the higher the child's occupational status tends to

be. In this "occupational mobility" table, people in the upper left-hand area above the diagonal are downwardly mobile (fathers had higher status occupations than the child) and people in the lower right-hand area under the diagonal are upwardly mobile (child has a higher status occupation than father).

4. *Nature of Association.* In this table, the nature of association (i.e., the pattern of concentration in the table) tends to be almost linear. There is a relatively uniform shift toward higher status child's occupation with shifts upward in the category of father's occupation. There are no "reversals" in the general trend of concentration. Because the variables are ordinal, it would be more appropriate to speak of this nature of association as "monotonic" rather than linear. If distances were defined then one could determine whether in fact there is a constant amount of shift in values of one variable, given a fixed amount of difference in the other. In ordinal variables one can only say that the value of one variable remained the same or shifted in a fixed direction with increases in the other variable—a monotonic nature of association. Contrasted with this type of nature are those such as the one shown in Table 6.9C.

TABLE 6.9C PERCENTAGE DISTRIBUTION OF BODY WEIGHT BY AGE FOR PERSONS 20 YEARS AND OLDER

Percentage Above or Below Desired Weight	*Age*					
	20–34	*35–44*	*45–54*	*55–64*	*65–74*	*75+*
30% or more above	10.7	16.7	21.8	21.3	21.5	11.6
20–29.9% above	7.8	11.0	13.3	13.9	13.3	11.2
10–19.9% above	16.5	21.5	23.1	23.3	22.0	20.1
5–9.9% above	12.6	13.1	13.1	12.9	12.3	15.2
Plus or minus 4.9%	27.6	22.5	18.1	18.9	18.1	*21.5*
5–9.9% below	11.5	8.0	5.7	5.0	6.2	*8.4*
10% or more below	13.3	7.2	4.9	4.7	6.6	*12.0*
Total	100.0	100.0	100.0	100.0	100.0	100.0
N (1000's)	(59.9)	(28.9)	(22.2)	(22.0)	(16.4)	(9.5)

Source: National Center for Health Statistics (1986).

1. *Existence of Association.* In this table, age is treated as the independent variable and the amount by which one's weight is above or below the desired weight is the dependent variable. Comparing across, there is a percentage difference, thus there is an association shown in the table.

2. *Strength of Association.* Among all the possible percentage comparisons, the strongest percentage difference should be evident in comparing the extreme categories of age for the extreme categories of percentage above or below desired weight. Taking the top row, the overall epsilon is only 1.1 and in the bottom row it is only 1.3. These are indeed small percentage differences. Yet, there are larger percentage differences in the table; for example, the difference between the 20–34 and 45–54 age categories for the top row of the table. As we shall see, the pattern of association in this table is irregular, making the assessment of strength of association more complex. Even at its best, however, the percentage differences are rather small, suggesting a weak association between the two variables.

3. *Direction of Association.* As before, we will underline the highest percentages in each comparison, underlining more than one where percentages are very close in magnitude. Table 6.9C shows a pattern that moves generally from upper right to lower left, between overweight associated with older ages and underweight with younger ages, a generally "positive" association.

4. *Nature of Association.* Although the overall pattern of association indicated by drawing a diagonal line through the underlined percentages (or the middle of several underlined percentages in a given comparison) is linear, there are other patterns that need to be examined. Notice that percentages are essentially tied from 45–74 years old for the top two rows but there is a clearer concentration of high percentage in one column for the bottom three rows. There is a broader area of high percentages for overweight rows than for underweight rows. Notice too that the percentages along the bottom three or four rows drop down as one moves across from low to high age categories and then the percentages begin to rise again. If the second "high" is italicized for each of the bottom three rows one can see a more complex pattern emerging from the table. The nature of association begins to appear "curvilinear." Underweight is concentrated in the lowest age bracket and to some extent in the oldest age bracket while overweight is more likely found in middle age categories. It was this curved nature of association that made an assessment of strength of association difficult to determine if we made epsilon comparisons as if we expected a monotonic relationship. We will have more to say about this later but suffice it to say that one needs to be aware of the nature of association in selecting measures of strength of association.

LOETHER Herman J. and Donald G. McTAVISH. 1988. Descriptive and Inferencial Statistics. An Introduction. Boston, Allyn and Bacon.

NORUŠIS, Marija J. 1998. *Guide to Data Analysis*. Upper Saddle River, Prentice Hall.

Counting Responses for Combinations of Variables

How can you study the relationship between two or more variables that have a small number of possible values?

- Why is a frequency table not enough?
- What is a crosstabulation?
- What kinds of percentages can you compute for a crosstabulation, and how do you choose among them?
- What's a dependent variable? An independent variable?
- What if you want to examine more than two variables together?
- How can you use a chart to display a crosstabulation?

The Means and Explore procedures described in Chapter 5 and Chapter 6 are useful only when statistics such as the mean and standard deviation are appropriate measures for the variable whose values you want to summarize. You can't use Means or Explore to look for relationships between color of car driven and region of the country, since it doesn't make sense to compute an average color or region. When you want to look at the relationship between two variables that have a small number of values or categories (sometimes called **categorical variables**), you may want to use a **crosstabulation**, a table that contains counts of the number of times various combinations of values of two variables occur. For example, you can count how many men and how many women are in each of the job satisfaction categories, or you can see the distribution of car colors for various regions of the country.

In this chapter, you'll use a crosstabulation to look at the relationship between job satisfaction and total family income, measured on a four-point scale.

▶ This chapter continues to use the *gssft.sav* file. For instructions on how to obtain the crosstabulation output shown in this chapter, see "How to Obtain a Crosstabulation" on p. 122.

Income and Job Satisfaction

In the General Social Survey, respondents are asked to select the range of values into which their annual family income falls. There are 21 categories, ranging from under $1,000 (assigned a code of 1) to $75,000 and over (assigned a code of 21). To look at the relationship between income and job satisfaction for full-time employees, you'll use four income groups with roughly the same number of cases. That is, you will use quartiles of income. (The variable *income4* contains the income data recoded into quartile categories.) You see from Figure 7.1, which shows a frequency distribution of the four categories of income, that as expected, roughly 25% of the people fall into each of the income groupings.

To obtain this frequency table, select the variable income4 in the Frequencies dialog box. See Chapter 3 for information on frequency tables.

Figure 7.1 Frequency table for income quartiles

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	24,999 or less	174	23.3	23.3	23.3
	25,000 to 39,999	194	26.0	26.0	49.3
	40,000 to 59,999	156	20.9	20.9	70.1
	60,000 or more	223	29.9	29.9	100.0
	Total	747	100.0	100.0	

To examine the relationship between income and job satisfaction, you want to count how many *very satisfied, moderately satisfied, a little dissatisfied*, and *very dissatisfied* people there are in each of the income categories. Figure 7.2 contains this information. The income groups make up the columns of the table. The rows are the job satisfaction categories. A cell appears in the table for each combination of values of the two variables. The first cell, at the top left of the table, is for *very satisfied* people in the lowest income group. You see that 53 people fall into this cell. The cell in the second row of the first column is for *moderately satisfied* people in the lowest income category. There are 93 people in this cell. Similarly, the cell in the fourth row of the fourth column tells you that there are 7 *very dissatisfied* people in the highest income group.

NORUŠIS, Marija J. 1998. *Guide to Data Analysis*. Upper Saddle River, Prentice Hall.

To obtain this crosstabulation, from the menus choose:

Statistics ▶ Summarize ▶ Crosstabs...

In the Crosstabs dialog box, select the variables satjob and income4, as shown in Figure 7.9.

Figure 7.2 Crosstabulation of job satisfaction by income

Count

		Total Family Income in quartiles				
		24,999 or less	25,000 to 39,999	40,000 to 59,999	60,000 or more	Total
Job Satisfaction	Very satisfied	53	90	74	110	327
	Mod satisfied	93	79	61	87	320
	A little dissatisfied	24	17	14	19	74
	Very dissatisfied	4	8	7	7	26
Total		174	194	156	223	747

194 people are in the second income group

110 people with incomes of $60,000 or more are very satisfied

To the right and at the bottom of the table are totals—often called **marginal totals** because they are in the table's margin. The margins on the table show the same information as frequency tables for each of the two variables. In the right margin, labeled *Total*, you have the total number of people who gave each of the job satisfaction answers. Similarly, the first column total of 174 is the number of people in the lowest income category. The very last number, 747, is the total number of people in the table.

Will the marginal totals that I get in a crosstabulation table always be the same as those I would get from frequency tables for the variables individually? Not if you have missing values for either of the two variables in the crosstabulation. For example, the crosstabulation in Figure 7.2 includes only cases that have nonmissing values both for job satisfaction and for income. The marginal totals for income are therefore based on cases that have nonmissing values for both income and job satisfaction. When you make a frequency table for income, the only cases excluded from the valid percentages are those with missing values for income.

If you look at the counts in the crosstabulation, you see that 53 people from the lowest income category said they are *very satisfied* with their jobs, 90 from the second income category, 74 from the third income category, and 110 from the highest income category. Can you tell from the counts just what the relationship is between income and job satisfaction? Of course not, since you can't just compare the counts when there are different numbers of people in the four income groups. To compare the groups you must look at percentages instead of counts. That is, you must look at the percentage of people in each of the income groups who gave each of the job satisfaction responses.

Row and Column Percentages

Figure 7.3 contains both the counts and the column percentages. From the totals for each of the rows, you see that, overall, 43.8% of the sample are *very satisfied* with their jobs. You also see that 30.5% of the lowest income group, 46.4% of the second income group, 47.4% of the third income group, and 49.3% of the highest income group are *very satisfied* with their jobs. It appears that the lowest income people are less likely than average to be *very satisfied*, while the high income people are more likely than average to be *very satisfied*.

To obtain column percentages, select Cells in the Crosstabs dialog box. Then select Column, as shown in Figure 7.11.

Use the Pivot Table Editor to specify labels of your choice.

Figure 7.3 Crosstabulation showing column percentages

			Total Family Income in quartiles				
			24,999 or less	25,000 to 39,999	40,000 to 59,999	60,000 or more	Total
Job Satisfaction	Very satisfied	Count	53	90	74	110	327
		% within Total Family Income in quartiles	30.5%	46.4%	47.4%	49.3%	43.8%
	Mod satisfied	Count	93	79	61	87	320
		% within Total Family Income in quartiles	53.4%	40.7%	39.1%	39.0%	42.8%
	A little dissatisfied	Count	24	17	14	19	74
		% within Total Family Income in quartiles	13.8%	8.8%	9.0%	8.5%	9.9%
	Very dissatisfied	Count	4	8	7	7	26
		% within Total Family Income in quartiles	2.3%	4.1%	4.5%	3.1%	3.5%
Total		Count	174	194	156	223	747
		% within Total Family Income in quartiles	100.0%	100.0%	100.0%	100.0%	100.0%

You can change the label to indicate that the column percentages are shown

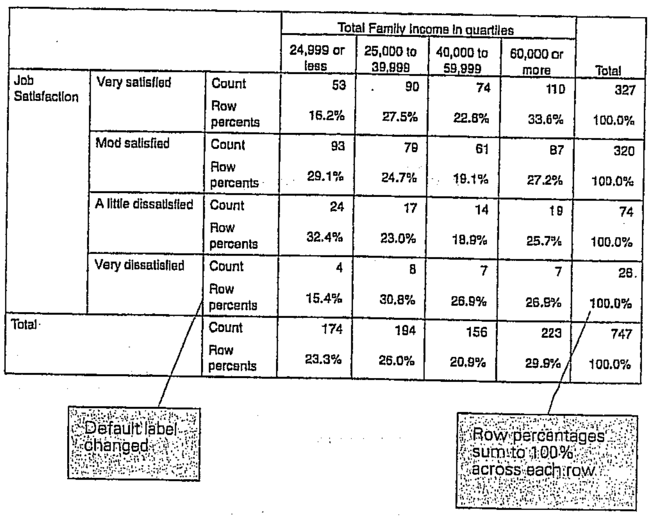
Column percentages sum to 100% in each column

The percentages you used to make comparisons are known as column percentages, since they express the number of cases in each cell of the table as a percentage of the column total. That is, for each income group, they tell you the distribution of job satisfaction. The column percentages sum up to 100% for each of the columns. In Figure 7.3, you can see that 30.5% of people with incomes less than $25,000 are very satisfied with their jobs, while 49.3% of people with incomes of $60,000 and up are very satisfied. The percentage of people who are very dissatisfied or a little dissatisfied is largest in the lowest income category (16.1%).

You can also calculate row percentages for the table. Row percentages tell you what percentage of the total cases of a row fall into each of the columns. For each job satisfaction category, they tell you the percentage of cases in each income category. (You can also compute what are called total percentages. The count in each cell of the table is expressed as a percentage of the total number of cases in the table.) Figure 7.4 contains counts and row percentages for our example.

To obtain row percentages, select Cells in the Crosstabs dialog box. Then select Row. (See Figure 7.11.)

Figure 7.4



			Total Family Income in quartiles				
			24,999 or less	25,000 to 39,999	40,000 to 59,999	60,000 or more	Total
Job Satisfaction	Very satisfied	Count	53	90	74	110	327
		Row percents	16.2%	27.5%	22.6%	33.6%	100.0%
	Mod satisfied	Count	93	79	61	87	320
		Row percents	29.1%	24.7%	19.1%	27.2%	100.0%
	A little dissatisfied	Count	24	17	14	19	74
		Row percents	32.4%	23.0%	18.9%	25.7%	100.0%
	Very dissatisfied	Count	4	8	7	7	26
		Row percents	15.4%	30.8%	26.9%	26.9%	100.0%
Total		Count	174	194	156	223	747
		Row percents	23.3%	26.0%	20.9%	29.9%	100.0%

Default label changed

Row percentages sum to 100% across each row

From the row percentages, you see that 16.2% of the *very satisfied* respondents are in the lowest income group, 27.5% are in the second income group, 22.6% are in the third income group, and 33.6% are in the fourth income group. The four row percentage values sum to 100 for each of the rows. In this example, the row percentages aren't very helpful, since you can't make much sense of them without taking into account the overall percentages of cases in each of the income categories. That is, you can't tell whether the percentage of high income cases in the *very satisfied* category is due to a large number of high income cases in your sample or to high satisfaction rates in that category.

How can I tell whether a table contains row or column percentages? If the column labeled *Total* shows all 100%, the table contains row percentages, which necessarily sum to a 100 for each row. If the row labeled *Total* contains 100%, the tables contains column percentages.

For a particular table, you must determine whether the row or column percentages answer the question of interest. This can be done easily if one of the variables can be thought of as an independent variable and the other as a dependent variable. An **independent variable** is a variable that is thought to influence another variable, the **dependent variable**. For example, if you are studying the incidence of lung cancer in smokers and nonsmokers, smoking is the independent variable. Smoking influences whether people get cancer, the dependent variable. Similarly, if you are studying the income categories of men and women, gender is the independent variable since it might influence how much you get paid.

If you can identify one of your variables as independent and the other as dependent, then you should compute percentages so that they sum to 100 for each category of the independent variable. In other words, what you want to see is the same number of people in each of the categories of the independent variable. Having the percentages sum to 100 for each category of the independent variable is the equivalent of having 100 cases in each category. For example, you want 100 smokers and 100 nonsmokers. Then you can compare the incidence of lung cancer in the two groups. In the current example, income category is the independent variable and job satisfaction is the dependent variable. That means you'd like to see 100 people in each of the income categories. Since income is the column variable in Figure 7.3, you use column percentages that sum to 100 for each category of income.

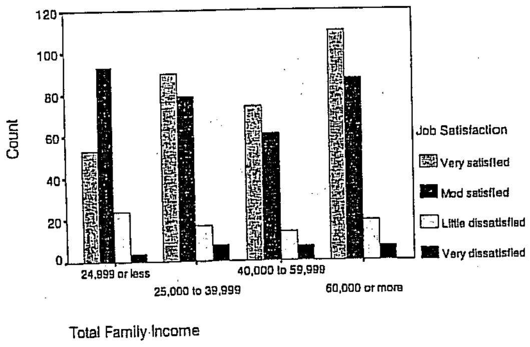
Can't you analyze these data using the Means procedure? The General Social Survey codes income in unequal intervals. For example, the interval from $8,000 to $9,999 is coded 8, but the interval $60,000 to $74,999 is coded 20. So you don't want to compute means for these codes. Instead, if you want to compute average family income, you must change the coding scheme so that the code for a case is the midpoint of the appropriate income interval. For example, an income anywhere in the range of $8,000 to $9,999 would be assigned a code of $9,000. Similarly, incomes in the range of $60,000 to $74,999 would be assigned a code of $67,500, the midpoint of the interval. You can then compute descriptive statistics for the recoded incomes. Of course, the means won't be the same as those you would get if you had the exact income for each person, but they're the best you can do given the limitations of the data.

Bar Charts

You can display the results of a crosstabulation in a clustered bar chart. Consider Figure 7.5, which is a bar chart of family income by job satisfaction. The length of a bar tells you the number of cases in a category.

Figure 7.5 Bar chart of income by job satisfaction

You can obtain this bar chart using the Graphs menu, as described in "Bar Charts" on p. 520 in Appendix A. In the Clustered Bar Summaries for Groups of Cases dialog box, select the variables income4 and satjob.



There is a cluster of bars for each of the four income categories. Within each cluster, there is a bar for each of the job satisfaction categories.

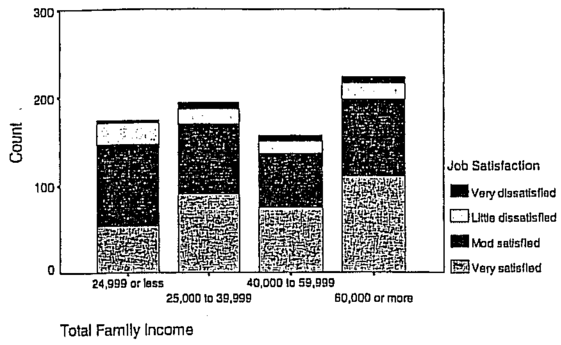
Since there are unequal numbers of people in the income categories, comparing bar lengths across income categories presents the same problem as looking at simple counts in a crosstabulation. All you can really do with this bar chart is compare bar lengths within a cluster and see whether the patterns are the same across clusters.

Figure 7.6 Stacked bar chart

You can obtain this chart by modifying Figure 7.5, as described in "Bar Charts" on p. 520 in Appendix A. From the Chart Editor menus choose:

Gallery
Bar...

In the Bar Charts dialog box, select Stacked.



Stacked Bar Charts

You can stack the bars in a clustered bar chart one on top of the other. The result is the stacked bar chart in Figure 7.6. Now it's easier to see for each income category the proportion of people in each of the job satisfaction categories. However, the lengths of the bars aren't equal for the four income categories, so that still gets in the way.

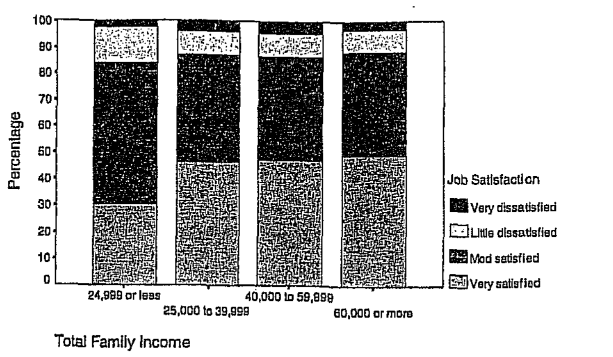
Ideally, you want each of the bars to be of the same length, so you can easily compare the areas across bars. What you'd really like to see is a plot of the column percentages from Figure 7.3. You can do this by turning the counts in each bar into percentages, as shown in Figure 7.7. Now each of the bars has the same length—100%—and you can easily compare the job satisfaction distributions across bars. You see that people in the lowest income group are least likely to be *very satisfied* with their jobs. They are also least likely to be *very dissatisfied*. The distribution of job satisfaction categories seems to be very similar for the other three income groups. The proportion of *very satisfied* people doesn't increase with income for these three groups. You can also see that the sum of the percentages for *very satisfied* and *moderately satisfied* is very similar for the four groups.

You can obtain this bar chart using the Graphs menu, as described in "Bar Charts" on p. 520 in Appendix A. In the Clustered Bar Summaries for Groups of Cases dialog box, select the variables income4 and satjob. Or you can select Display clustered bar charts in the Crosstabs dialog box. From the Chart Editor menus choose:

Series
Transpose Data

to cluster bars within income categories.

Figure 7.7 Stacked bar chart with percentage scale



Summary

How can you study the relationship between two or more variables that have a small number of possible values?

- A crosstabulation shows the numbers of cases that have particular combinations of values for two or more variables.
- The number of cases in each cell of a crosstabulation can be expressed as the percentage of all cases in that row (the row percentage) or the percentage of all cases in that column (the column percentage).
- A variable that is thought to influence the values of another variable is called an independent variable.
- The variable that is influenced is called the dependent variable.
- If there is an independent variable, percentages should be calculated so that they sum to 100% for each category of the independent variable.
- When you have more than two variables, you can make separate crosstabulations for each of the combinations of values of the other variables.
- Bar charts can be used to display a crosstabulation graphically.

NORUŠIS, Marija J. 1998. *Guide to Data Analysis*. Upper Saddle River, Prentice Hall.

16

Comparing Observed and Expected Counts

How can you test the null hypothesis that two variables are independent?

- What are observed and expected counts?
- How do you compute the chi-square statistic?
- What assumptions are needed for the chi-square test of independence?
- What is a one-sample chi-square test?
- Why is sample size important?

You know how to test a variety of hypotheses about population means. However, these tests are useful only when it makes sense to compute a mean for a variable. If you want to look at the relationship between preference among car colors and region of the country, or between type of treatment and remission of symptoms, you can't use a *t* test because it doesn't make sense to compare means. Rather, such variables are best summarized by a crosstabulation. In this chapter, you'll use the **chi-square test** to examine hypotheses about data that are best summarized by a crosstabulation.

▶ This chapter uses the *gss.sav* data file. The chi-square test output shown can be obtained using the SPSS Crosstabs procedure. (For more information on Crosstabs, see Chapter 7.)

Education and Anomia

The French sociologist Emile Durkheim introduced the concept of anomie to represent the feelings of alienation and rootlessness common in the modern world. The General Social Survey attempts to measure such feelings with a scale called *anomia*. One item on this scale asks respondents whether they agree or disagree with the following statement:

"In spite of what some people say, the lot of the average man is getting worse, not better." Let's consider whether education is related to the likelihood of agreeing with this statement.

You can obtain a crosstabulation using the Crosstabs procedure, as discussed in Chapter 7.

Select anomia5 and degree2 in the Crosstabs dialog box.

Figure 16.1 Crosstabulation of anomia and education

			College Degree		
			No College degree	College degree	Total
Lot of average man getting worse	Agree	Count	529	132	661
		Column %	**72.0%**	58.9%	68.9%
	Disagree	Count	206	92	298
		Column %	28.0%	41.1%	31.1%
Total		Count	735	224	959
		Column %	100.0%	100.0%	100.0%

72% of people without college degrees agreed that things are getting worse

Figure 16.1 is a crosstabulation of responses to the statement for those with and without college degrees. You see that 72% of respondents who have not completed college agree with the statement, while 58.9% of respondents with college degrees agree with this statement. Based on these results, do you think that, in the population, there is a difference between college graduates and non-college graduates in the perception of the lot of the average man? Certainly in this sample, college graduates are less pessimistic than nongraduates. But as usual, the sample results are not what you're interested in. You want to know what you can conclude about the population based on the observed sample results. You want to know whether you have enough evidence to reject the null hypothesis that, in the population, the same percentage of college graduates and nongraduates agree with the statement.

Observed and Expected Counts

The basic element of a crosstabulation table is the count of the number of cases in each cell of the table. The statistical procedure you'll use to test the null hypothesis is based on comparing the observed count in each of the cells to the expected count. The expected count is simply the number of cases you would expect to find in a cell if the null hypothesis is true. Here's how the expected counts are calculated.

Calculating Expected Counts

If the null hypothesis is true, you expect college graduates and nongraduates to answer the question in the same way. That is, you expect the *percentage* agreeing with the statement to be the same for the two groups of cases. You don't expect the same *number* of graduates and nongraduates to agree with the statement, since you don't have the same number of people in the two education categories.

From the row marginals in Figure 16.1, you see that in the sample, 68.9% of the respondents agreed with the statement and 31.1% disagreed. If the null hypothesis is true, these are the best estimates for the *percentages* you would expect for both graduates and nongraduates. To convert the percentages to the actual number of cases in each of the cells, multiply the expected percentages by the numbers of graduates and nongraduates. For example, the expected number of nongraduates agreeing with the statement is

$$68.93\% \times 735 = 506.6 \qquad \textbf{Equation 16.1}$$

Similarly, the expected number of nongraduates disagreeing with the statement is

$$31.07\% \times 735 = 228.4 \qquad \textbf{Equation 16.2}$$

For college graduates, the expected values are calculated in the same way, substituting the number of college graduates (224) for the number of nongraduates (735) in the above two equations.

Is there a simple way I can remember how to calculate expected values? Sure. The following rule is equivalent to what you've just done: To calculate the expected number of cases in any cell of a crosstabulation, multiply the number of cases in the cell's row by the number of cases in the cell's column and divide by the total number of cases in the table. Try it. You'll see it always works. ■■■

Figure 16.2 Observed and expected counts

To obtain observed and expected counts, select Cells in the Crosstabs dialog box.

See Chapter 7 for more information.

			College Degree		
			No College degree	College degree	Total
Lot of average man getting worse	Agree	Count	529	132	661
		Expected Count	506.6	154.4	661.0
		Residual	22.4	-22.4	
	Disagree	Count	206	92	298
		Expected Count	228.4	69.6	298.0
		Residual	-22.4	22.4	
Total		Count	735	224	959
		Expected Count	735.0	224.0	959.0

Residual is the difference between observed and expected counts

You see in Figure 16.2 the observed and expected counts for all four cells. The last entry in a cell is the residual, the difference between the observed and expected counts. A positive residual means that you observed more cases in a cell than you would expect if the null hypothesis were true. A negative residual indicates that you observed fewer cases than you would expect if the null hypothesis were true.

The sum of the expected counts for any row or column is the same as the observed count for that row or column. For example, the expected counts for college graduates add up to the observed number of college graduates. Similarly, the expected counts for the number agreeing add up to the observed number of cases agreeing. Another way of saying this is that the residuals add up to 0 across any row and any column.

The Chi-Square Statistic

When you test the null hypothesis that two population means are equal, you compute the *t* statistic, and then, using the *t* distribution, calculate how unusual the observed value is if the null hypothesis is true. To test hypotheses about data that are counts, you compute what's called a chi-square statistic and compare its value to the chi- square distribution to see how unlikely the observed value is if the null hypothesis is true.

What assumptions are needed to use the chi-square test? All of your observations must be independent. That implies that an individual can appear only once in a table. You can't let a person choose two favorite car colors and then make a table of color preference by gender. (Each person would appear twice in such a table.) It also means that the categories of a variable can't overlap. (For example, you can't use the age groups less than 30, 25–40, 35–90.) Also, most of the expected counts must be greater than 5, and none less than 1. ■■■

To compute the Pearson chi-square statistic, do the following:

1. For each cell, calculate the expected count by multiplying the number of cases in the cell's row by the number of cases in the cell's column and dividing the result by the total count.
2. Find the difference between the observed and expected counts.
3. Square the difference.
4. Divide the squared difference by the expected count for the cell.
5. Add up the results of the previous step for all of the cells.

In the current example, the value for the Pearson chi-square statistic is

$$\frac{(529-506.6)^2}{506.6}+\frac{(132-154.4)^2}{154.4}+\frac{(206-228.4)^2}{228.4}+\frac{(92-69.6)^2}{69.6} = 13.64 \qquad \textbf{Equation 16.3}$$

If the null hypothesis is true, the observed and expected values should be similar. Of course, even if the null hypothesis is true, the observed and expected values won't be identical, since the results you observe in a sample vary somewhat around the true population value. As before, you have to determine how often to expect a chi-square value at least as large as the one you've calculated, if the null hypothesis is true.

To determine whether a chi-square value of 13.64 is unusual, you compare it to the chi-square distribution. Like the *t* distribution, the chi-square distribution depends on the parameter called the degrees of freedom. The degrees of freedom for the chi-square statistic depend *not* on the number of cases in your sample, as they did for the *t* statistic, but on the number of rows and columns in your crosstabulation. The degrees of freedom for the chi-square statistic are

$$(\text{number of rows in the table} - 1) \times (\text{number of columns in the table} - 1) \qquad \textbf{Equation 16.4}$$

For this example, there is one degree of freedom, since there are two rows and two columns.

What's the logic behind the calculation of the degrees of freedom? For any row or column of a crosstabulation, the residuals sum to 0. That means that you can tell what the expected values must be for the last row and last column of a table without doing any calculations other than summing the expected values in the preceding rows or columns. The number of cells for which you have to calculate expected values is equal to the number of cells when you remove the last row and the last column from your table. The number of cells in a table when one row and one column are removed is the number of rows minus 1 multiplied by the number of columns minus 1, which is the formula for the degrees of freedom. ■■■

Figure 16.3 Pearson chi-square test for anomia by education

To obtain the Pearson chi-square test along with a crosstabulation, select Statistics in the Crosstabs dialog box.

See Chapter 7 for more information.

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	13.639	1	.000		
Continuity Correction	13.036	1	.000		
Likelihood Ratio	13.201	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	13.624	1	.000		
N of Valid Cases	959				

Pearson chi-square (from Equation 16.3)

In Figure 16.3, you see that the observed significance level for the Pearson chi-square value of 13.64 is less than 0.0005. This means that, if the null hypothesis is true, you expect to see a chi-square value at least as large as 13.64 less than five times out of 10,000. Since the observed significance level is small, you can reject the null hypothesis that college graduates and those who did not graduate from college give the same responses to the question. It appears that college graduates are more optimistic about the lot of the common man than high school graduates.

What's all that other stuff in Figure 16.3 along with the Pearson chi-square? The continuity-corrected chi-square is a modification of the Pearson chi-square for two-by-two tables. Most statisticians agree that the modification is unnecessary, so you can ignore it. The likelihood-ratio chi-square is a statistic very similar to the Pearson chi-square. For large sample sizes, the two statistics are close in value. The *Linear-by-Linear Association test* is a measure of the linear association between the row and column variables. It's useful only if both the row and column variables are ordered from smallest to largest. Ignore it in other situations.

If you have a table with two rows and two columns, you'll also find something labeled *Fisher's Exact Test* on your output. The advantage of Fisher's exact test is that it is appropriate for 2 × 2 tables in which the expected value in one or more cells is small. The disadvantage is that it requires a very restrictive assumption about the data: that you know in advance the number of cases in the margins. There's controversy among statisticians about the appropriateness of Fisher's exact test when this assumption is not met. In general, Fisher's exact test is less likely to find true differences than it should. Statistically, a test like this is called conservative. ■■■

College Degrees and Perception of Life

In the previous example, you tested whether college graduates and those who are not college graduates respond in the same way to the question about the lot of the average man. The null hypothesis can be stated in several equivalent ways. You can say the null hypothesis is that the percentage agreeing with the statement is the same for the two categories of education. Another way of stating the null hypothesis is that educational status and response are independent.

Independence means that knowing the value of one of the variables for a case tells you nothing about the value of the other variable. For example, if marital status and happiness with life are independent, knowing a person's marital status gives you no information about how happy they are with life. College education and perception of the lot of man, on the other hand, don't seem to be independent. If you know that a person is a college graduate, you know that he or she is less likely to agree with the pessimistic statement about the lot of the average man than is a person who is not a college graduate.

A Larger Table

The chi-square test can be used to test the hypothesis of independence for a table with any number of rows and columns. The idea is the same as for the two-row and two-column table. As an example, let's look at the relationship between highest degree earned and whether life is perceived as exciting, routine, or dull.

Figure 16.4 is a crosstabulation of highest degree earned and the response to the perception of life question.

Figure 16.4 Crosstabulation of education and life

			Is life exciting or dull			
			Dull	Routine	Exciting	Total
RS Highest Degree	Less than HS	Count	24	96	66	186
		Expected Count	12.0	85.8	88.2	186.0
		Row %	12.9%	51.6%	35.5%	100.0%
		Residual	12.0	10.2	-22.2	
	High school	Count	35	251	231	517
		Expected Count	33.3	238.5	245.3	517.0
		Row %	6.8%	48.5%	44.7%	100.0%
		Residual	1.7	12.5	-14.3	
	Junior college	Count	2	33	27	62
		Expected Count	4.0	28.6	29.4	62.0
		Row %	3.2%	53.2%	43.5%	100.0%
		Residual	-2.0	4.4	-2.4	
	Bachelor	Count	2	58	97	157
		Expected Count	10.1	72.4	74.5	157.0
		Row %	1.3%	36.9%	61.8%	100.0%
		Residual	-8.1	-14.4	**22.5**	
	Graduate	Count	1	21	51	73
		Expected Count	4.7	33.7	34.6	73.0
		Row %	1.4%	28.8%	69.9%	100.0%
		Residual	-3.7	-12.7	**16.4**	
Total		Count	64	459	472	995
		Expected Count	64.0	459.0	472.0	995.0
		Row %	6.4%	46.1%	47.4%	100.0%

College graduates have large positive residuals in the Exciting column

From the row percentages, you see that almost 70% of people with graduate degrees find life exciting. (They probably don't read or write statistics books!) Only 36% of people with less than a high school diploma find life exciting. In fact, as education increases, so does the likelihood of finding life exciting. (Don't be alarmed by the large number of missing observations. Not all people in the General Social Survey were asked the question.)

To test the null hypothesis that highest degree and perception of life are independent, you compute a chi-square statistic for this table the same way you did for a 2 × 2 table. For example, if the null hypothesis is true, the expected number of people without high school diplomas who find life exciting is 88.2. (That can be calculated by multiplying the overall percentage of people who find life exciting, 47.4%, by the number of people without high school diplomas, 186.)

The Pearson chi-square value for the table is shown in Figure 16.5. You see that the observed significance level is less than 0.0005, which leads you to reject the null hypothesis that degree and perception of life are independent. By looking at the residuals in Figure 16.4, you see that college graduates have large positive residuals for the response *Exciting*. That means that the observed number of college graduates in those cells is larger than that predicted by the independence hypothesis. By examining the residuals in a crosstabulation, you can tell where the departures from independence are.

Figure 16.5 Pearson chi-square for crosstabulation of education and life

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	53.962[1]	8	.000
Likelihood Ratio	55.874	8	.000
Linear-by-Linear Association	47.633	1	.000
N of Valid Cases	995		

[1] 2 cells (13.3%) have expected count less than 5. The minimum expected count is 3.99.

Check these values to be sure your test is valid

After the chi-square statistics are printed, SPSS tells you what the smallest expected count is in any cell of the table. In this example, the *Minimum Expected Frequency* is 3.99. This is important because, if too many of the expected values in a table are less than 5, the observed significance level based on the chi-square distribution may not be correct. As a general rule, you should not use the chi-square test if more than 20% of the cells have expected values less than 5, or if the minimum expected frequency is less than 1.

What should I do if one of these conditions is not satisfied? If your table has more than two rows and two columns, you can see if it makes sense to combine some of the rows or columns. For example, if you have few people with graduate degrees, you can combine them into a single category with bachelor's degrees. Similarly, if necessary, you can combine the junior college graduates with the high school graduates, since their responses appear to be similar. ■■■

A One-Sample Chi-Square Test

So far, you've used the chi-square test to test for independence in a crosstabulation of two variables. You can also use the chi- square test to test null hypotheses about the distribution of values of a single variable. That is, you can see whether the distribution of observed counts in a frequency table is compatible with a set of expected counts. The expected counts are specified by the null hypothesis that you want to test. For example, you can test the hypothesis that people are equally likely to find life exciting, routine, or dull. Or you can test the null hypothesis that there are twice as many people without college degrees as there are with college degrees.

You can obtain this output using the Chi-Square Test procedure, as described in "Chi-Square Test" on p. 337 in Chapter 17. Select the variable life and All categories equal in the Chi-Square dialog box.

Figure 16.6 Chi-square test for life

	Observed N	Expected N	Residual
Dull	65	332.3	-267.3
Routine	459	332.3	126.7
Exciting	473	332.3	140.7
Total	997		

	Is life exciting or dull
Chi-Square	322.865
df	2
Asymp. Sig.	.000

Expected counts if null hypothesis is true

Look at Figure 16.6, which shows counts of the number of people who find life exciting, routine, and dull. Before you looked at the data, you might have thought that people were equally likely to find life exciting, routine, or dull. To test the null hypothesis that the three responses are equally likely in the population, you have to determine the expected counts for each of the categories. That's easy to do. For this hypothesis, the expected count for each category is just the total number of cases divided by 3.

You calculate the chi-square statistic the same way as before. Square each of the residuals (difference between observed and expected), divide by the expected count, and sum up for all of the cells. In Figure 16.6, you see that the chi-square value is a whopping 322.86. Its degrees of freedom are 2, one less than the number of categories in the table. Based on the observed significance level, you can handily reject the null hypothesis.

Let's try another test, this time specifying unequal numbers of expected counts for the categories. You want to test the null hypothesis that there are twice as many people in the population without college degrees as there are people with college degrees. That means you expect two-thirds of the people not to have college degrees and one-third to have college degrees. The expected counts for the two cells are 997.3 and 498.7.

Figure 16.7 Chi-square test for degree

	Observed N	Expected N	Residual
No College degree	1149	997.3	151.7
College degree	347	498.7	-151.7
Total	1496		

	College Degree
Chi-Square	69.193
df	1
Asymp. Sig.	**.000**

Observed results are highly unlikely if null hypothesis is true

The results of this test are shown in Figure 16.7. You see that the expected count for people without a degree is twice as large as it is for people with a college degree. From the residuals, you see that the two-to-one-ratio hypothesis predicts more college graduates than you observe. In the sample, the ratio is slightly larger than three to one. Again the chi-square statistic is large and the observed significance level is small, so you reject the null hypothesis that in the population, non-college graduates are twice as common as college graduates.

Power Concerns

You know that your ability to reject the null hypothesis when it's false, the power of a test, depends not only on the size of the discrepancy from the null hypothesis, but also on the sample size. The same is true, of course, for chi-square tests. The value of the chi-square statistic depends on the number of observations in the sample. For example, if you leave the table percentages unchanged but multiply the number of cases in each cell by 10, the chi-square value will be multiplied by 10 as well. This means that if you have small sample sizes, you may not be able to reject the null hypothesis even when it's false. Similarly, for large sample sizes, you will find yourself rejecting the null hypothesis even when the departures from independence are quite small.

When one or both of the variables in your crosstabulation is measured on an ordinal scale (for example, good/better/best), the chi-square test is not as powerful as some other statistics for detecting departures from independence. These other statistics make use of the additional information available for ordinal variables to measure both the strength and the direction of the relationship between two variables. If examination of the residuals in such a table leads you to suspect that there are departures from independence, you should use one of the measures described in Chapter 18.

Summary

How can you test the null hypothesis that two variables are independent?

- In a crosstabulation, the observed count is the number of cases in a particular cell.
- An expected count is the number of cases predicted if the two variables are independent.
- The chi-square statistic is based on a comparison of observed and expected counts.
- To use the chi-square test, your observations must be independent, and most of the expected values must be at least 5.
- A one-sample chi-square test is used to test whether a sample comes from a population with specified probabilities for the occurrence of each value.

Table 11.4 The links between correlations and tests of significance

Correlation	Significance	N	Interpretation
0 35	0 27	100	Moderate association in sample but too likely to be due to sampling error Continue to assume correlation of 0 in the population
0 15	0 001	1500	Weak association but is very likely to hold in the population.
0 64	0 01	450	Strong relationship that is likely to hold in the population.
0 04	0 77	600	Negligible association Highly probable that the correlation differs from zero due only to sampling error Continue to assume correlation of 0 in the population.

Table 11.5 Guidelines for selecting measures of association

Level of measurement of variables		Appropriate methods	Appropriate descriptive summary statistics	Appropriate inferential statistic
1 Nominal/Nominal	'Shape' of variables 2 by 2	Crosstabulations	i Phi ii Yules Q iii Lambda iv Goodman & Kruskall's tau	chi square
2 Nominal/Nominal	3+ by 2+	Crosstabulations	i Lambda ii Goodman & Kruskall's tau iii Cramers V	chi square
3 Nominal/Ordinal	Nominal variable with 3+ categories	Crosstabulations	i Theta ii Any statistics in 2 above	Mann - Whitney U-test (dichotomous nominal independent variable) K-sample median test Kruskal · Wallis
4 Nominal/Interval	Nominal variable independent	a Crosstabulations (if interval variable has only a few categories) b Comparison of means (esp. if interval variable has many categories)	i Eta (also called correlation ratio) ii Any statistics in 2 or 3 above but not very wise. i Eta	F-test (one-way analysis of variance) chi square F-test (one-way analysis of variance)
5 Ordinal/Ordinal	Both with few categories	Crosstabulations	i Gamma ii Kendall's tau b (square tables) iii Kendall's tau c (any shape table)	Test for significance of gamma Test for significance of tau
6 Ordinal/Ordinal	One variable with many categories	Rank correlation	i Kendall's tau	Test for significance of tau
7 Ordinal/Ordinal	Both variables with many categories	Rank correlation	i Kendall's tau ii Spearman's rho	as above Test for significance of rho
8 Ordinal/Interval	Both with few categories	a Crosstabulations b Comparison of means (if dependent variable is interval)	i Eta ii Any statistics in 5 above i Eta	F-test F-test
9 Ordinal/Interval	Ordinal with low categories Interval with many	a Comparison of means b Rank order correlation	i Eta i Kendall's tau	F-test Test for significance of tau
10 Ordinal/Interval	Both with many categories	Rank correlation	i Kendall's tau ii Spearman's rho	as above Test for significance of rho

de VAUS D. A. 1990. *Survey in Social Research*. London: Unwin Hyman.

Table 11.5 Guidelines for selecting measures of association

Level of measurement of variables			Appropriate methods	Appropriate descriptive summary statistics	Appropriate inferential statistic
11	Interval/Interval	Both variable with small number of categories	Crosstabulations	i Pearson's r	Test for significance of r
12	Interval/Interval	At least one variable with many categories	Scattergram	i Pearson's r ii Regression	

Table 10.15 Characteristics of various measures of association

	Appropriate table size	Range	Directional	Symmetric	Linear only	Other features
Phi	2 × 2	0 - 1[2]	no	yes	no	Lower co-efficients than Yule's Q
Cramer's V	larger than 2 × 2	0 - 1	no	yes	no	More sensitive to a wider range of relationships than lambda
Yule's Q	2 × 2	0 – 1	no	yes	no	1. Higher co-efficients than phi 2. Same as gamma 2 by 2 case 3 Always 1 00 if an empty cell
Lambda	any size[1]	0 – 1	no	yes[4]	no	Insensitive and therefore not recommended
Goodman and Kruskal's tau	any size	0 - 1[3]	no	no	no	More sensitive than lambda but not available on SPSS
Gamma	any size	0 – 1	yes	yes	yes	Gives higher co-efficients than Kendall's Tau_b or Tau_c
Kendall's Tau_b	square tables only	0 – 1	yes	yes	yes	
Kendall's Tau_c	any size	0 – 1	yes	yes	yes	
Eta	any size	0 – 1	no	no	no	
Pearson's r	any size	0 1	yes	yes	yes	

Notes: (1) i.e. given the qualifications in section 10 1.5
(2) Under certain conditions the maximum may be less than 1 (see Guilford, 1965 336)
(3) Will only be if there is perfect association and if the independent variable has the same number of categories as the dependent variable
(4) There is both a symmetric and asymmetric version

9. lekce

MĚŘENÍ (SÍLY) ASOCIACE MEZI DVĚMA SPOJITÝMI PROMĚNNÝMI: KORELAČNÍ KOEFICIENTY A GRAFY - SCATTERPLOTS (modul GRAF: procedura Scatter) A KORELAČNÍ MATICE (modul CORRELATE: procedura bivariate) A.

WILLIAMSONN, J. B., David A. KARP, John R. DALPHIN, Paul S. GRAY. 1982. *The Research Craft. An Introduction to Social Research Methods.* Boston: Little, Brown and Company.

The scattergrams in Figure 16.5 illustrate several alternative relationships between the independent variable X and the dependent variable Y. The following observations can be made on the basis of the information in Figure 16.5:

(A) the data for this scattergram illustrate a moderately strong positive correlation that would be approximately .60. You will note that in this scattergram, as in the others, the X values increase from left to right, that is, from L (low) to H (high); and the Y values increase from bottom to top (also from low to high). As with all positive correlations, there is a tendency for the Y values to increase as the X values increase.

(B) Here all the data points fall along a straight line; this is what happens when there is a perfect positive correlation between X and Y (r = 1.00). The correlation is perfect only in the sense that it represents the upper limit for the correlation coefficient. In actual social research applications we do not get correlations of 1.00 unless we have somehow managed to correlate a variable with itself.

(C) Here there is no relationship between X and Y (r = .00).

(D) Here there is a weak positive correlation (r = +.20) between X and Y.

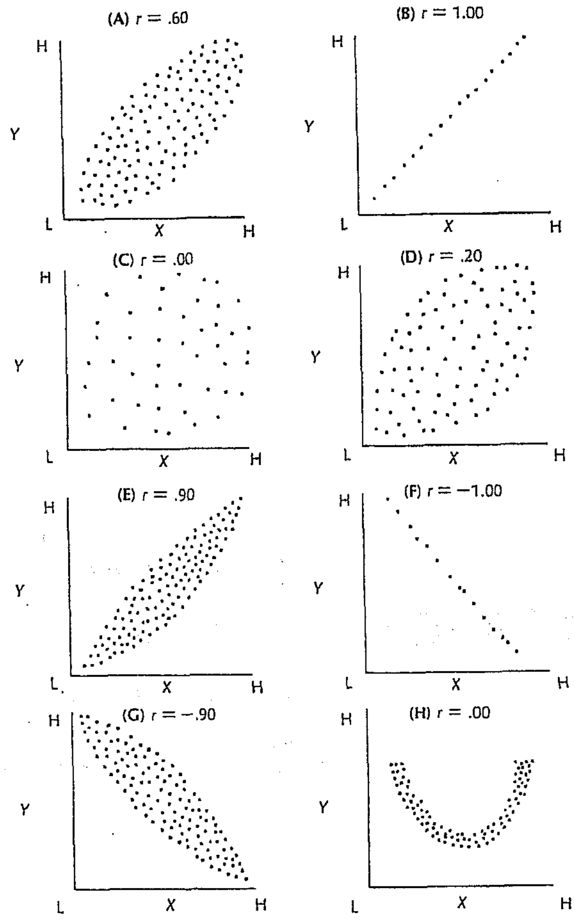
(E) Here there is a very strong positive correlation (r = +.90).

(F) Here there is a perfect negative correlation (r = −1.00). Note that for a negative correlation Y decreases as X increases.

(G) Here there is a strong negative correlation (r = −.90). An example of a negative correlation would be the relationship between cigarette consumption (X) and life expectancy (Y). As cigarette consumption increases, life expectancy decreases. (Undoubtedly, the actual correlation between these two variables is weaker than − .90.)

(H) Here there is a strong NONLINEAR RELATIONSHIP between X and Y (r = .00). It is not appropriate to use the correlation coefficient to summarize this relationship. The low correlation masks the evidence of its pronounced nonlinear shape.

FIGURE 16.5 Scattergrams for Alternative Correlations between X and Y



MORGAN George A. and Griego V. ORLANDO. 1998. *Easy Use and Interpretation of SPSS for Windows: Answering Research Questions With Statistics.* Mahwah: Lawrence Erlbaum Associates.

CHAPTER 9

Correlation and Scatterplots

In this assignment, you will learn about how to compute the basic associational statistics. The Pearson correlation is a parametric statistic used when both variables are at least interval scale. When you have ranked data or when other assumptions (such as normality of the data) are markedly violated, one should use a nonparametric equivalent of the Pearson correlation coefficient (such as Spearman's rho or Kendall's tau). The Kendall's tau is said to deal with ties in a better way than the Spearman rho. Here we ask you to compute all three correlations and compare them.

Chapter 7 is important background because it will help you understand when to compute/choose associational statistics, and it will remind you about what the significance test means and how to interpret it.

Problems/Research Questions

1. What is the association between grades in high school and math achievement? You will compute three bivariate (2 variable) **correlations** (Pearson, Spearman, and Kendall's tau-b) of *grades* and *mathach.*

2. What are the correlations among all of the variables, *mathach, visual, mosaic, mathcrs, pleasure, comptnc,* and *motivatn,* using Pearson correlations.

3. In this problem, you will compare **pairwise** and **listwise** exclusion of missing data.

4. Using the Graphs menu, you will request **Scatterplots** with the linear, quadratic, and cubic regression lines and r^2 printed on the scatterplot for *grades* and *mathach* and for some of the other correlations.

Lab Assignment E

Logon and Get Data

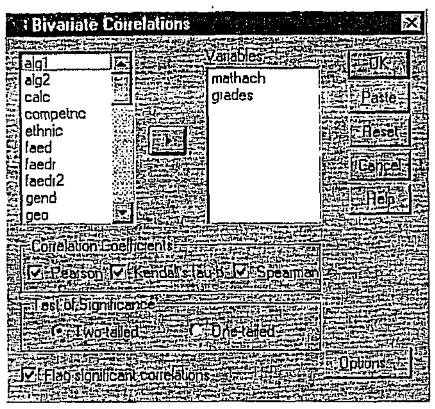
- Retrieve hsbdataD from your Data file.

Problem 1: Correlate Grades and Math Achievement

To do Pearson, Kendall, and Spearman correlations follow these commands:

- **Statistics => Correlate => Bivariate.**
- Move *mathach* and *grades* to the **Variables** box.
- Next, ensure that the **Pearson, Kendall's tau-b, and Spearman** boxes are checked.

- Make sure that the **Two-tailed** (under **Test of Significance**) and **Flag significant correlations** are checked (see Fig. 9.1).
- Now click on **Options** to get that dialog box.
- Click on **Means and standard deviations** and note that **Exclude cases pairwise** is checked. Does your screen look like Fig. 9.2?
- Click on **Continue** then on **OK**. What does your output file look like? Compare Output 9.1 to your output and syntax.



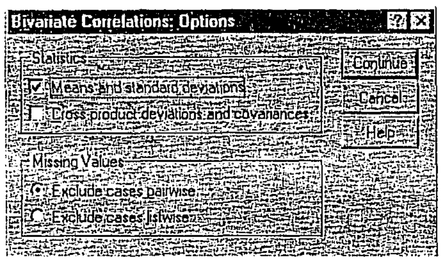
Variables: mathach, grades

Correlation Coefficients: ☑ Pearson ☑ Kendall's tau-b ☑ Spearman

Test of Significance: ● Two-tailed ○ One-tailed

☑ Flag significant correlations

Fig. 9.1. Bivariate correlations.



Statistics: ☑ Means and standard deviations □ Cross-product deviations and covariances

Missing Values: ● Exclude cases pairwise ○ Exclude cases listwise

Fig. 9.2. Bivariate correlations: Options.

Problem 2: Correlation Matrixes for Several Interval Scale Variables

Now, on your own, compute a **Pearson** correlation among all the following variables: *mathach, visual; mosaic, mathcrs, pleasure, competnc,* and *motivatn.* Follow similar procedures outlined previously except:

- Click *off* **Kendall's tau-b** and **Spearman** (under **Correlation Coefficients**).

It is usually best, except in exploratory research with small samples, to use two-tailed tests. The "flag" puts an asterisk beside the correlation coefficients that are statistically significant so that they can be identified quickly. The output also prints the exact significance level (p) which is redundant with the asterisk so you wouldn't report both in a thesis or paper.

- For **Options**, obtain **Means and standard deviations**, and **Exclude cases pairwise**.

This will produce Output 9.2, which was reduced in size to fit on the page. To see if you are doing the work right, compare your own syntax file and output to Output 9.2.

Problem 3: Correlations With Pairwise Exclusions

Next, rerun the same analysis, except:

- Click *off* **Means and standard deviation** (in the **Options** window).
- Change **Exclude cases pairwise** to **Exclude cases listwise** (under **Missing Values**).

Now, compare the correlations in Output 9.3 (listwise exclusion of participants with any missing data) to the Pearson correlations in Output 9.2 (pairwise deletion). Are they the same?

Problem 4: Scatterplots - Mathach With Grades

Let's now work on developing a scatterplot of the correlations of *mathach* with *grades*. Follow these commands:

- **Graphs => Scatter**. This will give you Fig. 9.3.
- Click on **Simple** then **Define** which will bring you to Fig. 9.4.

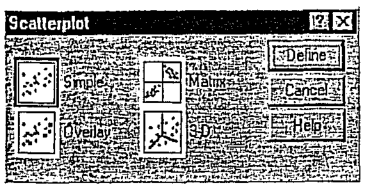
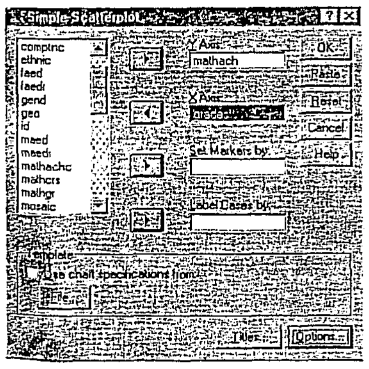


Fig. 9.3. Scatterplot.



Y Axis: mathach

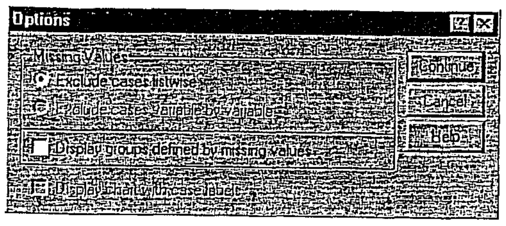
X Axis: grades

Set Markers by:

Label Cases by:

Fig. 9.4. Simple scatterplot.

- Now, move *mathach* to the **Y** axis and *grades* to the **X** axis (*the dependent variable goes on the Y axis*).
- Click on **Options** and make sure **Éxclude cases listwise** is highlighted (see Fig. 9.5).
- Click on **Continue.**
- Next, click on **Titles** (in Fig. 9.4) and type **"Correlation of math achievement with high school grades"** (see Fig. 9.6).
- Click on **Continue** then on **OK**. You will get an output chart which looks like Fig. 9.7.



Missing Values

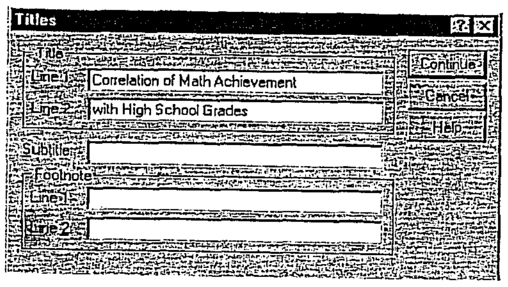
● Exclude cases listwise

○ Exclude cases variable by variable

□ Display groups defined by missing values

□ Display chart with case labels

Fig. 9.5. Options.



Title

Line 1: Correlation of Math Achievement

Line 2: with High School Grades

Subtitle:

Footnote

Line 1:

Line 2:

Fig. 9.6. Titles.

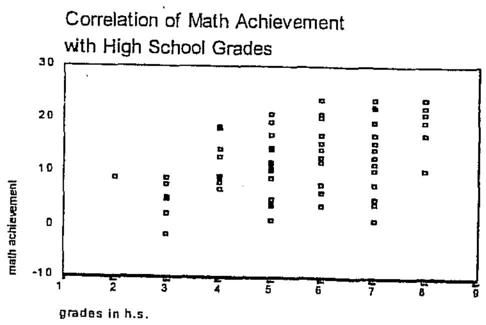


Fig. 9.7. Scatterplot output.

Now let's put the regression lines on the scatterplot so we can get a better sense of the correlation and how much scatter or deviation from the line there is.

- *Double click* on the chart in the output file. You will see a dialog box like Fig. 9.8.
- Select **Chart => Options** until you see Fig. 9.9.
- Click on **Total** in the **Fit Line** box and **Show sunflowers**; there is no need to change the **Sunflower Options**. The sunflowers indicate, by the number of petals, how many participants had essentially the same point on the scatterplot.
- Next, click on the **Fit Options** button, which will give you Fig. 9.10.
- Ensure that the **Linear Regression** box is highlighted.
- Then check the **Individual** box and **Display R-Square in legend** box. Check to be sure your window is like Fig. 9.10.
- Click on **Continue** then **OK.**

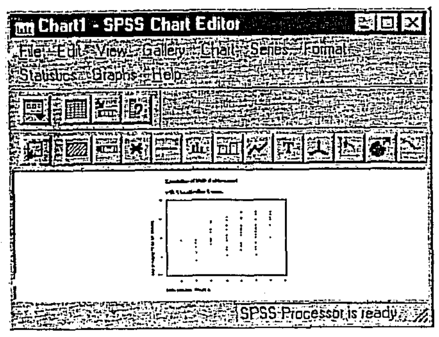
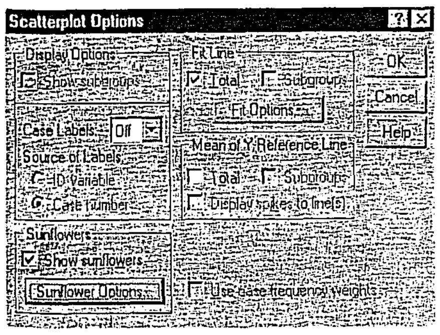


Fig. 9.8. SPSS chart editor.



Display Options

□ Show subgroups

Case Labels: Off

Source of Labels

○ ID Variable

● Case number

Sunflowers

☑ Show sunflowers

Fit Line

☑ Total □ Subgroups

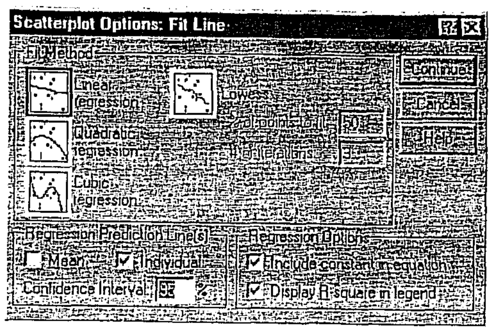
Mean of Y Reference Line

□ Total □ Subgroups

□ Display spikes to line(s)

□ Use case frequency weights

Fig. 9.9. Scatterplot options.



% of points to fit: 50

# of iterations: 3

Regression Prediction Line(s)

□ Mean ☑ Individual

Confidence Interval: 95 %

Regression Options

☑ Include constant in equation

☑ Display R-square in legend

Fig. 9.10. Chart: Scatterplot.

Now, if the points on the scatterplot do not lie close to the regression line, it could be that the data were curvilinear (better fit a curved line). If so, you could (in Fig. 9.10) click on **Quadratic** and possibly the **Cubic regression** boxes (one at a time) to and see what the fit and r^2 look like. If the quadratic and/or cubic r^2 are quite a bit higher, a linear Pearson correlation is not the best statistic to use. Output 9.4 shows the quadratic and cubic regression lines as well as the linear chart. Check your syntax and output against Output 9.4.

Now try the following scatterplots by doing the same steps as Problem 3. Don't forget to *change the title* before you run each scatterplot.

1. *Mosaic* (X) with *mathach* (Y).
2. *Mathcrs* (X) with *mathach* (Y).

Do your syntax and output look like the ones in Output 9.5 and 9.6?

Print, Save, and Exit

- Print your output if you want.
- Save your data file as **hsbdataE** (**File => Save As**).
- Save the SPSS log files as **hsblogE.**
- Exit SPSS.

Interpretation Questions

1. In Output 9.1: a) What do the correlation coefficients tell us? b) What is r^2 for the Pearson correlation? What does it mean? c) Compare the Pearson, Kendall, and Spearman correlations on both correlation size and significance level. d) When should you use which type?

2. In Output 9.2, how many of the Pearson correlation coefficients are significant?

3. In Output 9.3: a) How many Pearson correlations are there? b) How many are significant?

MORGAN George A. and Griego V. ORLANDO, 1998. *Easy Use and Interpretation of SPSS for Windows: Answering Research Questions With Statistics*. Mahwah: Lawrence Erlbaum Associates.

4. Write an interpretation of a) one of the significant and b) one of the nonsignificant correlations in Output 9.3. Include whether or not the correlation is significant, your decision about the null hypothesis, *and* a sentence or two describing the correlations in nontechnical terms.

5. What is the difference between the pairwise and listwise correlation matrixes?

6. Using Outputs 9.5, and 9.6, inspect the scatterplots. a) What is r^2? b) Is the linear relationship as good as a curvilinear (quadratic) one? c) Why should one do scatterplots?

Outputs and Interpretations

```
GET
FILE='A:\hsbdataD.sav'.
EXECUTE .
```

Output 9.1: Pearson, Spearman, and Kendall's Tau-b Correlations

Syntax for Pearson correlation of math achievement with grades in h.s.

```
CORRELATIONS
  /VARIABLES=mathach grades
  /PRINT=TWOTAIL SIG
  /STATISTICS DESCRIPTIVES
  /MISSING=PAIRWISE .
```

Interpretation of Output 9.1

The first table provides **descriptive statistics** for the variables to be correlated, in this case math achievement and grades. The two **correlations** tables are the key. Each has three parts, with the information in matrix form which, unfortunately, means that every number is presented twice. We have provided callout boxes to help you.

The Pearson correlation coefficient is .504; the significance level or *p* is .000 and the number of participants with both variables (*mathach* and *grades*) is 75. In a report, this would usually be written as *r* (73) = .50, *p* < .001. Note that the degrees of freedom (*N*-2 for correlations) is put in parentheses after the statistic (*r* for Pearson correlation) which is usually rounded to two decimal places. The significance or *p* value follows and is stated as less than .001 rather than .000. Note that the correlation values for Kendall's tau-b and Spearman's rho are different from *r*, but in this case, they have the same significance level (*p* < .001).

This correlation is significant, because the "sig" is less than .05, (*p* < .05) so we can reject the null hypothesis of no association and state that there *is an association* between grades and math achievement. Because the correlation is positive, students who have high grades generally have high math achievement scores and vice versa. This means that high grades generally are *associated* with high achievement, medium with medium, and low with low. If the correlation is significant and *negative* (e.g., -.50), high grades would be associated with low achievement and vice versa. If the correlation was not significant, there would be *no* systematic association between a student's grades and achievement. In that case you could not predict anything about math achievement from knowing someone's grades.

Descriptive Statistics

	Mean	Std. Deviation	N
math achievement	12.5645	6.6703	75
grades in h.s.	5.68	1.57	75

Correlations

		math achievement	grades in h.s.
Pearson Correlation	math achievement	1.000	.504
	grades in h.s.	.504	1.000
Sig. (2-tailed)	math achievement	.	.000
	grades in h.s.	.000	.
N	math achievement	75	75
	grades in h.s.	75	75

Note the diagonal is always 1.00

Correlation between math achievement and grades in high school

The statistical level of significance or *p*

Number of participants with both variables

These numbers are redundant. Ignore them.

Syntax for Kendall's Tau-b and Spearman Rho correlations of math achievement with grades in h.s.

```
NONPAR CORR
  /VARIABLES=mathach grades
  /PRINT=BOTH TWOTAIL NOSIG
  /MISSING=PAIRWISE .
```

Nonparametric Correlations

Correlations

			math achievement	grades in h.s.
Kendall's tau-b	Correlation Coefficient	math achievement	1.000	.370**
		grades in h.s.	.370**	1.000
	Sig. (2-tailed)	math achievement	.	.000
		grades in h.s.	.000	.
	N	math achievement	75	75
		grades in h.s.	75	75
Spearman's rho	Correlation Coefficient	math achievement	1.000	.481**
		grades in h.s.	.481**	1.000
	Sig. (2-tailed)	math achievement	.	.000
		grades in h.s.	.000	.
	N	math achievement	75	75
		grades in h.s.	75	75

**. Correlation is significant at the .01 level (2-tailed).

MORGAN George A. and Griego V. ORLANDO, 1998. *Easy Use and Interpretation of SPSS for Windows: Answering Research Questions With Statistics*. Mahwah: Lawrence Erlbaum Associates.

Output 9.2: Pearson Correlation Matrix (Pairwise Exclusion)

Syntax for Pearson correlation matrixes (pairwise exclusion of missing data)

```
CORRELATIONS
  /VARIABLES=mathach visual mosaic mathcrs pleasure competnc motivatn
  /PRINT=TWOTAIL NOSIG
  /STATISTICS DESCRIPTIVES
  /MISSING=PAIRWISE .
```

Interpretation of Output 9.2

Notice that after the descriptive statistics table, there is a large **correlations** table divided into three sections: Pearson correlation coefficients, significance, and *N*. These numbers are, as in Output 9.1, each given twice so you have to be careful in reading them. It is a good idea to look only at the numbers below the diagonal (1.00 as in the coefficients section, dots in the significance section, and 75s in the *N* section). There are 21 different correlations in the table. In the first column, there is the correlation of each of the other six variables with math achievement. In the second column, each of the other six variables is correlated with visualization score, but note that the .423 for *visual* and *mathach* is the same as the correlation of *mathach* and *visual* in the first column, so ignore it. The Pearson correlations on this table are interpreted similarly to the one in Output 9.1. However, because there are 21 correlations, the odds are that at least one could be statistically significant by chance (i.e., .05= 1/20). Thus, it would be prudent to use the .01 level of significance. The Bonferroni correction (.05/21= .002) would be a conservative approach designed to keep the significance level at .05 for the whole study.

Descriptive Statistics

	Mean	Std. Deviation	N
math achievement	12.5645	6.6703	75
visualization score	5.2433	3.9120	75
mosaic, pattern test	27.413	9.574	75
Math course taken	2.11	1.67	75
Pleasure scale	3.2267	.6300	75
Competence scale	3.2945	.6645	73
Motivation scale	2.8744	.6382	73

Correlations with math achievement.

Correlations with visualization score.

Use correlations above or below line.

Correlations

		math achievement	visualization score	mosaic, pattern test	Math course taken	Pleasure scale	Competence scale	Motivation scale
Pearson Correlation	math achievement	1.000	.423**	.213	.794**	.094	.332**	.316**
	visualization score	.423**	1.000	.030	.399**	-.160	.007	.047
	mosaic, pattern test	.213	.030	1.000	-.059	.085	.111	.083
	Math course taken	.794**	.399**	-.059	1.000	-.006	.309**	.298*
	Pleasure scale	.094	-.160	.085	-.006	1.000	.431**	.305**
	Competence scale	.332**	.007	.111	.309**	.431**	1.000	.570**
	Motivation scale	.316**	.047	.083	.298*	.305**	.570**	1.000
Sig. (2-tailed)	math achievement	.	.000	.067	.000	.421	.004	.006
	visualization score	.000	.	.798	.000	.171	.954	.695
	mosaic, pattern test	.067	.798	.	.616	.466	.349	.487
	Math course taken	.000	.000	.616	.	.958	.008	.010
	Pleasure scale	.421	.171	.466	.958	.	.000	.009
	Competence scale	.004	.954	.349	.008	.000	.	.000
	Motivation scale	.006	.695	.487	.010	.009	.000	.
N	math achievement	75	75	75	75	75	73	73
	visualization score	75	75	75	75	75	73	73
	mosaic, pattern test	75	75	75	75	75	73	73
	Math course taken	75	75	75	75	75	73	73
	Pleasure scale	75	75	75	75	75	73	73
	Competence scale	73	73	73	73	73	73	71
	Motivation scale	73	73	73	73	73	71	73

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Output 9.3: Pearson Correlation Matrix (Listwise Exclusion)

Syntax for Pearson correlation matrix

```
CORRELATIONS
  /VARIABLES=mathach visual mosaic mathcrs pleasure competnc motivatn
  /PRINT=TWOTAIL NOSIG
  /MISSING=LISTWISE .
```

Interpretation of Output 9.3

In this table there is not a separate section for *N* because, with **listwise** exclusion, only the same 71 subjects who have scores on all seven variables are used for all correlations. Note that the correlations are slightly different from those in Output 9.2 where the *N*'s varied depending on how many subjects had each pair of variables. Factor analysis, Cronbach's alpha, and multiple regression (Assignments F, G, and H) all use listwise deletion, so if you have one or more variables with quite a bit of missing data the *N* may be dramatically reduced.

MORGAN George A. and Griego V. ORLANDO. 1998. *Easy Use and Interpretation of SPSS for Windows: Answering Research Questions With Statistics*. Mahwah: Lawrence Erlbaum Associates.

Correlations[a]

		math achievement	visualization score	mosaic, pattern test	Math course taken	Pleasure scale	Competence scale	Motivation scale
Pearson Correlation	math achievement	1.000	.434**	.246*	.804**	.101	.335**	.316**
	visualization score	.434**	1.000	.035	.429**	-.191	.010	.047
	mosaic, pattern test	.246*	.035	1.000	.000	.067	.104	.083
	Math course taken	.804**	.429**	.000	1.000	.005	.318**	.301*
	Pleasure scale	.101	-.191	.067	.005	1.000	.443**	.309**
	Competence scale	.335**	.010	.104	.318**	.443**	1.000	.570**
	Motivation scale	.316**	.047	.083	.301*	.309**	.570**	1.000
Sig. (2-tailed)	math achievement	.	.000	.038	.000	.400	.004	.007
	visualization score	.000	.	.773	.000	.111	.931	.695
	mosaic, pattern test	.038	.773	.	.999	.576	.386	.489
	Math course taken	.000	.000	.999	.	.964	.007	.011
	Pleasure scale	.400	.111	.576	.964	.	.000	.009
	Competence scale	.004	.931	.386	.007	.000	.	.000
	Motivation scale	.007	.695	.489	.011	.009	.000	.

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

a. Listwise *N*=71

Cases with missing values for any variable are excluded from all correlations using listwise.

Output 9.4: Scatterplots, Grades With Math Achievement

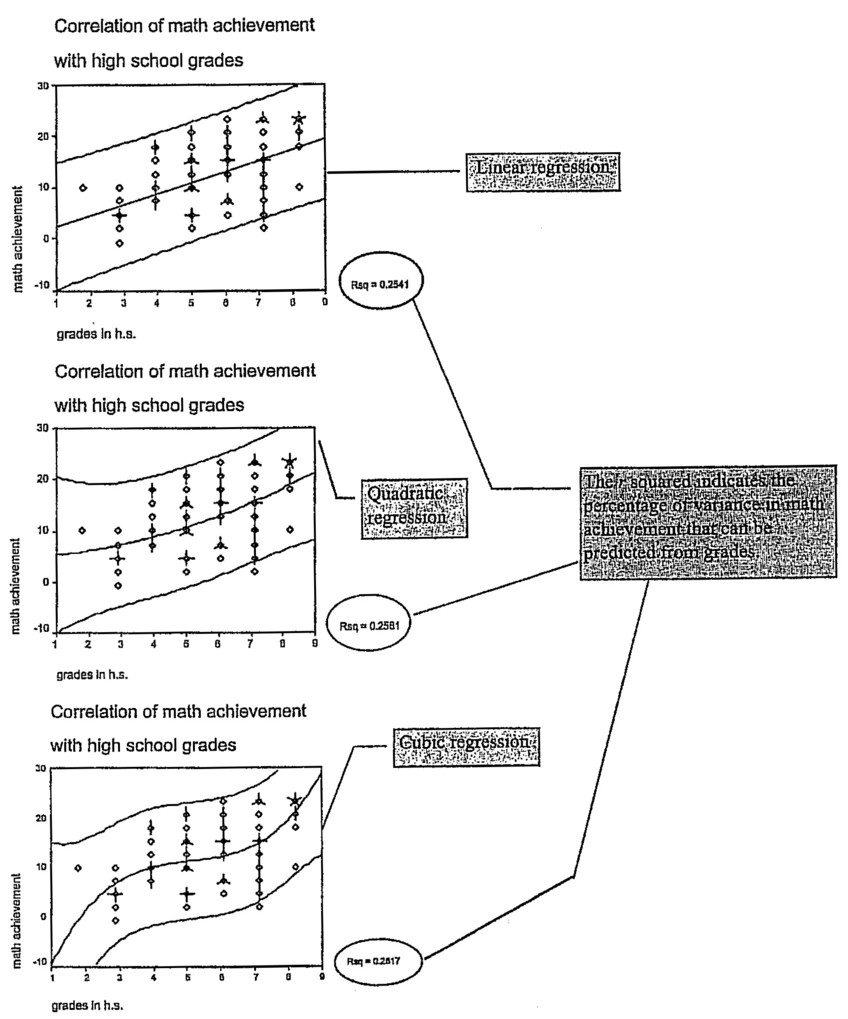
Syntax for Scatterplot of grades with math achievement with linear, quadratic (curved), and cubic (2 bend) regression

```
GRAPH
  /SCATTERPLOT(BIVAR)=grades WITH mathach
  /MISSING=LISTWISE
  /TITLE= 'Correlation of math achievement' 'with high school grades'.
```

Interpretation of Output 9.4

The three scatterplots shown in Output 9.4 are the same except that the first shows the best fit for a straight (linear) regression line (i.e., it minimizes the squared differences between the points and the line), and the second is the best fit for a line with one curve (quadratic) and the third is the best fit for a line with two bends or curves (cubic). To the right of each scatterplot is the r^2 which indicates the percentage of variance in the dependent variable (*mathach*) that can be predicted from the independent variable (*grades*). Note that the r^2 (.2541) for the linear (first) scatterplot means that $r = .504$, which is what Output 9.1 had for the same correlation.

Note also that the r^2 for the quadratic (.2581) and the r^2 for the cubic regression (.2817) are quite similar to that (.2541) for the linear regression line. Thus, there is little improvement in fit over the linear regression, which is what is used in computing the Pearson and Kendall correlations.



The *r* squared indicates the percentage of variance in math achievement that can be predicted from grades.

10. lekce
JAK ODHALIT VLIV TŘETÍ PROMĚNNÉ: PRÁCE S PODSOUBORY NEBOLI TŘÍDĚNÍ VYŠŠÍCH STUPŇŮ A PARCIÁLNÍ KOEFICIENTY.

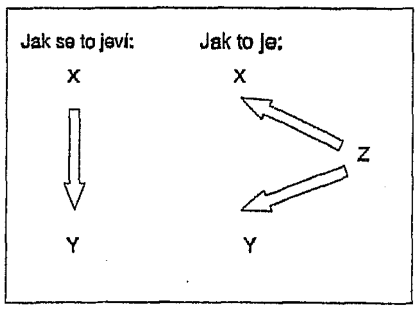
DISMAN Miroslav. 1993. *Jak se vyrábí sociologická znalost.* Praha: Karolinum.

Nepravá korelace

V řadě evropských regionů bylo zjištěno, že čím více čápů žije v určité krajině, tím vyšší je tam porodnost. Korelační koeficienty byly tak významné, že je velice nepravděpodobné, že zjištěná souvislost je náhodná. Jsme tedy ochotni přijmout hypotézu, že čápi přece jen nosí děti? Asi sotva. Ale pak je naší povinností navrhnout hypotézu, která by uspokojivě vysvětlovala naměřenou souvislost.

Graf 1.2.

Nepravá korelace



Toto je klasický příklad nepravé korelace ("spurious correlation"). Zkreslení vzniká tehdy, když třetí nepozorovaná nebo neanalyzovaná proměnná ovlivňuje nějak obě proměnné X a Y, které studujeme.

Cvičení 1.3.

Podívejte se pečlivě na graf 1.2., popisující nepravou korelaci. Navrhněte, co může být to tajemné Z.

Jistě nám nehrozí nebezpečí, že bychom přijali hypotézu, že čápi nosí děti. Ale představme si, že nepravá korelace se zdá potvrzovat naši oblíbenou hypotézu. Potom výzkumník musí mít objektivnost anděla a trpělivost nerostného krystalu, aby pracně zabil to, co se po měsíce pokoušel dokázat.

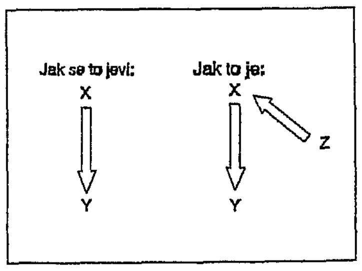
Nepravá korelace je skutečným nebezpečím ve výzkumu. Není to ani tak technický problém analýzy, ale spíše problém lidské kvality výzkumníka.

Vývojová sekvence

Tak nazýváme zkreslení, způsobené faktem, že proměnná X, která ovlivňuje Y, je určována předcházející, ale nepozorovanou proměnnou Z.

Graf 1.3.

Vývojová sekvence



Taková situace je skutečně naprosto nevyhnutelná. Každá příčina má totiž jinou příčinu, ta zase jinou příčinu, která má opět svoji příčinu, a tak bychom mohli pokračovat až k aktu stvoření, nebo k tomu, co astronomové nazývají Big Bang. To je problém velmi dobře známý filozofům, kteří ho obvykle nazývají "regresus ad infinitivum"

Nicméně, někdy může předčasné přerušení kauzálního řetězce vést k mylné interpretaci. Některé studie tvrdí, že četba pornografické literatury vyvolává násilné chování mužů k ženám. Nelze však vyloučit, že je zde nějaký předcházející činitel, jako kupř. autoritativní metoda socializace respondenta v dětství, který vyvolal silný zájem jedince o pornografii. Takový omyl je závažný zejména tehdy, když cílem výzkumu je sociální intervence.

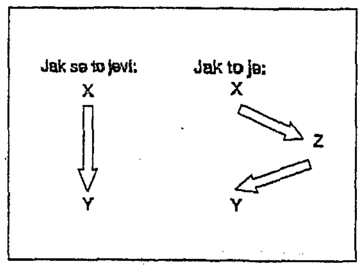
DISMAN Miroslav. 1993. *Jak se vyrábí sociologická znalost.* Praha: Karolinum.

Chybějící střední člen

Tak je označována situace, kde mezi nezávisle proměnnou X a závislou Y je ještě proměnná Z, kterou jsme nezahrnuli do analýzy. Graf 1.3. tuto situaci jasně popisuje. Je to opět konfigurace, která je téměř všudypřítomná. Kdybychom se jakousi sociologickou lupou podívali, co se děje mezi nějakou příčinou a jejím následkem, existuje ještě řada mezikroků. Často můžeme tyto elementy ignorovat bez rizika zkreslení. Ne však vždycky.

Graf 1.4.

Chybějící střední člen



Řekněme, že X reprezentuje pohlaví respondentů a Y jejich skóre v testu inteligence. Je možné, že výsledky žen, a to zejména žen příslušejících k nižším sociálním třídám, by byly signifikantně nižší, než výsledky mužů.

Cvičení 1.4.

Zamyslete se, prosím, nad předchozím odstavcem a navrhněte alternativní hypotézu, ukazující mužským šovinistům, že takové výsledky nepotvrzují superioritu nás, pánů tvorstva.

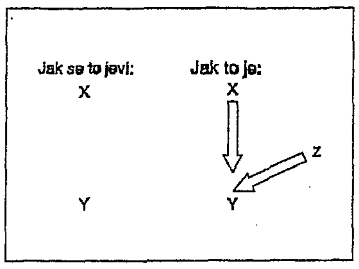
Zkreslení tohoto typu může být nebezpečné. Můžeme je často najít v quasivědeckých pracích, podporujících rasismus, v některých politických pamfletech atd. Mnohé noviny se dopouštějí tohoto hříchu z nevědomosti, když publikují výsledky statistických šetření.

Dvojí příčina

Takto můžeme označit situaci, kdy závislá proměnná Y má dvě příčiny, ale jenom jedna z nich, X, byla zahrnuta do výzkumu. Toto je asi nejčastější problém výzkumu v sociálních vědách. Pravděpodobně neexistuje žádný sociální jev, který by měl jedinou příčinu. I v našem nesmírně zmenšeném vesmíru, složeném jenom ze tří proměnných, si můžeme představit, jaké zkreslení může vyvolat, není-li tato další příčina zahrnuta do analýzy.

Graf 1.5.

Dvojí příčina



Cvičení 1.5.

Představme si třeba, že X je vzdělání jedince a že Y je jeho příjem. Pokud přepokládáme, že se vzděláním příjem poroste, mohli bychom zjistit, že souvislost je velice nízká, nebo dokonce nulová. Co mohlo vyvolat toto zkreslení?

Teoreticky by bylo možné namítnout, že v některých situacích neměřená proměnná Z může posilovat vliv příčiny X. Ale to je krajně nepravděpodobné. Jak vidíme z grafu 1.4., mezi X a Y není žádný příčinný vztah. I naše cvičení 1.5. je platné jenom uvnitř našeho nerealisticky miniaturizovaného systému tří proměnných. Je jasné, že realisticky by bylo třeba zahrnout další proměnnou "věk", která ovlivňuje vzdělání, a prostřednictvím "zkušenosti", "seniority" i příjem.

9.1. Proč tabulka nemusí být placatá

Někdy není těžké zapochybovat, že s určitou zjištěnou souvislostí není všechno v pořádku. Zejména je snadné pozastavit se nad některými nálezy, publikovanými v denním tisku. *Toronto Star* jsou dobré a seriózní noviny, jedny z nejlepších v Kanadě a možná i v celé Americe. Je proto vysoce pravděpodobné, že zpráva, ze které uvádíme výňatek, není plodem novinářovy fantazie, ale je založena na poněkud svérázné interpretaci skutečně existujícího výzkumu.

> TEMPE, Ariz. (AFP)
>
> Ve studii sponzorované US vládou se uvádí, že osoby, které rády jedí hamburgy, milují své rodiny, svoji práci a náboženství... Labužníci, kteří dávají přednost ústřicím a kaviáru, mají obecně ateistické a liberální postoje..
>
> *Toronto Star*, 9.listopadu 1981.

Nesnažili jsme se obdržet původní data, ale ta mohla mít třeba takovou distribuci, jako v tabulce 9.1.

Tabulka 9.1.

	Preferované jídlo:		Celkem:
Zbožnost:	HAMBURGY	KAVIÁR	
vysoká	78% 780	25% 125	905
nízká	22% 220	75% 375	595
Celkem: N	100% 1000	100% 500	1500

LAMBDA = .420

Tahle data ukazují, že mezi oběma proměnnými existuje povážlivá souvislost. Kdybychom tuto souvislost interpretovali jako kauzální, znamenalo by to v sociologii úplnou revoluci. Sotva by někdo z nás by byla připraven obhajovat teorii o biochemických determinantách postojů nebo experimentálně testovat možnost, jak změny v dietě změní jednotlivcovu morálku. Všichni si včas vzpomeneme na klasický příklad s čápy a porodností, na koncept nepravé souvislosti. Teď jde jen o to navrhnout, co je ten třetí faktor, který vyvolal souběžné změny v obou našich proměnných, a hlavně **dokázat**, že souvislost naměřená v naší tabulce je v nějaké podstatnější míře vyvolána tímto faktorem.

Nalézt něco, co ovlivňuje právě tak vzorce preferencí ve stravě, jako postoje k náboženství, rodině atd., nebude tak těžké navrhnout. Bude to pravděpodobně životní styl. Operacionalizovat životní styl by bylo obtížné. Tak si zjednodušíme situaci tím, že proměnnou "životní styl" necháme zastupovat proměnnou "vzdělání". Osoby s vyšším vzděláním - alespoň v severoamerické společnosti - jsou spíše ochotny přiznat, že to s jejich náboženstvím není tak žhavé. Výše vzdělané osoby mají většinou také vyšší plat, takže si mohou dovolit - alespoň občas - kaviár, nebo ústřice. A aby naše diskuse byle opravdu přehledná, předstírejme, že proměnná "vzdělání" má jen dvě kategorie.

Teď zbývá už jen jedno: dokázat, že souvislost pozorovaná v tabulce 9.1. je nepravá, že je vyvolána vlivem vzdělání na obojí, na "zbožnost" a na "preferovanou stravu". Jednoduchá technika, kterou můžeme použít, se nazývá **kontrola dalším faktorem** ("control for test factor"). Tady je návod:

Kontrola pro další faktor:

(1) Náš vzorek rozdělíme do tolika podsouborů, kolik kategorií má proměnná, jejíž vliv kontrolujeme. Všichni jedinci v každém podsouboru budou mít v této proměnné stejnou hodnotu. (V našem případě získáme jeden podsoubor, ve kterém budou všichni jedinci mít jenom "nízké" vzdělání a v druhém podsouboru budou jenom jedinci s "vysokým" vzděláním.)

(2) Pak zkonstruujeme pro každý podsoubor tabulku, která má stejnou formu jako původní tabulka popisující souvislost, o jejíž platnosti pochybujeme. (V našem případě budeme mít dvě tabulky, formou shodné s tabulkou 9.1. V jedné budou jen osoby s nízkým vzděláním, v druhé jen s vysokým.)

(3) Porovnáme intenzitu souvislosti v původní tabulce se souvislostí zjištěnou v nových tabulkách. Je-li souvislost v původní tabulce funkcí třetího faktoru, v nových tabulkách souvislost mezi původními daty zmizí, nebo je alespoň podstatně oslabena.

A teď to můžeme vyzkoušet. Tabulka 9.2. obsahuje data o osobách s "nízkým" vzděláním, tabulka 9.3. data o osobách s vysokým vzděláním:

Tabulka 9.2.

Osoby s nízkým vzděláním

	Preferované jídlo:		Celkem:
Zbožnost:	HAMBURGY	KAVIÁR	
vysoká	78% 624	78% 156	780
nízká	22% 176	22% 44	220
Celkem: N	100% 800	100% 200	1000

LAMBDA = 0

Tabulka 9.3.

Osoby s vysokým vzděláním

	Preferované jídlo:		Celkem:
Zbožnost:	HAMBURGY	KAVIÁR	
vysoká	25% 50	25% 75	125
nízká	75% 150	75% 325	475
Celkem: N	100% 200	100% 400	600

LAMBDA = 0

Výsledky jsou docela jasné: souvislost mezi původními dvěma proměnnými úplně zmizela. Lambda v obou nových tabulkách klesla na nulu. Vidíme to i bez jakéhokoliv měření síly souvislosti: procento zbožných je v obou zcela shodné pro ty, kteří dávají přednost kaviáru, jako pro ty, kteří mají raději hamburgy. Můžeme tedy uzavřít, že původní souvislost mezi zbožností a jídlem byla jen a jen funkcí třetí proměnné, funkcí vzdělání.

Jistě už víte, že to tak pěkně vyšlo jen proto, že jsme připravili data tak, aby bylo všechno jasné a průhledné. Ve skutečnosti by nám kontrola dalších faktorů v nových tabulkách nedala nulovou souvislost. Životní styl může být nadto ovlivněn dalšími faktory, třeba tím, zda respondent žije ve velkém městě nebo na venkově. Ale i takové situace může kontrolu dalšího faktoru zvládnout. Tady je návod, jak kontrolovat souvislost mezi stravou a zbožností pro dva další faktory (vzdělání a místo bydliště) současně:

<u>Kontrola dalších dvou faktorů</u>

(1) Nejdříve rozdělíme náš vzorek na dvě skupiny. V jedné budou jenom respondenti žijící ve městě, ve druhé jenom ti, kteří žijí na venkově.

(2) Pak každou skupinu rozdělíme do dvou podsouborů podle vzdělání respondentů. Výsledkem budou čtyři skupiny dat.

(3) Pro každou skupinu zkonstruujeme tabulku shodnou v její formě s původní tabulkou 9.1. Budeme mít tedy čtyři tabulky. V jedné budou data o jedincích, žijících ve městě a majících vysoké vzdělání. Ve druhé budou data o těch, kdo žijí ve městě, ale mají nízké vzdělání. Ve třetí tabulce budou data o respondentech žijících na venkově a majících vysoké vzdělání. v poslední tabulce pak budou respondenti, kteří žijí na venkově a mají nízké vzdělání.

(4) Poslední krok bude stejný jako předtím: porovnáme souvislosti v nových tabulkách se souvislostí zjištěnou v původní tabulce.

A nyní by nám už měla být jasná **logika testování dalšího faktoru:**

Vytvořením nových tabulek je <u>testovaný faktor udržován na konstantní hodnotě.</u> Tím je souvislost mezi původními proměnnými očištěna od zkreslujícího vlivu této další proměnné.

Teoreticky nemáme důvod, proč omezit tuto kontrolu na jednu nebo dvě další proměnné. Popsaná logika může být aplikována i na vyšší počet testovaných faktorů.

Dr.Watson:

Výborně! Mě vždycky zajímalo, jak vzdělání ovlivňuje rozhodnutí, pro koho budou lidé hlasovat ve volbách. a teď to mohu zjistit mnohem jasněji. Budu kontrolovat typ vzdělání, povolání, příjem a velikost obce. a také to, jak respondent hlasoval v minulých volbách a ovšem pohlaví a věk.

Teoreticky má dr. Watson pravdu. Prakticky je v tom háček. Podívejme se, co všechno by náš přítel musel pro navrženou kontrolu udělat. Řekněme, že by proměnná "volební preference" měla jenom 6 kategorii, a proměnná "vzdělání" jenom pět; původní tabulka by měla tedy 30 polí. Abychom mohli kontrolovat pro "typ vzdělání" museli bychom původní tabulku opakovat pro každou, řekněme ze 4 kategorií této proměnné. Máme teď 4 tabulky se 120 poli. Pak musíme tuto sérii 4 tabulek opakovat pro každou z kategorií proměnné "povolání". Ale teď se to už stává trochu nepřehledné. Shrneme si to tedy do tabelární formy:

Proměnná:	**Počet kategorií:**	**Počet polí:**
preferovaná strana	6	6
vzdělání	5	30
typ vzdělání	4	120
povolání	5	600
příjem	5	3.000
velikost obce	4	12.000
strana volená v minulých volbách	6	72.000
věk	3	216.000
pohlaví	2	432.000

Tak tohle by opravdu nešlo. Co bychom si počali s 7.200 tabulek? Takové množství tabulek by vůbec nebylo možné interpretovat. Ale hlavně, pro takovéhle cvičení **nemáme dost lidí!** I kdyby dr. Watson měl hodně štědrého sponzora a mohl si dovolit vzorek s dvěma tisíci jedinci, více než **99% polí v jeho tabulkách by bylo prázdných.**

Zkusme tedy podstatně skromnější přístup. Budeme kontrolovat jen ty proměnné, které jsou snad nejdůležitější: povolání a pohlaví. Teď bychom skončili s 300 poli v deseti tabulkách. I zde by vzorek 2.000 jedinců sotva stačil. Teoreticky by na každé pole v tabulkách připadlo o něco méně než 7 pozorování. To by nemuselo být dost. **Prázdná pole v tabulce, jakož i pole s velice nízkým počtem pozorování, mohou podstatně zkreslit význam koeficientů, měřících souvislost.**

Počet faktorů, které si můžeme dovolit kontrolovat nezávisí ovšem jen na počtu proměnných, ale i na počtu kategorií každé z nich. Tak kdybychom kontrolovali povolání a pohlaví naši tabulku 9.1. a dostali bychom ve výsledných deseti tabulkách jen 80 polí a při stejné velikosti vzorku by na každé pole připadalo v průměru 25 pozorování. To už je podstatně lepší, ale kdo si kdy může dovolit dvoutisícový vzorek?

Někdy nám však i kontrola jediného dalšího faktoru může podstatně prospět. Například jsme zapojeni do výzkumu trhu a studujeme, zda balení typu A je atraktivnější, než balení B. První výsledky jsou v tabulce 9.4.:

Tabulka 9.4.

	Balení A	Balení B	Celkem:
asi by koupil	40% 80	40% 160	240
asi ne	60% 120	60% 240	360
Celkem:	100% 200	100% 400	600

LAMBDA = 0

Data zřejmě ukazují, že typy balení nemá vliv na úmysl zakoupit výrobek. Přesně stejné procento respondentů vyjádřilo úmysl koupit, ať již byl výrobek uveden v balení a nebo B. Ale je tomu opravdu tak? Podívejme se, co se stane, když budeme kontrolovat pohlaví respondentů.

Tabulka 9.5. shrnuje údaje pro muže. Na prvý pohled se zdá, že se nic nezměnilo: proměnná "typ balení" a proměnná "úmysl koupit" jsou vzájemně naprosto nezávislé:

Tabulka 9.5.

Muži:

	Balení A	Balení B	Celkem:
asi by koupil	40% 40	40% 40	80
asi ne	60% 60	60% 60	120
Celkem:	100% 100	100% 100	100

LAMBDA = 0

Ale tabulka 9.6. nám podává docela jiný obraz:

Tabulka 9.6.

Ženy:

	Balení A	Balení B	Celkem:
asi by koupil	100% 100	20% 60	160
asi ne	0% 0	80% 240	240
Celkem:	100% 100	100% 300	400

LAMBDA = .625

DISMAN Miroslav. 1993. *Jak se vyrábí sociologická znalost*. Praha: Karolinum.

Všechny ženy ve vzorku vyjádřily úmysl zakoupit výrobek v balení A a jen pětina z nich v balení B. Pro marketing to je jistě velice užitečná informace, která byla zcela neviditelná v původní tabulce 9.4.

Pro nás je tenhle poznatek také pěkně důležitý: ukazuje nám, jak statistická analýza více proměnných **současně** může odhalit nejen nepravou souvislost, produkovanou nějakým faktorem, ale může i odkrýt nepravou nezávislost mezi proměnnými. Příčinou tohoto typu zkreslení může být fakt, že souvislost existuje pouze v určité části vzorku, v našem specifickém případě jen mezi ženami. Zde může být kontrola dalších faktorů velice účinným nástrojem.

Nicméně, musíme-li pracovat najednou s mnoha proměnnými, kontrola dalšího faktoru brzy ztrácí dech. To jsme si už demonstrovali. Musíme se tedy porozhlédnout po jiných postupech, které by umožnily dr. Watsonovi realizovat jeho volební projekt.

9.2. Výprava do čtyřrozměrného prostoru

> Z poselství Vogona Jetze, člena Plánovacího výboru pro galaktický nadprostor:
> "Bohužel, vaše planeta je jednou z těch, které byly určeny pro demolici. Tento proces započne za necelé dvě vaše pozemské minuty. **Nepodléhejte panice!** Děkuji vám.
>
> Douglas Adams, *The Hitch-Hikers Guide to the Gallaxy*, 1979, p.30

On to bude víc než čtyřrozměrný prostor, ale nepanikařte. i když operace, o kterých budeme hovořit, mají vznešená a lehce hrozivá jména, jako vícerozměrná regresní analýza, "path analysis", faktorová analýza atd. Jejich **logika a jejich interpretace** není složitá. Složitá je jen logika jejich výpočtu a logika zdůvodnění těchto výpočtů. Ale tím se zde nebudeme zabývat. s tím se setkáte, až budete studovat skutečnou statistiku.

V podstatě značnou část toho, o čem tu budeme mluvit, už znáte. Zde to jen trošku rozšíříme. Např. už dovedeme pomocí regrese odhadnout jednotlivcův příjem, když známe jeho vzdělání. Ale příjem nezávisí jenom na vzdělání, ale i na délce odborné praxe, povolání atd. Zkusme, zda je možné aplikovat postup, který už známe, i na více proměnných.

1.
Pro každého jedince jsou pozorované hodnoty násobeny koeficienty odpovídajícími koeficienty prvé diskriminantní funkce. (To jsou ta čísla v prvém sloupci tabulky 9.9.) Součet těchto násobků reprezentuje jednotlivcovu pozici na prvé diskriminační funkci.

2.
Stejnou operaci opakujeme s koeficienty druhé funkce a obdržíme jeho pozici na této funkci.

3.
Teď máme pro jedince dvě souřadnice a můžeme ho zakreslit do mapy, odpovídající mapě v grafu 9.7.

4.
Pak už zbývá jen jedno, rozhodnout ke kterému z centroidů je jednotlivcova pozice nejblíže a zařadit ho do té skupiny. (Ve skutečnosti jsou i obě poslední operace prováděny matematicky.)

Tabulka 9.11. shrnuje výsledky takové klasifikace:

Tabulka 9.11.
Výsledky klasifikace

Skutečné členství	Odhadnuté členství			
	Amerika	Střední Evropa	Itálie	N
Amerika	66%	14%	20%	616
Střední Evropa	13%	69%	18%	1146
Itálie	26%	21%	53%	512

Z celkového počtu 2273 jedinců bylo správně klasifikováno 65%. Uvážíme-li charakter použitých prediktorů, je to opravdu docela pozoruhodný výsledek. Sociálně psychologické rozdíly podmíněné rozdílnou kulturou a rozdílným politicko-ekonomickým systémem jsou daleko silnější, než jsme očekávali. Ale tím se opět dostáváme do obsahové oblasti, a ta, bohužel, nepatří do naší knížky.

Tak nezbývá než poznamenat, že diskriminační analýza je nejen zajímavá, ale i velice užitečná hračka a přejít k poslední ze statistických technik, které zde budeme probírat.

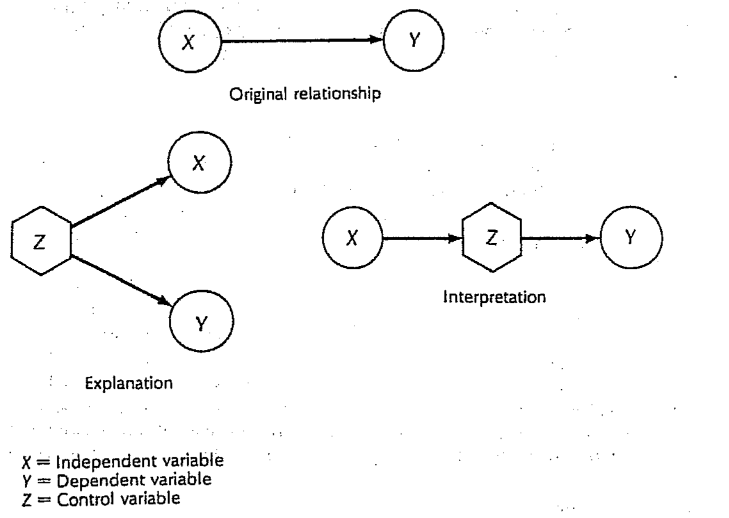
WILLIAMSONN, J. B., David A. KARP, John R. DALPHIN, Paul S. GRAY. 1982. *The Research Craft. An Introduction to Social Research Methods*. Boston: Little, Brown and Company.

FIGURE 16.8 The Bivariate Relationship between Abortion Attitude and Size of Birth Place

		Size of Birthplace	
		Town	City
"Should it be possible for a woman to obtain an abortion on demand?"	No	82% (410)	18% (90)
	Yes	18% (90)	82% (410)
	Total	100% (500)	100% (500)

INTERPRETATION. Figure 16.8, a cross-tabulation of the relationship between *attitudes toward abortion* and *size of one's birthplace*, shows persons from cities much more likely (82 percent) to endorse the right of women to obtain an abortion than are persons from towns (18 percent). Suppose, as in the previous example, we try to explain away the relationship but fail to discover any control variable that meets both requirements (i.e., associated with and causally prior to both original variables). When explanation fails to reduce such a nonobvious relationship between two variables, there still exists the possibility that we can uncover a third factor to help clarify the chain of circumstances that connects the two variables to one another.

FIGURE 16.9 Models Illustrating the Distinction Between Explanation and Interpretation



INTERPRETATION, the second part of the elaboration paradigm, is the search for a control variable (Z) that *causally intervenes* between the independent variable (X) and the dependent variable (Y). Figure 16.9 diagrams the differences between explanation and interpretation as they modify the original relationship between the independent and dependent variables. An INTERVENING VARIABLE must be related to both the independent and the dependent variable, and it must be plausible to think of it as somehow a result of the independent variable that, in turn, affects the dependent variable. Figure 16.10 illustrates the effects of an intervening variable.

Searching for an intervening variable that might explain the relationship between abortion attitude and size of birthplace (Figure 16.8), one might hypothesize that towns and cities promote very different kinds of political and social ideologies, which in turn might account for the city/town differences in abortion attitudes. In effect, people born in towns are more likely to be conservative than are people born in cities, and conservatives are more likely than liberals to oppose abortion. Note in Figure 16.10 that there are no longer any differences in abortion attitudes between town people and city people in either subtable; all town/city differences have been accounted for by subdividing the sample into *conservatives* and *liberals*. Hence we have successfully interpreted the relationship by locating an intervening variable.

Compare Figure 16.7 with Figure 16.10. Notice that the results have the same statistical form; that is, the introduction of a control variable makes the original relationship disappear. Hence the difference between explanation and interpretation rests in the underlying logic, not in the statistics. We now turn to a third form of elaboration, referred to as SPECIFICATION, in which the objective is not to make the original relationship disappear, but rather to specify the conditions under which the strength of the original relationship varies in intensity.

SPECIFICATION. Figure 16.11 reexamines the relationship between size of birthplace and attitudes toward abortion while controlling for a third variable, the *region of the country in which a person was born*. Here the original relationship changes (compare with Figure 16.8) but does not disappear; instead, it takes on a different form from one subtable to the next. The original relationship disappears for persons born in the South, where town and city people show identical attitudes

FIGURE 16.10 The Relationship Between Abortion Attitude and Size of Birthplace Controlling for Political Ideology: an Example of Interpretation

		Political Ideology			
		Conservative		Liberal	
		Size of Birthplace		Size of Birthplace	
		Town	City	Town	City
"Should it be possible for a woman to obtain an abortion on demand?"	No	90% (405)	90% (45)	10% (5)	10% (45)
	Yes	10% (45)	10% (5)	90% (45)	90% (405)
	Total	100% (450)	100% (50)	100% (50)	100% (450)

WILLIAMSON, J. B., David A. KARP, John R. DALPHIN, Paul S. GRAY. 1982. *The Research Craft. An Introduction to Social Research Methods*. Boston: Little, Brown and Company.

toward abortion; it remains strong in the West, where town people are more likely than city people to oppose abortion (86 percent versus 21 percent); and it intensifies in the North, where differences between town and city people regarding abortion attitudes are most pronounced (89 percent versus 0 percent oppose abortion). Introducing a control variable has enabled us to analyze the relationship between size of birthplace and attitude toward abortion more precisely, pinpointing the circumstances under which the association holds. This is an example of specification.

It is entirely possible that the use of a control variable for specification of a relationship, as in Figure 16.11, may produce fundamentally different relationships in different subtables. It is conceivable that town persons might favor abortion more than city persons in one region, and yet the opposite might be true in another area. When this occurs, there is good reason to suspect that other, undiscovered factors are affecting the relationship. A specification that results in such markedly different subtables is an invitation to pursue the analysis further, as the following case illustrates.

Suppressor Variables. Suppose we have a table in which no relationship appears, even though we had good reason to expect to find an association. In Figure 16.11 the data for the West and the North indicate a strong association between size of birthplace and abortion attitude; yet the association disappears in data for the South. Why? It is possible that some hidden third factor is *suppressing* the true relationship between the two original variables. Such a factor is referred to as a suppressor variable, because it hides the actual relationship until it is controlled.

Figure 16.12 reanalyzes this data for the South, controlling for another variable, *percentage of persons in the community who are black*. Whereas the original data showed no relationship between size of birthplace and abortion attitude, these two subtables each show strong (but opposite) associations. Subtable 2 shows data that are consistent with the overall findings presented in Figure 16.11, while subtable 1 isolates the deviant cases. When the two subtables are combined, as they were in Figure 16.11, the relationship is no longer discernible.

FIGURE 16.11 The Relationship Between Abortion Attitudes and Size of Birthplace, Controlling for Region of Birthplace: Example of Specification

Region of Birthplace

		South		West		North	
		Size of Birthplace		*Size of Birthplace*		*Size of Birthplace*	
		Town	City	Town	City	Town	City
"Should it be possible for a woman to obtain an abortion on demand?"	No	50% (40)	50% (40)	86% (160)	21% (50)	89% (210)	0% (0)
	Yes	50% (40)	50% (40)	14% (25)	79% (190)	11% (25)	100% (180)
	Total	100% (80)	100% (80)	100% (185)	100% (240)	100% (235)	100% (180)
		(Subtable 1)		(Subtable 2)		(Subtable 3)	

FIGURE 16.12 A Three-Way Table Illustrating the Effect of Introducing a Suppressor Variable

Percent Black in Community of Birth for Respondents Born in South

		High		Low	
		Size of Birthplace		*Size of Birthplace*	
		Town	City	Town	City
"Should it be possible for a woman to obtain an abortion on demand?"	No	100% (40)	0% (0)	0% (0)	100% (40)
	Yes	0% (0)	100% (40)	100% (40)	0% (0)
	Total	100% (40)	100% (40)	100% (40)	100% (40)
		(Subtable 1)		(Subtable 2)	

The data we have presented in this discussion of various methods of elaboration (Figures 16.6 to 16.8 and 16.10 to 16.12) are hypothetical and exaggerated to illustrate points of analysis. In actual research, relationships are seldom so strong, nor are distinctions between types of elaboration so clear. However, the logic that underlies these idealized examples embodies the range of possibilities for analysis that you will encounter in real research, and a thorough knowledge of these classifications will serve as a useful guide.

For the sake of simplicity we have developed elaborations around *dichotomies*—variables with only two values. The same logic applies to more complex variables, but when tables get larger, the elaborations soon become unwieldy. Indeed, it is often desirable to control for the effects of more than one variable, but we find ourselves confronted with the same practical difficulty. Just as correlation analysis was introduced to solve the analogous problem for two-variable tables with many cells, a technique called partial correlation exists to aid in the analysis of more complex elaborations.

NORUŠIS, Marija J. 1998. *Guide to Data Analysis*. Upper Saddle River, Prentice Hall.

Adding Control Variables

So far, you've considered the relationship between income and job satisfaction for the entire sample. It's possible that if you consider additional variables, the relationship you've seen between the two variables may change. For example, it may be that the relationship between income and job satisfaction is different for men and women. To test this, you can make separate tables of income and job satisfaction for men and for women. Gender is then called a control variable, since its effect is removed, or "controlled" for, in each of the separate tables. Figure 7.8 shows separate crosstabulation tables for men and women.

To obtain these separate—or layered—crosstabulations, select sex as a layer variable, as shown in Figure 7.10.

Figure 7.8 Job satisfaction by income for men and women

Respondent's Sex			Total Family Income in quartiles				Total
			24,999 or less	25,000 to 39,999	40,000 to 59,999	60,000 or more	
Male	Very satisfied	Count	30	51	41	57	179
		Column %	34.9%	47.2%	45.6%	46.0%	43.9%
	Mod satisfied	Count	44	44	36	49	173
		Column %	51.2%	40.7%	40.0%	39.5%	42.4%
	A little dissatisfied	Count	10	10	7	14	41
		Column %	11.6%	9.3%	7.8%	11.3%	10.0%
	Very dissatisfied	Count	2	3	6	4	15
		Column %	2.3%	2.8%	6.7%	3.2%	3.7%
	Total	Count	86	108	90	124	408
		Column %	100.0%	100.0%	100.0%	100.0%	100.0%
Female	Very satisfied	Count	23	39	33	53	148
		Column %	26.1%	45.3%	50.0%	53.5%	43.7%
	Mod satisfied	Count	49	35	25	38	147
		Column %	55.7%	40.7%	37.9%	38.4%	43.4%
	A little dissatisfied	Count	14	7	7	5	33
		Column %	15.9%	8.1%	10.6%	5.1%	9.7%
	Very dissatisfied	Count	2	5	1	3	11
		Column %	2.3%	5.8%	1.5%	3.0%	3.2%
	Total	Count	88	86	66	99	339
		Column %	100.0%	100.0%	100.0%	100.0%	100.0%

Each cell contains counts and column percentages. An interesting difference emerges between the two tables. For women, job satisfaction seems to increase with income. Almost twice as many women in the highest income category (53.5%) are *very satisfied* compared to women in the lowest income category (26.1%). For men, the difference is not as striking. Almost 35% of men in the lowest income category are *very satisfied*, compared to 46% of the men in the highest income category.

Case Selection

Sometimes you want to analyze only part of your cases. For example, some of the analyses described in this book look only at full-time workers or only at college graduates.

The Select Cases dialog box allows you to restrict your analysis to a specific group of cases. There are a number of options for selecting cases:

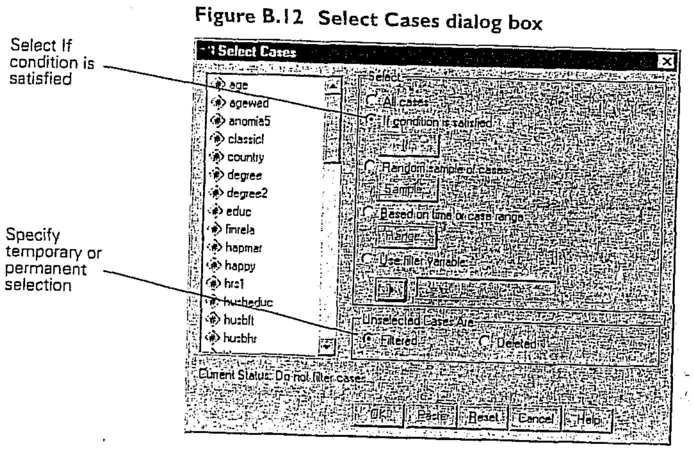
- You can choose cases according to a logical condition based on their data values.
- You can select a random sample of the cases in your file.
- You can select a range of cases according to their order in the file.
- You can select those cases that are marked with a non-zero value for a "filter variable."

▶ From the menus choose:

Data
 Select Cases...

This opens the Select Cases dialog box, as shown in Figure B.12.

Figure B.12 Select Cases dialog box



Temporary or Permanent Selection

The Select Cases dialog box offers a choice between filtering cases (selecting temporarily) or deleting cases (selecting permanently). The distinction between temporary and permanent case selection is important to understand.

- When you filter cases, or select a temporary subset, the unselected cases remain in the working data file. You can regain all the original cases at any time.
- When you delete cases, or select a permanent subset, SPSS deletes them forever from your working data file. If you save the working data file, replacing the copy on your disk, the deleted cases are gone forever from that file, too. This can be useful because it allows you to save a smaller data file.

If you haven't saved the working data file, you can often "undo" a permanent case selection by reopening the original data file. If you have saved the working data file there is no way to get back cases that have been deleted, unless you have a backup copy of the data file.

Example: Selecting Full-Time Employees

Many of the analyses in this book that use the GSS data restrict the analysis to full-time workers. In the *gss.sav* file, respondents who are employed full time are coded 1 for the variable *wrkstat*. To select full-time employees:

▶ From the Data Editor menus choose:

Data
 Select Cases...

▶ Select If condition is satisfied alternative in the Select Cases dialog box (see Figure B.12).

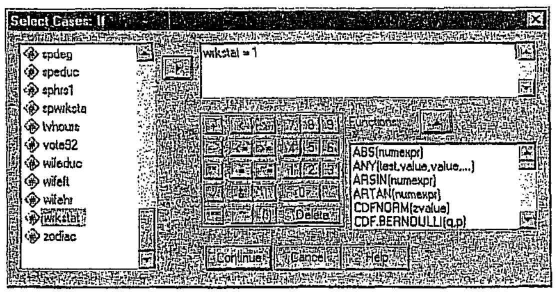
▶ Select Filtered in the Unselected Cases Are group.

This assures that the unselected cases will remain in the working data file if you want to use them in future analyses.

▶ Click If.

This opens the Select Cases If dialog box, as shown in Figure B.13. This dialog box, which strongly resembles the Compute Variable If Cases dialog box shown in Figure B.9, allows you to specify a conditional expression.

Figure B.13 Select Cases If dialog box



▶ Enter the expression wrkstat = 1.

▶ Click Continue to return to the Select Cases dialog box.

▶ Click OK.

Cases for people who work full time are now selected. In the Data Editor, unselected cases are indicated by a slash mark over the row number.

▶ To turn off case selection, open the Select Cases dialog box again, select All cases, and click OK.

Example: Selecting College Graduates

In the *gss.sav* and *gssft.sav* data files, the variable *degree* indicates the highest degree earned by each respondent. Four-year college graduates are coded 3 (for bachelor's degree) or 4 (for advanced degree). To select people with bachelor's or advanced degrees:

▶ Open the Select Cases dialog box as described above.

▶ Select Filtered in the Unselected Cases Are group.

▶ Select If condition is satisfied and click If.

▶ Enter degree >= 3 in the Numeric Expression box.

This expression specifies that cases should be selected "if degree is greater than or equal to 3."

▶ Click Continue to return to the Select Cases dialog box and click OK.

Other Selection Methods

Other options available in the Select Cases dialog box include:

Random sample. Sometimes you want a random subset of cases. You have no particular criterion for choosing which cases to process, but you don't want the whole data file.

Based on time or case range. Under some circumstances, it is desirable to select a range of cases according to the order of cases, as displayed in the Data Editor. This can be useful for time series data files.

Use filter variable. A filter variable is simply a variable that indicates whether or not a particular case should be selected. Cases for which the specified filter variable has a valid non-zero value are retained. Cases for which it is 0 or missing are dropped.

Bryman Alan and Duncan Cramer. 1997. *Quantitative Data Analysis with SPSS for Windows*. London, Routledge.

Testing for intervening variables

The quest for intervening variables is different from the search for potentially spurious relationships. An intervening variable is one that is both a product of the independent variable and a cause of the dependent variable. Taking the data examined in Table 10.1, the sequence depicted in Figure 10.2 might be imagined. The analysis presented in Table 10.4 strongly suggests that the level of people's interest in their work is an intervening variable. As with Tables 10.2 and 10.3, we partition the sample into two groups (this time those who report that they are interested and those reporting no interest in their work) and examine the relationship between work variety and job satisfaction for each group. Again, we can compare d_1 in Table 10.1 with d_1 and d_2 in Table 10.4. In Table 10.1 d_1 is 56 per cent, but in Table 10.4 d_1 and d_2 are 13 per cent and 20 per cent respectively. Clearly, d_1 and d_2 in Table 10.3 have not been reduced to zero (which would suggest that the whole of the relationship was through interest in work), but they are also much lower than the 56 per cent difference in Table 10.1. If d_1 and d_2 in Table 10.4 had remained at or around 56 per cent, we would conclude that interest in work is not an intervening variable.

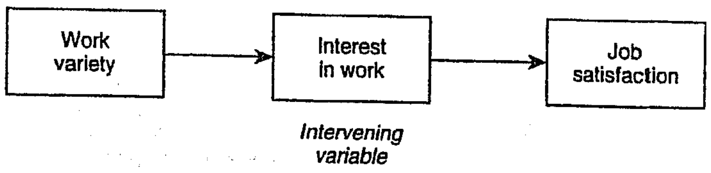


Figure 10.2 Is the relationship between work variety and job satisfaction affected by an intervening variable?

The sequence in Figure 10.2 suggests that variety in work affects the degree of interest in work that people experience, which in turn affects their level of job satisfaction. This pattern differs from that depicted in Figure 10.1 in that if the analysis supported the hypothesized sequence, it suggests that there is a relationship between amount of variety in work and job satisfaction, but the relationship is not direct. The search for intervening variables is often referred to as *explanation* and it is easy to see why. If we find that a test variable acts as an intervening variable, we are able to gain some explanatory leverage on the bivariate relationship. Thus, we find that there is a relationship between amount of variety in work and job satisfaction and then ask why that relationship might exist. We speculate that it may be because those who have varied work become more interested in their work, which heightens their job satisfaction.

It should be apparent that the computation of a test for an intervening variable is identical to a test for spuriousness. How, then, do we know which is which? If we carry out an analysis like those shown in Tables 10.2, 10.3 and 10.4, how can we be sure that what we are taking to be an intervening variable is not in fact an indication that the relationship is spurious? The answer is that there should be only one logical possibility, that is, only one that makes sense. If we take the trio of variables in Figure 10.1, to argue that the test variable – size of firm – could be an intervening variable would mean that we would have to suggest that a person's level of work variety affects the size of the firm in which he or she works – an unlikely scenario. Similarly, to argue that the trio in Figure 10.2 could point to a test for spuriousness, would mean that we would have to accept that the test variable – interest in work – can affect the amount of variety in a person's work. This too makes much less sense than to perceive it as an intervening variable.

One further point should be registered. It is clear that controlling for interest in work in Table 10.4 has not totally eliminated the difference between those reporting varied work and those whose work is not varied, in terms of job satisfaction. It would seem, therefore, that there are aspects of the relationship between amount of variety in work and job satisfaction that are not totally explained by the test variable, interest in work.

Testing for moderated relationships

A moderated relationship occurs when a relationship is found to hold for some categories of a sample but not others. Diagrammatically this can be displayed as in Figure 10.3. We may even find the character of a relationship can differ for categories of the test variable. We might find that for one category those who report varied work exhibit greater job satisfaction, but for another category of people the reverse may be true (i.e. varied work seems to engender lower levels of job satisfaction than work that is not varied).

Bryman Alan and Duncan Cramer. 1997. *Quantitative Data Analysis with SPSS for Windows*. London, Routledge.

Table 10.3 A non-spurious relationship: the relationship between work variety and job satisfaction, controlling for size of firm (imaginary data)

		Large firms		Small firms	
		Varied work	*Not varied work*	*Varied work*	*Not varied work*
Job satisfaction	*Satisfied*	1 (166) 83% d_1 =	2 (14) 28% 55%	3 (34) 68% d_2 =	4 (46) 23% 45%
	Not satisfied	5 (34) 17% d_3 =	6 (36) 72% 55%	7 (16) 32% d_4 =	8 (154) 77% 45%

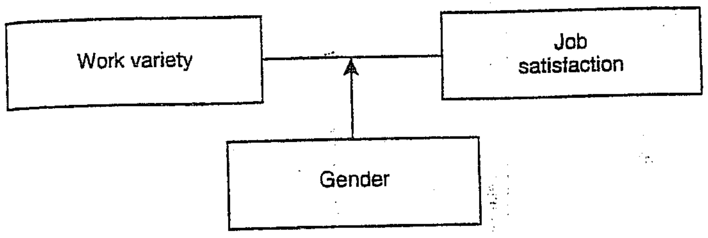


Figure 10.3 Is the relationship between work variety and job satisfaction moderated by gender?

Table 10.5 looks at the relationship between variety in work and job satisfaction for men and women. Once again, we can compare d_1 (56 per cent) in Table 10.1 with d_1 and d_2 in Table 10.5, which are 85 per cent and 12 per cent respectively. The bulk of the 56 percentage point difference between those reporting varied work and those reporting that work is not varied in Table 10.1 appears to derive from the relationship between variety in work and job satisfaction being far stronger for men than women and there being more men (300) than women (200) in the sample. Table 10.5 demonstrates the importance of searching for moderated relationships in that they allow the researcher to avoid inferring that a set of findings pertains to a sample as a whole, when in fact it only really applies to a portion of that sample. The term *interaction effect* is often employed to refer to the situation in which a relationship between two variables differs substantially for categories of the test variable. This kind of occurrence was also addressed in Chapter 9. The discovery of such an effect often inaugurates a new line of inquiry in that it stimulates reflection about the likely reasons for such variations.

The discovery of moderated relationships can occur by design or by chance. When they occur by design, the researcher has usually anticipated the possibility that a relationship may be moderated (though he or she may be wrong of course). They can occur by chance when the researcher conducts a test for an intervening variable or a test for spuriousness and finds a marked contrast in findings for different categories of the test variable.

Multiple causation

Dependent variables in the social sciences are rarely determined by one variable alone, so that two or more potential independent variables can usefully be considered in conjunction. Figure 10.4 suggests that whether someone is allowed participation in decision-making at work also affects their level of job satisfaction. It is misleading to refer to participation in decision-making as a test variable in this context, since it is really a second independent variable. What, then, is the impact of amount of variety in work on job satisfaction when we control the effects of participation?

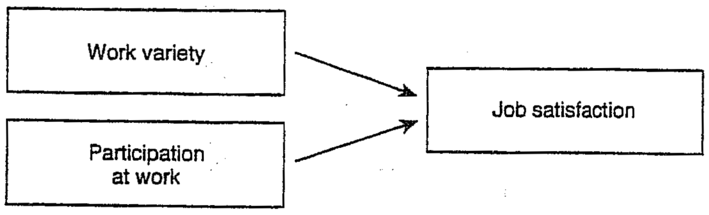


Figure 10.4 Work variety and participation at work

Again, we compare d_1 in Table 10.1 (56 per cent) with d_1 and d_2 in Table 10.6. The latter are 19 and 18 per cent respectively. This suggests that although the effect of amount of variety in work has not been reduced to zero or nearly zero, its impact has been reduced considerably. Participation in decision-making appears to be a more important cause of variation in job satisfaction. For example, compare the percentages in cells 1 and 3 in Table 10.6: among those respondents who report that they perform varied work, 93 per cent of those who experience participation exhibit job satisfaction, whereas only 30 per cent of those who do not experience participation are satisfied.

One reason for this pattern of findings is that most people who experience participation in decision-making also have varied jobs, that is (cell1 + cell5) − (cell2 + cell6). Likewise, most people who do not experience participation have work which is not varied, that is (cell4 + cell8) − (cell3 + cell7). Could this mean that the relationship between variety in work and job satisfaction is really spurious, when participation in decision-making is employed as the test variable? The answer is that this is unlikely, since it would mean that participation in decision-making would have to cause variation in the amount of variety in work, which is a less likely possibility (since technological conditions tend to be the major influence on variables like work variety). Once again, we have to resort to a combination of intuitive logic and theoretical reflection in order to discount such a possibility. We will return to this kind of issue in the context of an examination of the use of multivariate analysis through correlation and regression.

Table 10.4 An intervening variable: the relationship between work variety and job satisfaction, controlling for interest in work (imaginary data)

		Interested		Not interested	
		Varied work	*Not varied work*	*Varied work*	*Not varied work*
Job satisfaction	*Satisfied*	1 (185) 93%	2 (40) 80%	3 (15) 30%	4 (20) 10%
		d_1 = 13%		d_2 = 20%	
	Not satisfied	5 (15) 7%	6 (10) 20%	7 (35) 70%	8 (180) 90%
		d_3 = 13%		d_4 = 20%	

Table 10.5 A moderated relationship: the relationship between work variety and job satisfaction, controlling for gender (imaginary data)

		Men		Women	
		Varied work	*Not varied work*	*Varied work*	*Not varied work*
Job satisfaction	*Satisfied*	1 (143) 95%	2 (15) 10%	3 (57) 57%	4 (45) 45%
		d_1 = 85%		d_2 = 12%	
	Not satisfied	5 (7) 5%	6 (135) 90%	7 (43) 43%	8 (55) 55%
		d_3 = 85%		d_4 = 12%	

Table 10.6 Two independent variables: the relationship between work variety and job satisfaction, controlling for participation at work (imaginary data)

		Participation		Little or no participation	
		Varied work	*Not varied work*	*Varied work*	*Not varied work*
Job satisfaction	*Satisfied*	1 (185) 93%	2 (37) 74%	3 (15) 30%	4 (23) 12%
		d_1 = 19%		d_2 = 18%	
	Not satisfied	5 (15) 7%	6 (13) 26%	7 (35) 70%	8 (177) 88%
		d_3 = 19%		d_4 = 18%	

3.2.4. Partial Correlation

3.2.4.1. The Theory behind Part and Partial Correlation

I mentioned earlier that there is a type of correlation that can be done that allows you to look at the relationship between two variables when the effects of a third variable are held constant. For example, analyses of the exam anxiety data (in the file **ExamAnx.sav**) showed that exam performance was negatively related to exam anxiety, but positively related to revision time, and revision time itself was negatively related to exam anxiety. This scenario is complex, but given that we know that revision time is related to both exam anxiety and exam performance, then if we want a pure measure of the relationship between exam anxiety and exam performance we need to take account of the influence of revision time. Using the values of R^2 for these relationships, we know that exam anxiety accounts for 19.4% of the variance in exam performance, that revision time accounts for 15.7% of the variance in exam performance, and that revision time accounts for 50.2% of the variance in exam anxiety. If revision time accounts for half of the variance in exam anxiety, then it seems feasible that at least some of the 19.4% of variance in exam performance that is accounted for by anxiety is the same variance that is accounted for by revision time. As such, some of the variance in exam performance explained by exam anxiety is not *unique* and can be accounted for by revision time. A correlation between two variables in which the effects of other variables are held constant is known as a partial correlation.

Figure 3.13 illustrates the principle behind partial correlation. In part 1 of the diagram there is a box labelled exam performance that represents the total variation in exam scores (this value would be the variance of exam performance). There is also a box that represents the variation in exam anxiety (again, this is the variance of that variable). We know already that exam anxiety and exam performance share 19.4% of their variation (this value is the correlation coefficient squared). Therefore, the variations of these two variables overlap (because they share variance) creating a third box (the one with diagonal lines). The overlap of the boxes representing exam performance and exam anxiety is the common variance. Likewise, in part 2 of the diagram the shared variation between exam performance and revision time is illustrated. Revision time shares 15.7% of the variation in exam scores. This shared variation is represented by the area of overlap (filled with diagonal lines). We know that revision time and exam anxiety also share 50% of their variation: therefore, it is very probable that some of the variation in exam performance shared by exam anxiety is the same as the variance shared by revision time.

Part 3 of the diagram shows the complete picture. The first thing to note is that the boxes representing exam anxiety and revision time have a large overlap (this is because they share 50% of their variation). More important, when we look at how revision time and anxiety contribute to exam performance we see that there is a portion of exam performance that is shared by both anxiety and revision time (the dotted area). However, there are still small chunks of the variance in exam performance that are unique to the other two variables. So, although in part 1 exam anxiety shared a large chunk of variation in exam performance, some of this overlap is also shared by revision time. If we remove the portion of variation that is also shared by revision time, we get a measure of the unique relationship between exam performance and exam anxiety. We use partial correlations to find out the size of the unique portion of variance. Therefore, we could conduct a partial correlation between exam anxiety and exam performance while 'controlling' the effect of revision time. Likewise, we could carry out a partial correlation between revision time and exam performance 'controlling' for the effects of exam anxiety.

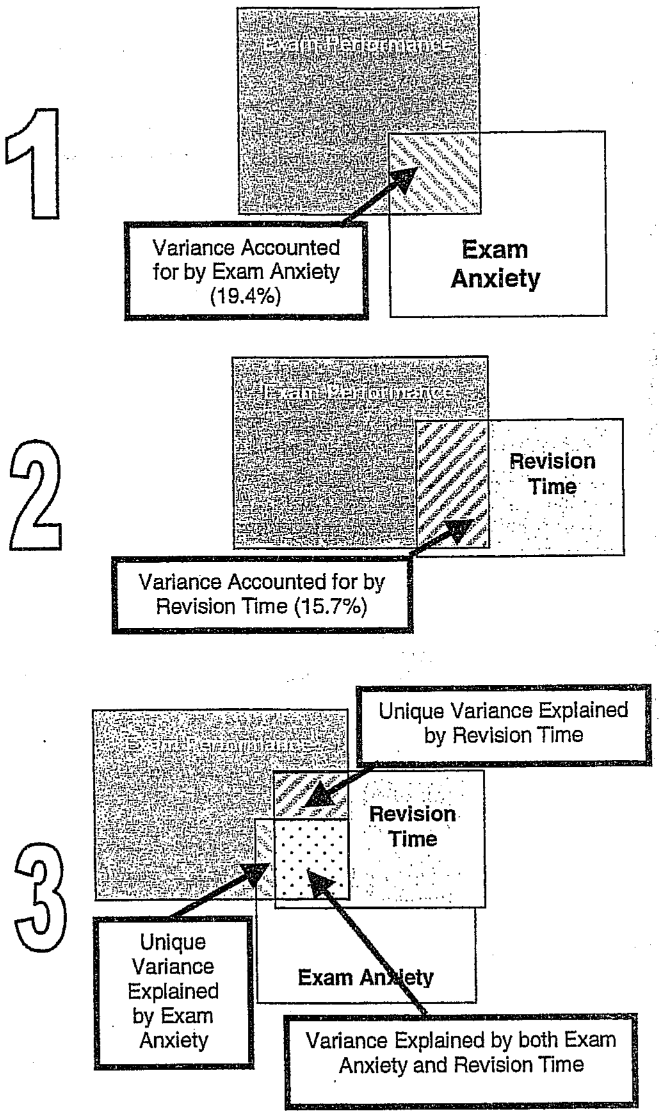
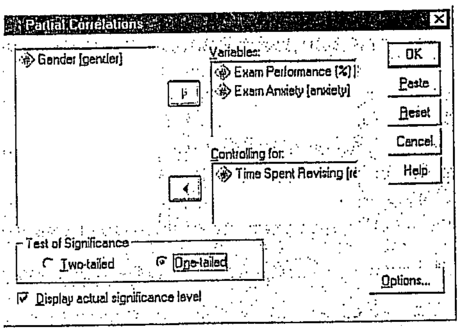

1
Exam Performance
Variance Accounted for by Exam Anxiety (19.4%)
Exam Anxiety
2
Exam Performance
Revision Time
Variance Accounted for by Revision Time (15.7%)
3
Exam Performance
Unique Variance Explained by Revision Time
Revision Time
Unique Variance Explained by Exam Anxiety
Exam Anxiety
Variance Explained by both Exam Anxiety and Revision Time

Figure 3.13: Diagram showing the principle of partial correlation

3.2.4.2. Partial Correlation Using SPSS

To conduct a partial correlation on the exam performance data select the *Correlate* option from the *Analyze* menu and then select *Partial* (**Analyze⇒Correlate⇒Partial**) and the dialog box in Figure 3.14 will be activated. This dialog box lists all of the variables in the data editor on the left-hand side and there are two empty spaces on the right-hand side. The first space is for listing the variables that you want to correlate and the second is for declaring any variables the effects of which you want to control. In the example I have described, we want to look at the unique effect of exam anxiety on exam performance and so we want to correlate the variables **exam** and **anxiety**, while controlling for **revise**. Figure 3.14 shows the completed dialog box. If you click on Options then another dialog box appears as shown in Figure 3.15.


Partial Correlations

Gender [gender]

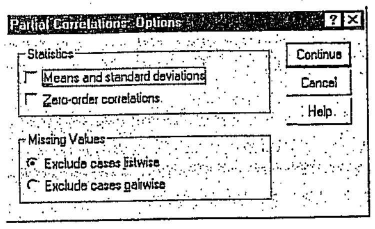
Variables:
- Exam Performance (%)
- Exam Anxiety [anxiety]

Controlling for:
- Time Spent Revising [re

Test of Significance
○ Two-tailed ● One-tailed

☑ Display actual significance level

Figure 3.14: Main dialog box for conducting a partial correlation


Partial Correlations: Options

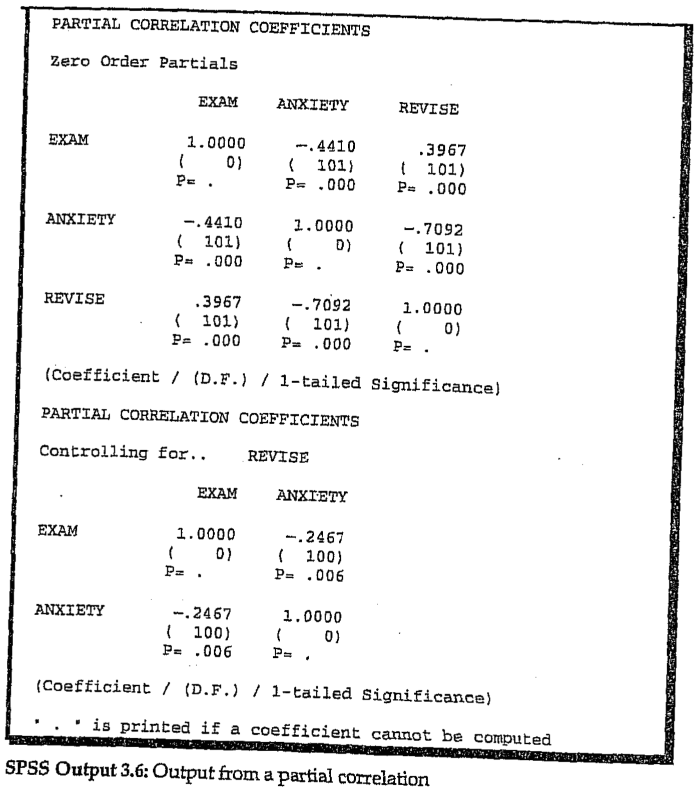
Statistics
□ Means and standard deviations
□ Zero-order correlations

Missing Values
● Exclude cases listwise
○ Exclude cases pairwise

Figure 3.15: Options for partial correlation

These further options are similar to those in bivariate correlation except that you can choose to display zero-order correlations. Zero-order correlations are the bivariate correlation coefficients without controlling for any other variables. So, in our example, if we select the tick-box for zero-order correlations SPSS will produce a correlation matrix of

anxiety, **exam** and **revise**. If you haven't conducted bivariate correlations before the partial correlation then this is a useful way to compare the correlations that haven't been controlled against those that have. This comparison gives you some insight into the contribution of different variables. Tick the box for zero-order correlations but leave the rest of the options as they are.



```
PARTIAL CORRELATION COEFFICIENTS

Zero Order Partials

                EXAM      ANXIETY     REVISE

EXAM           1.0000     -.4410       .3967
              (     0)   (  101)     (  101)
              P= .       P= .000     P= .000

ANXIETY        -.4410     1.0000     -.7092
              (  101)    (     0)    (  101)
              P= .000    P= .        P= .000

REVISE          .3967     -.7092     1.0000
              (  101)    (  101)     (     0)
              P= .000    P= .000     P= .

(Coefficient / (D.F.) / 1-tailed Significance)

PARTIAL CORRELATION COEFFICIENTS

Controlling for..    REVISE

                EXAM      ANXIETY

EXAM           1.0000     -.2467
              (     0)   (  100)
              P= .       P= .006

ANXIETY        -.2467     1.0000
              (  100)    (     0)
              P= .006    P= .

(Coefficient / (D.F.) / 1-tailed Significance)

" . " is printed if a coefficient cannot be computed
```

SPSS Output 3.6: Output from a partial correlation

SPSS Output 3.6 shows the output for the partial correlation of exam anxiety and exam performance controlling for revision time. The first thing to notice is the matrix of zero-order correlations, which we asked for using the *options* dialog box. The correlations displayed here are identical to those obtained from the Pearson correlation procedure (compare this matrix with the one in SPSS Output 3.2). Underneath the zero-order correlations is a matrix of correlations for the variables

anxiety and **exam** but controlling for the effect of revision. In this instance we have controlled for one variable and so this is known as a first-order partial correlation. It is possible to control for the effects of two variables at the same time (a second-order partial correlation) or control three variables (a third-order partial correlation) and so on. First, notice that the partial correlation between exam performance and exam anxiety is −0.2467, which is considerably less than the correlation when the effect of revision time is not controlled for ($r = -0.4410$). In fact, the correlation coefficient is nearly half what it was before. Although this correlation is still statistically significant (its p value is still below 0.05), the relationship is diminished. In terms of variance, the value of R^2 for the partial correlation is 0.06, which means that exam anxiety can now account for only 6% of the variance in exam performance. When the effects of revision time were not controlled for, exam anxiety shared 19.4% of the variation in exam scores and so the inclusion of revision time has severely diminished the amount of variation in exam scores shared by anxiety. As such, a truer measure of the role of exam anxiety has been obtained. Running this analysis has shown us that exam anxiety alone does explain much of the variation in exam scores, and we have discovered a complex relationship between anxiety and revision that might otherwise have been ignored. Although causality is still not certain, because relevant variables are being included, the third variable problem is, at least, being addressed in some form.

3.2.4.3. Semi-Partial (or Part) Correlations

In the next chapter, we come across another form of correlation known as a semi-partial correlation (also referred to as a part correlation). While I'm babbling on about partial correlations it is worth me explaining the difference between this type of correlation and a semi-partial correlation. When we do a partial correlation between two variables, we control for the effects of a third variable. Specifically, the effect that the third variable has on *both* variables in the correlation is controlled. In a semi-partial correlation we control for the effect that the third variable has on only one of the variables in the correlation. Figure 3.16 illustrates this principle for the exam performance data. The partial correlation that we calculated took account not only of the effect of revision on exam performance, but also of the effect of revision on anxiety. If we were to calculate the semi-partial correlation for the same data, then this would control for only the effect of revision on exam performance (the effect of revision on exam anxiety is ignored). Partial correlations are most useful for looking at the unique relationship between two variables when other variables are ruled out. Semi-partial correlations are, therefore, useful when trying to explain the variance in one particular variable (an outcome) from a set of predictor variables. This idea leads us nicely toward Chapter 4 ...

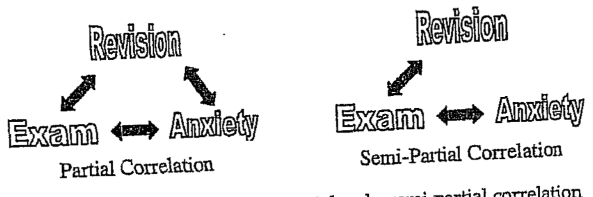


Figure 3.16: The difference between a partial and a semi-partial correlation

11. lekce

ZÁKLADY LINEÁRNÍ REGRESE - VZTAH SPOJITÝCH PROMĚNNÝCH

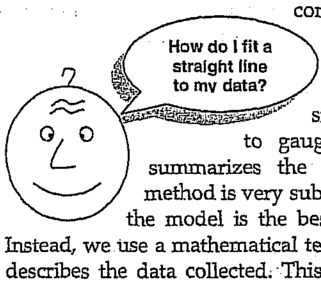
FIELD Andy. 2000. *Discovering Statistics Using SPSS for Windows. Advanced Techniques for the Beginner*. London: Sage.

4 Regression

4.1. An Introduction to Regression

Correlations can be a very useful research tool but they tell us nothing about the predictive power of variables. In regression analysis we fit a predictive model to our data and use that model to predict values of the dependent variable (DV) from one or more independent variables (IVs).[1] Simple regression seeks to predict an outcome from a single predictor whereas multiple regression seeks to predict an outcome from several predictors. This is an incredibly useful tool because it allows us to go a step beyond the data that we actually possess. The model that we fit to our data is a linear one and can be imagined by trying to summarize a data set with a straight line (think back to Figure 1.5).

With any data set there are a number of lines that could be used to summarize the general trend and so we need a way to decide which of many possible lines to chose. For the sake of drawing accurate conclusions we want to fit a model that *best* describes the data. There are several ways to fit a straight line to the data you have collected. The simplest way would be to use your eye to gauge a line that looks as though it summarizes the data well. However, the 'eyeball' method is very subjective and so offers no assurance that the model is the best one that could have been chosen. Instead, we use a mathematical technique to establish the line that best describes the data collected. This method is called the *method of least squares*.



[1] Unfortunately, you will come across people (and SPSS for that matter) referring to regression variables as dependent and independent variables (as in controlled experiments). However, correlational research by its nature seldom controls the independent variables to measure the effect on a dependent variable. Instead, variables are measured simultaneously and without strict control. It is, therefore, inaccurate to label regression variables in this way. For this reason I label 'independent variables' as *predictors*, and the 'dependent variable' as the *outcome*.

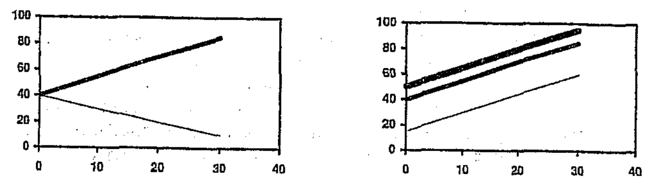
4.1.1. Some Important Information about Straight Lines

To use linear regression it is important that you know a few algebraic details of straight lines. Any straight line can be drawn if you know two things: (1) the slope (or gradient) of the line, and (2) the point at which the line crosses the vertical axis of the graph (known as the *intercept* of the line). The equation of a straight line is defined in equation (4.1), in which *Y* is the outcome variable that we want to predict and X_i is the *i*th subject's score on the predictor variable. β_1 is the gradient of the straight line fitted to the data and β_0 is the intercept of that line. There is a residual term, ε_i, which represents the difference between the score predicted by the line for subject *i* and the score that subject *i* actually obtained. The equation is often conceptualized without this residual term (so, ignore it if it's upsetting you); however, it is worth knowing that this term represents the fact our model will not fit perfectly the data collected.

$$Y = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{4.1}$$

A particular line has a specific intercept and gradient. Figure 4.1 shows a set of lines that have the same intercept but different gradients, and a set of lines that have the same gradient but different intercepts. Figure 4.1 also illustrates another useful point: that the gradient of the line tells us something about the nature of the relationship being described. In Chapter 3 we saw how relationships can be either positive or negative (and I don't mean the difference between getting on well with your girlfriend and arguing all the time!). A line that has a gradient with a positive value describes a positive relationship, whereas a line with a negative gradient describes a negative relationship. So, if you look at the graph in Figure 4.1 in which the gradients differ but the intercepts are the same, then the thicker line describes a positive relationship whereas the thinner line describes a negative relationship.

If it is possible to describe a line knowing only the gradient and the intercept of that line, then we can use these values to describe our model (because in linear regression the model we use is a straight line). So, the model that we fit to our data in linear regression can be conceptualized as a straight line that can be described mathematically by equation (4.1). With regression we strive to find the line that best describes the data collected, then estimate the gradient and intercept of that line. Having defined these values, we can insert different values of our predictor variable into the model to estimate the value of the outcome variable.



Same intercept, different slopes

Same slope, different intercepts

Figure 4.1: Shows lines with the same gradients but different intercepts, and lines that share the same intercept but have different gradients

4.1.2. *The Method of Least Squares*

I have already mentioned that the method of least squares is a way of finding the line that best fits the data (i.e. finding a line that goes through, or as close to, as many of the data points as possible). This 'line of best fit' is found by ascertaining which line, of all of the possible lines that could be drawn, results in the least amount of difference between the observed data points and the line. Figure 4.2 shows that when any line is fitted to a set of data, there will be small differences between the values predicted by the line, and the data that were actually observed. We are interested in the vertical differences between the line and the actual data because we are using the line to predict values of *Y* from values of the *X* variable. Although some data points fall exactly on the line, others lie above and below the line, indicating that there is a difference between the model fitted to these data and the data collected. Some of these differences are positive (they are above the line, indicating that the model underestimates their value) and some are negative (they are below the line, indicating that the model overestimates their value). These differences are usually called *residuals*. In the discussion of variance in section 1.1.3.1 I explained that if we sum positive and negative differences then they tend to cancel each other out. To avoid this problem we square the differences before adding them up. These squared differences provide a gauge of how well a particular line fits the data: if the squared differences are large, the line is not representative of the data; if the squared differences are small then the line is representative. The sum of squared differences (or sum of squares for short) can be calculated for any line that is fitted to some data; the 'goodness-of-fit' of each line can then be compared by looking at the sum of squares for each. The method of least squares works by selecting the line that has the lowest sum of squared differences (so it chooses the line that best represents the observed data). One way to select this optimal line would be to fit every possible line to a set of data, calculate the sum of squared differences for each line, and then choose the line for which this value is smallest. This would take quite a long time to do! Fortunately, there is a mathematical technique for finding maxima and minima and this technique (calculus) is used to find the line that minimizes the sum of squared differences. The end result is that the value of the slope and intercept of the 'line of best fit' can be estimated. Social scientists generally refer to this line of best fit as a regression line.

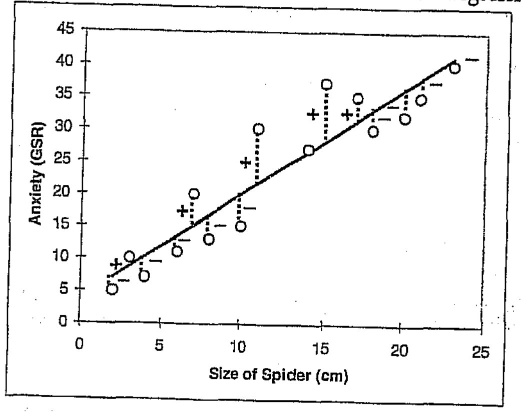
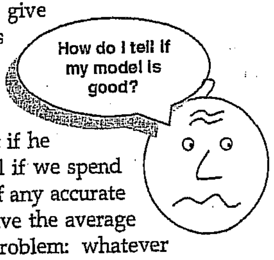


Figure 4.2: This graph shows a scatterplot of some data with a line representing the general trend. The vertical lines (dotted) represent the differences (or residuals) between the line and the actual data

4.1.3. *Assessing the Goodness-of-Fit: Sums of Squares, R and R^2*

Once we have found the line of best fit it is important that we assess how well this line fits the actual data (we assess the *goodness-of-fit* of the model). In section 1.1.3.1 we saw that one measure of the adequacy of a model is the sum of squared differences. Sticking with this theme, there are several sums of squares that can be calculated to help us gauge the contribution of our model to predicting the outcome. Imagine that I was interested in predicting record sales (*Y*) from the amount of money spent advertising that record (*X*). One day my boss came in to my office and said 'Andy, how many records will we sell if we spend £100,000 on advertising?' If I didn't have an accurate model of the relationship between record sales and advertising, what would my best guess be? Well, probably the best answer I could give would be the mean number of record sales (say, 200,000) because on average that's how many records we expect to sell. This response might well satisfy a brainless record company executive. However, what if he had asked 'How many records will we sell if we spend £1 on advertising?' Again, in the absence of any accurate information, my best guess would be to give the average number of sales (200,000). There is a problem: whatever amount of money is spent on advertising I always predict the same levels of sales. It should be pretty clear then that the mean is fairly useless as a model of a relationship between two variables—but it is the simplest model available.



So, as a basic strategy for predicting the outcome, we might choose to use the mean, because on average (*sic*) it will be a fairly good guess of an outcome. Using the mean as a model, we can calculate the difference between the observed values, and the values predicted by the mean. We saw in section 1.1.3.1 that we square all of these differences to give us the sum of squared differences. This sum of squared differences is known as the *total sum of squares* (denoted SS_T) because it is the total amount of differences present when the most basic model is applied to the data. This value represents how good the mean is as a model of the observed data. Now, if we fit the more sophisticated model to the data, such as a line of best fit, we can again work out the differences between this new model and the observed data. In the previous section we saw that the method of least squares finds the best possible line to describe a set of data by minimizing the difference between the model fitted to the data and the data themselves. However, even with this optimal model there is still some inaccuracy, which is represented by the differences between each observed data point and the value predicted by the regression line. As before, these differences are squared before they are added up so that the directions of the differences do not cancel out. The result is known as the *sum of squared residuals* (SS_R). This value represents the degree of inaccuracy when the best model is fitted to the data. We can use these two values to calculate how much better the regression line (the line of best fit) is than just using the mean as a model (i.e. how much better is the best possible model than the worst model?). The improvement in prediction resulting from using the regression model rather than the mean is calculated by calculating the difference between SS_T and SS_R. This difference shows us the reduction in the inaccuracy of the model resulting from fitting the regression model to the data. This improvement is the *model sum of squares* (SS_M). Figure 4.3 shows each sum of squares graphically.

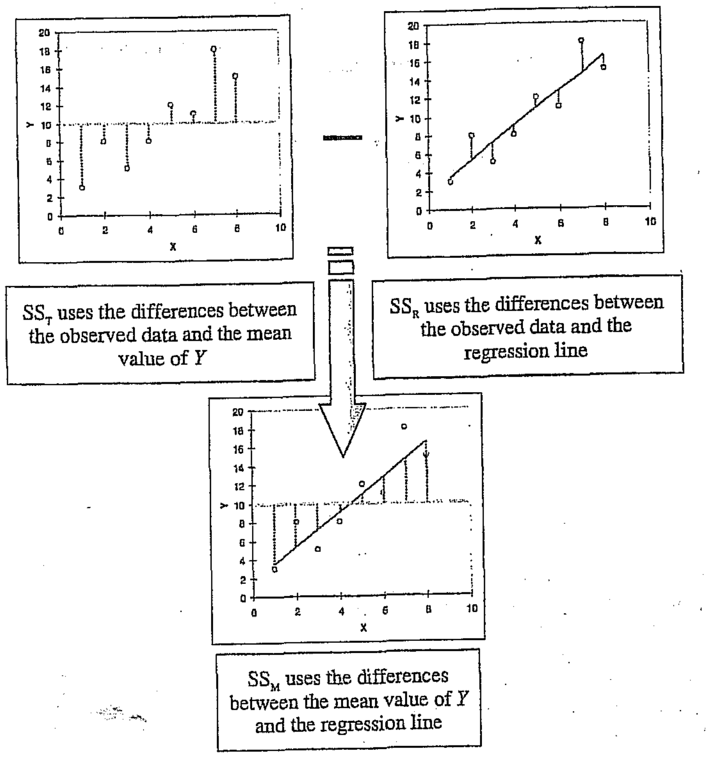

20
18
16
14
12
Y 10
8
6
4
2
0
0 2 4 6 8 10
X
SS_T uses the differences between the observed data and the mean value of *Y*
SS_R uses the differences between the observed data and the regression line
SS_M uses the differences between the mean value of *Y* and the regression line

Figure 4.3: Diagram showing from where the regression sums of squares derive

If the value of SS_M is large then the regression model is very different from using the mean to predict the outcome variable. This implies that the regression model has made a big improvement to how well the outcome variable can be predicted. However, if SS_M is small then using the regression model is little better than using the mean (i.e. the regression model is no better than taking our 'best guess'). A useful measure arising from these sums of squares is the proportion of improvement due to the model. This is easily calculated by dividing the sum of squares for the model by the total sum of squares. The resulting value is called R^2 and to express this value as a percentage you should multiply it by 100. So, R^2 represents the amount of variance in the outcome explained by the model (SS_M) relative to how much variation there was to explain in the first place (SS_T). Therefore, as a percentage, it represents the percentage of the variation in the outcome that can be explained by the model.

$$R^2 = \frac{SS_M}{SS_T} \qquad (4.2)$$

Interestingly, this value is the same as the R^2 we met in Chapter 3 (section 3.2.3.3) and you'll notice that it is interpreted in the same way. Therefore, in simple regression we can take the square root of this value to obtain the Pearson correlation coefficient. As such, the correlation coefficient provides us with a good estimate of the overall fit of the regression model, and R^2 provides us with a good gauge of the substantive size of the relationship.

A second use of the sums of squares in assessing the model is through the *F*-test. The *F*-test is something we will cover in greater depth in Chapter 7, but briefly this test is based upon the ratio of the improvement due to the model (SS_M) and the difference between the model and the observed data (SS_R). In fact, rather than using the sums of squares themselves, we take the mean sums of squares (referred to as the *mean squares* or MS). To work out the mean sums of squares it is necessary to divide by the degrees of freedom (this is comparable to calculating the variance from the sums of squares—see section 1.1.3.1). For SS_M the degrees of freedom are simply the number of variables in the model, and for SS_R they are the number of observations minus the number of parameters being estimated (i.e. the number of beta coefficients including the constant). The result is the mean squares for the model (MS_M) and the residual mean squares (MS_R). At this stage it isn't essential that you understand how the mean squares are derived (it is explained in Chapter 7). However, it is important that you understand that the *F*-ratio (equation (4.3)) is a measure of how much the model has improved the prediction of the outcome compared to the level of inaccuracy of the model.

$$F = \frac{MS_M}{MS_R} \qquad (4.3)$$

If a model is good, then we expect the improvement in prediction due to the model to be large (so, MS_M will be large) and the difference between the model and the observed data to be small (so, MS_R will be small). In short, a good model should have a large *F*-ratio (greater than one at least) because the top half of equation (4.3) will be bigger than the bottom. The exact magnitude of this *F*-ratio can be assessed using critical values for the corresponding degrees of freedom.

4.1.4. Simple Regression on SPSS

So far, we have seen a little of the theory behind regression, albeit restricted to the situation in which there is only one predictor. To help clarify what we have learnt so far, we will go through an example of a simple regression on SPSS. Earlier on I asked you to imagine that I worked for a record company and that my boss was interested in predicting record sales from advertising. There are some data for this example in the file **Record1.sav**. This data file has 200 rows, each one representing a different record. There are also two columns, one representing the sales of each record in the week after release and the other representing the amount (in pounds) spent promoting the record before release. This is the format for entering regression data: the outcome variable and any predictors should be entered in different columns, and each row should represent independent values of those variables. The pattern of the data is shown in Figure 4.4 and it should be clear that a positive relationship exists: so, the more money spent advertising the record, the more it is likely to sell. Of course there are some records that sell well regardless of advertising (top left of scatterplot), but there are none that sell badly when advertising levels are high (bottom right of scatterplot). The scatterplot also shows the line of best fit for these data: bearing in mind that the mean would be represented by a flat line at around the 200,000 sales mark, the regression line is noticeable different.

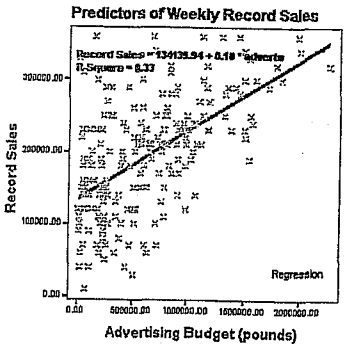


Figure 4.4: Scatterplot showing the relationship between record sales and the amount spent promoting the record

To find out the parameters that describe the regression line, and to see whether this line is a useful model, we need to run a regression analysis.

To do the analysis you need to access the main dialog box by using the <u>A</u>nalyze⇒<u>R</u>egression⇒<u>L</u>inear... menu path. Figure 4.5 shows the resulting dialog box. There is a space labelled *<u>D</u>ependent* in which you should place the outcome variable (in this example **sales**). So, select **sales** from the list on the left-hand side, and transfer it by clicking on ▶. There is another space labelled *<u>I</u>ndependent(s)* in which any predictor variable should be placed. In simple regression we use only one predictor (in this example **adverts**) and so you should select **adverts** from the list and click on ▶ to transfer it to the list of predictors. There are a variety of options available, but these will be explored within the context of multiple regression (see section 4.2). For the time being just click on OK to run the basic analysis.

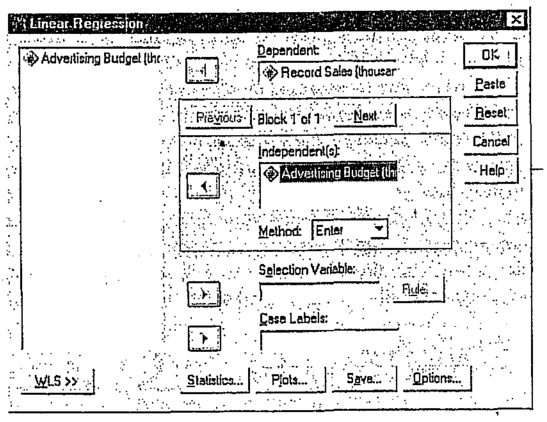


Figure 4.5: Main dialog box for regression

4.1.5. Output from SPSS

4.1.5.1. Overall Fit of the Model

The first table provided by SPSS is a summary of the model. This summary table provides the value of R and R^2 for the model that has been derived. For these data, R has a value of 0.578 and because there is only one predictor, this value represents the simple

Model Summary

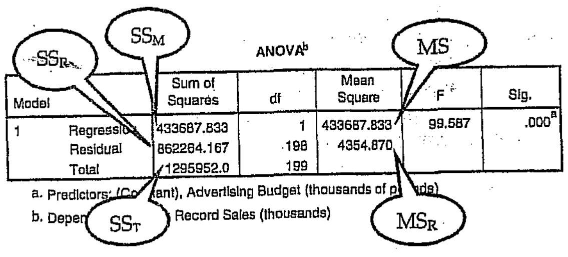
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.578[a]	.335	.331	65.9914

a. Predictors: (Constant), Advertising Budget (thousands of pounds)

SPSS Output 4.1

correlation between advertising and record sales (you can confirm this by running a correlation using what you were taught in Chapter 3). The value of R^2 is 0.335, which tells us that advertising expenditure can account for 33.5% of the variation in record sales. In other words, if we are trying to explain why some records sell more than others, we can look at the variation in sales of different records. There might be many factors that can explain this variation, but our model, which includes only advertising expenditure, can explain 33% of it. This means that 66% of the variation in record sales cannot be explained by advertising alone. Therefore, there must be other variables that have an influence also.

The next part of the output reports an analysis of variance (ANOVA—see Chapter 7). The summary table shows the various sums of squares described in Figure 4.3 and the degrees of freedom associated with each. From these two values, the average sums of squares (the mean squares) can be calculated by dividing the sums of squares by the associated degrees of freedom. The most important part of the table is the *F*-ratio, which is calculated using equation (4.3), and the associated significance value of that *F*-ratio. For these data, *F* is 99.59, which is significant at $p < 0.001$ (because the value in the column labelled *Sig.* is less than 0.001). This result tells us that there is less than a 0.1% chance that an *F*-ratio this large would happen by chance alone. Therefore, we can conclude that our regression model results in significantly better prediction of record sales than if we used the mean value of record sales. In short, the regression model overall predicts record sales significantly well.


SS_R
SS_M
MS
SS_T
MS_R

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	433687.833	1	433687.833	99.587	.000[a]
	Residual	862264.167	198	4354.870		
	Total	1295952.0	199			

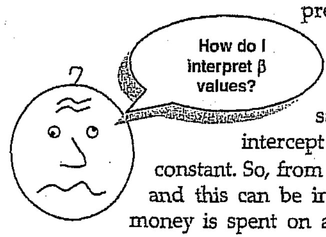
a. Predictors: (Constant), Advertising Budget (thousands of pounds)

b. Dependent Variable: Record Sales (thousands)

SPSS Output 4.2

4.1.5.2. Model Parameters

The ANOVA tells us whether the model, overall, results in a significantly good degree of prediction of the outcome variable. However, the ANOVA doesn't tell us about the individual contribution of variables in the model (although in this simple case there is only one variable in the model and so we can infer that this variable is a good



predictor). The table in SPSS Output 4.3 provides details of the model parameters (the beta values) and the significance of these values. We saw in equation (4.1) that β_0 was the Y intercept and this value is the value B for the constant. So, from the table, we can say that β_0 is 134.14, and this can be interpreted as meaning that when no money is spent on advertising (when $X = 0$), the model predicts that 134,140 records will be sold (remember that our unit of measurement was thousands of records). We can also read off the value of β_1 from the table and this value represents the gradient of the regression line. It is 9.612 E–02, which in unabbreviated form is 0.09612.[2] Although this value is the slope of the regression line, it is more useful to think of this value as representing *the change in the outcome associated with a unit change in the predictor*. Therefore, if our predictor variable is increased by 1 unit (if the advertising budget is increased by 1), then our model predicts that 0.096 extra records will be sold. Our units of measurement were thousands of pounds and thousands of records sold, so we can say that for an increase in advertising of £1000 the model predicts 96 (0.096 × 1000 = 96) extra record sales. As you might imagine, this investment is pretty bad for the record company: they invest £1000 and get only 96 extra sales! Fortunately, as we already know, advertising accounts for only one-third of record sales!

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	134.140	7.537		17.799	.000
	Advertising Budget (thousands of pounds)	9.612E-02	.010	.578	9.979	.000

a. Dependent Variable: Record Sales (thousands)

SPSS Output 4.3

[2] You might have noticed that this value is reported by SPSS as 9.612 E–02 and many students find this notation confusing. Well, this notation simply means 9.61×10^{-2} (which might be a more familiar notation). OK, some of you are still confused. Well think of E–02 as meaning 'move the decimal place 2 steps to the left', so 9.612 E–02 becomes 0.09612. If the notation read 9.612 E–01, then that would be 0.9612, and if it read 9.612 E–03, that would be 0.009612. Likewise, think of E+02 (notice the minus sign has changed) as meaning 'move the decimal place 2 places to the right'. So 9.612 E+02 becomes 961.

The values of β represent the change in the outcome resulting from a unit change in the predictor. If the model was useless at predicting the outcome, then if the value of the predictor changes, what might we expect the change in the outcome to be? Well, if the model is very bad then we would expect the change in the outcome to be zero. Think back to Figure 4.3 (see the panel representing SS_T) in which we saw that using the mean was a very bad way of predicting the outcome. In fact, the line representing the mean is flat, which means that as the predictor variable changes, the value of the outcome does *not* change (because for each level of the predictor variable, we predict that the outcome will equal the mean value). The important point here is that a bad model (such as the mean) will have regression coefficients of zero for the predictors. A regression coefficient of zero means: (a) a unit change in the predictor variable results in no change in the predicted value of the outcome (the predicted value of the outcome does not change at all), and (b) the gradient of the regression line is zero, meaning that the regression line is flat. Hopefully, what should be clear at this stage is that if a variable significantly predicts an outcome, then it should have a β value significantly different from zero. This hypothesis is tested using a *t*-test (see Chapter 6). The *t*-statistic tests the null hypothesis that the value of β is zero: therefore, if it is significant we accept the hypothesis that the β value is significantly different from zero and that the predictor variable contributes significantly to our ability to estimate values of the outcome.

One problem with testing whether the β values are different from zero is that their magnitude depends on the units of measurement (for example advertising budget has a very small β value, yet it seems to have a strong relationship to record sales). Therefore, the *t*-test is calculated by taking account of the standard error. The standard error tells us something about how different β values would be if we took lots and lots of samples of data regarding record sales and advertising budgets and calculated the β values for each sample. We could plot a frequency distribution of these samples to discover whether the β values from all samples would be relatively similar, or whether they would be very different. We can use the standard deviation of this distribution (known as the *standard error*) as a measure of the similarity of β values across samples. If the standard error is very small, then it means that most samples are likely have a β value similar to the one in the sample collected (because there is little variation across samples). The *t*-test tells us whether the β value is different from zero relative to the variation in β values for similar samples. When the standard error is small even a small deviation from zero can reflect a meaningful difference because β is representative of the majority of possible samples.

Equation (4.4) shows how the *t*-test is calculated and you'll find a general version of this equation in Chapter 6 (equation (6.1)). The β_{expected} is simply the value of β that we would expect to obtain if the null hypothesis were true. I mentioned earlier that the null hypothesis is that β is zero and so this value can be replaced by zero. The equation simplifies to become the observed value of β divided by the standard error with which it is associated.[3]

$$\begin{aligned} t &= \frac{\beta_{\text{observed}} - \beta_{\text{expected}}}{\text{SE}_\beta} \\ &= \frac{\beta}{\text{SE}_\beta} \end{aligned} \tag{4.4}$$

The values of t can then be compared to the values that we would expect to find by chance alone: if t is very large then it is unlikely to have occurred by chance. SPSS provides the exact probability that the observed value of t is a chance result, and as a general rule, if this observed significance is less than 0.05, then social scientists agree that the result reflects a genuine effect. For these two values, the probabilities are 0.000 (zero to 3 decimal places) and so we can say that the probability of these t values occurring by chance is less than 0.001. Therefore, they reflect genuine effects. We can, therefore, conclude that advertising budget makes a significant contribution ($p < 0.001$) to predicting record sales.

4.1.5.3. Using the Model

So far, we have discovered that we have a useful model, one that significantly improves our ability to predict record sales. However, the next stage is often to use that model to make some predictions. The first stage is to define the model by replacing the β values in equation (4.1) with the values from SPSS Output 4.3. In addition, we can replace the X and Y with the variable names so that the model becomes:

$$\begin{aligned} \text{Record Sales} &= \beta_0 + \beta_1 \text{Advertising Budget}_i \\ &= 134.14 + (0.09612 \times \text{Advertising Budget}_i) \end{aligned} \tag{4.5}$$

It is now possible to make a prediction about record sales, by replacing the advertising budget with a value of interest. For example, imagine a record executive wanted to spend £100,000 on advertising a new record. Remembering that our units are already in thousands of pounds, we can simply replace the advertising budget with 100. He would discover that record sales should be around 144,000 for the first week of sales.

[3] To see that this is true you can use the values from SPSS Output 4.3 to calculate t for the constant. For advertising budget, the standard error has been rounded to 3 decimal places, so to verify how t is calculated you should use the unrounded value. This value is obtained by double-clicking the table in the SPSS output and then double-clicking the value that you wish to see in full. You should find that $t = 0.096124 / 0.009632 = 9.979$.

$$\begin{aligned} \text{Record Sales} &= 134.14 + (0.09612 \times \text{Advertising Budget}_i) \\ &= 134.14 + (0.09612 \times 100) \\ &= 143.75 \end{aligned} \tag{4.6}$$

12. lekce

FAKTOROVÁ ANALÝZA - REDUKCE DAT A VSTUP DO MULTIVARIAČNÍ ANALÝZY (Modul ANALYZE: pocedura: Data reduction-factor analysis).

DISMAN Miroslav. 1993. *Jak se vyrábí sociologická znalost.* Praha: Karolinum.

Faktorová analýza

To je technika, která byla a někde ještě je přijímána s určitou nedůvěrou. Domníváme se, že neprávem. Důvodem pro tuto nedůvěru je zřejmě fakt, že funkce faktorové analýzy je velice odlišná od ostatních statistických technik. Většina statistických operací v sociologickém výzkumu je používána pro **testování hypotéz**, faktorová analýza je používána pro tento účel spíše výjimečně. Faktorová analýza je spíše nástrojem pro explorativní výzkum. Většinou netestuje hypotézy, ale je nástrojem pro jejich formulování a upřesňování. Je ovšem i neobyčejně účinným nástrojem zjednodušování dat.

Funkci faktorové analýzy si můžeme vysvětlit - velice nespisovně, ale jasně a v podstatě správně - asi takto: faktorová analýza je schopna **nalézt seskupení proměnných, které patří nějakým způsobem k sobě.** Jaký je to způsob, to nám faktorová analýza neřekne. Na to musí odpověď nalézt výzkumník sám, na základě své odborné znalosti.

A teď si ukažme, až v neslušně zjednodušené a zkrácené formě, jak tato analýza funguje. Vstup do faktorové analýzy je korelační matice, tabulka korelačních koeficientů mezi všemi proměnnými, které hodláme analyzovat. Výstupem z faktorové analýzy jsou sloupce čísel, z nichž každý představuje jeden extrahovaný **faktor**. Faktor je ono dosud nepojmenované "něco" co sdružuje proměnné s vysokými čísly v daném sloupci. Čísla v sloupcích jsou nazývána "factor loadings", **faktorová zátěž,** míra spojení proměnné v daném řádku s tímto faktorem. Je možné říci, že faktorová zátěž je korelace mezi proměnnou a faktorem.

Ideální by bylo, kdyby skupiny proměnných měly vysokou zátěž v jednom faktoru a téměř nulovou ve všech ostatních faktorech. Faktorová analýza se snaží dosáhnout tohoto cíle *značně složitými matematickými postupy, které bychom snad mohli nejlépe popsat* jako rotaci souřadnic v mnohodimenzionálním prostoru. Tyto postupy nejsou bez problémů a pouštět se do faktorové analýzy bez spolupráce statistika může být někdy velice riskantní. Nicméně ani tyto velice chytré postupy nemohou izolovat zcela čisté faktory. Nezapomeňte na první Dismanův zákon ("data jsou potvory") nebo jinými slovy, na rozsáhlé sítě souvislostí v společenských vědách. Tak v tabulce 9.12. vidíme, že proměnná (5), strava, má vysokou zátěž v prvém faktoru, který, jak uvidíme, se zdá zachycovat etnickou symboliku stravy, ale i nezanedbatelnou náložv druhém faktoru, odrážejícím materiální utilitu stravy.

A jak interpretovat obsahový význam faktorů? Na to nemáme jednoduchý recept kromě jednoho. Použít svou odbornou znalost a zdravý rozum. Někdy to může být velice jednoduché. Řekněme, že jsme třeba připravili sérii otázek, o nichž se domníváme, že měří týž koncept. Faktorová analýza nám dává možnost testovat tento předpoklad. V ideálním případě by měla být schopna extrahovat jediný faktor, t.j. koeficienty by měly být nejsilnější v prvém faktoru. Proměnné, které by měly vysoké koeficienty v jiném než prvém faktoru, měří zřejmě něco jiného a měly by být ze souboru vyloučeny. To je ovšem případ, kdy je faktorová analýza použita pro testování hypotéz. Tato analýza je také výhodným nástrojem na zjednodušení souboru proměnných.

Rozsáhlá faktorová analýza, používající data z desítek velice rozdílných populací byla s to redukovat stovky stimulů, používaných v technice sémantického diferenciálu. V běžné praxi se nyní dosti standardně používá kolem deseti párů podnětů, a to bez jakékoliv podstatné ztráty informace. Faktorová analýza jako nástroj pomáhající ustavit validitu jiných výzkumných technik má značně široké pole použití. Slyšeli jsme *dokonce o tom*, jak tato technika byla použita pro odhalení tazatele, který falšoval rozhovory; nenavštěvoval respondenty, ale *vymýšlel si odpovědi* u *svého* psacího stolu. Byl dosti chytrý, aby odpovědi v jednotlivých rozhovorech byly konzistentní, ale nebyl - a nemohl být - chytrý natolik, aby se faktory vyvozené z jeho dat shodovaly s faktory extrahovanými z dat ostatních tazatelů.

Ovšem důležitou oblastí aplikace této techniky je explorativní výzkum a zde se interpretace významů faktorů stává základním úkolem. Klíčem pro identifikaci faktoru je v podstatě odpověď na tuto otázku: "Co mají proměnné s vysokými koeficienty v daném faktoru společného?" Forslund (1980) studoval delikvenci mládeže v malém městě ve Wyomingu. (Citováno v Babbie, 1989.) Rozdal středoškolským studentům dotazník, ve kterém se dotazoval na řadu lehce delikventních aktivit. Ve *faktorové analýze* pak identifikoval 4 zřetelné faktory.

Kupř. následující proměnné měly vysoké koeficienty ve faktoru 1:

- rozbíjení pouličních světel
- rozbíjení oken
- vypouštění vzduchu z pneumatik automobilů
- drobné krádeže atd.

Tento faktor byl identifikován jako přestupky proti majetku.

Druhý faktor byl identifikován jako nezvládnutelnost a měl vysokou zátěž kupř. v následujících proměnných:

- neuposlechnutí rodičů
- psaní a malování po stěnách a lavicích
- odmlouvání učitelům
- anonymní telefonické rozhovory

Poslední faktor byl zcela jednoznačný, měl vysokou zátěž jen ve dvou proměnných, týkajících se rvaček a bitek.

Třetí faktor jsme si nechali na konec. Je totiž zajímavý z hlediska metodologie faktorové analýzy a proto zde uvedeme všechny proměnné, ve kterých měl významnou zátěž, i spolu s koeficienty:

-	kouření marihuany	.755
-	používání jiných drog	.669
-	falšování podpisů na omluvenkách	.395
-	pití alkoholu v nepřítomnosti rodičů	.358
-	chození za školu	.319

Vidíme zřetelně, že tento faktor shrnuje dva různé problémy: drogy a chození za školu. Forslund také použil pro tento faktor toto dvojité jméno. Pro nás je tato podvojnost faktoru ilustrací jedné důležité a nebezpečné vlastnosti faktorové analýzy. V naší interpretaci jsou obě složky faktoru sdruženy jinak, než tomu bylo u zbývajících faktorů. U nich proměnné byly sdruženy tím způsobem, že všechny proměnné byly ukazateli shodného podloženého faktoru. V případě faktoru 3. jeho obě složky mohou patřit k dvěma poněkud odlišným podloženým konceptům, "chození za školu" a "drogy". Faktorová analýza je prezentovala jako jediný faktor prostě proto, že tyto dvě složky jsou v nějakém, patrně silném, kauzálním vztahu. Je jedno, zda nedostatek dohledu v nekontrolované situaci "za školou" zvyšuje pravděpodobnost přístupu k drogám, nebo zda snaha získat "marjanku" vede k chození za školu.

Tohle není útok na validitu Forslundova výzkumu, ta byla dobře potvrzena i tím, že faktorová analýza byla opakována odděleně pro subpopulace chlapců a děvčat a v obou případech extrahovala faktory srovnatelné s výsledky původní analýzy. Jde nám jen o to zdůraznit technické aspekty, které jsou spojeny právě s jedinečnou schopností faktorové analýzy nalézt struktury v datech, která "nějak" spolu souvisejí. Ono "nějak" může být opravdu odrazem obsahu proměnných, a tak umožní identifikovat struktury, které mohou být teoreticky významné a které jsou neviditelné jiným analytickým postupům. Ale to "nějak" může mít někdy velice prozaický a mechanický charakter. Faktorová analýza může prostě identifikovat jako samostatný faktor skupinu proměnných jenom proto, že informace byla získána shodnou technikou sběru dat. Řekněme, že existují určité sociální typy osobnosti, charakterizované jednak postoji jedince, jednak jeho demografickými charakteristikami. Kdybychom chtěli "objevit" tyto typy s pomocí faktorové analýzy, dočkali bychom se pravděpodobně značného zklamání. Analýza by asi objevila jenom (nebo hlavně) dva faktory: jeden, obsahující demografické proměnné, druhý postojové proměnné. Jsou cesty, jak takové problémy překonat, ale obecně bychom se měli vyvarovat toho, používat jako vstup do faktorové analýzy proměnné, které jsou formálně různorodé.

Pokud není faktorová analýza použita jako pouhé testování hypotéz, představuje jenom prvý krok v poznávacím procesu. Jejím výsledkem může být třeba jen nová formulace problému, navržení nových hypotéz. Ale to není málo.

V našem torontském výzkumu o postojích starých Italů, Portugalců a anglicky mluvících osob, narozených v Kanadě jsme se zajímali o struktury obsahu strachu, který staří lidé pociťují téměř univerzálně, uvažují-li o možnosti, že budou nuceni vstoupit do domova důchodců. Zjišťovali jsme mimo jiné, jak by pro respondenta byly důležité - v případě institucionalizace - následující body:

1. být odříznut od sousedství, na které je respondent zvyklý;
2. možnost pokračovat v koníčcích, které nyní má;
3. mít pokoj pro sebe;
4. přinést si do instituce nábytek a jiné věci;
5. dostávat stravu, na kterou je respondent zvyklý;
6. bydlet s někým, kdo je s respondentem sociálně srovnatelný a
7. bydlet s osobou mluvící respondentovým jazykem.

Výsledky byly podrobeny faktorové analýze a zde jsou její výsledky pro italskou subpopulaci:

Tabulka 9.12.

FAKTOR:	(1)	(2)	(3)
Proměnná:			
(6) osoba jako respondent	.901*	-.071	-.047
(7) jazyk	.896*	.015	.007
(5) strava	.604*	.335	.100
(4) přinést věci	.026	.813*	-.123
(3) pokoj pro sebe	.067	.772*	.323
(2) koníček	-.258	.180	.818*
(1) sousedství	.397	-.070	.677*

Zdá se, že prvý faktor reprezentuje **etnickou identitu** respondentů: jazyk jako symbol národní identity, potřeba partnera "jako já" jako symbol kulturní a třídní sounáležitosti, a konečně strava, která je v literatuře zcela shodně považována za jeden z nejdéle přežívajících prvků etnické kultury.

Druhý faktor se zdá reprezentovat jedincovo **pragmatické, materiální okolí** a konečně třetí faktor může reprezentovat **sféru soukromých, osobních zájmů**. Plausibilita této interpretace zdá se být podporována i následujícím faktem. Zcela shodné faktory byly extrahovány i pro portugalský vzorek, ne však pro vzorek rozených, anglicky mluvících Kanaďanů. Pro ně, jako členy hlavního kulturního proudu není etnická identita vůbec problémem. Zde faktorová analýza izolovala pouze dva slabé faktory, odpovídající přibližně faktorům 2. a 3. ve zbývajících etnických skupinách. Další teoretickou podporou pro naši interpretaci jsou výsledky analýzy jiného souboru proměnných ze stejného výzkumu. Tyto proměnné měřily míru obav asociovaných s různými prvky spojenými se vstupem do instituce pro staré. Faktorová analýza těchto dat extrahovala opět faktory významem podobné těm, které jsme zde právě představili.

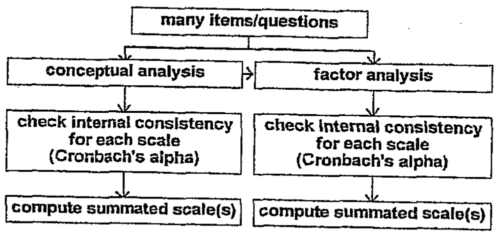
Tyto nálezy samy o sobě neznamenají mnoho. Ale otevřely pro nás docela zajímavou cestu do studia obsahu strachu a jeho sociálních a zejména kulturních determinant. Ale to je už zase jiná povídka, které musí být teprve napsána.

CHAPTER 10

Factor Analysis: Data Reduction With Principal Components Analysis

Factor analysis, a complex associational technique, is used for several purposes, but the main one is data reduction.[1] When you have a number of questions about the same general topic (e.g., attitudes about mathematics), you may want to ask whether the questions could be grouped into a smaller number of composite variables. Table 10.1 shows that *either* factor analysis or conceptual analysis (i.e., thinking based on theory and/or literature) can be used to reduce the number of variables to a more manageable and meaningful number of summated scales. The table also shows that you should check the internal consistency reliability of these new scales with Cronbach's alpha (see Assignment G) before actually computing the scales. You may want to check your conceptual analysis with factor analysis as well as Cronbach alphas. That is what we are doing here for our conceptualization of the three math attitude scales.

Table 10.1. ***Two Strategies for Reducing Many Related Items to Fewer Composite Variables***


many items/questions

conceptual analysis

factor analysis

check internal consistency for each scale (Cronbach's alpha)

check internal consistency for each scale (Cronbach's alpha)

compute summated scale(s)

compute summated scale(s)

After doing your conceptual or factor analysis, you have several decisions:

1. If you identify only one conceptual scale or one factor and it is supported by an alpha above .70, compute one overall summary scale score. However,
2. If there is more than one conceptual scale or more than one factor and they have good alphas, compute the several summated scale scores. Also, compute an overall scale score *if* it makes sense conceptually. However,
3. If the factor analysis results do not make good conceptual sense, do not use them. In this case, use the conceptual factors, rethink the conceptualization, or use each item separately

[1] Statisticians call what we have done in this chapter principal components analysis (PCA) rather than exploratory factor analysis (EFA). In SPSS, principal components analysis is done with the factor analysis program using the **principal components extraction** method. This is consistent with common usage, but there are technical differences between PCA and EFA (see Grimm & Yarnold, 1994).

Problems/ Research Questions

1. Can variables *q01* to *q13* be grouped into a smaller number of composite variables called components or factors? Using the principal components extraction method of the factor analysis program, you will have the computer sort the variables and suppress printing of values if the factor loading is less than .30. You will use a Varimax rotation and allow all factors with eigenvalues over 1.00 to be computed.
2. Rerun the factor analysis but specify that you want the number of factors to be three because our conceptualization is that there are three math attitude scales or factors: motivation, competence, and pleasure.
3. Run a factor analysis to see how the four "achievement" variables, *mathach, visual, mosaic,* and *grades*, cluster or factor.

Lab Assignment F

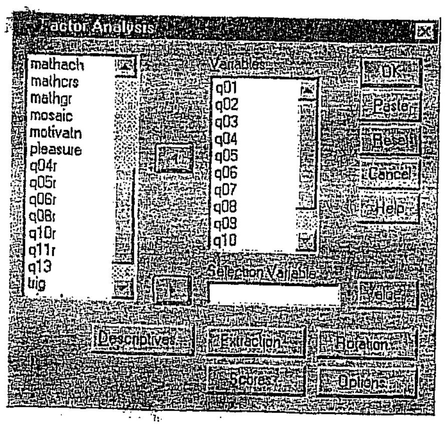
Logon and Get Data

- Retrieve your most recent data file: **hsbdataE.**

Problem 1: Factor Analysis on Math Attitude Variables

To begin factor analysis use these commands:

- **Statistics => Data Reduction => Factor** to get Fig. 10.1.
- Next select the variables *q01* through *q13*. *Do not* include q04r or any of the other reversed questions.


Factor Analysis

mathach
mathcrs
mathgr
mosaic
motivatn
pleasure
q04r
q05r
q06r
q08r
q10r
q11r
q13
trig

Variables:
q01
q02
q03
q04
q05
q06
q07
q08
q09
q10

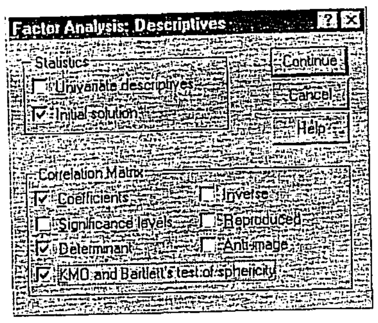
Selection Variable:

Fig. 10.1. Factor analysis.

MORGAN George A. and Griego V. ORLANDO. 1998. *Easy Use and Interpretation of SPSS for Windows: Answering Research Questions With Statistics*. Mahwah: Lawrence Erlbaum Associates.

Now click on **Descriptives** to produce Fig. 10.2.

- Then click on the following: **Initial solution** (under **Statistics**), **Coefficients, Determinant, KMO and Bartlett's test of sphericity** (under **Correlation Matrix**).
- Click on **Continue.**


Factor Analysis: Descriptives

Statistics
☐ Univariate descriptives
☑ Initial solution

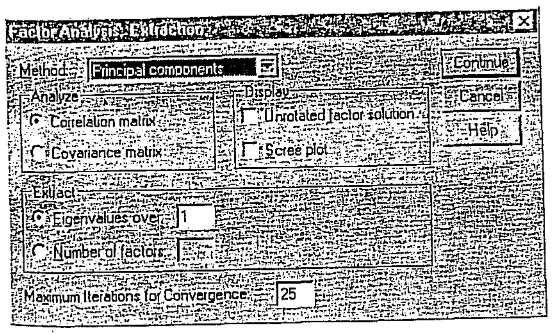
Correlation Matrix
☑ Coefficients
☐ Significance levels
☑ Determinant
☑ KMO and Bartlett's test of sphericity
☐ Inverse
☐ Reproduced
☐ Anti-image

Fig. 10.2. Factor analysis: Descriptives.

- Next, click on **Extraction** at the bottom of Figure 10.1. This will give you Fig. 10.3.
- Make sure **Eigenvalues over 1** is checked.

This default setting will allow the *computer to decide* how many math attitude factors to compute; i.e., as many as have eigenvalues (a measure of variability explained) greater than 1.0. If you have a clear theory or conceptualization about how many factors or scales there should be, you can set the **number of factors** to that number, as we will in Problem 2.

- *Unclick* display **Unrotated factor solution.**
- Click on **Continue.**


Factor Analysis: Extraction

Method: Principal components

Analyze
● Correlation matrix
○ Covariance matrix

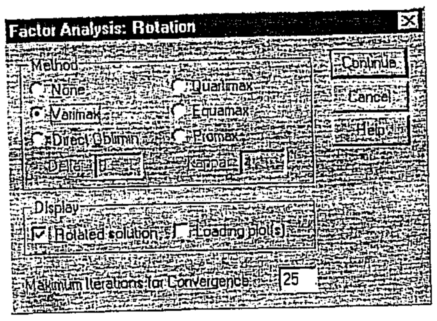
Display
☐ Unrotated factor solution
☐ Scree plot

Extract
● Eigenvalues over: 1
○ Number of factors:

Maximum Iterations for Convergence: 25

Fig. 10.3. Extraction method to produce principal components analysis (PCA).

- Now click on **Rotation,** which will give you Fig. 10.4.
- Click on **Varimax.**
- Then click on **Continue.**


Factor Analysis: Rotation

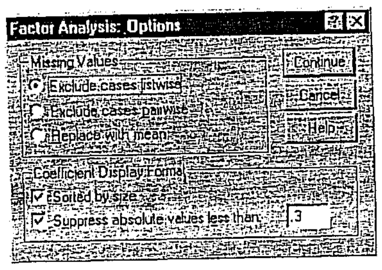
Method
○ None
● Varimax
○ Direct Oblimin
○ Quartimax
○ Equamax
○ Promax
Delta: 0
Kappa: 4

Display
☑ Rotated solution
☐ Loading plot(s)

Maximum Iterations for Convergence: 25

Fig. 10.4. Factor analysis: Rotation.

- Next, click on **Options** which will give you Fig. 10.5.
- Then click on **Sorted by size.**
- Click on **Suppress absolute values less.than** and type .3 (point 3) in the box (see Fig. 10.5). Suppressing small factor loadings makes the output easier to read.
- Click on **Continue** then **OK**. Compare Output 10.1 to your output and syntax.


Factor Analysis: Options

Missing Values
● Exclude cases listwise
○ Exclude cases pairwise
○ Replace with mean

Coefficient Display Format
☑ Sorted by size
☑ Suppress absolute values less than: .3

Fig. 10.5. Factor analysis: Options.

Problem 2: Factor Analysis on Math Attitude Variables With Three Factors Specified

Now try doing factor analysis yourself with the same variables, rotation, and options. This time, however, *click off* everything on the descriptive screen except **Initial solution.** Also, use a different **Extraction** subcommand; click on **Extract Number of factors** and then type 3 because our conceptualization is that there are 3 factors. Compare Output 10.2 to your output and syntax. Note that the **Initial Statistics** table is the same, but the **Rotated Component Matrix** now shows three somewhat different factors.

Problem 3: Factor Analysis on Achievement Variables

Now try doing another factor analysis yourself on the "achievement variables," *mathach, visual, mosaic,* and *grades.*

- First press **Reset**.
- Use the default settings (i.e., the boxes that are already checked) for **Extraction.**
- In addition, under **Rotation,** check **Varimax,** display **Rotated solution,** and **Loading plots.**
- Under **Descriptives** check **Univariate descriptives, Initial solution, Coefficients, Determinant,** and **KMO** and **Bartlett's test of sphericity.**

We have requested a more extensive, less simplified output for contrast with the earlier ones. Compare Output 10.3 to your syntax and output.

Print, Save, and Exit

- **Print** your lab assignment results if you want.
- **Save** your data file as **hsbdataF (File => Save As).**
- **Save** the SPSS log files as **hsblogF.**
- **Exit** SPSS.

Interpretation Questions

1. Using Output 10.1: a) Make a table of the five highest correlations and the five lowest. Indicate whether the variables for the highest and lowest correlations are in the same or in different conceptual clusters (i.e., competence, motivation, and pleasure) as indicated on page 19 for each question. b) What might you name each component or factor in the rotated factor matrix? c) How do these statistical components differ from the three conceptual math attitude composite variables (competence, motivation, and pleasure) computed in Assignment C and shown in chapter 2 and the codebook (Appendix D)?

2. Using Output 10.2: a) How do the rotated components in Output 10.2 differ from those in Output 10.1? b) Are the factors in Output 10.2 closer to the conceptual composites in the codebook? c) How might you name the three factors in Output 10.2?

3. Using Output 10.3: a) Are the assumptions that were tested violated? Explain. b) How many components or factors are there with eigenvalues greater than 1.0, and what total/cumulative percent of variance is accounted for by them? c) Describe the main aspects of the correlation matrix, rotated component matrix, and plot in Output 10.3.

```
GET
  FILE='A:\hsbdataE.sav'.
EXECUTE .
```

Output 10.1: Factor Analysis for Math Attitude Questions

```
Syntax for factor analysis of math attitude questions

FACTOR
  /VARIABLES q01 q02 q03 q04 q05 q06 q07 q08 q09 q10 q11 q12 q13  /MISSING
  LISTWISE /ANALYSIS q01 q02 q03 q04 q05 q06 q07 q08 q09 q10 q11 q12 q13
  /PRINT INITIAL CORRELATION DET KMO ROTATION
  /FORMAT SORT BLANK(.3)
  /CRITERIA MINEIGEN(1) ITERATE(25)
  /EXTRACTION PC
  /CRITERIA ITERATE(25)
  /ROTATION VARIMAX
  /METHOD=CORRELATION .
```

Interpret Output 10.1

The factor analysis program generates a number of tables depending on which options you have chosen. The first table in Output 10.1 is a **correlation matrix** showing how each of the 13 questions is associated with each of the other 12. Note some of the correlations are high (e.g., + or - .60 or greater) and some are low (i.e., near zero). The high correlations indicate that two items are associated and will probably be grouped together by the factor analysis.

Next, several assumptions are tested. The **determinant** (located under the correlation matrix) should be more than .00001. For instance, 2.316E-3 is the same as .002316 so this assumption is met. The **KMO** should be greater than .70, and is inadequate if less than .50. The **Bartlett** test should be significant (i.e., significance less than .05); these assumptions also are met.

The **Total Variance Explained** table shows how the variance is divided among the 13 possible components/factors. Note that four factors have **eigenvalues** (a measure of explained variance) greater than 1.0, which is a common criterion for a factor to be useful. Thus, unless you specify otherwise, as we will in Problem 2, the computer will look for the best four-factor solution.

In this case, the computer tried seven iterations before converging on the solution shown in the **Rotated Component Matrix** table. This table is the key one for understanding the results of the analysis. Note that the computer has sorted the 13 math attitude questions (Q01 to Q13) into four groups of 5, 3, 3, and 2 items, respectively. Within each component, the items are sorted from the one with the highest factor weight or loading (i.e., Q05 for factor 1, with a loading of -.88) to the one with the lowest (q02) that was still loaded *the most* on that factor. We have enclosed these items in circles for easy identification. Loadings are correlation coefficients of each item with the component so they range from -1.0 through 0 to +1.0. A negative loading just means that the question needs to be reversed when interpreting that factor.

The investigator should examine the content of the items that load high on each factor to see if they fit together conceptually and can be named. Items 5, 3, and 11 were intended to reflect an attitude or perception of competence at math (see page 19). Item 1 was intended to measure motivation for doing math, but in retrospect one can imagine that the phrase "until I can do them well" could be interpreted as competence. Likewise, Item 2, " I feel happy after solving a hard problem," although intended to measure pleasure at doing math, might also reflect competence at doing math. Every item has a weight or loading on every factor, but in a "clean" factor analysis almost all of the loadings that are not in the circles that we have drawn will be quite low (less than .40). We asked the computer to print only loadings of .30 or above so all the blanks in the table are low loadings. Note that Item 11 and, especially, Item 2 load above .40 on both Components 1 and 4. The latter component could be labeled pleasure at math, which was conceptually composed of Items 2, 6 and 10.

For our purposes, we will ignore the Factor Transformation Matrix; it was used to convert the initial factor matrix into the rotated factor matrix.

Factor Analysis

High correlations mean these items probably will be in the same factor.

Indicates how each question is associated with each other.

Correlation Matrix[a]

		question 1	question 2	question 3	question 4	question 5	question 6	question 7	question 8	question 9	question 10	question 11	question 12	question 13
Correlation	question 1	1.000	.484	.626	-.305	-.745	-.165	.461	-.340	.209	.071	-.441	.166	.187
	question 2	.484	1.000	.389	-.166	-.547	-.312	.381	-.176	.219	-.389	-.401	.116	.028
	question 3	.626	.389	1.000	-.348	-.743	-.209	.423	-.248	.328	.027	-.513	.165	.170
	question 4	-.305	-.166	-.348	1.000	.363	.323	-.596	.576	-.120	.102	.398	-.391	-.334
	question 5	-.745	-.547	-.743	.363	1.000	.260	-.536	.276	-.351	.130	.605	-.187	-.169
	question 6	-.165	-.312	-.209	.323	.260	1.000	-.268	.192	-.131	.217	.418	-.044	.001
	question 7	.461	.381	.423	-.596	-.536	-.268	1.000	-.608	.228	-.169	-.331	.347	.381
	question 8	-.340	-.176	-.248	.576	.276	.192	-.608	1.000	-.243	.087	.370	-.392	-.308
	question 9	.209	.219	.328	-.120	-.351	-.131	.228	-.243	1.000	-.109	-.407	.406	.286
	question 10	.071	-.389	.027	.102	.130	.217	-.169	.087	-.109	1.000	.250	-.059	-.082
	question 11	-.441	-.401	-.513	.398	.605	.418	-.331	.370	-.407	.250	1.000	-.148	-.006
	question 12	.166	.116	.165	-.391	-.187	-.044	.347	-.392	.406	-.059	-.148	1.000	.507
	question 13	.187	.028	.170	-.334	-.169	.001	.381	-.308	.286	-.082	-.006	.507	1.000

a. Determinant = 2.316E-03

Low correlations will not be in the same factor.

Should be greater than .0001. If close to zero, colinearity is too high.

Test of assumptions

Should be greater than .70

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.787
Bartlett's Test of Sphericity	Approx. Chi-Square	393.413
	df	78
	Sig.	.000

Should be significant (less than .05)

Communalities

	Initial
question 1	1.000
question 2	1.000
question 3	1.000
question 4	1.000
question 5	1.000
question 6	1.000
question 7	1.000
question 8	1.000
question 9	1.000
question 10	1.000
question 11	1.000
question 12	1.000
question 13	1.000

Extraction Method: Principal Component Analysis.

Notice these are greater than one. Eigenvalues refer to the variance explained or accounted for.

Total Variance Explained

Component	Initial Eigenvalues			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.805	36.963	36.963	3.207	24.666	24.666
2	1.826	14.049	51.011	2.327	17.898	42.564
3	1.333	10.255	61.267	1.887	14.514	57.078
4	1.133	8.718	69.985	1.678	12.907	69.985
5	.883	6.791	76.776			
6	.666	5.120	81.895			
7	.541	4.159	86.055			
8	.453	3.481	89.536			
9	.380	2.920	92.456			
10	.299	2.299	94.755			
11	.285	2.193	96.948			
12	.241	1.853	98.801			
13	.156	1.199	100.000			

Extraction Method: Principal Component Analysis.

Over 2/3 of the variance is accounted for by these four factors.

Percent of variance for each component after rotation.

Component Matrix[a]

a. 4 components extracted.

Rotated Component Matrix [a]

	Component			
	1	2	3	4
question 5	-.878			
question 3	.839			
question 1	.833			
question 11	-.596			.447
question 2	.559			-.551
question 4		.838		
question 8		.763		
question 7	.371	-.699		
question 12		-.323	.805	
question 13		-.315	.768	
question 9	.368		.662	
question 10				.853
question 6		.382		.560

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 7 iterations.

Factor weights or loadings are interpreted similarly to correlations.

Items in each component are sorted from highest factor weight to lowest.

Questions 11 and 2 "load" above .40 on factor 4 as well as factor 1.

Component Transformation Matrix

Component	1	2	3	4
1	.711	-.540	.343	-.292
2	-.495	-.410	.681	.352
3	.484	.193	.028	.853
4	.123	.709	.647	-.251

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

Output 10.2: Factor Analysis for Math Attitude Questions Limited to Three Factors

Syntax for factor analysis of math attitude questions limited to three factors

```
FACTOR
  /VARIABLES q01 q02 q03 q04 q05 q06 q07 q08 q09 q10 q11 q12 q13  /MISSING
  LISTWISE /ANALYSIS q01 q02 q03 q04 q05 q06 q07 q08 q09 q10 q11 q12 q13
  /PRINT INITIAL ROTATION
  /FORMAT SORT BLANK(.3)
  /CRITERIA FACTORS(3) ITERATE(25)
  /EXTRACTION PC
  /CRITERIA ITERATE(25)
  /ROTATION VARIMAX
  /METHOD=CORRELATION .
```

Factor Analysis

Communalities

	Initial
question 1	1.000
question 2	1.000
question 3	1.000
question 4	1.000
question 5	1.000
question 6	1.000
question 7	1.000
question 8	1.000
question 9	1.000
question 10	1.000
question 11	1.000
question 12	1.000
question 13	1.000

Extraction Method: Principal Component Analysis.

Total Variance Explained

	Initial Eigenvalues			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.805	36.963	36.963	3.289	25.300	25.300
2	1.826	14.049	51.011	2.843	21.869	47.169
3	1.333	10.255	61.267	1.833	14.097	61.267
4	1.133	8.718	69.985			
5	.883	6.791	76.776			
6	.666	5.120	81.895			
7	.541	4.159	86.055			
8	.453	3.481	89.536			
9	.380	2.920	92.456			
10	.299	2.299	94.755			
11	.285	2.193	96.948			
12	.241	1.853	98.801			
13	.156	1.199	100.000			

Extraction Method: Principal Component Analysis.

Component Matrix [a]

a. 3 components extracted.

Rotated Component Matrix [a]

	Component		
	1	2	3
question 5	-.878		
question 1	.851		
question 3	.845		
question 11	-.598		.484
question 12		.815	
question 13		.786	
question 8		-.682	
question 4		-.651	
question 7	.421	.611	
question 9	.303	.397	
question 10			.807
question 6			.651
question 2	.534		-.537

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization

a. Rotation converged in 5 iterations.

Component Transformation Matrix

Component	1	2	3
1	.729	.585	-.356
2	-.477	.806	.350
3	.492	-.086	.867

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

Output 10.3: Factor Analysis for Achievement Scores

Syntax for factor analysis of achievement scores

```
FACTOR
  /VARIABLES mathach visual mosaic grades  /MISSING LISTWISE /ANALYSIS
  mathach visual mosaic grades
  /PRINT UNIVARIATE INITIAL CORRELATION DET KMO EXTRACTION ROTATION
  /PLOT ROTATION
  /CRITERIA MINEIGEN(1) ITERATE(25)
  /EXTRACTION PC
  /CRITERIA ITERATE(25)
  /ROTATION VARIMAX
  /METHOD=CORRELATION .
```

Interpret Output 10.3

Compare Output 10.3 to your output and syntax in Output 10.1. Note that in addition to the tables in Output 10.1 you have: a) a table of descriptive statistics for each variable; it also provides the listwise *N* which you would not know otherwise; b) a table of communalities; c) a component matrix, which is unrotated and is used for purposes beyond the scope of this book; and d) plots of the factor loadings. Note that the default setting we used does not sort the variables by factors and does not suppress low loadings in the **rotated factor matrix**. Thus, you have to organize the table yourself; i.e., *mathach*, *grades*, and *visual* in that order are factor 1, and *mosaic* is factor 2.

Factor Analysis

Descriptive Statistics

	Mean	Std. Deviation	Analysis N
math achievement	12.5645	6.6703	75
visualization score	5.2433	3.9120	75
mosaic, pattern test	27.413	9.574	75
grades in h.s.	5.68	1.57	75

Correlation Matrix [a]

		math achievement	visualization score	mosaic, pattern test	grades in h.s.
Correlation	math achievement	1.000	.423	.213	.504
	visualization score	.423	1.000	.030	.127
	mosaic, pattern test	.213	.030	1.000	-.012
	grades in h.s.	.504	.127	-.012	1.000

a. Determinant = .562

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.468
Bartlett's Test of Sphericity	Approx. Chi-Square	41.414
	df	6
	Sig.	.000

This is not adequate because there is only one score (mosaic) to represent the second component. You should have several for each component.

Communalities

	Initial	Extraction
math achievement	1.000	.801
visualization score	1.000	.401
mosaic, pattern test	1.000	.959
grades in h.s.	1.000	.603

Extraction Method: Principal Component Analysis.

MORGAN George A. and Griego V. ORLANDO. 1998. *Easy Use and Interpretation of SPSS for Windows: Answering Research Questions With Statistics.* Mahwah: Lawrence Erlbaum Associates.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.755	43.873	43.873	1.755	43.873	43.873	1.717	42.928	42.928
2	1.009	25.237	69.110	1.009	25.237	69.110	1.047	26.183	69.110
3	.872	21.795	90.906						
4	.364	9.094	100.000						

Extraction Method: Principal Component Analysis.

Component Matrix [a]

	Component	
	1	2
math achievement	.894	3.309E-02
visualization score	.629	-6.96E-02
mosaic, pattern test	.267	.943
grades in h.s.	.699	-.339

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

Rotated Component Matrix [a]

	Component	
	1	2
math achievement	.864	.234
visualization score	.629	7.393E-02
mosaic, pattern test	4.740E-02	.978
grades in h.s.	.757	-.173

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

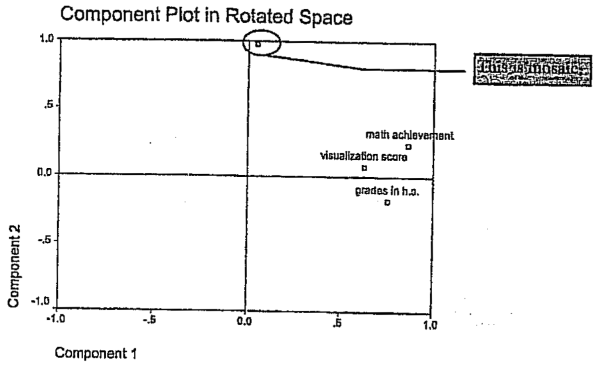
a. Rotation converged in 3 iterations.

Refer to the questions at the end of the assignment

Component Transformation Matrix

Component	1	2
1	.974	.225
2	-.225	.974

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.


Component Plot in Rotated Space
Component 2
Component 1
math achievement
visualization score
grades in h.s.
This is mosaic