to back with horizontal bars pointing to opposite directions, where each bar represents a five-year span. This type of histogram is called a **population pyramid**. In a population pyramid, the last class is left open. Usually it is the '80 years and more' class, as shown in Figure 3.10.

FREQUENCY POLYGONS AND DENSITY CURVES

If we join all the midpoints at the top of the columns in a histogram, we get what is called a **frequency polygon**. The polygon shows the general pattern of the distribution. Imagine now a frequency polygon drawn on a histogram with very large number of columns. We could redraw it as a smooth curve, called a **density curve** (Figure 3.11). A density curve is drawn in such a way that its surface is equal to 1. And if we look at the surface under the curve between any two values, it tells us the exact proportion of data that falls within these two values. We can now be more specific about the definition of the mode for a quantitative variable.

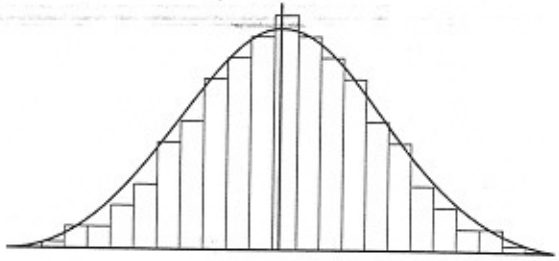


Figure 3.11 **A density curve can be thought of as the curve resulting from joining the midpoints at the top of the various bars of a histogram with a large number of classes**

If the variable is represented by a histogram, the **mode** is *the class with the highest frequency*. If it is represented by a density curve, the mode is *the x-value that corresponds to the highest point on the density curve*.

HISTOGRAM OR BAR CHART?

When we have a quantitative variable that has been grouped into a small number of categories, we can represent it either by a histogram or by a bar chart. But which of the two representations is better? It depends on what we want to convey. To explain this point, consider a situation where we have the variable Age represented by seven categories as shown in Figure 3.4. If we want to convey how the ages of the sample studied are distributed over the whole range of ages, the histogram shown in Figure 3.9 is better. But if we want to show how the various age groups are divided among men and women, or among married vs. unmarried individuals, a clustered bar chart allows us to do that, as shown in Figure 3.12. A histogram would not permit us to juxtapose corresponding categories of age groups for men and women.

In Figure 3.12, we see the distribution separately for men and women, and we can determine that women are more represented in the older categories, as they tend to live

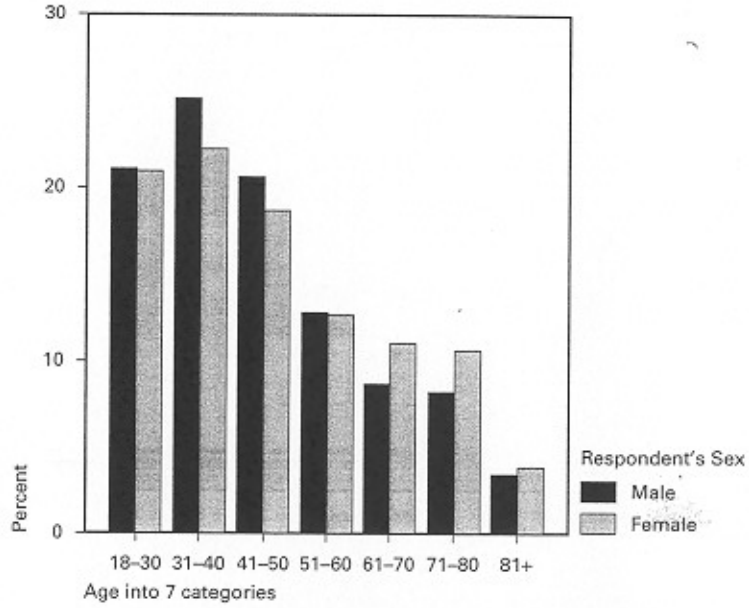


Figure 3.12 **A clustered bar chart allows us to show the pattern of ages separately for men and for women. It is appropriate for a quantitative variable grouped into a small number of categories**

longer than men. Notice that the vertical axis represents the percentages, not the frequencies. If we made it represent the frequencies instead, we would still get the same general shape, but we would not be able to determine whether men or women are more represented in a given class, as the overall number of women in this sample is greater than the overall number of men. In almost every age category, we would therefore find more women, not because a higher percentage of women (as opposed to men) fall into that category, but because there are more women in the sample as a whole.

Box plots

Box plots are very useful to show how the values of a *quantitative* variable are distributed. The box plot indicates the minimum and maximum values, and the three quartiles. The central 50% of the data (the 2nd and 3rd quarters) are represented as a shaded solid box, whereas the first and last quarters are represented by thin lines.

The box plot gives automatically the *five-number summary* of the data: the minimum, the 1st quartile, the median (which is the 2nd quartile), the 3rd quartile, and the maximum.

In symbols the **five-number summary** is given by: **Min**, $\mathbf{Q_1}$, **Median**, $\mathbf{Q_3}$, **Max**. The box plot is shown in Figure 3.13.

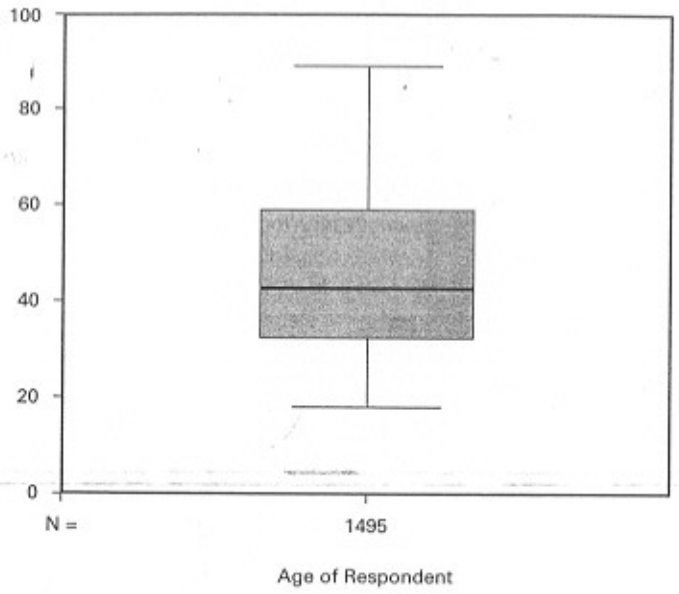


Figure 3.13 **The box plot representing the variable Age of respondent. We can read off directly the five-number summary of the distribution**

Box plots can also be used to represent several similar variables on the same graph, allowing comparisons. You could also split a population into several separate groups (such as men and women) and have a separate box plot for each group, drawn in the same graph, next to each other, to permit comparisons. This is illustrated in Figure 3.14 where five box plots of respondents' income are drawn, for the various groups defined by educational level. This figure illustrates clearly how the income varies with the highest level of education attained. We should note, however, that this data comes from a file where the income is not measured as a continuous scale variable, but is coded into 21 categories, and that the 22nd category is made up, as explained earlier in this chapter. More details are found in Lab 5.

Line Charts

Line charts are most useful to represent the variation of a quantitative variable over time. The X-axis represents the time line, and the Y-axis represents some quantitative variable. For example, the variable could be the number of students enrolled in a given program, or the inflation rate, or the market value of a given portfolio of stocks. The line chart would show how the variable increases or decreases as time goes by. A common mistake sometimes made intentionally consists in not showing

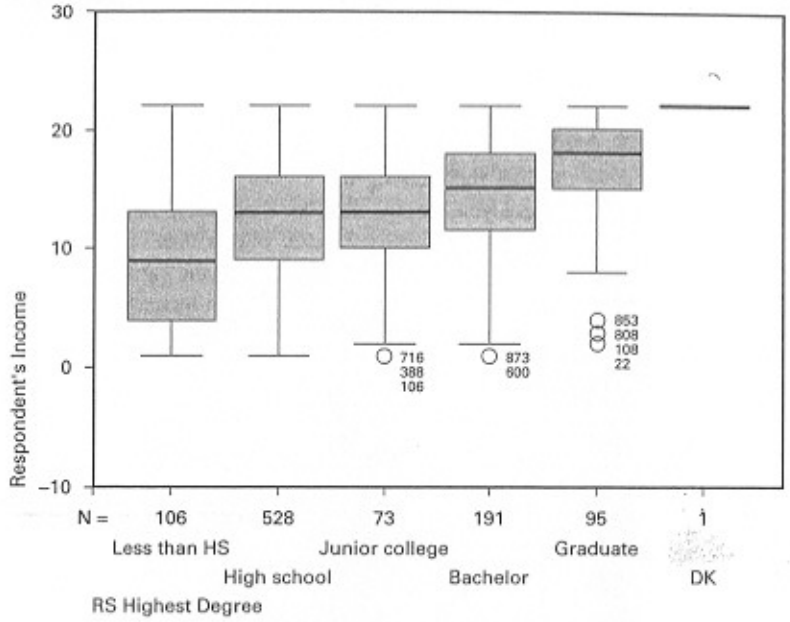


Figure 3.14 **Although the income is coded into 22 categories and not given as a dollar amount, the comparisons of the income for each educational level gives us a good idea of how incomes vary as a function of education**

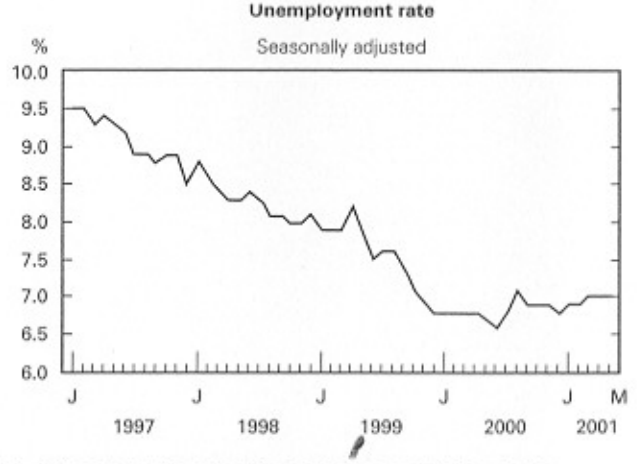


Figure 3.15 **A line chart showing the variation over time of the unemployment rate in Canada. The reader should be aware of the fact that the Y-axis does not start at zero, which may give the impression that the variations are greater than they really are. An awareness of this can guard us against misinterpretations. (Source: Statistics Canada)**

the zero level of the quantity, or in drawing the Y-axis shorter than it should be. This procedure has the effect of giving the impression that the variations are bigger than they really are, but at the same time it allows us to see the variations in the graph in greater detail. When it is necessary to show a shorter Y-axis, this should be indicated by an interruption in the line representing the Y-axis. Figure 3.15 provides an example of a line chart. Here you can see that the Y-axis does not start at zero, giving the impression that the variations are much bigger than they really are. However, this is justified by the fact that it allows us to see the variation in unemployment rates in great detail, and by the fact that this is an increasingly standard practice, which means that readers should be aware of the resulting distortion and interpret what they see accordingly.

The General Shape of a Distribution

In addition to the measures explained above, we could describe the general shape of the distribution of a quantitative variable by looking at two of its features: *symmetry* and *kurtosis*.

Symmetry

The first characteristic to look at is *symmetry*. A distribution is said to be **symmetric** if the mean splits its histogram into two equal halves, which are mirror images of each other. A typical symmetric distribution is the *normal distribution*. It is a bell-shaped distribution that follows a very specific pattern, and occurs in a wide range of situations. It is represented by the curve of Figure 3.16. It will be studied later on.

In a symmetric distribution, the mean and the median are equal. If the distribution is also unimodal, then the mean, the median, and the mode are all equal. This is true of normal distributions.

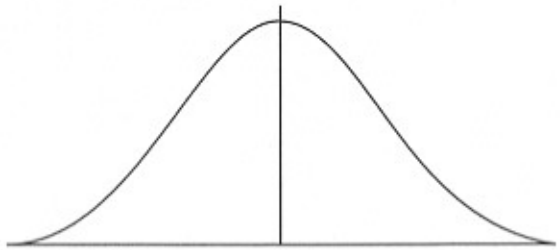


Figure 3.16 **An example of a symmetric distribution. This one is the normal distribution, which will be studied in Chapter 5**

If a distribution is symmetric and unimodal, the mean is a good representative of the center. However, it often happens that a distribution is not symmetric. We then say

that it is **skewed**. That means that one side of the graph of the distribution is stretched more than the other. We say that it is **positively skewed** if it is stretched on the right side, and **negatively skewed** if it is stretched on the left side. Figure 3.17 illustrates the difference between symmetric distributions and skewed distributions. SPSS allows you to compute a statistic called **skewness**, which is a measure of how skewed a distribution is. A normal curve has a skewness of 0. If the skewness is larger than 1, the shape starts to look significantly different from that of a normal curve.

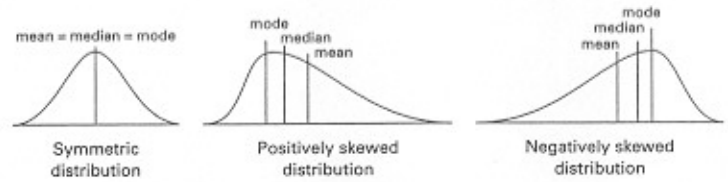


Figure 3.17 **Symmetric and skewed distributions**

How can we know that a distribution is skewed? The first indication is the histogram: the tail end of the histogram is longer on one side than on the other. We can also see that a distribution is skewed through its numerical features: the mean is different from the median. When the distribution is positively skewed, the mean is larger than the median, as it is pushed by the extreme values toward the longer tail. For negatively skewed distributions, the mean is smaller than the median. Therefore, a mean larger than the median tells us that the extreme values on the higher end of the distribution are much larger than the bulk of the data in the distribution, pulling the mean toward the positive side. This is illustrated by the numerical example given in the section on the median, where one extreme value (60) pulls the mean up but does not affect the median. Therefore, when a distribution is highly skewed, the median is usually a better representative of the center of the data than the mean.

Kurtosis

This is a measure of the degree of peakedness of the curve. It tells you whether the curve representing the distribution tends to be very peaked, with a high proportion of data entries clustered near the center, or rather flat, with data spread out over a wide range. A normal distribution has a kurtosis equal to 0. A positive value indicates that the data is clustered around the center, and that the curve is highly peaked. A negative value indicates that the data is spread out, and that the curve is flatter than a normal curve. Figure 3.18 shows three curves with zero, positive, and negative kurtosis respectively.

Methodological Issues

Although they seem to be simple, descriptive measures can be tricky to use. We would like to point out here some of the pitfalls and difficulties associated with their use.

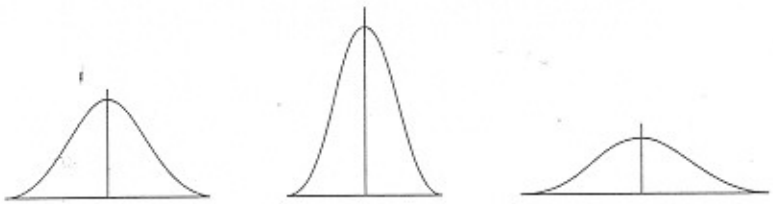


Figure 3.18 **Illustration of zero, positive and negative kurtosis**

The Definition of the Categories over which the Counting is Done

Suppose I say that the passing rate in a given class is 82%. In another college, a colleague tells me that his passing rate is 95%. Before concluding that his passing rate is much higher, I have to make sure that we are defining the passing rate in the same way. I may define the passing rate as the number of students who pass a course compared to those who were registered at the beginning of the semester. If he defines it the same way, we can make meaningful comparisons. But if he defines it as the number of students who pass the course compared to the number registered at the end of the semester, we cannot make a meaningful comparison. This is so because all the students who dropped out would not be taken into account in his calculation, whereas they would be taken into account in mine. A careful definition of the categories used to define a concept is therefore important. Such problems arise when we define the unemployment rate in various countries, or even wealth. The conclusion is that careful attention should be given to the way categories are defined when comparing the statistics that refer to different populations.

Outliers

Outliers are values that are unusually large or unusually small in a distribution. They have to be examined carefully to determine if they are the result of an error of measurement, or a typing error, or whether they actually represent an extreme case. For instance, the value 69 in the column of the variable **age** for college students could be a typing error, but it could also represent the interesting case of a retired person who decided to pursue a college program. Even if they represent an extreme case, it may be desirable to disregard extreme values in some of the statistical computations.

When producing a Box Plot diagram, SPSS excludes the outliers from the computation, and prints them above or below the box plot. An option allows users to have the case number printed next to the dot representing the outlier, so as to be able to identify the case and examine it more closely.

Summary

We have seen in this chapter the various measures used to summarize the data pertaining to a single variable as well as the various types of charts that could be used to illustrate the distribution. You should keep in mind one fundamental point: *the level of measurement used for the variable determines which measures and graphs are appropriate*. It does not make sense, for example, to compute the mean of the variable when the level of measurement is nominal, that is, when the variable is qualitative.

There are three types of univariate descriptive measures:

- measures of central tendency,
- measures of dispersion, and
- measures of position.

Measures of central tendency, also called measures of the center, tell us the values around which most of the data is found. They give us an order of magnitude of the data, allowing comparisons across populations and subgroups within a population. They include the mean, the median, and the mode. The mean should not be used when the variable is qualitative.

Measures of dispersion are an indication of how spread out the data is. They are mostly used for quantitative data. The most important ones are the range, the interquartile range, the variance, and the standard deviation.

Measures of position tell us how one particular data entry is situated in comparison to the others. The percentile rank is one such measure. Other measures include the quartiles and the deciles.

In addition to these measures, we have seen the weighted mean. When calculating it, the various entries are multiplied by a weight, which is a positive number between 0 and 1. All the weights add up to 1. The weighted mean is used when the numbers that are averaged have been calculated over populations of unequal size. For instance, if you have the birth rates in all Canadian provinces and you want to find the average birth rate for Canada as a whole, you must weight these numbers by the demographic importance of every province. The weighted mean is also used when you want to increase or decrease the relative importance of the numbers you are averaging, as is done when finding the average grade over exams that do not count for the same percentage in the final grade.

When categories are involved (either because the variable is qualitative, or when quantitative values have been grouped) we can find ratios, percentages, and proportions of the groups corresponding to the categories.

The general shape of a distribution is analyzed in terms of symmetry or skewness, and in terms of kurtosis (the degree to which the curve is peaked).

The comparison of the mean and the median is very useful. Recall the following:

> If the distribution is very skewed, the median is a better representative of the center of the data, as the extreme values tend to pull the mean towards one side of the curve. The median is not affected by extreme values.
>
> If the mean is larger than the median, the distribution is positively skewed. If the mean is smaller than the median, the distribution is negatively skewed.

As for the graphical representation of a distribution, recall again that the level of measurement of the variable determines what kind of chart is appropriate. Bar charts and pie charts are appropriate when the data is qualitative, or measured at the nominal or ordinal levels. Quantitative data (whether measured at the ordinal of numerical scale levels) could also be represented by bar charts or pie charts if the values have been grouped into a small number of categories.

The essential difference between pie charts and bar charts is that in the former, the emphasis is on the relative importance of each category as compared to the other categories, whereas in the latter, the emphasis is on the size of each category. However, there is no clear-cut distinction between the two, and if one is appropriate, the other is usually appropriate also, even if the emphasis is slightly different. The great advantage of bar charts is that it allows making comparisons between the distributions of subgroups, with the help of clustered bar charts.

Quantitative variables are better represented through histograms. A specific type of histogram is the population pyramid, which is a standard tool in demography.

Line charts are most suited to represent the variation of a quantity across time.

In all kinds of charts, truncating the Y-axis is sometimes done to zoom in on the variations of the variables and to represent them in a more detailed way. However, we should be aware of the fact that truncating the Y-axis may also convey a mistaken impression that the variations of the variable are more important than they are in reality.

Keywords

Univariate	Frequencies	Ratios
Bivariate	Cumulative frequencies	Proportions
Measures of central tendency	Valid percent	Bar graph
Measures of dispersion	Range	Clustered bar graph
Measures of position	Trimmed range	Pie chart
Mean	Interquartile range	Histogram
Trimmed mean	Deviation from the mean	Frequency polygon
Weighted mean	Standard deviation	Line chart
Median	Variation ratio	Box plot
Mode	Coefficient of variation	Five-number summary
Modal category	Quartiles	Symmetry
Majority	Deciles	Skewness
Plurality	Percentiles	Kurtosis
	Percentile rank	Outliers

Suggestions for Further Reading

Devore, Jay and Peck, Roxy (1997) ***Statistics, the Exploration and Analysis of Data*** (3rd Edn). Belmont, Albany: Duxbury Press.

Harnett, Donald H. and Murphy, James L. (1993) ***Statistical Analysis for Business and Economics***. Don Mills, Ontario: Addison-Wesley Publishers.

Trudel, Robert and Antonius, Rachad (1991) ***Méthodes quantitatives appliquées aux sciences humaines***, Montréal: CEC.

Wonnacott, Thomas H. and Wonnacott, Ronald J. (1977) ***Introductory Statistics*** (3rd edn). New York: John Wiley and Sons.

EXERCISES

3.1 Complete the following sentences:

(a) Three types of measures are useful to summarize a numerical distribution. They are

_____,
_____ and
_____.

(b) The most frequent value in a distribution is called _____.

(c) When the values of the distribution are grouped into classes, the mode is the _____ with the highest frequency.

(d) When there are two classes that are bigger than the ones immediately next to them, the distribution is called _____.

(e) If the modal class includes more than 50% of the population, we say that it constitutes the _____. Otherwise, we simply talk of a _____.

(f) The median falls _____ of the ordered list of entries. _____% of the data are less than or equal to the median, and _____ % are larger than or equal to it.

(g) The mean of a numerical distribution is equal to the _____ of all entries divided by _____.

(h) The mathematical measure used to find the mean when the entries do not have the same relative importance is called _____.

4 WRITING A DESCRIPTIVE SUMMARY

The purpose of this chapter is to explain how to proceed in order to write a good descriptive report, and how to analyze a frequency table beyond a first-level reading of the percentages, in order to identify the numerical features of the data and to highlight them.

After studying this chapter, the student should know:

- how to proceed when writing a descriptive report to summarize data;
- which measures and charts are appropriate, depending on the measurement level of the variable;
- how to summarize a set of variables that measure a given concept;
- how to analyze a frequency table in detail and identify its important features;
- the difference between a first-level description and an analytical description;
- the criteria for a good descriptive summary.

In Chapter 3, we have seen how to produce simple descriptive statistical measures, as well as simple tables and graphs. We have also seen that the statistical measures to be used depend on the level of measurement of the variable. Now, we would like to see how we can integrate all these elements and produce a synthetic report that describes certain features of a population. For the time being, we will restrict these explanations to univariate descriptions of variables. Later on, you will have to include bivariate descriptions, that is, descriptions of the statistical associations between variables, as well as confidence statements, that is, generalizations from the observed sample to the population as a whole, two statistical topics studied later on in this book. We will also learn how to report the result of a hypothesis testing.

How to Write a Descriptive Report

We will consider two types of report. Basic reports consist in a direct reading of the tables produced by SPSS, and a reformulation in direct, plain language of what the tables say, with accompanying charts as illustrations. There is very little interpretation

in this case. A second level in sophistication consists in writing analytical reports: such reports would highlight the outstanding tendencies that can be seen in the data, and may include a greater degree of interpretation. We will now explore both kinds of reports.

Basic, Direct Reports

Suppose you want to describe the educational level of the individuals included in the **GSS93 subset** data file supplied with the SPSS package. This means that you would like to have some global description that tells you whether the people in your sample tend to have a high level of education or not (this is a description of the central tendency), and whether there is a big polarization, with some people having a lot of education and many others very little (this is a description of the dispersion).

The first thing to do is to see which variables concern education. You will find three such variables in the **GSS93 subset** data file. List them, and list the level of measurement of each.

In this data file, you will find that the three variables are:

- Highest year of schooling completed (scale),
- Highest degree obtained (ordinal, 5 categories), and
- Possession or not of a college degree (ordinal, 2 categories).

Determine what kind of descriptive measures you would use for each. Would you use a frequency table? For which of the variables? Which charts would be more appropriate?

Sometimes you will feel that you are not too sure which type of chart is appropriate. Get SPSS to produce several charts, examine them carefully to see which ones convey a better representation of the distribution of the variable, then select one of them, and paste it into your report.

One of the important pitfalls that you should avoid is to give a lot of tables or charts that are not very useful. You may want to be selective here: select the relevant information, and try to write it in a clear and concise way. For example, SPSS produces tables giving you the number of valid answers. You do not need to include the table itself. You could simply write in brackets ($n = 1500$) when describing the sample, to indicate that your sample contains 1500 individuals. Whenever you discuss or describe the results that relate to one of the variables, if you see that there are a lot of missing answers, add a phrase about the number of valid answers, such as (valid $n = \ldots$) and fill in the number of valid answers. Although the number of people in the sample is the same throughout the analysis of this data file ($n = 1500$), the number of valid answers varies a lot. This is why

you have to specify how many *valid* answers you have to a particular question. You do not have to do that for every single question: you report the number of valid answers only when there is a lot of missing data, and the valid percentages differ by several points from the total percentages. It is advisable in this case to report the valid percentages. In some cases it may be relevant to report both the valid and total percentages.

What follows is a set of criteria that define a good descriptive report.

Criteria for a Good Report

THE GENERAL PRESENTATION

Make sure the text is clear, well organized, and concise. If the analysis is long, a cover page may be desirable. Make sure that all the relevant information is in it: a title, your name, the name of the course and the course number, the name of the instructor to which you are presenting it, and the date.

Some of this information, such as your name and the assignment number, could be written in the header of your document (refer to Lab 2 for explanations on the header). The tables and graphs must be printed with the correct identification: a title must be given to every table or graph. If you copy the tables from SPSS with the **Copy**... command (rather than the **Copy Object**... command), you can edit the table, and delete the rows or columns that are not useful or relevant. Also avoid grammatical mistakes: a spell check may be useful, but rely always on a careful reading of your report.

Include in your report a **description of the data file** you are using: its source, the year the survey was conducted, the kind of variables that are found in it, the institution under which it was conducted, etc.

DESCRIPTION OF THE VARIABLES UNDER STUDY

Make sure to include in your study all the variables that are relevant for your subject. If there are several variables that address a given topic, use them all to analyze this topic. For instance, 'education' can be measured in several ways. If there are several variables that deal with education, examine the distribution of each.

To describe a variable properly, you must select the appropriate measures. Do not compute the mean of a qualitative variable, because it is meaningless. You may want to use some of the recoded variables, or recode some variables yourself. Do not include a table of frequencies if the variable is quantitative. Such tables are usually quite long, and they are not useful to the reader. If the quantitative variable has been grouped into a small number of categories, a frequency table may be useful, in addition to the descriptive measures used for quantitative variables. Finally, formulate your conclusions in full, grammatically correct sentences that highlight the meaning of your numerical results. An example of a very concise description of the educational level of the people in our sample is given in Insert 4.1.

The appropriate measures to be used are summarized in Table 4.1.

Table 4.1 **Appropriate descriptive measures for the various levels of measurement**

Level of Measurement	Appropriate Statistical Measures	Appropriate Charts
Nominal (categories)	Frequencies, percentages, mode. Ratios, proportions and rates.	Bar charts, pie charts
Ordinal	Frequencies; mode; median. Cumulative frequencies. (If there are many categories, you may compute the mean and median, but the interpretation of the numerical results may be problematic.)	Bar charts; histograms
Numerical scale, ungrouped	Mean, median, mode, range, minimum, maximum standard deviation, interquartile range. (Frequency tables are not useful for this type of measure.)	Histograms, frequency polygons, box plots, time lines
Numerical scale, grouped	Frequency tables, mode. If there are a large number of groups: mean and standard deviation. The mean is usually the mean code of the categories. It can be used for comparative purposes if other samples are grouped in the same way, but it should not be mistaken for the mean of the variable itself. If grouped into a small number of categories, it should be treated like ordinal data.	Histograms, bar charts, pie chart. Box plots may be misleading if the number of categories is small.

Examples of Concise Descriptive Reports

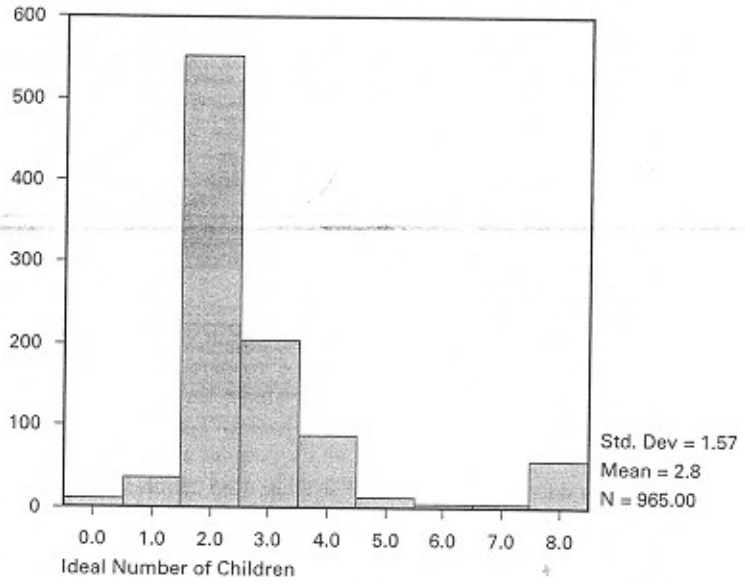
What follows (Insert 4.1) is an example of a short descriptive report, which answers the question: Describe the educational level of the sample given in the file **GSS93 subset** that comes with the SPSS program.

INSERT 4.1 Descriptive report of the educational level of the sample

The data set used here is a subset of the General Social Survey conducted in the US in 1993 ($n = 1500$). There are three variables in this data set that address the issue of education: the highest year of schooling completed (scale), the highest degree obtained (ordinal, 5 categories) and the possession or not of a college degree (ordinal, 2 categories).

The *average* highest year of schooling completed is 13 years with a standard deviation close to 3 years. The graph below shows the distribution of this variable.

If we compare that situation with the Ideal number of children, we see that the mean for that variable is 2.76 children, but the comparison with the actual number of children is difficult to make, as there are 535 missing answers for that variable (we can assume that only those who had children were asked that question). It is better to examine the histogram of the ideal number of children. Here we see that the mode, or most desirable situation, is by far the situation with two children. Very few people think that one child is the ideal situation.



6. Spanking Children

We have answers for 66% of the respondents, and the rest of the answers are missing. Of those who answered, about three-quarters (73.3%) indicated they either agree or strongly agree with spanking children as a disciplinary measure, while the rest (26.7%) disagree or strongly disagree.

7. Number of Siblings

We see here that the average is 3.7 brothers and/or sisters. If we examine the cumulative frequencies, we see that 60.2% of the respondents come from families of 4 children or less (the respondent plus 3 brothers or sisters), the rest (almost 40%) coming from families with 5 children or more. Comparing that with the number of children people currently have, we see that in general, individuals come from families that are larger than the families they themselves establish, since the average number of children in this sample tends to be much smaller than the number of brothers or sisters respondents have.

Analytical Descriptive Reports

The examples shown above are quite direct, and consist essentially in reporting, almost as is, the information provided in the frequency tables. But a more analytical view would permit a richer reading of such tables. To illustrate what is meant by that we will go into a more detailed – and more analytical – reading of frequency tables.

EXAMPLES OF HOW TO ANALYZE A FREQUENCY TABLE

To make our point clear, we are going to analyze four cases of the same situation, represented by the tables below. They all deal with the frequencies of the variable Political Party Affiliation, taken from the **GSS93 subset** file. The first table is the one that we get from the actual data in this file. The other three have been modified to illustrate how the analysis can highlight the distribution pattern.

Table 4.2 **Political Party Affiliation A**

	Frequency	Percent	Valid Percent
Strong Democrat	213	14.2	14.3
Not Str Democrat	298	19.9	20.0
Ind, Near Democrat	180	12.0	12.1
Independent	187	12.5	12.5
Ind, Near Republican	148	9.9	9.9
Not Str Republican	280	18.7	18.8
Strong Republican	168	11.2	11.3
Other Party	17	1.1	1.1
Total valid	1491	99.4	100.0
NA	9	.6	
Total	1500	100.0	

Case A Analysis of Case A (Table 4.2). We see from the table that those who are affiliated with the Democrats (strongly or not strongly) add up to 34.3%, or slightly more than a third. Those who are affiliated with the Republicans add up to 30.1%, or slightly less than a third. The independents add up to 34.5, again a little more than a third. It is interesting to note that the population is almost evenly divided into three groups, and that those who affiliate to neither party are as numerous (or a little more numerous) than those who affiliate with either of the two main parties. We can also notice that, within each of the two main parties, those who do not have a strong affiliation with the party are more numerous than those who have a strong affiliation (for the Republicans: 280:168, or about 7:4, and for the Democrats, 298:213, or about 3:2). The bar chart shown in Figure 4.1 illustrates this situation.

Case B Analysis of Case B (Table 4.3). We see from the table that those who affiliate with the Democrats add up to 42.1%. Those who are affiliated with the Republicans add up to 39.1%, or slightly less than the Democrats. The independents add up only to 17.6%, indicating that there is a strong polarization between the two

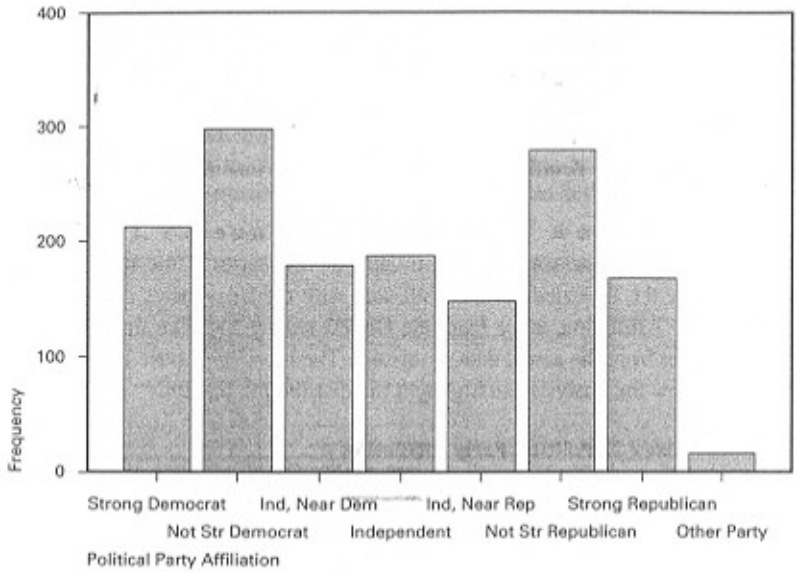


Figure 4.1 **Political Party Affiliation**

Table 4.3 **Political Party Affiliation B**

	Frequency	Percent	Valid Percent
Strong Democrat	272	18.2	18.2
Not Str Democrat	356	23.7	23.9
Ind, Near Democrat	122	8.1	8.2
Independent	57	3.8	3.8
Ind, Near Republican	84	5.6	5.6
Not Str Republican	351	23.4	23.5
Strong Republican	232	15.5	15.6
Other Party	17	1.1	1.1
Total valid	1491	99.4	100.0
NA	9	.6	
Total	1500	100.0	

parties, with less than 1 person out of 5 not affiliated to one of these two parties. We can also notice that, within a party, those who are not strongly affiliated with the party are more numerous than those who are (for the Republicans 23.4% vs. 15.5%, or a ratio of about 3:2, and for the Democrats 23.7% vs. 18.1%, or a ratio of about 4:3). The bar chart in Figure 4.2 illustrates this situation, and the polarization between the two parties is clearly visible.

Case C Analysis of case C (Table 4.4). We see from the table that those who are affiliated with the Democrats add up to 35.6%, or slightly more than a third. Those

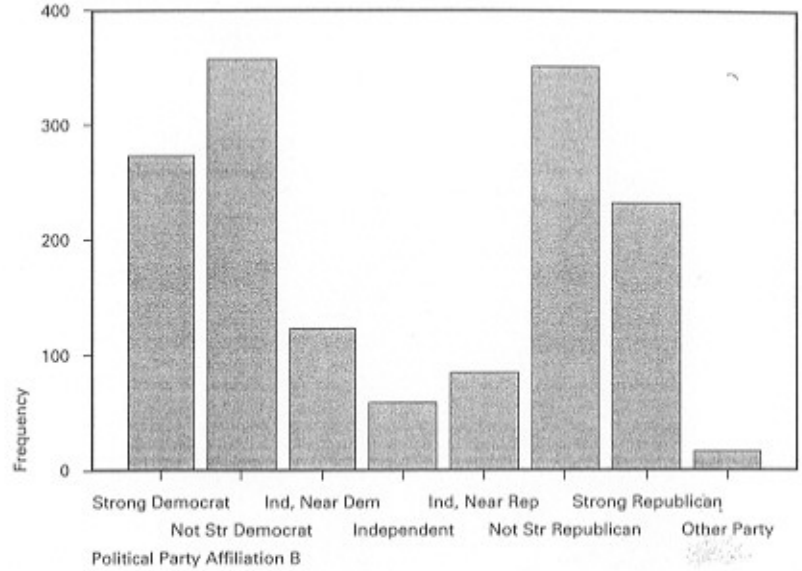


Figure 4.2 **Political Party Affiliation B**

Table 4.4 **Political Party Affiliation C**

	Frequency	Percent	Valid Percent
Strong Democrat	292	19.5	19.6
Not Str Democrat	236	15.7	15.8
Ind, Near Democrat	188	12.5	12.6
Independent	93	6.2	6.2
Ind, Near Republican	165	11.0	11.1
Not Str Republican	233	15.5	15.6
Strong Republican	267	17.8	17.9
Other Party	17	1.1	1.1
Total valid	1491	99.4	100.0
NA	9	.6	
Total	1500	100.0	

who are affiliated with the Republicans add up to 33.5%, or about a third. The independents add up to 29.9%. Thus, the population is almost evenly split between the three groups, with the Democrats only slightly ahead of the Republicans. Notice that, within each party, those who are strongly affiliated with the party are more numerous than those who are not (a ratio of 4:3 for the Democrats, and a ratio of 6:5 for the Republicans). This is illustrated in Figure 4.3.

Case D Analysis of case D (Table 4.5). We see from the table that this is a situation of weak polarization between the Republicans and the Democrats. The Democrats attract 42.8% of the population, while the Republicans only get 30% of the

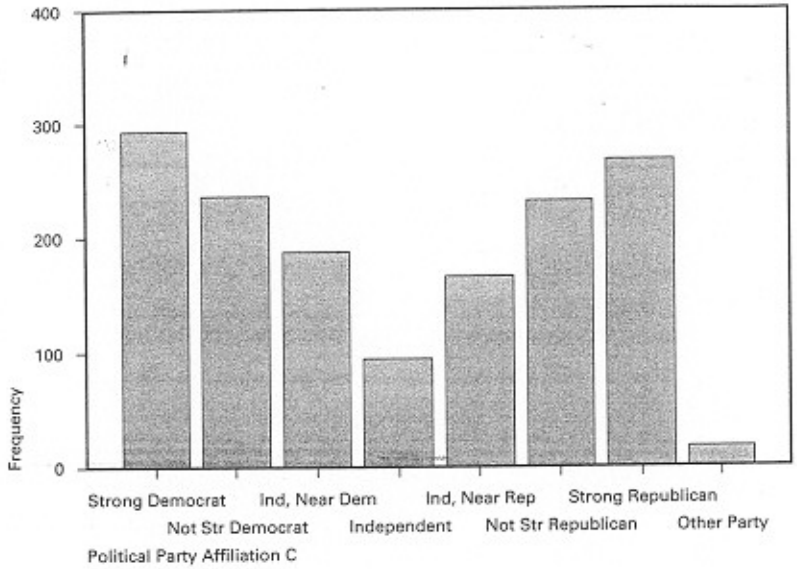


Figure 4.3 **Political Party Affiliation C**

support, almost 13 points behind the Democrats. The independents add up to 26.0% of the population. Notice that, within each party, those who are strongly affiliated with the party are the majority, with a ratio of about 4:3 for the Democrats and about 5:4 for the Republicans, a situation illustrated by Figure 4.4.

Table 4.5 **Political Party Affiliation D**

	Frequency	Percent	Valid Percent
Strong Democrat	356	23.7	23.9
Not Str Democrat	282	18.8	18.9
Ind, Near Democrat	188	12.5	12.6
Independent	116	7.7	7.8
Ind, Near Republican	84	5.6	5.6
Not Str Republican	202	13.5	13.5
Strong Republican	246	16.4	16.5
Other Party	17	1.1	1.1
Total valid	1491	99.4	100.0
NA	9	.6	
Total	1500	100.0	

As we have seen, the short descriptive paragraphs that follow each table do not simply report the frequencies. We have tried to highlight the specific features of each situation by answering the following questions: Is there a polarization

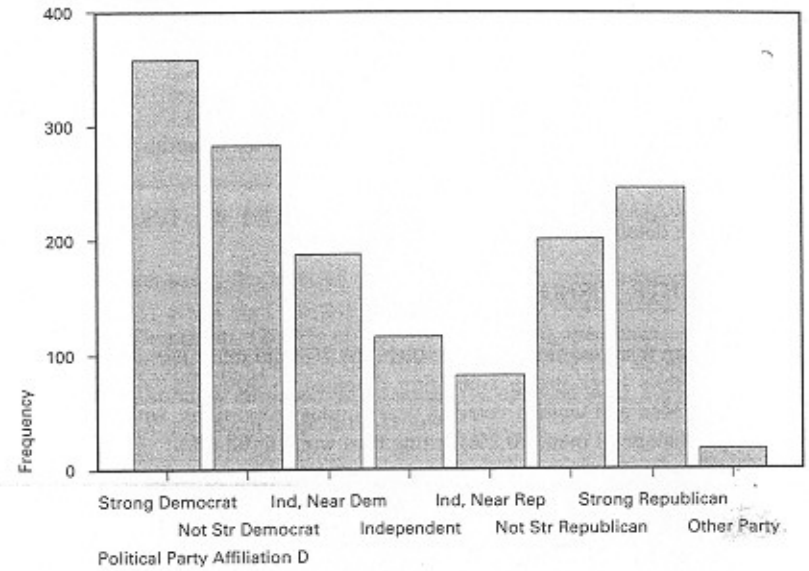


Figure 4.4 **Political Party Affiliation D**

between the two parties? Is one of them clearly more popular than the other? Is there a large proportion of independents? How is the level of mobilization *within* each party? We answered that last question by providing the ratio of those who feel a strong affiliation to the party compared to those who do not feel a strong affiliation.

A descriptive report that does that systematically is more analytical than one where the percentages are flatly reported as is. Insert 4.3 illustrates such a report.

INSERT 4.3 Description of the Voting Behaviour and of the Political Tendencies of a Sample of US Residents

The data summarized here come from a (non-representative) sample of 1500 individuals, which is a subset of the General Social Survey conducted in the US in 1993.

Four variables deal with our topic: Voting in 1992 Election, Political Party Affiliation, Think of self as Liberal or Conservative, and Political outlook. All four variables are measured at the nominal level. An examination of the frequency tables shows that the last variable is a recode of the third one, as explained below.