STATISTICAL METHODS FOR CATEGORICAL DATA ANALYSIS: 2ND EDITION

DANIEL A. POWERS

*Department of Sociology and Population Research Center,
University of Texas at Austin,
Austin, Texas, USA*

YU XIE

*Department of Sociology, Department of Statistics, and
Institute for Social Research,
University of Michigan,
Ann Arbor, Michigan, USA*

Emerald

United Kingdom • North America • Japan
India • Malaysia • China

Chapter 1

Introduction

1.1. Why Categorical Data Analysis?

What is common about birth, marriage, schooling, employment, occupation, migration, divorce, and death? The answer: they are all categorical variables commonly studied in social science research. In fact, most observed outcomes in social science research are measured categorically. If you are a practicing social scientist, chances are good that you have studied a phenomenon involving a categorical variable. (This is true even if you have not used any special statistical method for handling categorical data.) If you are in a graduate program to become a social scientist, you will soon, if not already, encounter a categorical variable. Notice that even our statement of whether or not you have encountered a categorical variable in your career is itself a categorical measurement!

Statistical methods and techniques for categorical data analysis have undergone rapid development in the past 25 years or so. Their applications in applied research have become commonplace in recent years, due in large part to the availability of commercial software and inexpensive computing. Since some of the material is rather new and dispersed among several disciplines, we believe that there is a need for a systematic treatment of the subject in a single book. This book is aimed at helping applied social scientists use special tools that are well suited for analyzing categorical data. In this chapter, we will first define categorical variables and then introduce our approach to the subject.

1.1.1. *Defining Categorical Variables*

We define categorical variables as those variables that can be measured using only a limited number of values or categories. This definition distinguishes categorical variables from continuous variables, which, in principle, can assume an infinite number of values.

Although this definition of categorical variables is clear, its application to applied work is far more ambiguous. Many variables of long-lasting interest to social scientists are clearly categorical. Such variables include: race, gender, immigration status, marital status, employment, birth, and death. However, conceptually continuous variables are sometimes treated as continuous and other times as categorical. When a continuous variable is treated as a categorical variable, it is called *categorization* or *discretization* of the continuous variable. Categorization is often necessary in practice because either the substantive meaning or the actual measurement of a continuous variable is categorical. Age is a good example. Although conceptually continuous, age is often treated as categorical in actual research for substantive and practical reasons. Substantively, age serves as a proxy for qualitative states for some research purposes, qualitatively transforming an individual's status at certain key points. Changes in legal and social status occur first during the transition into adulthood and later during the transition out of the labor force. For practical reasons, age is usually reported in single-year or five-year intervals.[1]

Indeed, our usual instruments in social science research are crude in the sense that they typically constrain possible responses to a limited number of possible values. It is for this reason that we earlier stated that most, if not all, observed outcomes in social science are categorical.

What variables should then be considered categorical as opposed to continuous in empirical research? The answer depends on many factors, two of which are their substantive meaning in the theoretical model and their measurement precision. One requirement for treating a variable as categorical is that its values are repeated for at least a significant portion of the sample.[2] As will be shown later, the distinction between continuous and categorical variables is far more consequential for response variables than for explanatory variables.

1.1.2. Dependent and Independent Variables

A *dependent* (also called response, outcome, or endogenous) variable represents a population characteristic of interest being explained in a study. *Independent* (also called explanatory, predetermined, or exogenous) variables are variables that are used to explain the variation in the dependent variable. Typically, the characteristic of interest is the population mean of the dependent variable (or its transformation) *conditional* on values of an independent variable or set of independent variables. It is in this sense that we mean that the dependent variable depends on, is explained by, or is a function of independent variables in regression-type statistical models.

By *regression-type statistical models*, we mean models that predict either the expected value of the dependent variable or some other characteristic of the dependent variable, as a regression function of independent variables. Although in principle we could design our models to best predict any population parameter (e.g., the median) of the dependent variable or its transformation, in practice we commonly use the term *regression* to denote the problem of predicting conditional means. When the regression function is a linear combination of independent variables, we have so-called linear regressions, which are widely used for continuous dependent variables.

1.1.3. Categorical Dependent Variables

Although categorical and continuous variables share many properties in common, we wish to highlight some of the differences here. The distinction between categorical and continuous variables as dependent variables requires special attention. In contrast, the distinction is of relatively minor significance when they are used as independent variables in regression-type statistical models. Our definition of regression-type statistical models includes statistical methods for the analysis of variance and covariance, which can be represented by regressing the dependent variable on a set of dummy variables and, in the case of the analysis of covariance, other continuous covariates. Hence, including categorical variables as independent variables in regression-type models does not present any particular difficulties, as it mainly involves constructing dummy variables corresponding to different categories of the independent variable; all known properties of regression models are directly generalizable to models for the analysis of variance and covariance. As we will show later in this book, the situation changes drastically when we treat categorical variables as dependent variables, as much of our knowledge derived from linear regressions is simply inapplicable. In brief, special statistical methods are required for categorical data analysis (i.e., analysis involving categorical *dependent* variables).

Although the methods for analyzing categorical variables as independent variables in regression-type models have been a part of the standard statistical knowledge base that is now required for most advanced degrees in social science, methods for the analysis of categorical dependent variables are much less widely known. Much of the fundamental research on the methodology of analyzing categorical data has been developed only recently. We aim to give a systematic treatment of several important topics on categorical data analysis in this book so as to facilitate the integration of the material into social science research.

Unlike methods for continuous variables, methods for categorical data require close attention to the type of measurement of the dependent variable. Methods for analyzing one type of categorical dependent variable may be inappropriate for analyzing another type of variable.

1.1.4. Types of Measurement

The type of measurement plays a key role in determining the appropriate method of analysis when a variable is used as a dependent variable. We present a typology for

[1] Education is another example. The substantive distinctions among "less than 12 years of schooling," "high-school diploma," "college degree," or "graduate degree" cannot be captured without categorization. A few categories offer a concise representation of the important points in the distribution of education.

[2] Note that a continuous variable can be truncated, meaning that it has zero probability of yielding a value beyond a particular threshold or cut-off point. When a continuous variable is truncated, the untruncated part is still continuous, whereas the part that is truncated resembles a categorical variable.

four types of measurement based on three distinctions.[3] First, let us distinguish between *quantitative* and *qualitative* measurements. The distinction between the two is that quantitative measurements closely index the substantive meanings of a variable with numerical values, whereas numerical values for qualitative measurements are substantively less meaningful, sometimes merely as classifications to denote mutually exclusive categories of characteristics (or attributes) uniquely. Qualitative variables are categorical variables.

Within the class of *quantitative* variables, it is often useful to distinguish further between *continuous* and *discrete* variables. Continuous variables, also called interval variables, may assume any real value. Variables such as income and socioeconomic status are typically treated as continuous over their plausible range of values. Discrete variables may assume only integer values and often represent event counts. Variables such as the number of children per family, the number of delinquent acts committed by a juvenile, and the number of accidents per year at a particular intersection are examples of discrete variables. According to our earlier definition, discrete (but quantitative) variables are also categorical variables.

Qualitative measurements can be further distinguished between ordinal and nominal. Ordinal measurements give rise to ordered qualitative variables, or *ordinal* variables. It is quite common to use numerical values to denote the ordering information in an ordered qualitative variable. However, numerical values corresponding to categories of ordinal variables reflect only the ranking order in a particular attribute; therefore, distances between two adjacent values are not the same. Attitudes toward gun control (strongly approve, approve, neutral, disapprove, and strongly disapprove), occupational skill level (highly skilled, medium skilled, low skilled, and unskilled), and the classification of levels of education as (grade school, high school, college, and graduate) are examples of ordinal variables.

Nominal measurements yield unordered qualitative variables, often referred to as *nominal* variables. Nominal variables possess no inherent ordering, nor numerical distance, between category levels. Classifications of race and ethnicity (white, black, Hispanic, and other), gender (male and female), and marital status (never married, married, divorced, and widowed) are examples of unordered qualitative variables. It is worth noting at this point, however, that the distinction between ordinal and nominal variables is not always clear-cut. Much of the distinction depends on the research questions. The same variable may be ordinal for some researchers but nominal for others.

To further illustrate the last point, let us use occupation as an example. Distinct occupations are often measured by open-ended questions and then manually coded into a classification system with three-digit numerical codes that do not represent magnitudes in substantive dimensions. Since the number of potential occupations is large (usually at least a few hundred in a coding scheme for a modern society), it is desirable, and indeed necessary, to reduce the amount of detail in an occupational measure through data reduction. One method of data reduction is to collapse detailed occupational codes into major occupational categories and treat them as constituting either an ordinal or even a nominal measurement (Duncan, 1979; Hauser, 1978). Another method of data reduction is to scale occupations along the dimension of a socioeconomic index (SEI) (Duncan, 1961) — thus into an interval variable. More recently, Hauser and Warren (1997) challenged Duncan's approach and suggested instead that to measure occupational socioeconomic status, occupations are best scaled into two separate dimensions of occupational income and occupational education. Hauser and Warren's work illustrates the importance of considering multiple dimensions when nominal measures are scaled into interval measures.

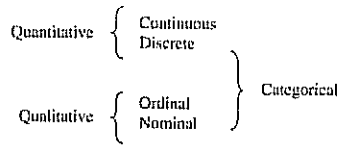

Figure 1.1: Typology of the four types of measurements.

Figure 1.1 summarizes our typology scheme for the four types of measurements. According to this typology, there are three types of categorical variables: discrete, ordinal, and nominal, all of which will be discussed in this book. This distinction among the three types of categorical variables is useful only when the number of possible values equals or exceeds three. When the number of possible values is two, we have a special case called a binary variable. A *binary* variable can be discrete, ordinal, or nominal, depending on the researcher's interpretation. For example, if a researcher is interested in studying compliance with the one-child policy in China, the dependent variable is whether a couple has given birth to more than one child. For simplicity, assume that in a particular sample a woman has at least one child and no more than two children. Let us code y so that $y = 0$ if a woman has one child, and $y = 1$ if she has two children. In this case, the dependent variable can be interpreted as discrete (number of children − 1), ordinal (one child or more than one child), or nominal (compliance vs. noncompliance). Fortunately, the researcher may apply the same statistical methods for all three cases. It is the substantive understanding of the results that varies from one interpretation to another.

1.2. Two Philosophies of Categorical Data

The development of methods for the analysis of categorical data has benefitted greatly from contributions by scholars in such diverse fields as statistics, biostatistics, economics, psychology, and sociology. This multidisciplinary origin has given categorical data analysis multiple approaches to similar problems and multiple interpretations for similar methodologies. As a result, categorical data analysis is an intellectually rich and expanding field. However, this interdisciplinary nature has also

[3] For an historical background, see Duncan's (1984) important book *Notes on Social Measurement*.

made synthesizing and consolidating available techniques difficult due to the diverse applications and differing terminology across disciplines.

Part of this difficulty stems from two fundamentally different "philosophies" concerning the nature of categorical data. One philosophy views categorical variables as being inherently categorical and relies on transformations of the data to derive regression-type models. The other philosophy presumes that categorical variables are conceptually continuous but are observed, or measured, as categorical. In the one-child policy example, a researcher may view "compliance" as a behavioral continuum. However, he/she can only *observe* two distinct values of this dependent variable. This approach relies on latent variables to derive regression-type models. These very different philosophies can be traced back to the acrimonious debate between Karl Pearson and G. Udny Yule between 1904 and 1913 (Agresti, 2002, pp. 619–622). Although these two approaches can be found in any single discipline, the first is more closely identified with statistics and biostatistics, and the second with econometrics and psychometrics. For simplicity, we will refer to the first approach as *statistical* or *transformational* and to the second as *econometric* or *latent variable*. We intend the terms *statistical* and *econometric* here as short-hand labels rather than as descriptions of the two disciplines.

1.2.1. The Transformational Approach

In the *transformational*, or statistical, approach, categorical data are considered as inherently categorical and should be modeled as such. In this approach, there is a direct one-to-one correspondence between population parameters of interest and sample statistics. The focus is on estimating population parameters that correspond to their sample analogs. No latent, or unobserved, variable is invoked.

In the transformational approach, statistical modeling means that the expected value of the categorical dependent variable, after some transformation, is expressed as a linear function of the independent variables. Given the categorical nature of the dependent variable, the regression function cannot be linear. The problem of nonlinearity is handled through nonlinear functions that transform the expected value of the categorical variable into a linear function of the independent variables. Such transformation functions are now commonly referred to as *link* functions.[4]

For example, in the analysis of discrete (count) data, the expected frequencies (or cell counts) must be nonnegative. To ensure that the predicted values from regression models fit these constraints, the natural logarithm function (or *log* link) is used to transform the expected value of the dependent variable so that a model for the logged count can be expressed as a linear function of independent variables. This *loglinear* transformation serves two purposes: it ensures that the fitted values are appropriate

[4] Models that can be transformed to linear models via link functions are referred to as *generalized linear models*. McCullagh and Nelder (1989) provide an extensive treatment of these types of models.

for count data (i.e., nonnegative), and it permits the unknown regression parameters to lie within the entire real space (parameter space).

In binomial response models, estimated probabilities must lie in the interval [0,1], a range that is violated by any linear function if independent variables are allowed to vary freely. Instead of directly modeling probabilities in this range, we can model a transformation of probability that lies in the interval $(-\infty, +\infty)$. There are a number of ways to transform probabilities. The *logit* transformation, $\log[p/(1-p)]$, can be used to transform the probability scale so that it can be expressed as a linear function of independent variables. A *probit* transformation, $\Phi^{-1}(p)$, can be used in a similar fashion to re-scale probabilities. The probit link utilizes the inverse of the cumulative standard normal distribution function to transform the expected probability to the range $(-\infty, +\infty)$ (i.e., by transforming probabilities to Z-scores). As in the logit model, the probit link transforms the probability so that it can be expressed as a linear function of independent variables. Both the logit and probit transformations ensure that the predicted probabilities are in the proper range for all possible values of parameters and independent variables.

1.2.2. The Latent Variable Approach

The latent variable, or econometric, approach provides a somewhat different view of categorical data. The key to this approach is to assume the existence of a continuous unobserved or *latent* variable underlying an observed categorical variable. When the latent variable crosses a threshold, the observed categorical variable takes on a different value. According to the latent variable approach, what makes categorical variables different from usual continuously distributed variables is partial observability. That is, we can infer from observed categorical values only the intervals within which latent variables lie but not the actual values themselves. For this reason, econometricians commonly refer to categorical variables as limited-dependent variables (Maddala, 1983).

In the latent variable approach, the researcher's theoretical interest lies more in how independent variables affect the latent continuous variables (called structural analysis) than in how independent variables affect the observed categorical variable. From the latent variable perspective, it is thus convenient to think of the sample data as actual *realizations* of population quantities that are *unobservable*. For instance, the observed response categories may reflect the actual choices made by individuals in a sample, but underlying each choice at the population level is a latent variable representing the difference between the cost and the benefit of a particular choice made by an individual decision maker. Similarly, a binary variable may be thought of as the sample realization of a continuous variable representing an unobserved *propensity*. For example, in studies of college admissions, we may assume the existence of a continuous latent variable — qualification — such that applicants whose qualifications exceed the required threshold are admitted, and those whose qualifications fall short of the threshold are rejected (Manski & Wise, 1983).

In studies of women's labor force participation, economic reasoning holds that a woman will participate in the labor force if her market wage exceeds her reservation wage (Heckman, 1979). In practice, it is not possible for the researcher to observe applicants' qualifications, nor the difference between the market and reservation wages. We can, however, observe admission decisions and labor force participation status, which can be taken as *observed* realizations of the underlying population-level latent variable representing likelihood of admission or labor force participation.

Experimental studies in the biological sciences have also made good use of latent variables. In studies of the effectiveness of pesticides, for example, whether an insect dies depends on its *tolerance* to a level of dosage of an insecticide. It is assumed that an insect will die if a dosage level exceeds the insect's tolerance. The binary variable (lives/dies) is the realization of a continuous unobservable variable, the difference between dosage and tolerance.

The latent variable concept has been extended to the construction of latent *categorical* variables. A prime example is the latent class model, which capitalizes on independence conditional on membership in latent classes. This is analogous to factor analysis for continuously distributed variables. Heckman and Singer's (1984) nonparametric method of handling unobserved heterogeneity in survival analysis is also rooted in this fundamental idea.

1.3. An Historical Note

The development of techniques for the analysis of categorical data has been motivated in part by particular substantive concerns in fields such as sociology, economics, epidemiology, and demography (for an historical account in social science, see Camic & Xie, 1994). For example, several innovations in loglinear modeling had their origins in the study of social mobility (e.g., Duncan, 1979; Goodman, 1979; Hauser, 1978); the literature on sample selection models emerged from economic analyses of women's earnings (Heckman, 1979); and problems in the analysis of consumer choices led to the development of many of the techniques for multicategory response variables (McFadden, 1974). Methodological advances in survival analysis arose as extensions of the life-tables technique in demography by statisticians and biostatisticians to incorporate covariates in modeling hazard rates (Cox, 1972; Laird & Oliver, 1981). McCullagh and Nelder's (1989) theory of generalized linear models provided a unified framework which can be applied to most of these models.

Today's latent variable approach grew out of the early psychophysics tradition, where observed frequency distributions of qualitative "judgments" were used to scale the intensity of continuously distributed stimuli (e.g., Thurstone, 1927). In the experimental framework of psychophysics, the "latent" variables were unobservable only to the subjects under an experiment, since the stimuli were manipulated by and thus known to the researcher. For illustration, imagine that a group of subjects are asked to rank the relative weights of two similar objects given by the experimenter.

It is reasonable to assume that the probability of giving the correct answer is positively associated with the actual difference in weight. Thurstone (1927) explicitly assumed a normal distribution for the psychological stimulus and related it to the distribution of "judgments," thus paving the way to today's probit analysis. With time, social scientists have expanded this approach to uncover properties of latent variables from observed data, through such techniques as latent trait models and latent class models. For a treatment of sociologists' contributions to the latent variable approach, see Clogg (1992).

1.4. Approach of This Book

Two features distinguish this book from other texts on the analysis of categorical data. First, this book presents both the transformational and latent variable approaches and, in doing so, synthesizes similar methods in statistical and econometric literatures. Whenever possible, we shall show how the two approaches are similar and in what ways they are different. Second, this book has an applied as opposed to theoretical orientation. We shall draw examples from applied social science research and use data sets constructed for pedagogical purposes. In keeping with the applied orientation of this book, we shall also present actual programming examples for the models discussed, while keeping theoretical discussions at a minimum. We shall provide our data sets, program code, and computer outputs through a website.[5]

1.4.1. *Combining the Statistical and Latent Variable Approaches*

In many instances, the transformational and latent variable approaches are simply two parallel ways of looking at the same phenomena. More often than not, the two approaches yield exactly the same statistical procedures except for minor differences due to the manner in which the model is specified or parameterized. When this is the case, one's viewpoint about the underlying nature of observed categorical variables does not affect specific statistical techniques that we will cover but simply alters the substantive interpretations of results.

1.4.2. *Organization of the Book*

This book begins by considering the simplest models for categorical data and proceeds to more complex models and methods. We begin with a review of the

[5] Our website is continuously updated with new examples utilizing several computer packages. The URL is http://webspace.utexas.edu/dpowers/www/, linkable through YuXie.com and Powers-Xie.com.

general concepts behind regression models for continuous dependent variables. This is a natural starting point since many of the familiar ideas and principles used in the analysis of covariance and regression for continuous variables will carry over to the analysis of categorical dependent variables. These concepts are described in Chapter 2, along with a general orientation to regression models. Chapter 3 discusses models for binary data and issues pertaining to estimation, model building, and the interpretation of results. Chapter 4 provides an overview of measures of association and models for contingency tables. Chapter 5 builds on material in Chapter 3 to introduce multilevel (or hierarchical) models for binary data. Chapter 6 presents methods for event occurrences in time. Chapters 7 and 8 outline various methods for the analysis of polytomous (or multinomial) response variables that assume ordinal or nominal measures.

Chapter 2

Review of Linear Regression Models

2.1. Regression Models

This chapter reviews the classic linear regression model for continuous dependent variables. We assume the reader's familiarity with the linear regression model and thus will not delve into its details. Instead, we will highlight some general concepts and principles underlying the linear regression model that will be useful in later chapters focused on categorical dependent variables.

Regression is one of the most widely used statistical techniques for analyzing observational data. As mentioned in Chapter 1, the analysis of observational data typically requires a structural and multivariate approach. Regression models are used in this context to uncover net relationships between an outcome, or response, variable and a few key explanatory variables while controlling for confounding factors.

Regression models are used to meet different research goals. Sometimes, regression modeling is aimed at learning the causal effect of one variable, or a set of variables, on a dependent variable. Other times, regression models are used to predict the value of a response variable. Finally, regression models are often intended as short-hand summaries providing a description linking a dependent variable and independent variables.

2.1.1. *Three Conceptualizations of Regression*

A researcher faced with a large amount of raw data will want to summarize it in a way that presents essential information without too much distortion. Examples of data reduction include frequency tables or group-specific means and variances. Like most methods in statistics, regression is also a data-reduction technique. In regression analysis, the objective is to predict, as closely as possible, an array of observed values of the dependent variable based on a simple function of independent variables. Obviously, predicted values from regression models are not exactly the same as observed ones. Characteristically, regression partitions an observation into two parts:

$$\boxed{\text{observed}} = \boxed{\text{structural}} + \boxed{\text{stochastic}}$$

The observed part represents the actual values of the dependent variable at hand. The structural part denotes the relationship between the dependent and independent