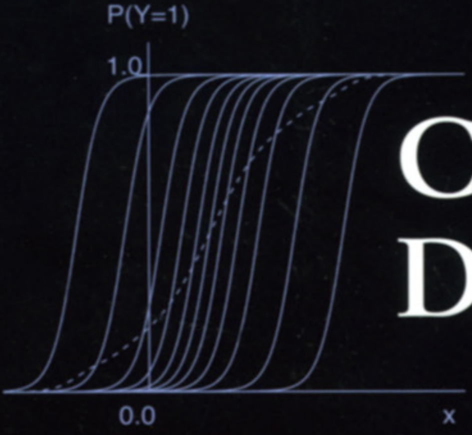
Wiley Series in Probability and Statistics



Categorical Data Analysis

Third Edition

ALAN AGRESTI

WILEY

Categorical Data Analysis

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

A complete list of the titles in this series appears at the end of this volume.

Categorical Data Analysis

Third Edition

ALAN AGRESTI

Department of Statistics
University of Florida
Gainesville, Florida

WILEY-INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

Cover Image: (background) Peter Firus/iStockphoto, (line art) courtesy of the author

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at 800-762-2974, outside the United States at 317-572-3993 or fax 317-572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data

Agresti, Alan.
Categorical data analysis / Alan Agresti. – 3rd ed.
p. cm. – (Wiley series in probability and statistics; 792)
Includes bibliographical references and index.
ISBN 978-0-470-46363-5 (hardback)
1. Multivariate analysis. I. Title.
QA278.A353 2013
519.5′35–dc23
2012009792

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

To Jacki

Contents

Preface

The explosion in the development of methods for analyzing categorical data that began in the 1960s has continued apace in recent years. This book provides an overview of these methods, as well as older, now standard, methods. It gives special emphasis to generalized linear modeling techniques, which extend linear model methods for continuous variables, and their extensions for multivariate responses.

OUTLINE OF TOPICS

Chapters 1–10 present the core methods for categorical response variables. Chapters 1–3 cover distributions for categorical responses and traditional methods for two-way contingency tables. Chapters 4-8 introduce logistic regression and related models such as the probit model for binary and multicategory response variables. Chapters 9 and 10 cover loglinear models for contingency tables.

In the past quarter century, a major area of new research has been the development of methods for repeated measurement and other forms of clustered categorical data. Chapters 11–14 present these methods, including marginal models and generalized linear mixed models with random effects. Chapter 15 introduces non-model-based methods for classification and clustering. Chapter 16 presents theoretical foundations as well as alternatives to the maximum likelihood paradigm that this text adopts. Chapter 17 is devoted to a historical overview of the development of the methods. It examines contributions of noted statisticians, such as Pearson and Fisher, whose pioneering efforts—and sometimes vocal debates—broke the ground for this evolution.

Appendices illustrate the use of statistical software for analyzing categorical data. The website for the text, `www.stat.ufl.edu/~aa/cda/cda.html`, contains an appendix with detailed examples of the use of software (especially R, SAS, and Stata) for performing the analyses in this book, solutions to many of the exercises, extra exercises, and corrections.

CHANGES IN THIS EDITION

Given the explosion of research in the past 50 years on categorical data methods, it is an increasing challenge to write a comprehensive book covering all the commonly used methods. The second edition of this book already exceeded 700 pages. In including much new material without letting the book grow much, I have necessarily had to make compromises in depth and use relatively simple examples. I try to present a broad overview, while presenting bibliographic notes with many references in which the reader can find more details. In attempting to make the book relatively comprehensive while presenting substantive new material, every chapter of the first two editions has been extensively rewritten. The major changes are:

- A new Chapter 7 presents alternative methods for binary response data, including some regularization methods that are becoming popular in this age of massive data sets with enormous numbers of variables.
- A new Chapter 15 introduces non-model-based methods of classification, such as linear discriminant analysis and classification trees, and cluster analysis.
- Many chapters now include a section describing the Bayesian approach for the methods of that chapter. We also have added material (e.g., Sections 6.5 and 7.4) about ways that frequentist methods can deal with awkward situations such as infinite maximum likelihood estimates.
- The use of various software for categorical data methods is discussed at a much expanded website for the text, `www.stat.ufl.edu/~aa/cda/cda.html`. Examples are shown of the use of R, SAS, and Stata for most of the examples in the text, and there is discussion also about SPSS, StatXact, and other software. That website also contains many of the text's data sets, some of which have only excerpts shown in the text itself, as well as solutions for many exercises and corrections of errors found in early printings of the book. I recommend that you refer to this appendix (or specialized software manuals) while reading the text, perhaps printing the pages about the software you prefer, as an aid to implementing the methods. This material was placed at the website partly because the text is already so long without it and also because it is then easier to keep the presentation up-to-date.

In this text, I interpret *categorical data analysis* to refer to methods for categorical response variables. For most methods, explanatory variables can be categorical or quantitative, as in ordinary regression. Thus, the focus is intended to be more general than contingency table analysis, although for simplicity of data presentation, most examples use contingency tables. These examples are simplistic, but should help you focus on understanding the methods themselves and make it easier for you to replicate results with your favorite software.

Other special features of the text include:

- More than 100 analyses of data sets.
- About 600 exercises, some directed toward theory and methods and some toward applications and data analysis.
- Notes at the end of each chapter that provide references for recent research and many topics not covered in the text, linked to a bibliography of more than 1200 sources.

INTENDED AUDIENCE AND USE AS A TEXTBOOK

I intend this book to be accessible to the diverse mix of students who take graduate-level courses in categorical data analysis. But I have also written it with practicing statisticians and biostatisticians in mind. I hope it enables them to catch up with recent advances and learn about methods that sometimes receive inadequate attention in the traditional statistics curriculum.

The development of new methods has influenced—and been influenced by—the increasing availability of data sets with categorical responses in the social, behavioral, and biomedical sciences, as well as in public health, genetics, ecology, education, marketing and the financial industry, and industrial quality control. And so, although this book is directed mainly to statisticians and biostatisticians, I also aim for it to be helpful to methodologists in these fields.

Readers should possess a background that includes regression and analysis of variance models, as well as maximum likelihood methods of statistical theory. Those not having much theory background should be able to follow most methodological discussions. Those with mainly applied interests can skip most of Chapter 4 on the theory of generalized linear models and proceed to other chapters. However, the book has a distinctly higher technical level and is more thorough and complete than my lower-level text, *An Introduction to Categorical Data Analysis, Second Edition* (Wiley, 2007).

Today, because of the ubiquity of categorical data in applications, most statistics and biostatistics departments offer courses on categorical data analysis or on generalized linear models with strong emphasis on methods for discrete data. This book can be used as a text for such courses. The material in Chapters 1–6 forms the heart of most courses. There is too much material in this book for a single course, but a one-term course can be based on the following outline:

- Basic contingency table analysis, covering Chapters 1–3, perhaps skipping some tangential sections such as 1.5.7, 1.6, 2.4, 3.4–3.7.
- Logistic regression and related methods for binary data, covering Chapters 4–6, perhaps skipping some tangential sections such as 4.4–4.7 and 6.4–6.6.
- Multinomial response models, covering at least Sections 8.1 and 8.2.
- Matched pairs and clustered data, covering at least Sections 11.1–11.2.

Courses with biostatistical orientation may want to include bits from Chapters 12 and 13 on marginal and random effects models. Courses with social science emphasis may want to include some topics on loglinear modeling from Chapters 9 and 10. Some courses may want to select specialized topics from Chapter 7, such as probit modeling, conditional logistic regression, Bayesian binary data modeling, smoothing, and issues in the analysis of high-dimensional data.

ACKNOWLEDGMENTS

I thank those who commented on parts of the manuscript or provided help of some type. Special thanks to Anna Gottard, David Hoaglin, Maria Kateri, Bernhard Klingenberg, Keli Liu, and Euijung Ryu, who gave insightful comments on some chapters and made many helpful suggestions, and Brett Presnell for his advice and resources about R software and his comments about some of the material. Thanks to people who made suggestions about new material for this edition, including Jonathan Bischof, James Booth, Brian Caffo, Tianxi Cai, Brent Coull, Nicholas Cox, Ralitza Gueorguieva, Debashis Ghosh, John Henretta,

David Hitchcock, Galin Jones, Robert Kushler, Xihong Lin, Jun Liu, Gianfranco Lovison, Giovanni Marchetti, David Olive, Art Owen, Alessandra Petrucci, Michael Radelet, Gerard Scallan, Maura Stokes, Anestis Touloumis, and Ming Yang. Thanks to those who commented on aspects of the second edition, including pointing out errors or typos, such as Pat Altham, Roberto Bertolusso, Nicholas Cox, David Firth, Rene Gonin, David Hoaglin, Harry Khamis, Bernhard Klingenberg, Robert Kushler, Gianfranco Lovison, Theo Nijsse, Richard Reyment, Misha Salganik, William Santo, Laura Thompson, Michael Vock, and Zhongming Yang. Thanks also to Laura Thompson for preparing her very helpful manual on using R and S-Plus for examples in the second edition. Thanks to the many who reviewed material or suggested examples for the first two editions, mentioned in the Prefaces of those editions. Thanks also to Wiley Executive Editor Steve Quigley and Associate Editor Jacqueline Palmieri for their steadfast encouragement and facilitation of this project. Finally, thanks to my wife Jacki Levine for continuing support of all kinds.

ALAN AGRESTI

Gainesville, Florida and Brookline, Massachusetts
February 2012

CHAPTER 1

Introduction: Distributions and Inference for Categorical Data

From helping to assess the value of new medical treatments to evaluating the factors that affect our opinions and behaviors, analysts today are finding myriad uses for categorical data methods. In this book we introduce these methods and the theory behind them.

Statistical methods for categorical responses were late in gaining the level of sophistication achieved early in the twentieth century by methods for continuous responses. Despite influential work around 1900 by the British statistician Karl Pearson, relatively little development of models for categorical responses occurred until the 1960s. In this book we describe the early fundamental work that still has importance today but place primary emphasis on more recent modeling approaches.

1.1 CATEGORICAL RESPONSE DATA

A *categorical variable* has a measurement scale consisting of a set of categories. For instance, political philosophy is often measured as liberal, moderate, or conservative. Diagnoses regarding breast cancer based on a mammogram use the categories normal, benign, probably benign, suspicious, and malignant.

The development of methods for categorical variables was stimulated by the need to analyze data generated in research studies in both the social and biomedical sciences. Categorical scales are pervasive in the social sciences for measuring attitudes and opinions. Categorical scales in biomedical sciences measure outcomes such as whether a medical treatment is successful.

Categorical data are by no means restricted to the social and biomedical sciences. They frequently occur in the behavioral sciences (e.g., type of mental illness, with the categories schizophrenia, depression, neurosis), epidemiology and public health (e.g., contraceptive method at last sexual intercourse, with the categories none, condom, pill, IUD, other), genetics (type of allele inherited by an offspring), botany and zoology (e.g., whether or not a particular organism is observed in a sampled quadrat), education (e.g., whether a student response to an exam question is correct or incorrect), and marketing (e.g., consumer

Categorical Data Analysis, Third Edition. Alan Agresti.

preference among the three leading brands of a product). They even occur in highly quantitative fields such as engineering sciences and industrial quality control. Examples are the classification of items according to whether they conform to certain standards, and subjective evaluation of some characteristic: how soft to the touch a certain fabric is, how good a particular food product tastes, or how easy a worker finds it to perform a certain task.

Categorical variables are of many types. In this section we provide ways of classifying them.

1.1.1 Response–Explanatory Variable Distinction

Statistical analyses distinguish between *response* (or *dependent*) *variables* and *explanatory* (or *independent*) *variables*. This book focuses on methods for categorical response variables. As in ordinary regression modeling, explanatory variables can be any type. For instance, a study might analyze how opinion about whether same-sex marriages should be legal (yes or no) changes according to values of explanatory variables, such as religious affiliation, political ideology, number of years of education, annual income, age, gender, and race.

1.1.2 Binary–Nominal–Ordinal Scale Distinction

Many categorical variables have only two categories. Such variables, for which the two categories are often given the generic labels "success" and "failure," are called *binary variables*. A major topic of this book is the modeling of binary response variables.

When a categorical variable has more than two categories, we distinguish between two types of categorical scales. Variables having categories without a natural ordering are said to be measured on a *nominal scale* and are called *nominal variables*. Examples are mode of transportation to get to work (automobile, bicycle, bus, subway, walk), favorite type of music (classical, country, folk, jazz, rock), and choice of residence (apartment, condominium, house, other). For nominal variables, the order of listing the categories is irrelevant to the statistical analysis.

Many categorical variables *do* have ordered categories. Such variables are said to be measured on an *ordinal scale* and are called *ordinal variables*. Examples are social class (upper, middle, lower), political philosophy (very liberal, slightly liberal, moderate, slightly conservative, very conservative), patient condition (good, fair, serious, critical), and rating of a movie for Netflix (1 to 5 stars, representing hated it, didn't like it, liked it, really liked it, loved it). For ordinal variables, distances between categories are unknown. Although a person categorized as very liberal is more liberal than a person categorized as slightly liberal, no numerical value describes *how much more* liberal that person is.

An *interval variable* is one that *does* have numerical distances between any two values. For example, systolic blood pressure level, length of prison term, and annual income are interval variables. For most such variables, it is also possible to compare two values by their ratio, in which case the variable is also called a *ratio variable*.

The way that a variable is measured determines its classification. For example, "education" is only nominal when measured as (public school, private school, home schooling); it is ordinal when measured by highest degree attained, using the categories (none, high school, bachelor's, master's, doctorate); it is interval when measured by number of years of education completed, using the integers $0, 1, 2, 3, \ldots$.

A variable's measurement scale determines which statistical methods are appropriate. It is usually best to apply methods appropriate for the actual scale. In the measurement hierarchy, interval variables are highest, ordinal variables are next, and nominal variables are lowest. Statistical methods for variables of one type can also be used with variables at higher levels but not at lower levels. For instance, statistical methods for nominal variables can be used with ordinal variables by ignoring the ordering of categories. Methods for ordinal variables cannot, however, be used with nominal variables, since their categories have no meaningful ordering. The distinction between ordered and unordered categories is not important for binary variables, because ordinal methods and nominal methods then typically reduce to equivalent methods.

In this book, we present methods for the analysis of binary, nominal, and ordinal variables. The methods also apply to interval variables having a small number of distinct values (e.g., number of times married, number of distinct side effects experienced in taking some drug) or for which the values are grouped into ordered categories (e.g., education measured as ≤ 12 years, >12 but <16 years, ≥ 16 years).

1.1.3 Discrete–Continuous Variable Distinction

Variables are classified as *discrete* or *continuous*, according to whether the number of values they can take is countable. Actual measurement of all variables occurs in a discrete manner, due to precision limitations in measuring instruments. The discrete–continuous classification, in practice, distinguishes between variables that take few values and variables that take lots of values. For instance, statisticians often treat discrete interval variables having a large number of values (such as test scores) as continuous, using them in methods for continuous responses.

This book deals with certain types of discretely measured responses: (1) binary variables, (2) nominal variables, (3) ordinal variables, (4) discrete interval variables having relatively few values, and (5) continuous variables grouped into a small number of categories.

1.1.4 Quantitative–Qualitative Variable Distinction

Nominal variables are *qualitative*—distinct categories differ in quality, not in quantity. Interval variables are *quantitative*—distinct levels have differing amounts of the characteristic of interest. The position of ordinal variables in the qualitative–quantitative classification is fuzzy. Analysts often treat them as qualitative, using methods for nominal variables. But in many respects, ordinal variables more closely resemble interval variables than they resemble nominal variables. They possess important quantitative features: Each category has a *greater* or *smaller* magnitude of the characteristic than another category; and although not possible to measure, an underlying continuous variable is often present. The political ideology classification (very liberal, slightly liberal, moderate, slightly conservative, very conservative) crudely measures an inherently continuous characteristic.

Analysts often utilize the quantitative nature of ordinal variables by assigning numerical scores to the categories or assuming an underlying continuous distribution. This requires good judgment and guidance from researchers who use the scale, but it provides benefits in the variety of methods available for data analysis.

1.1.5 Organization of Book and Online Computing Appendix

The models for categorical response variables discussed in this book resemble regression models for continuous response variables; however, they assume binomial or multinomial response distributions instead of normality. One type of model receives special attention—*logistic regression.* Ordinary logistic regression models apply with *binary* responses and assume a binomial distribution. Generalizations of logistic regression apply with multicategory responses and assume a multinomial distribution.

The book has four main units. In the first, Chapters 1 through 3, we summarize descriptive and inferential methods for univariate and bivariate categorical data. These chapters cover discrete distributions, methods of inference, and measures of association for contingency tables. They summarize the non-model-based methods developed prior to about 1960.

In the second and primary unit, Chapters 4 through 10, we introduce models for categorical responses. In Chapter 4 we describe a class of *generalized linear models* having models of this text as special cases. Chapters 5 and 6 cover the most important model for binary responses, logistic regression. Chapter 7 presents alternative methods for binary data, including the probit, Bayesian fitting, and smoothing methods. In Chapter 8 we present generalizations of the logistic regression model for nominal and ordinal multicategory response variables. In Chapters 9 and 10 we introduce the modeling of multivariate categorical response data, in terms of association and interaction patterns among the variables. The models, called *loglinear models*, apply to counts in the table that cross-classifies those responses.

In the third unit, Chapters 11 through 14, we discuss models for handling repeated measurement and other forms of clustered data. In Chapter 11 we present models for a categorical response with matched pairs; these apply, for instance, with a categorical response measured for the same subjects at two times. Chapter 12 covers models for more general types of repeated categorical data, such as longitudinal data from several times with explanatory variables. In Chapter 13 we present a broad class of models, *generalized linear mixed models*, that use random effects to account for dependence with such data. In Chapter 14 further extensions of the models from Chapters 11 through 13 are described, unified by treating the response as having a mixture distribution of some type.

The fourth and final unit has a different nature than the others. In Chapter 15 we consider non-model-based classification and clustering methods. In Chapter 16 we summarize large-sample and small-sample theory for categorical data models. This theory is the basis for behavior of model parameter estimators and goodness-of-fit statistics. Chapter 17 presents a historical overview of the development of categorical data methods.

Maximum likelihood methods receive primary attention throughout the book. Many chapters, however, contain a section presenting corresponding Bayesian methods.

In Appendix A we review software that can perform the analyses in this book. The website `www.stat.ufl.edu/~aa/cda/cda.html` for this book contains an appendix that gives more information about using R, SAS, Stata, and other software, with sample programs for text examples. In addition, that site has complete data sets for many text examples and exercises, solutions to some exercises, extra exercises, corrections, and links to other useful sites. For instance, a manual prepared by Dr. Laura Thompson provides examples of how to use R and S-Plus for all examples in the second edition of this text, many of which (or very similar ones) are also in this edition.

In the rest of this chapter, we provide background material. In Section 1.2 we review the key distributions for categorical data: the binomial and multinomial, as well as another that is important for discrete data, the Poisson. In Section 1.3 we review the primary mechanisms for statistical inference using maximum likelihood. In Sections 1.4 and 1.5 we illustrate these by presenting significance tests and confidence intervals for binomial and multinomial parameters. In Section 1.6 we introduce Bayesian inference for these parameters.

1.2 DISTRIBUTIONS FOR CATEGORICAL DATA

Inferential data analyses require assumptions about the random mechanism that generated the data. For regression models with continuous responses, the normal distribution plays the central role. In this section we review the three key distributions for categorical responses: *binomial*, *multinomial*, and *Poisson*.

1.2.1 Binomial Distribution

Many applications refer to a fixed number n of binary observations. Let $y_1, y_2, \ldots, y_n$ denote observations from n independent and identical trials such that $P(Y_i = 1) = \pi$ and $P(Y_i = 0) = 1 - \pi$. We refer to outcome 1 as "success" and outcome 0 as "failure." *Identical trials* means that the probability of success π is the same for each trial. *Independent trials* means that the $\{Y_i\}$ are independent random variables. These are often called *Bernoulli trials*. The total number of successes, $Y = \sum_{i=1}^{n} Y_i$, has the *binomial distribution* with index n and parameter π, denoted by bin(n, π).

The probability mass function for the possible outcomes y for Y is

$$p(y) = \binom{n}{y} \pi^y (1-\pi)^{n-y}, \quad y = 0, 1, 2, \ldots, n, \tag{1.1}$$

where the binomial coefficient $\binom{n}{y} = n!/[y!(n-y)!]$. Since $E(Y_i) = E(Y_i^2) = 1 \times \pi + 0 \times (1 - \pi) = \pi$,

$$E(Y_i) = \pi \quad \text{and} \quad \text{var}(Y_i) = \pi(1-\pi).$$

The binomial distribution for $Y = \sum_i Y_i$ has mean and variance

$$\mu = E(Y) = n\pi \quad \text{and} \quad \sigma^2 = \text{var}(Y) = n\pi(1-\pi).$$

The skewness is described by $E(Y - \mu)^3/\sigma^3 = (1 - 2\pi)/\sqrt{n\pi(1-\pi)}$. The distribution is symmetric when $\pi = 0.50$ but becomes increasingly skewed as π moves toward either boundary. The binomial distribution converges to normality as n increases, for fixed π, the approximation being reasonable[1] when $n[\min(\pi, 1-\pi)]$ is as small as about 5.

There is no guarantee that successive binary observations are independent or identical. Thus, occasionally, we will utilize other distributions. One such case is sampling binary outcomes without replacement from a finite population, such as observations on whether a homework assignment was completed for 10 students sampled from a class of size 20. The

[1] See `www.stat.tamu.edu/~west/applets/binomialdemo2.html`.

hypergeometric distribution, studied in Section 3.5.1, is then relevant. In Section 1.2.4 we discuss another case that violates the binomial assumptions.

1.2.2 Multinomial Distribution

Some trials have more than two possible outcomes. Suppose that each of n independent, identical trials can have outcome in any of c categories. Let $y_{ij} = 1$ if trial i has outcome in category j and $y_{ij} = 0$ otherwise. Then $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{ic})$ represents a multinomial trial, with $\sum_j y_{ij} = 1$; for instance, (0, 0, 1, 0) denotes outcome in category 3 of four possible categories. Note that y_{ic} is redundant, being linearly dependent on the others. Let $n_j = \sum_i y_{ij}$ denote the number of trials having outcome in category j. The counts $(n_1, n_2, \ldots, n_c)$ have the *multinomial distribution*.

Let $\pi_j = P(Y_{ij} = 1)$ denote the probability of outcome in category j for each trial. The multinomial probability mass function is

$$p(n_1, n_2, \ldots, n_{c-1}) = \left(\frac{n!}{n_1!n_2!\cdots n_c!}\right)\pi_1^{n_1}\pi_2^{n_2}\cdots\pi_c^{n_c}. \tag{1.2}$$

Since $\sum_j n_j = n$, this is $(c-1)$-dimensional, with $n_c = n - (n_1 + \cdots + n_{c-1})$. The binomial distribution is the special case with $c = 2$.

For the multinomial distribution,

$$E(n_j) = n\pi_j, \qquad \text{var}(n_j) = n\pi_j(1-\pi_j), \qquad \text{cov}(n_j, n_k) = -n\pi_j\pi_k. \tag{1.3}$$

We derive the covariance in Section 16.1.4. The marginal distribution of each n_j is binomial.

1.2.3 Poisson Distribution

Sometimes, count data do not result from a fixed number of trials. For instance, if $Y =$ number of automobile accidents today on motorways in Italy, there is no fixed upper bound n for Y (as you are aware if you have driven in Italy!). Since Y must take a nonnegative integer value, its distribution should place its mass on that range. The simplest such distribution is the *Poisson*. Its probabilities depend on a single parameter, the mean μ. The Poisson probability mass function (Poisson 1837, p. 206) is

$$p(y) = \frac{e^{-\mu}\mu^y}{y!}, \qquad y = 0, 1, 2, \ldots. \tag{1.4}$$

It satisfies $E(Y) = \text{var}(Y) = \mu$. It is unimodal with mode equal to the integer part of μ. Its skewness is described by $E(Y-\mu)^3/\sigma^3 = 1/\sqrt{\mu}$. The Poisson distribution approaches normality as μ increases, the normal approximation being quite good when μ is at least about 10.

The Poisson distribution is used for counts of events that occur randomly over time or space, when outcomes in disjoint periods or regions are independent. It also applies as an approximation for the binomial when n is large and π is small, with $\mu = n\pi$. For example, suppose $Y =$ *number of deaths* today in auto accidents in Italy (rather than the *number of accidents*). Then, Y has an upper bound. If each of the 50 million people driving in Italy is an independent trial with probability 0.0000003 of dying today in an auto accident, the number of deaths Y is a bin(50000000, 0.0000003) variate. This is approximately Poisson with $\mu = n\pi = 50000000(0.0000003) = 15$.

A key feature of the Poisson distribution is that its variance equals its mean. Sample counts vary more when their mean is higher. When the mean number of daily fatal accidents equals 15, greater variability occurs from day to day than when the mean equals 2.

1.2.4 Overdispersion

In practice, count observations often exhibit variability exceeding that predicted by the binomial or Poisson. This phenomenon is called *overdispersion.* We assumed above that each person has the same probability each day of dying in a fatal auto accident. More realistically, these probabilities vary from day to day according to the amount of road traffic and weather conditions and vary from person to person according to factors such as the amount of time spent in autos, whether the person wears a seat belt, how much of the driving is at high speeds, gender, and age. Such variation causes fatality counts to display more variation than predicted by the Poisson model.

Suppose that Y is a random variable with variance $\text{var}(Y\,|\,\mu)$ for given μ, but μ itself varies because of unmeasured factors such as those just described. Let $\theta = E(\mu)$. Then unconditionally,

$$E(Y) = E[E(Y\,|\,\mu)], \qquad \text{var}(Y) = E[\text{var}(Y\,|\,\mu)] + \text{var}[E(Y\,|\,\mu)].$$

When Y is conditionally Poisson (given μ), then $E(Y) = E(\mu) = \theta$ and $\text{var}(Y) = E(\mu) + \text{var}(\mu) = \theta + \text{var}(\mu) > \theta$.

Assuming a Poisson distribution for a count variable is often too simplistic, because of factors that cause overdispersion. The *negative binomial* is a related distribution for count data that has a second parameter and permits the variance to exceed the mean. We introduce it in Section 4.3.4.

Analyses assuming binomial (or multinomial) distributions are also sometimes invalid because of overdispersion. This might happen because the true distribution is a mixture of different binomial distributions, with the parameter varying because of unmeasured variables. To illustrate, suppose that an experiment exposes pregnant mice to a toxin and then after a week observes the number of fetuses in each mouse's litter that show signs of malformation. Let n_i denote the number of fetuses in the litter for mouse i. The pregnant mice also vary according to other factors, such as their weight, overall health, and genetic makeup. Extra variation then occurs because of the variability from litter to litter in the probability π of malformation. The distribution of the number of fetuses per litter showing malformations might cluster near 0 and near n_i, showing more dispersion than expected for binomial sampling with a single value of π. Overdispersion could also occur when π varies among fetuses in a litter according to some distribution (Exercise 1.17). In Chapters 4, 13, and 14 we introduce methods for data that are overdispersed relative to binomial and Poisson assumptions.

1.2.5 Connection Between Poisson and Multinomial Distributions

For adult residents of Britain who visit France this year, let Y_1 = number who fly there, Y_2 = number who travel there by train without a car (Eurostar), Y_3 = number who travel there by ferry without a car, and Y_4 = number who take a car (by Eurotunnel Shuttle or a ferry). A Poisson model for (Y_1, Y_2, Y_3, Y_4) treats these as independent Poisson random variables, with parameters $(\mu_1, \mu_2, \mu_3, \mu_4)$. The joint probability mass function for $\{Y_i\}$ is the product of the four mass functions of form (1.4). The total $n = \sum_i Y_i$ also has a Poisson distribution, with parameter $\sum_i \mu_i$.

With Poisson sampling the total count n is random rather than fixed. If we assume a Poisson model but condition on n, $\{Y_i\}$ no longer have Poisson distributions, since each Y_i cannot exceed n. Given n, $\{Y_i\}$ are also no longer independent, since the value of one affects the possible range for the others.

For c independent Poisson variates, with $E(Y_i) = \mu_i$, the conditional probability of a set of counts $\{n_i\}$ satisfying $\sum_i Y_i = n$ is

$$P\left[(Y_1 = n_1, Y_2 = n_2, \ldots, Y_c = n_c) \middle| \sum_j Y_j = n\right]$$
$$= \frac{P(Y_1 = n_1, Y_2 = n_2, \ldots, Y_c = n_c)}{P\left(\sum_j Y_j = n\right)}$$
$$= \frac{\prod_i [\exp(-\mu_i)\mu_i^{n_i}/n_i!]}{\exp\left(-\sum_j \mu_j\right)\left(\sum_j \mu_j\right)^n/n!} = \frac{n!}{\prod_i n_i!}\prod_i \pi_i^{n_i}, \tag{1.5}$$

where $\left\{\pi_i = \mu_i/\left(\sum_j \mu_j\right)\right\}$. This is the multinomial $(n, \{\pi_i\})$ distribution, characterized by the sample size n and the probabilities $\{\pi_i\}$.

Many categorical data analyses assume a multinomial distribution. Such analyses usually have the same inferential results as those of analyses assuming a Poisson distribution, because of the similarity in the likelihood functions.

1.2.6 The Chi-Squared Distribution

Another distribution of fundamental importance for categorical data is the *chi-squared*, not as a distribution for the data but rather as a sampling distribution for many statistics. Because of its importance, we summarize here a few of its properties.

The chi-squared distribution with degrees of freedom denoted by df has mean df, variance 2(df), and skewness $\sqrt{8/\text{df}}$. It converges (slowly) to normality as df increases, the approximation being reasonably good when df is at least about 50.

Let Z denote a standard normal random variable (mean 0, variance 1). Then Z^2 has a chi-squared distribution with df $= 1$. A chi-squared random variable with df $= \nu$ has representation $Z_1^2 + \cdots + Z_\nu^2$, where $Z_1, \ldots, Z_\nu$ are independent standard normal variables. Thus, a chi-squared statistic having df $= \nu$ has partitionings into independent chi-squared components—for example, into ν components each having df $= 1$. Conversely, the *reproductive property* states that if X_1^2 and X_2^2 are independent chi-squared random variables having degrees of freedom ν_1 and ν_2, then $X^2 = X_1^2 + X_2^2$ has a chi-squared distribution with df $= \nu_1 + \nu_2$.

1.3 STATISTICAL INFERENCE FOR CATEGORICAL DATA

In practice, the probability distribution assumed for the response variable has unknown parameter values. In this section we review methods of using sample data to make inferences about the parameters. Sections 1.4 and 1.5 illustrate these methods for binomial and multinomial parameters.

1.3.1 Likelihood Functions and Maximum Likelihood Estimation

In this book we use *maximum likelihood* for parameter estimation. Maximum likelihood estimators have desirable properties: They have large-sample normal distributions; they are asymptotically consistent, converging to the parameter as n increases; and they are asymptotically efficient, producing large-sample standard errors no greater than those from other estimation methods. These results hold under weak regularity conditions, mainly that the number of parameters remains constant as n increases and that the true values of those parameters fall in the interior (rather than on the boundary) of the parameter space.

Given the data, for a chosen probability distribution the *likelihood function* is the probability of those data, treated as a function of the unknown parameter. The maximum likelihood (ML) estimate is the parameter value that maximizes this function. This is the parameter value under which the data observed have the highest probability of occurrence. We denote a parameter for a generic problem by β and its ML estimate by $\hat{\beta}$. We denote the likelihood function by $\ell(\beta)$. The β value that maximizes $\ell(\beta)$ also maximizes $L(\beta) = \log[\ell(\beta)]$. It is simpler to maximize $L(\beta)$ since it is a sum rather than a product of terms. For many models, $L(\beta)$ has concave shape and $\hat{\beta}$ is the point at which the derivative equals 0. The ML estimate is then the solution of the likelihood equation, $\partial L(\beta)/\partial\beta = 0$. Often, β is multidimensional, denoted by $\boldsymbol{\beta}$, and $\hat{\boldsymbol{\beta}}$ is the solution of a set of likelihood equations.

Let $\text{cov}(\hat{\boldsymbol{\beta}})$ denote the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$. Under regularity conditions (Rao 1973, p. 364), $\text{cov}(\hat{\boldsymbol{\beta}})$ is the inverse of the *information matrix*. The (j, k) element of the information matrix is

$$-E\left(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_j\partial\beta_k}\right). \tag{1.6}$$

The standard errors are the square roots of the diagonal elements for the inverse of the information matrix. The greater the curvature of the log likelihood function, the smaller the standard errors. This is reasonable, since large curvature implies that the log likelihood drops quickly as $\boldsymbol{\beta}$ moves away from $\hat{\boldsymbol{\beta}}$; hence, the data would have been much more likely to occur if $\boldsymbol{\beta}$ took a value near $\hat{\boldsymbol{\beta}}$ rather than a value far from $\hat{\boldsymbol{\beta}}$.

1.3.2 Likelihood Function and ML Estimate for Binomial Parameter

The part of a likelihood function involving the parameters is called the *kernel*. Since the maximization of the likelihood is done with respect to the parameters, the rest is irrelevant.

To illustrate, consider the binomial distribution (1.1). The binomial coefficient $n!/[y!(n-y)!]$ has no influence on where the maximum occurs with respect to π. Thus, we ignore it and treat the kernel as the likelihood function. The binomial log likelihood function is then

$$L(\pi) = \log[\pi^y(1-\pi)^{n-y}] = y\log(\pi) + (n-y)\log(1-\pi). \tag{1.7}$$

Differentiating with respect to π yields

$$\partial L(\pi)/\partial\pi = y/\pi - (n-y)/(1-\pi) = (y - n\pi)/\pi(1-\pi). \tag{1.8}$$

Equating this to 0 gives the likelihood equation, which has solution $\hat{\pi} = y/n$, the sample proportion of successes for the n trials.

Calculating $\partial^2 L(\pi)/\partial\pi^2$, taking the expectation, and combining terms, we get

$$-E[\partial^2 L(\pi)/\partial\pi^2] = E[y/\pi^2 + (n-y)/(1-\pi)^2] = n/[\pi(1-\pi)]. \tag{1.9}$$

Thus, the asymptotic variance of $\hat{\pi}$ is $\pi(1-\pi)/n$. This is no surprise. Since $E(Y) = n\pi$ and $\text{var}(Y) = n\pi(1-\pi)$, the distribution of $\hat{\pi} = Y/n$ has mean and standard deviation

$$E(\hat{\pi}) = \pi, \quad \sigma(\hat{\pi}) = \sqrt{\frac{\pi(1-\pi)}{n}}.$$

1.3.3 Wald–Likelihood Ratio–Score Test Triad

There are three standard ways to use the likelihood function to perform large-sample inference. We introduce these for a significance test of a null hypothesis H_0: $\beta = \beta_0$ and then discuss their relation to interval estimation. They all exploit the large-sample normality of ML estimators.

Standard errors obtained from the inverse of the information matrix depend on the unknown parameter values. When we substitute the unrestricted ML estimates (i.e., not assuming the null hypothesis) we obtain an estimated standard error of $\hat{\beta}$, which we denote by SE. Denote $-E[\partial^2 L(\beta)/\partial\beta^2]$ (i.e., the information) evaluated at $\hat{\beta}$ by $\iota(\hat{\beta})$. The first large-sample inference method has test statistic using this estimated standard error,

$$z = (\hat{\beta} - \beta_0)/SE, \quad \text{where} \quad SE = 1/\sqrt{\iota(\hat{\beta})}.$$

This statistic has an approximate standard normal distribution when $\beta = \beta_0$. We refer z to the standard normal table to obtain one- or two-sided P-values. Equivalently, for the two-sided alternative, z^2 has an approximate chi-squared null distribution with df $= 1$; the P-value is then the right-tailed chi-squared probability above the observed value. This type of statistic, using the nonnull estimated standard error, is called a *Wald statistic* (Wald 1943).

The multivariate extension[2] for the Wald test of H_0: $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ has test statistic

$$W = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T [\text{cov}(\hat{\boldsymbol{\beta}})]^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0).$$

The nonnull covariance is based on the curvature (1.6) of the log-likelihood function at $\hat{\boldsymbol{\beta}}$ and typically itself requires estimation. The asymptotic multivariate normal distribution for $\hat{\boldsymbol{\beta}}$ implies an asymptotic chi-squared distribution for W. The df equal the rank of $\text{cov}(\hat{\boldsymbol{\beta}})$, which is the number of nonredundant parameters in $\boldsymbol{\beta}$.

[2] The ^{T} superscript on a vector or matrix denotes the transpose.

A second general-purpose method uses the likelihood function through the ratio of two maximizations: (1) the maximum over the possible parameter values under H_0, and (2) the maximum over the larger set of parameter values permitting H_0 or an alternative H_a to be true. Let ℓ_0 denote the maximized value of the likelihood function under H_0, and let ℓ_1 denote the maximized value generally (i.e., under $H_0 \cup H_a$). For instance, for parameters $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1)$ and H_0: $\boldsymbol{\beta}_0 = \mathbf{0}$, ℓ_1 is the likelihood function calculated at the $\boldsymbol{\beta}$ value for which the data would have been most likely; ℓ_0 is the likelihood function calculated at the $\boldsymbol{\beta}_1$ value for which the data would have been most likely, when $\boldsymbol{\beta}_0 = \mathbf{0}$. Then ℓ_1 is always at least as large as ℓ_0, since ℓ_0 results from maximizing over a restricted set of the parameter values.

The ratio $\Lambda = \ell_0/\ell_1$ of the maximized likelihoods cannot exceed 1. Wilks (1935, 1938) showed that $-2\log\Lambda$ has a limiting null chi-squared distribution, as $n \to \infty$. The df equal the difference in the dimensions of the parameter spaces under $H_0 \cup H_a$ and under H_0. The *likelihood-ratio test statistic* equals

$$-2\log\Lambda = -2\log(\ell_0/\ell_1) = -2(L_0 - L_1),$$

where L_0 and L_1 denote the maximized log-likelihood functions. [In this book, we use the natural logarithm throughout, for which its inverse is the exponential function; so, if $a = \log(b)$, then $b = \exp(a) = e^a$.]

The third method uses the *score statistic*, due to R. A. Fisher and C. R. Rao. The score test, referred to in some literature as the *Lagrange multiplier test*, is based on the slope and expected curvature of the log-likelihood function $L(\beta)$ at the null value β_0. It utilizes the size of the *score function*

$$u(\beta) = \partial L(\beta)/\partial\beta,$$

evaluated at β_0. The value $u(\beta_0)$ tends to be larger in absolute value when $\hat{\beta}$ is farther from β_0. Denote $-E[\partial^2 L(\beta)/\partial\beta^2]$ evaluated at β_0 by $\iota(\beta_0)$. The score statistic is the ratio of $u(\beta_0)$ to its null SE, which is $[\iota(\beta_0)]^{1/2}$. This has an approximate standard normal null distribution. The chi-squared form of the score statistic is

$$\frac{[u(\beta_0)]^2}{\iota(\beta_0)} = \frac{[\partial L(\beta)/\partial\beta_0]^2}{-E[\partial^2 L(\beta)/\partial\beta_0^2]},$$

where the notation reflects derivatives with respect to β that are evaluated at β_0. In the multiparameter case, the score statistic is a quadratic form based on the vector of partial derivatives of the log likelihood with respect to $\boldsymbol{\beta}$ and the inverse information matrix, both evaluated at the H_0 estimates (i.e., assuming that $\boldsymbol{\beta} = \boldsymbol{\beta}_0$).

Figure 1.1 shows a plot of a generic log-likelihood function $L(\beta)$ for the univariate case. It illustrates the three tests of H_0: $\beta = 0$. The Wald test uses the behavior of $L(\beta)$ at the ML estimate $\hat{\beta}$, having chi-squared form $(\hat{\beta}/SE)^2$. The SE of $\hat{\beta}$ depends on the curvature of $L(\beta)$ at $\hat{\beta}$. The score test is based on the slope and curvature of $L(\beta)$ at $\beta = 0$. The likelihood-ratio test combines information about $L(\beta)$ at both $\hat{\beta}$ and $\beta_0 = 0$. It compares the log-likelihood values L_1 at $\hat{\beta}$ and L_0 at $\beta_0 = 0$ using the chi-squared statistic $-2(L_0 - L_1)$. In Figure 1.1, this statistic is twice the vertical distance between values of $L(\beta)$ at $\hat{\beta}$ and at 0.

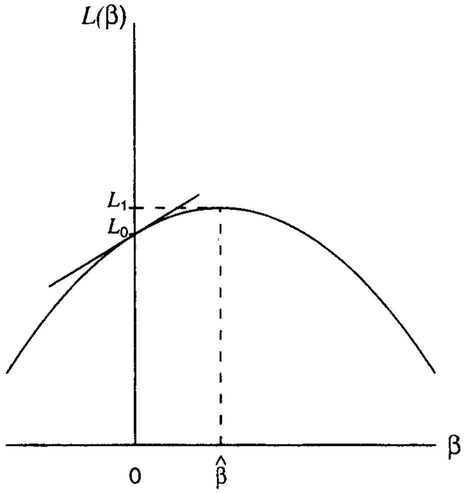


Figure 1.1 Log-likelihood function and information used in three tests of H_0: $\beta = 0$.

Section 1.4.1 illustrates the Wald, likelihood-ratio, and score tests for inference about a binomial parameter. As $n \to \infty$, the three tests have certain asymptotic equivalences (Cox and Hinkley 1974, Sec. 9.3). For small to moderate sample sizes, the likelihood-ratio and score tests are usually more reliable than the Wald test, having actual error rates closer to the nominal level.

1.3.4 Constructing Confidence Intervals by Inverting Tests

In practice, it is more informative to construct confidence intervals for parameters than to test hypotheses about their values. For any of the three test methods, we can construct a confidence interval by inverting the test. For instance, a 95% confidence interval for β is the set of β_0 for which the test of H_0: $\beta = \beta_0$ has P-value exceeding 0.05.

Let z_a denote the z-score from the standard normal distribution having right-tailed probability a; this is the $100(1 - a)$ percentile of that distribution. A $100(1 - \alpha)\%$ confidence interval based on asymptotic normality uses $z_{\alpha/2}$, for instance $z_{0.025} = 1.96$ for 95% confidence. The Wald confidence interval is the set of β_0 for which $|\hat{\beta} - \beta_0|/SE < z_{\alpha/2}$. This gives the interval $\hat{\beta} \pm z_{\alpha/2}(SE)$. Let $\chi^2_{\text{df}}(a)$ denote the $100(1 - a)$ percentile of the chi-squared distribution with degrees of freedom df. The likelihood-ratio-based confidence interval is the set of β_0 for which $-2[L(\beta_0) - L(\hat{\beta})] < \chi^2_1(\alpha)$. [Note that $\chi^2_1(\alpha) = z^2_{\alpha/2}$.]

When $\hat{\beta}$ has a normal distribution, the log-likelihood function has a parabolic shape. For small samples with categorical data, $\hat{\beta}$ may be far from normality and the log-likelihood function can be far from a symmetric, parabolic-shaped curve. This can also happen with moderate to large samples when β falls near the boundary of the parameter space, such as a population proportion that is near 0 or near 1. In such cases, inference based on asymptotic normality of $\hat{\beta}$ may have inadequate performance. A marked divergence in results of Wald and likelihood-ratio inference indicates that the distribution of $\hat{\beta}$ may not be close to normality. The example in Section 1.4.3 illustrates.

The Wald confidence interval is commonly used in practice, because it is simple to construct using ML estimates and standard errors reported by statistical software. The