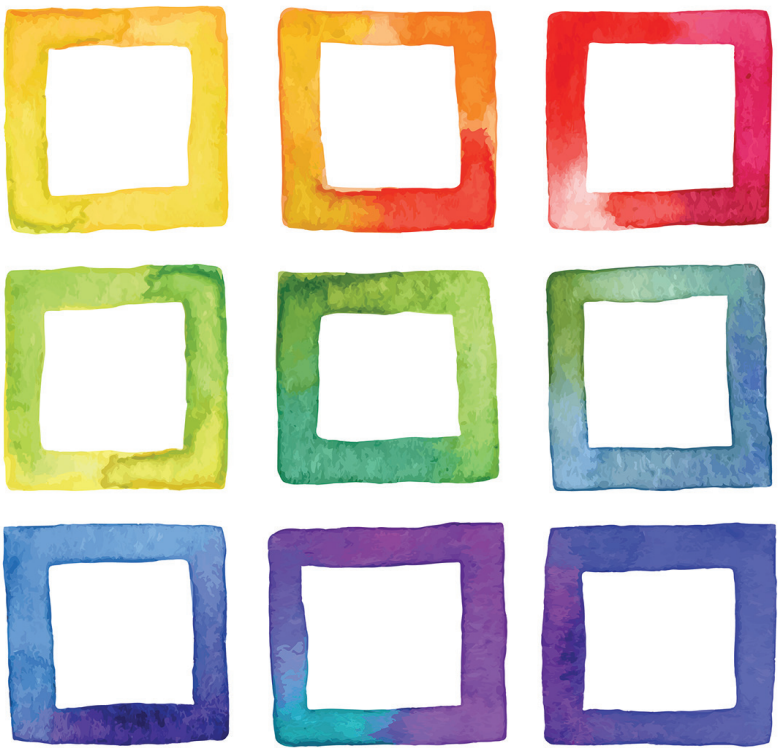
Wiley Series in Probability and Statistics

AN INTRODUCTION TO CATEGORICAL DATA ANALYSIS

THIRD EDITION



ALAN AGRESTI

with website

WILEY

AN INTRODUCTION TO CATEGORICAL DATA ANALYSIS

WILEY SERIES IN PROBABILITY AND STATISTICS
Established by *Walter A. Shewhart and Samuel S. Wilks*

The ***Wiley Series in Probability and Statistics*** is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches. This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

A complete list of titles in this series can be found at
http://www.wiley.com/go/wsps

AN INTRODUCTION TO CATEGORICAL DATA ANALYSIS

Third Edition

Alan Agresti
University of Florida, Florida, United States

WILEY

This third edition first published 2019

Edition History

(1e, 1996); John Wiley & Sons, Inc. (2e, 2007); John Wiley & Sons, Inc.

Registered Office
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

Editorial Office
111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Library of Congress Cataloging-in-Publication Data
Names: Agresti, Alan, author.
Title: An introduction to categorical data analysis / Alan Agresti.
Description: Third edition. | Hoboken, NJ : John Wiley & Sons, 2019. | Series: Wiley series in probability and statistics | Includes bibliographical references and index. |
Identifiers: LCCN 2018026887 (print) | LCCN 2018036674 (ebook) | ISBN 9781119405276 (Adobe PDF) | ISBN 9781119405283 (ePub) | ISBN 9781119405269 (hardcover)
Subjects: LCSH: Multivariate analysis.
Classification: LCC QA278 (ebook) | LCC QA278 .A355 2019 (print) | DDC 519.5/35–dc23
LC record available at https://lccn.loc.gov/2018026887

Cover Design: Wiley
Cover Image: © iStock.com/Anna_Zubkova

Set in 10/12.5pt Nimbus by Aptara Inc., New Delhi, India

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTENTS

PREFACE

In recent years, the use of specialized statistical methods for categorical data has increased dramatically, particularly for applications in the biomedical and social sciences. Partly this reflects the development during the past few decades of sophisticated methods for analyzing categorical data. It also reflects the increasing methodological sophistication of scientists and applied statisticians, most of whom now realize that it is unnecessary and often inappropriate to use methods for continuous data with categorical responses.

This third edition of the book is a substantial revision of the second edition. The most important change is showing how to conduct all the analyses using `R` software. As in the first two editions, the main focus is presenting the most important methods for analyzing categorical data. The book summarizes methods that have long played a prominent role, such as chi-squared tests, but gives special emphasis to modeling techniques, in particular to logistic regression.

The presentation in this book has a low technical level and does not require familiarity with advanced mathematics such as calculus or matrix algebra. Readers should possess a background that includes material from a two-semester statistical methods sequence for undergraduate or graduate nonstatistics majors. This background should include estimation and significance testing and exposure to regression modeling.

This book is designed for students taking an introductory course in categorical data analysis, but I also have written it for applied statisticians and practicing scientists involved in data analyses. I hope that the book will be helpful to analysts dealing with categorical response data in the social, behavioral, and biomedical sciences, as well as in public health, marketing, education, biological and agricultural sciences, and industrial quality control.

The basics of categorical data analysis are covered in Chapters 1 to 7. Chapter 2 surveys standard descriptive and inferential methods for contingency tables, such as odds ratios, tests

of independence, and conditional versus marginal associations. I feel that an understanding of methods is enhanced, however, by viewing them in the context of statistical models. Thus, the rest of the text focuses on the modeling of categorical responses. I prefer to teach categorical data methods by unifying their models with ordinary regression models. Chapter 3 does this under the umbrella of generalized linear models. That chapter introduces generalized linear models for binary data and count data. Chapters 4 and 5 discuss the most important such model for binary data, logistic regression. Chapter 6 introduces logistic regression models for multicategory responses, both nominal and ordinal. Chapter 7 discusses loglinear models for contingency tables and other types of count data.

I believe that logistic regression models deserve more attention than loglinear models, because applications more commonly focus on the relationship between a categorical response variable and some explanatory variables (which logistic regression models do) than on the association structure among several response variables (which loglinear models do). Thus, I have given main attention to logistic regression in these chapters and in later chapters that discuss extensions of this model.

Chapter 8 presents methods for matched-pairs data. Chapters 9 and 10 extend the matched-pairs methods to apply to clustered, correlated observations. Chapter 9 does this with marginal models, emphasizing the generalized estimating equations (GEE) approach, whereas Chapter 10 uses random effects to model more fully the dependence. Chapter 11 is a new chapter, presenting classification and smoothing methods. That chapter also introduces regularization methods that are increasingly important with the advent of data sets having large numbers of explanatory variables. Chapter 12 provides a historical perspective of the development of the methods. The text concludes with an appendix showing the use of `R`, SAS, Stata, and SPSS software for conducting nearly all methods presented in this book. Many of the chapters now also show how to use the Bayesian approach to conduct the analyses.

The material in Chapters 1 to 7 forms the heart of an introductory course in categorical data analysis. Sections that can be skipped if desired, to provide more time for other topics, include Sections 1.5, 2.5–2.7, 3.3 and 3.5, 5.4–5.6, 6.3–6.4, and 7.4–7.6. Instructors can choose sections from Chapters 8 to 12 to supplement the topics of primary importance. Sections and subsections labeled with an asterisk can be skipped for those wanting a briefer survey of the methods.

This book has lower technical level than my book *Categorical Data Analysis* (3rd edition, Wiley 2013). I hope that it will appeal to readers who prefer a more applied focus than that book provides. For instance, this book does not attempt to derive likelihood equations, prove asymptotic distributions, or cite current research work.

Most methods for categorical data analysis require extensive computations. For the most part, I have avoided details about complex calculations, feeling that statistical software should relieve this drudgery. The text shows how to use `R` to obtain all the analyses presented. The Appendix discusses the use of SAS, Stata, and SPSS. The full data sets analyzed in the book are available at the text website `www.stat.ufl.edu/~aa/cat/data`. That website also lists typos and errors of which I have become aware since publication. The data files are also available at `https://github.com/alanagresti/categorical-data`.

Brief solutions to odd-numbered exercises appear at the end of the text. An instructor's manual will be included on the companion website for this edition: `www.wiley.com/go/Agresti/CDA_3e`. The aforementioned data sets will also be available on the companion website. Additional exercises are available there and at `www.stat.ufl.edu/`

`~aa/cat/Extra_Exercises`, some taken from the 2nd edition to create space for new material in this edition and some being slightly more technical.

I owe very special thanks to Brian Marx for his many suggestions about the text over the past twenty years. He has been incredibly generous with his time in providing feedback based on teaching courses based on the book. I also thank those individuals who commented on parts of the manuscript or who made suggestions about examples or material to cover or provided other help such as noticing errors. Travis Gerke, Anna Gottard, and Keramat Nourijelyani gave me several helpful comments. Thanks also to Alessandra Brazzale, Debora Giovannelli, David Groggel, Stacey Handcock, Maria Kateri, Bernhard Klingenberg, Ioannis Kosmidis, Mohammad Mansournia, Trevelyan McKinley, Changsoon Park, Tom Piazza, Brett Presnell, Ori Rosen, Ralph Scherer, Claudia Tarantola, Anestis Touloumis, Thomas Yee, Jin Wang, and Sherry Wang. I also owe thanks to those who helped with the first two editions, especially Patricia Altham, James Booth, Jane Brockmann, Brian Caffo, Brent Coull, Al DeMaris, Anna Gottard, Harry Khamis, Svend Kreiner, Carla Rampichini, Stephen Stigler, and Larry Winner. Thanks to those who helped with material for my more advanced text (*Categorical Data Analysis*) that I extracted here, especially Bernhard Klingenberg, Yongyi Min, and Brian Caffo. Many thanks also to the staff at Wiley for their usual high-quality help.

A truly special by-product for me of writing books about categorical data analysis has been invitations to teach short courses based on them and spend research visits at many institutions around the world. With grateful thanks I dedicate this book to my hosts over the years. In particular, I thank my hosts in Italy (Adelchi Azzalini, Elena Beccalli, Rino Bellocco, Matilde Bini, Giovanna Boccuzzo, Alessandra Brazzale, Silvia Cagnone, Paula Cerchiello, Andrea Cerioli, Monica Chiogna, Guido Consonni, Adriano Decarli, Mauro Gasparini, Alessandra Giovagnoli, Sabrina Giordano, Paolo Giudici, Anna Gottard, Alessandra Guglielmi, Maria Iannario, Gianfranco Lovison, Claudio Lupi, Monia Lupparelli, Maura Mezzetti, Antonietta Mira, Roberta Paroli, Domenico Piccolo, Irene Poli, Alessandra Salvan, Nicola Sartori, Bruno Scarpa, Elena Stanghellini, Claudia Tarantola, Cristiano Varin, Roberta Varriale, Laura Ventura, Diego Zappa), the UK (Phil Brown, Bianca De Stavola, Brian Francis, Byron Jones, Gillian Lancaster, Irini Moustaki, Chris Skinner, Briony Teather), Austria (Regina Dittrich, Gilg Seeber, Helga Wagner), Belgium (Hermann Callaert, Geert Molenberghs), France (Antoine De Falguerolles, Jean-Yves Mary, Agnes Rogel), Germany (Maria Kateri, Gerhard Tutz), Greece (Maria Kateri, Ioannis Ntzoufras), the Netherlands (Ivo Molenaar, Marijte van Duijn, Peter van der Heijden), Norway (Petter Laake), Portugal (Francisco Carvalho, Adelaide Freitas, Pedro Oliveira, Carlos Daniel Paulino), Slovenia (Janez Stare), Spain (Elias Moreno), Sweden (Juni Palmgren, Elisabeth Svensson, Dietrich van Rosen), Switzerland (Anthony Davison, Paul Embrechts), Brazil (Clarice Demetrio, Bent Jörgensen, Francisco Louzada, Denise Santos), Chile (Guido Del Pino), Colombia (Marta Lucia Corrales Bossio, Leonardo Trujillo), Turkey (Aylin Alin), Mexico (Guillermina Eslava), Australia (Chris Lloyd), China (I-Ming Liu, Chongqi Zhang), Japan (Ritei Shibata), and New Zealand (Nye John, I-Ming Liu). Finally, thanks to my wife, Jacki Levine, for putting up with my travel schedule in these visits around the world!

ALAN AGRESTI

Gainesville, Florida and Brookline, Massachusetts
March 2018

ABOUT THE COMPANION WEBSITE

This book comes with a companion website of other material, including all data sets analyzed in the book and some extra exercises.

www.wiley.com/go/Agresti/CDA_3e 

CHAPTER 1

INTRODUCTION

From helping to assess the value of new medical treatments to evaluating the factors that affect our opinions on controversial issues, scientists today are finding myriad uses for categorical data analyses. It is primarily for these scientists and their collaborating statisticians – as well as those training to perform these roles – that this book was written.

This first chapter reviews the most important probability distributions for categorical data: the *binomial* and *multinomial* distributions. It also introduces *maximum likelihood*, the most popular method for using data to estimate parameters. We use this type of estimate and a related *likelihood function* to conduct statistical inference. We also introduce the *Bayesian* approach to statistical inference, which utilizes probability distributions for the parameters as well as for the data. We begin by describing the major types of categorical data.

1.1 CATEGORICAL RESPONSE DATA

A *categorical* variable has a measurement scale consisting of a set of categories. For example, political ideology might be measured as liberal, moderate, or conservative; choice of accommodation might use categories house, condominium, and apartment; a diagnostic test to detect e-mail spam might classify an incoming e-mail message as spam or legitimate. Categorical variables are often referred to as *qualitative*, to distinguish them from *quantitative* variables, which take numerical values, such as age, income, and number of children in a family.

An Introduction to Categorical Data Analysis, Third Edition. Alan Agresti.

Companion Website: www.wiley.com/go/Agresti/CDA_3e

Categorical variables are pervasive in the social sciences for measuring attitudes and opinions, with categories such as (agree, disagree), (yes, no), and (favor, oppose, undecided). They also occur frequently in the health sciences, for measuring responses such as whether a medical treatment is successful (yes, no), mammogram-based breast diagnosis (normal, benign, probably benign, suspicious, malignant with cancer), and stage of a disease (initial, intermediate, advanced). Categorical variables are common for service-quality ratings of any company or organization that has customers (e.g., with categories excellent, good, fair, poor). In fact, categorical variables occur frequently in most disciplines. Other examples include the behavioral sciences (e.g., diagnosis of type of mental illness, with categories schizophrenia, depression, neurosis), ecology (e.g., primary land use in satellite image, with categories woodland, swamp, grassland, agriculture, urban), education (e.g., student responses to an exam question, with categories correct, incorrect), and marketing (e.g., consumer cell-phone preference, with categories Samsung, Apple, Nokia, LG, Other). They even occur in highly quantitative fields such as the engineering sciences and industrial quality control, when items are classified according to whether or not they conform to certain standards.

1.1.1 Response Variable and Explanatory Variables

Most statistical analyses distinguish between a *response* variable and *explanatory* variables. For instance, ordinary regression models describe how the mean of a quantitative response variable, such as annual income, changes according to levels of explanatory variables, such as number of years of education and number of years of job experience. The response variable is sometimes called the *dependent variable* and the explanatory variable is sometimes called the *independent variable*. When we want to emphasize that the response variable is a random variable, such as in a probability statement, we use upper-case notation for it (e.g., Y). We use lower-case notation to refer to a particular value (e.g., $y = 0$).

This text presents statistical models that relate a categorical response variable to explanatory variables that can be categorical or quantitative. For example, a study might analyze how opinion about whether same-sex marriage should be legal (yes or no) is associated with explanatory variables such as number of years of education, annual income, political party affiliation, religious affiliation, age, gender, and race.

1.1.2 Binary–Nominal–Ordinal Scale Distinction

Many categorical variables have only two categories, such as (yes, no) for possessing health insurance or (favor, oppose) for legalization of marijuana. Such variables are called *binary variables*.

When a categorical variable has more then two categories, we distinguish between two types of categorical scales. Categorical variables having *unordered* scales are called *nominal* variables. Examples are religious affiliation (categories Christian, Jewish, Muslim, Buddhist, Hindu, none, other), primary mode of transportation to work (automobile, bicycle, bus, subway, walk), and favorite type of music (classical, country, folk, jazz, pop, rock). Variables having naturally *ordered* categories are called *ordinal* variables. Examples are perceived happiness (not too happy, pretty happy, very happy), frequency of feeling anxiety (never, occasionally, often, always), and headache pain (none, slight, moderate, severe).

A variable's measurement scale determines which statistical methods are appropriate. For nominal variables, the order of listing the categories is arbitrary, so methods designed for them give the same results no matter what order is used. Methods designed for ordinal variables utilize the category ordering.

1.1.3 Organization of this Book

Chapters 1 and 2 describe basic non model-based methods of categorical data analysis. These include analyses of proportions and of association between categorical variables.

Chapters 3 to 7 introduce models for categorical response variables. These models resemble regression models for quantitative response variables. In fact, Chapter 3 shows they are special cases of a class of *generalized linear models* that also contains the ordinary normal-distribution-based regression models. *Logistic regression* models, which apply to binary response variables, are the focus of Chapters 4 and 5. Chapter 6 extends logistic regression to multicategory responses, both nominal and ordinal. Chapter 7 introduces *loglinear* models, which analyze associations among multiple categorical response variables.

The methods in Chapters 1 to 7 assume that observations are independent. Chapters 8 to 10 introduce logistic regression models for observations that are correlated, such as for matched pairs or for repeated measurement of individuals in longitudinal studies. Chapter 11 introduces some advanced methods, including ways of classifying and clustering observations into categories and ways of dealing with data sets having huge numbers of variables. The book concludes (Chapter 12) with a historical overview of the development of categorical data methods.

Statistical software packages can implement methods for categorical data analysis. We illustrate throughout the text for the free software `R`. The Appendix discusses the use of SAS, Stata, and SPSS. A companion website for the book, `www.stat.ufl.edu/~aa/cat`, has additional information, including complete data sets for the examples. The data files are also available at `https://github.com/alanagresti/categorical-data`.

1.2 PROBABILITY DISTRIBUTIONS FOR CATEGORICAL DATA

Parametric inferential statistical analyses require an assumption about the probability distribution of the response variable. For regression models for quantitative variables, the normal distribution plays a central role. This section presents the key probability distributions for categorical variables: the *binomial* and *multinomial* distributions.

1.2.1 Binomial Distribution

When the response variable is binary, we refer to the two outcome categories as *success* and *failure*. These labels are generic and the *success* outcome need not be a preferred result.

Many applications refer to a fixed number n of independent and identical trials with two possible outcomes for each. *Identical trials* means that the probability of success is the same for each trial. *Independent trials* means the response outcomes are independent random variables. In particular, the outcome of one trial does not affect the outcome of another. These are often called *Bernoulli trials*. Let π denote the probability of success for

each trial. Let Y denote the number of successes out of the n trials. Under the assumption of n independent, identical trials, Y has the *binomial distribution* with index n and parameter π. The probability of a particular outcome y for Y equals

$$P(y) = \frac{n!}{y!(n-y)!}\pi^y(1-\pi)^{n-y}, \quad y = 0, 1, 2, \ldots, n. \tag{1.1}$$

To illustrate, suppose a quiz has ten multiple-choice questions, with five possible answers for each. A student who is completely unprepared randomly guesses the answer for each question. Let Y denote the number of correct responses. For each question, the probability of a correct response is 0.20, so $\pi = 0.20$ with $n = 10$. The probability of $y = 0$ correct responses, and hence $n - y = 10$ incorrect ones, equals

$$P(0) = \frac{10!}{0!10!}(0.20)^0(0.80)^{10} = (0.80)^{10} = 0.107.$$

The probability of 1 correct response equals

$$P(1) = \frac{10!}{1!9!}(0.20)^1(0.80)^9 = 10(0.20)(0.80)^9 = 0.268.$$

Table 1.1 shows the binomial distribution for all the possible values, $y = 0, 1, 2, \ldots, 10$. For contrast, it also shows the binomial distributions when $\pi = 0.50$ and when $\pi = 0.80$.

Table 1.1 Binomial distributions with $n = 10$ and $\pi = 0.20$, 0.50, and 0.80. The binomial distribution is symmetric when $\pi = 0.50$.

y	$P(y)$ when $\pi = 0.20$ ($\mu = 2.0$, $\sigma = 1.26$)	$P(y)$ when $\pi = 0.50$ ($\mu = 5.0$, $\sigma = 1.58$)	$P(y)$ when $\pi = 0.80$ ($\mu = 8.0$, $\sigma = 1.26$)
0	0.107	0.001	0.000
1	0.268	0.010	0.000
2	0.302	0.044	0.000
3	0.201	0.117	0.001
4	0.088	0.205	0.005
5	0.027	0.246	0.027
6	0.005	0.205	0.088
7	0.001	0.117	0.201
8	0.000	0.044	0.302
9	0.000	0.010	0.268
10	0.000	0.001	0.107

The binomial distribution for n trials with parameter π has mean and standard deviation

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

The binomial distribution with $\pi = 0.20$ in Table 1.1 has $\mu = 10(0.20) = 2.0$. The standard deviation is $\sigma = \sqrt{10(0.20)(0.80)} = 1.26$, which σ also equals when $\pi = 0.80$.

The binomial distribution is symmetric when $\pi = 0.50$. For fixed n, it becomes more bell-shaped as π gets closer to 0.50. For fixed π, it becomes more bell-shaped as n increases.

When n is large, it can be approximated by a normal distribution with $\mu = n\pi$ and $\sigma = \sqrt{n\pi(1-\pi)}$. A guideline[1] is that the expected number of outcomes of the two types, $n\pi$ and $n(1-\pi)$, should both be at least about 5. For $\pi = 0.50$ this requires only $n \geq 10$, whereas $\pi = 0.10$ (or $\pi = 0.90$) requires $n \geq 50$. When π gets nearer to 0 or 1, larger samples are needed before a symmetric, bell shape occurs.

1.2.2 Multinomial Distribution

Nominal and ordinal response variables have more than two possible outcomes. When the observations are independent with the same category probabilities for each, the probability distribution of counts in the outcome categories is the *multinomial*.

Let c denote the number of outcome categories. We denote their probabilities by $(\pi_1, \pi_2, \ldots, \pi_c)$, where $\sum_j \pi_j = 1$. For n independent observations, the multinomial probability that y_1 fall in category 1, y_2 fall in category 2, ... , y_c fall in category c, where $\sum_j y_j = n$, equals

$$P(y_1, y_2, ..., y_c) = \left(\frac{n!}{y_1!y_2!\cdots y_c!}\right)\pi_1^{y_1}\pi_2^{y_2}\cdots\pi_c^{y_c}.$$

The binomial distribution is the special case with $c = 2$ categories. We will not need to use this formula, because our focus is on inference methods that use *sampling distributions* of statistics computed from the multinomial counts, and those sampling distributions are approximately *normal* or *chi-squared*.

1.3 STATISTICAL INFERENCE FOR A PROPORTION

In practice, the parameter values for binomial and multinomial distributions are unknown. Using sample data, we estimate the parameters. This section introduces the *maximum likelihood* estimation method and illustrates it for the binomial parameter.

1.3.1 Likelihood Function and Maximum Likelihood Estimation

The parametric approach to statistical modeling assumes a family of probability distributions for the response variable, indexed by an unknown parameter. For a particular family, we can substitute the observed data into the formula for the probability function and then view how that probability depends on the unknown parameter value. For example, in $n = 10$ trials, suppose a binomial count equals $y = 0$. From the binomial formula (1.1) with parameter π, the probability of this outcome equals

$$P(0) = \frac{10!}{0!10!}\pi^0(1-\pi)^{10} = (1-\pi)^{10}.$$

This probability is defined for all the potential values of π between 0 and 1.

[1] You can explore this with the binomial distribution applet at `www.artofstat.com/webapps.html`.

The probability of the observed data, expressed as a function of the parameter, is called the *likelihood function*. With $y = 0$ successes in $n = 10$ trials, the binomial likelihood function is $\ell(\pi) = (1 - \pi)^{10}$, for $0 \leq \pi \leq 1$. If $\pi = 0.40$, for example, the probability that $y = 0$ is $\ell(0.40) = (1 - 0.40)^{10} = 0.006$. Likewise, if $\pi = 0.20$ then $\ell(0.20) = (1 - 0.20)^{10} = 0.107$, and if $\pi = 0.0$ then $\ell(0.0) = (1 - 0.0)^{10} = 1.0$. Figure 1.1 plots this likelihood function for all π values between 0 and 1.

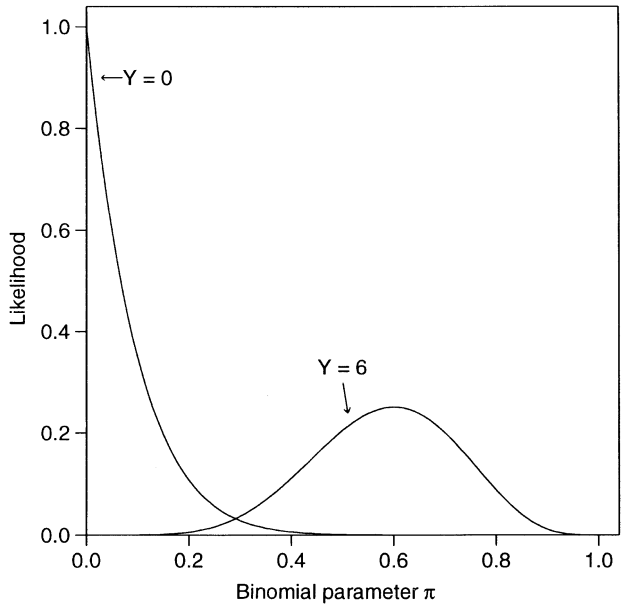


Figure 1.1 Binomial likelihood functions for $y = 0$ successes and for $y = 6$ successes in $n = 10$ trials.

The *maximum likelihood estimate* of a parameter is the parameter value at which the likelihood function takes its maximum. That is, it is the parameter value for which the probability of the observed data takes its greatest value. Figure 1.1 shows that the likelihood function $\ell(\pi) = (1 - \pi)^{10}$ has its maximum at $\pi = 0.0$. Therefore, when $n = 10$ trials have $y = 0$ successes, the maximum likelihood estimate of π equals 0.0. This means that the result $y = 0$ in $n = 10$ trials is more likely to occur when $\pi = 0.00$ than when π equals any other value.

We use the abbreviation *ML* to symbolize *maximum likelihood*. The ML estimate is often denoted by the parameter symbol with a ^ (a *hat*) over it. We denote the ML estimate of the binomial parameter π by $\hat{\pi}$, called *pi-hat*. In general, for the binomial outcome of y successes in n trials, the maximum likelihood estimate of π is $\hat{\pi} = y/n$. This is the sample proportion of successes for the n trials. If we observe $y = 6$ successes in $n = 10$ trials, then the maximum likelihood estimate of π is $\hat{\pi} = 6/10 = 0.60$. Figure 1.1 also plots the likelihood function when $n = 10$ with $y = 6$, which from formula (1.1) equals $\ell(\pi) = [10!/(6!4!)]\pi^6(1 - \pi)^4$. The maximum value occurs at $\hat{\pi} = 0.60$. The result $y = 6$ in $n = 10$ trials is more likely to occur when $\pi = 0.60$ than when π equals any other value.

If we denote each success by a 1 and each failure by a 0, then the sample proportion equals the sample mean of the data. For instance, for 4 failures followed by 6 successes in 10 trials, the data are $(0, 0, 0, 0, 1, 1, 1, 1, 1, 1)$ and the sample mean is

$$\hat{\pi} = (0 + 0 + 0 + 0 + 1 + 1 + 1 + 1 + 1 + 1)/10 = 0.60.$$

Thus, results that apply to sample means with random sampling apply also to sample proportions. These include the *Central Limit Theorem*, which states that the sampling distribution of the sample proportion $\hat{\pi}$ is approximately normal for large n, and the *Law of Large Numbers*, which states that $\hat{\pi}$ converges to the population proportion π as n increases.

Before we observe the data, the value of the ML estimate is unknown. The estimate is then a random variable having some sampling distribution. We refer to it as an *estimator* and its value for observed data as an *estimate*. Estimators based on the method of maximum likelihood are popular because they have good large-sample behavior. Sampling distributions of ML estimators are typically approximately normal and no other "good" estimator has a smaller standard error.

1.3.2 Significance Test About a Binomial Parameter

For the binomial distribution, we now use the ML estimator in statistical inference for the parameter π. The ML estimator $\hat{\pi}$ is the sample proportion. Its sampling distribution has mean and standard error

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

Consider the null hypothesis H_0: $\pi = \pi_0$ that the parameter equals some fixed value, π_0, such as 0.50. When H_0 is true, the standard error of $\hat{\pi}$ is $SE_0 = \sqrt{\pi_0(1-\pi_0)/n}$, which we refer to as the *null standard error*. The test statistic

$$z = \frac{\hat{\pi} - \pi_0}{SE_0} = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \tag{1.2}$$

divides the difference between the sample proportion $\hat{\pi}$ and the null hypothesis value π_0 by the null standard error. The z test statistic measures the number of standard errors that $\hat{\pi}$ falls from the H_0 value. For large samples, the null sampling distribution of z is the standard normal, which has mean $= 0$ and standard deviation $= 1$.

1.3.3 Example: Surveyed Opinions About Legalized Abortion

Do a majority, or minority, of adults in the United States believe that a pregnant woman should be able to obtain an abortion? Let π denote the proportion of the American adult population that responds *yes* when asked, "Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if she wants it for any reason." We test H_0: $\pi = 0.50$ against the two-sided alternative hypothesis, H_a: $\pi \neq 0.50$.

This item was one of many about legalized abortion included in the 2016 General Social Survey (GSS). This survey, conducted every other year by the National Opinion Research Center (NORC) at the University of Chicago, asks a sample of adult Americans their opinions about a wide variety of issues.[2] The GSS is a multi-stage sample, but it has characteristics similar to a simple random sample. Of 1810 respondents to this item in 2016, 837 replied *yes* and 973 replied *no*. The sample proportion of *yes* responses was $\hat{\pi} = 837/1810 = 0.4624$.

[2] You can view responses to surveys since 1972 at `sda.berkeley.edu/archive.htm`.

The test statistic for H_0: $\pi = 0.50$ is

$$z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0.4624 - 0.50}{\sqrt{\frac{0.50(0.50)}{1810}}} = -3.20.$$

The two-sided P-value is the probability that the absolute value of a standard normal variate exceeds 3.20, which is $P = 0.0014$. The evidence is very strong that, in 2016, $\pi < 0.50$, that is, that fewer than half of Americans favored unrestricted legal abortion. In some other situations, such as when the mother's health was endangered, an overwhelming majority favored legalized abortion. Responses depended strongly on the question wording.

1.3.4 Confidence Intervals for a Binomial Parameter

A significance test merely indicates whether a particular value for a parameter (such as 0.50) is plausible. We learn more by constructing a confidence interval to determine the range of plausible values. Let $SE = \sqrt{\hat{\pi}(1-\hat{\pi})/n}$ denote the estimated standard error of $\hat{\pi}$. This formula obtains SE by substituting the ML estimate $\hat{\pi}$ for the unknown parameter π in $\sigma(\hat{\pi}) = \sqrt{\pi(1-\pi)/n}$. One way to form a $100(1-\alpha)\%$ confidence interval for π uses the formula

$$\hat{\pi} \pm z_{\alpha/2}(SE), \text{ with } SE = \sqrt{\hat{\pi}(1-\hat{\pi})/n}, \tag{1.3}$$

where $z_{\alpha/2}$ denotes the standard normal percentile having right-tail probability equal to $\alpha/2$; for example, for 95% confidence, $\alpha = 0.05$, $z_{\alpha/2} = z_{0.025} = 1.96$.

For the opinion about legalized abortion example just discussed, $\hat{\pi} = 0.462$ for $n = 1810$ observations. The 95% confidence interval equals

$$0.462 \pm 1.96\sqrt{0.462(0.538)/1810}, \text{ which is } 0.462 \pm 0.023, \text{ or } (0.439, 0.485).$$

We can be 95% confident that the population proportion of Americans in 2016 who favored unrestricted legalized abortion is between 0.439 and 0.485.

The significance test and confidence interval for π, as well as other confidence intervals presented next, are readily available in software and at web sites.[3]

1.3.5 Better Confidence Intervals for a Binomial Proportion *

Formula (1.3) is simple. When π is near 0 or near 1, however, it performs poorly unless n is very large. Its *actual* coverage probability, that is, the probability that the method produces an interval that captures the true parameter value, may be much less than the nominal value (such as 0.95).

A better way to construct confidence intervals uses a duality with significance tests. The confidence interval consists of all H_0 values π_0 that are judged plausible in the z test of Section 1.3.2. A 95% confidence interval contains all values π_0 for which the two-sided P-value exceeds 0.05. That is, it contains all values that are *not rejected* at the 0.05

[3] For instance, see `https://istats.shinyapps.io/Inference_prop`. The confidence interval (1.3) is the *Wald* type listed in the menu.

significance level. These are the H_0 values for π_0 that have test statistic z less than 1.96 in absolute value. This alternative method, called the *score confidence interval*, has the advantage that we do not need to estimate π in the standard error, because the standard error $SE_0 = \sqrt{\pi_0(1-\pi_0)/n}$ in the test statistic uses the null value π_0.

To illustrate, suppose that a clinical trial to evaluate a new treatment has 9 successes in the first 10 trials. For a sample proportion of $\hat{\pi} = 0.90$ based on $n = 10$, the value $\pi_0 = 0.596$ for the H_0 parameter value yields the test statistic value

$$z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0.90 - 0.596}{\sqrt{\frac{0.596(0.404)}{10}}} = 1.96$$

and a two-sided P-value of $P = 0.05$. The value $\pi_0 = 0.982$ yields

$$z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0.90 - 0.982}{\sqrt{\frac{0.982(0.018)}{10}}} = -1.96$$

and also a two-sided P-value of $P = 0.05$. All π_0 values between 0.596 and 0.982 have $|z| < 1.96$ and P-value > 0.05. Therefore, the 95% score confidence interval for π is (0.596, 0.982). For particular values of $\hat{\pi}$ and n, the π_0 values that have test statistic value $z = \pm 1.96$ are the solutions to the equation

$$\frac{|\,\hat{\pi} - \pi_0\,|}{\sqrt{\pi_0(1-\pi_0)/n}} = 1.96$$

for π_0. We will not deal here with how to solve this equation, as this confidence interval is readily available in software and at web sites.[4]

The simple formula (1.3) using estimated standard error fails spectacularly when $\hat{\pi} = 0$ or when $\hat{\pi} = 1$, regardless of how large n is. To illustrate, suppose the clinical trial had 10 successes in the 10 trials. Then, $\hat{\pi} = 10/10 = 1.0$ and $SE = \sqrt{\hat{\pi}(1-\hat{\pi})/n} = \sqrt{1.0(0.0)/10} = 0$, so the 95% confidence interval $1.0 \pm 1.96(SE)$ is 1.0 ± 0.0. This interval (1.0, 1.0) is completely unrealistic. When a sample estimate is at or near the boundary of the parameter space, having that estimate in the middle of the confidence interval results in poor performance of the method. By contrast, the 95% score confidence interval based on the corresponding significance test with null standard error SE_0 is (0.72, 1.0).

The score confidence interval itself has actual coverage probability a bit too small when π is very close to 0 or 1. A simple alternative confidence interval approximates the score interval but is a bit wider and has better coverage probability when π is near 0 or 1. It uses the simple formula (1.3) with the estimated standard error after adding 2 to the number of successes and 2 to the number of failures (and thus 4 to n). With 10 successes in 10 trials, you apply formula (1.3) to 12 successes in 14 trials and get (0.68, 1.0). This simple method,[5] called the *Agresti–Coull confidence interval*, has adequate coverage probability for small n even when π is very close to 0 or 1.

[4] Such as the "Wilson score" option at `https://istats.shinyapps.io/Inference_prop`.

[5] More precisely, software and the website `https://istats.shinyapps.io/Inference_prop` adds $(z^2_{\alpha/2})/2$ to each count (e.g., $(1.96)^2/2 = 1.92$ for 95% confidence); the CI then performs well because it has the same midpoint as the score CI but is a bit wider.

1.4 STATISTICAL INFERENCE FOR DISCRETE DATA

In summary, two methods we have presented for constructing a confidence interval for a proportion (1) use $\hat{\pi} \pm z_{\alpha/2}(SE)$ with the estimated standard error, $SE = \sqrt{\hat{\pi}(1-\hat{\pi})/n}$, or (2) invert results of a significance test using test statistic $z = (\hat{\pi} - \pi_0)/SE_0$ with the null standard error, $SE_0 = \sqrt{\pi_0(1-\pi_0)/n}$. These methods apply two of the three standard ways of conducting statistical inference (confidence intervals and significance tests) about parameters. We present the methods in a more general context in this section and also introduce a third standard inference method that uses the likelihood function.

1.4.1 Wald, Likelihood-Ratio, and Score Tests

Let β denote an arbitrary parameter, such as a linear effect of an explanatory variable in a model. Consider a significance test of H_0: $\beta = \beta_0$, such as H_0: $\beta = 0$ for which $\beta_0 = 0$. The simplest test statistic exploits the large-sample normality of the ML estimator $\hat{\beta}$. Let SE denote the unrestricted standard error of $\hat{\beta}$, evaluated by substituting the ML estimate for the unknown parameter in the expression for the true standard error. (For example, for the binomial parameter π, $SE = \sqrt{\hat{\pi}(1-\hat{\pi})/n}$.) When H_0 is true, the test statistic

$$z = (\hat{\beta} - \beta_0)/SE$$

has approximately a standard normal distribution. Equivalently, z^2 has approximately a chi-squared distribution with $df = 1$. This type of statistic, which uses the standard error evaluated at the ML estimate, is called a *Wald statistic*. The z test using this test statistic, or the corresponding chi-squared test that uses z^2, is called a *Wald test*.[6]

We can refer z to the standard normal distribution to get one-sided or two-sided P-values. For the two-sided alternative H_a: $\beta \neq \beta_0$, the P-value is also the right-tail chi-squared probability with $df = 1$ above the observed value of z^2. That is, the two-tail probability beyond $\pm z$ for the standard normal distribution equals the right-tail probability above z^2 for the chi-squared distribution with $df = 1$. For example, the two-tail standard normal probability of 0.05 that falls below -1.96 and above 1.96 equals the right-tail chi-squared probability above $(1.96)^2 = 3.84$ when $df = 1$. With the software `R` and its functions `pnorm` and `pchisq` for *cumulative probabilities* (i.e., probabilities *below* fixed values) for normal and chi-squared distributions, we find (with comments added following the # symbol):

```
------------------------------------------------------------------------------
> 2*pnorm(-1.96) # 2(standard normal cumulative probability below -1.96)
[1] 0.0499958    # essentially equals 0.05
> pchisq(1.96^2, 1) # pchisq gives chi-squared cumulative probability
[1] 0.9500042       # here, cumul. prob. at (1.96)(1.96) = 3.84 when df=1
> 1 - pchisq(1.96^2, 1) # right-tail prob. above (1.96)(1.96) when df=1
[1] 0.0499958    # same as normal two-tail probability
> # can also get this by  pchisq(1.96^2, 1, lower.tail=FALSE)
------------------------------------------------------------------------------
```

You can also find chi-squared and normal tail probabilities with applets on the Internet.[7]

[6] Proposed by the statistician Abraham Wald in 1943.

[7] See, for example, the applets at `www.artofstat.com/webapps.html#Distributions`.

A second possible test is called the *score test*.[8] This test uses standard errors that are valid when H_0 is true, rather than estimated more generally. For example, the z test (1.2) for a binomial parameter that uses the null standard error $SE_0 = \sqrt{\pi_0(1-\pi_0)/n}$ of $\hat{\pi}$ is a score test. The z test that uses $SE = \sqrt{\hat{\pi}(1-\hat{\pi})/n}$ instead of SE_0 is the Wald test.

A third possible test of H_0: $\beta = \beta_0$ uses the likelihood function through the ratio of two of its values. For a single parameter β, these are (1) the value ℓ_0 when H_0 is true (so $\beta = \beta_0$), (2) the maximum ℓ_1 over all possible parameter values, which is the likelihood function calculated at the ML estimate $\hat{\beta}$. Then ℓ_1 is always at least as large as ℓ_0, because ℓ_1 refers to maximizing over the entire parameter space rather than just at β_0. The *likelihood-ratio* test statistic[9] equals

$$2\log(\ell_1/\ell_0).$$

The reason for taking the log transform and doubling is that it yields an approximate chi-squared sampling distribution. Under H_0: $\beta = \beta_0$, this test statistic has a large-sample chi-squared distribution with $df = 1$. The test statistic $2\log(\ell_1/\ell_0)$ is nonnegative, and the P-value is the chi-squared right-tail probability. Larger values of (ℓ_1/ℓ_0) yield larger values of $2\log(\ell_1/\ell_0)$ and smaller P-values and stronger evidence against H_0.

For ordinary regression models that assume a normal distribution for Y, the Wald, score, and likelihood-ratio tests provide identical test statistics and P-values. For parameters in other statistical models, they have similar behavior when the sample size n is large and H_0 is true. When n is small to moderate, the Wald test is the least reliable of the three tests. The likelihood-ratio inference and score-test based inference are better in terms of actual inferential error probabilities, coming closer to matching nominal levels.

For any of the three tests, the P-value that software reports is an approximation for the true P-value. This is because the normal (or chi-squared) sampling distribution used is a large-sample approximation for the actual sampling distribution. Thus, when you report a P-value, it is overly optimistic to use many decimal places. If you are lucky, the P-value approximation is good to the second decimal place. Therefore, for a P-value that software reports as 0.028374, it makes more sense to report it as 0.03 (or, at best, 0.028) rather than 0.028374. An exception is when the P-value is zero to many decimal places, in which case it is sensible to report it as $P < 0.001$ or $P < 0.0001$. A P-value merely summarizes the strength of evidence against H_0, and accuracy to two or three decimal places is sufficient for this purpose.

Each significance test method has a corresponding confidence interval. The 95% confidence interval for β is the set of β_0 values for the test of H_0: $\beta = \beta_0$ such that the P-value is larger than 0.05. For example, the 95% *Wald confidence interval* is the set of β_0 values for which $z = (\hat{\beta} - \beta_0)/SE$ has $|z| < 1.96$. It is $\hat{\beta} \pm 1.96(SE)$.

1.4.2 Example: Wald, Score, and Likelihood-Ratio Binomial Tests

We illustrate the Wald, score, and likelihood-ratio tests by testing H_0: $\pi = 0.50$ against H_a: $\pi \neq 0.50$ for the toy example mentioned on page 16 of a clinical trial to evaluate a new treatment that has 9 successes in $n = 10$ trials. The sample proportion is $\hat{\pi} = 0.90$.

[8] Proposed by the statistician Calyampudi Radhakrishna Rao in 1948.

[9] Proposed by the statistician Sam Wilks in 1938; in this text, we use the *natural log*, which has $e = 2.718...$ as the base. It is often denoted on calculators by LN.

For the Wald test, the estimated standard error is $SE = \sqrt{\hat{\pi}(1-\hat{\pi})/n} = \sqrt{0.90(0.10)/10} = 0.095$. The z test statistic is

$$z = (\hat{\pi} - \pi_0)/SE = (0.90 - 0.50)/0.095 = 4.22.$$

The corresponding chi-squared statistic is $(4.22)^2 = 17.78$ ($df = 1$). The P-value < 0.001.

For the score test, the null standard error is $SE_0 = \sqrt{\pi_0(1-\pi_0)/n} = \sqrt{0.50(0.50)/10} = 0.158$. The z test statistic is

$$z = (\hat{\pi} - \pi_0)/SE_0 = (0.90 - 0.50)/0.158 = 2.53.$$

The corresponding chi-squared statistic is $(2.53)^2 = 6.40$ ($df = 1$). The P-value $= 0.011$.

The likelihood function is the binomial probability of the observed result of 9 successes in 10 trials, viewed as a function of the parameter,

$$\ell(\pi) = \frac{10!}{9!1!}\pi^9(1-\pi)^1 = 10\pi^9(1-\pi).$$

The likelihood-ratio test compares this when H_0: $\pi = 0.50$ is true, for which $\ell_0 = 10(0.50)^9(0.50) = 0.00977$, to the value at the ML estimate of $\hat{\pi} = 0.90$, for which $\ell_1 = 10(0.90)^9(0.10) = 0.3874$. The likelihood-ratio test statistic equals

$$2\log(\ell_1/\ell_0) = 2[\log(0.3874/0.00977)] = 7.36.$$

From the chi-squared distribution with $df = 1$, this statistic has P-value $= 0.007$.

A marked divergence in the values of the three statistics, such as often happens when n is small and the ML estimate is near the boundary of the parameter space, indicates that the sampling distribution of the ML estimator may be far from normality and an estimate of the standard error may be poor. In that case, special small-sample methods are more reliable.

1.4.3 Small-Sample Binomial Inference and the Mid P-Value *

For statistical inference about a binomial parameter, the large-sample likelihood-ratio and two-sided score tests and the confidence intervals based on those tests perform reasonably well when $n\pi \geq 5$ and $n(1-\pi) \geq 5$. Otherwise, it is better to use the binomial distribution directly. With modern software, we can use this direct approach with any n.

To illustrate using the binomial directly, consider testing H_0: $\pi = 0.50$ against H_a: $\pi > 0.50$ for the toy example of a clinical trial, with $y = 9$ successes in $n = 10$ trials. The exact P-value, based on the right tail of the null binomial distribution with $\pi = 0.50$, is the binomial probability

$$P(Y \geq 9) = P(9) + P(10) = \frac{10!}{9!1!}(0.50)^9(0.50)^1 + \frac{10!}{10!0!}(0.50)^{10}(0.50)^0 = 0.011.$$

For the two-sided alternative H_a: $\pi \neq 0.50$, the P-value is

$$P(Y \geq 9 \text{ or } Y \leq 1) = 2[P(Y \geq 9)] = 0.021.$$

With discrete probability distributions, small-sample inference using the ordinary P-value is *conservative*. This means that when H_0 is true, the P-value is ≤ 0.05 (thus leading to rejection of H_0 at the 0.05 significance level) not *exactly* 5% of the time, but *no more* than 5% of the time. Then, the actual P(Type I error) is not exactly 0.05, but may be much less than 0.05. For example, for testing H_0: $\pi = 0.50$ against H_a: $\pi > 0.50$ with $y = 9$ successes in $n = 10$ trials, from the binomial probabilities with $\pi = 0.50$ in Table 1.1 in Section 1.2.1, the right-tail P-value is ≤ 0.05 only when $y = 9$ or 10. This happens with probability $0.010 + 0.001 = 0.011$. Thus, the probability of rejecting H_0 (i.e., getting a P-value ≤ 0.05) is only 0.011. That is, the actual P(Type I error) $= 0.011$, much smaller than the intended significance level of 0.05.

This illustrates an awkward aspect of small-sample significance testing when the test statistic has a discrete distribution. Imagine how a P-value, regarded as a random variable, may vary from study to study. For test statistics having a *continuous* distribution, the P-value has a *uniform* null distribution over the interval [0, 1]. That is, when H_0 is true, the P-value is equally likely to fall anywhere between 0 and 1. Then, the probability that the P-value falls below 0.05 equals exactly 0.05. The expected value of the P-value, that is, its long-run average value, is exactly 0.50. By contrast, for a test statistic having a *discrete* distribution, the null distribution of the P-value is discrete and has an expected value greater than 0.50 (e.g., it can equal 1.00 but never exactly 0.00). In this average sense, ordinary P-values for discrete distributions tend to be too large.

To address the conservatism difficulty, with discrete data we recommend using a different type of P-value. Called the *mid P-value*, it adds only *half* the probability of the observed result to the probability of the more extreme results. To illustrate, with $y = 9$ successes in $n = 10$ trials, the ordinary P-value for H_a: $\pi > 0.50$ is $P(9) + P(10) = 0.010 + 0.001 = 0.011$. The mid P-value is $[P(9)/2] + P(10) = (0.010/2) + 0.001 = 0.006$. The two-sided mid P-value for H_a: $\pi \neq 0.50$ is 0.012. The mid P-value has a null expected value of 0.50, the same as the regular P-value for test statistics that have a continuous distribution. Also, the two separate one-sided mid P-values sum to 1.0. By contrast, the observed result has probability counted in each tail for the ordinary one-sided P-values, so the two one-sided P values have a sum exceeding 1.

Inference based on the mid P-value compromises between the conservativeness of small-sample methods and the potential inadequacy of large-sample methods. It is also possible to construct a confidence interval for π from the set of π_0 values not rejected in the corresponding binomial test using the mid P-value. We shall do this with software in Section 1.6. In that section, we will see that it is straightforward to use software to obtain all the results for the examples in this chapter.

1.5 BAYESIAN INFERENCE FOR PROPORTIONS *

This book mainly uses the traditional, so-called *frequentist*, approach to statistical inference. This approach treats parameter values as fixed and data as realizations of random variables that have some assumed probability distribution. That is, probability statements refer to possible values for the data, given the parameter values. Recent years have seen increasing popularity of the *Bayesian* approach, which also treats parameters as random variables and therefore has probability distributions for them as well as for the data. This yields inferences

in the form of probability statements about possible values for the parameters, given the observed data.

1.5.1 The Bayesian Approach to Statistical Inference

The Bayesian approach assumes a *prior distribution* for the parameters. This probability distribution may reflect subjective prior beliefs, or it may reflect information about the parameter values from other studies, or it may be relatively non-informative so that inferential results are more objective, based almost entirely on the data. The prior distribution combines with the information that the data provide through the likelihood function to generate a *posterior distribution* for the parameters. The posterior distribution reflects the information about the parameters based both on the prior distribution and the data observed in the study.

For a parameter β and data denoted by y, let $f(\beta)$ denote the probability function[10] for the prior distribution of β. For example, when β is the binomial parameter π, this is a probability distribution over the interval [0, 1] of possible values for the probability π. Also, let $p(y \mid \beta)$ denote the probability function for the data, given the parameter value. (The vertical slash $\mid$ symbolizes "given" or "conditional on.") An example is the binomial formula (1.1), treating it as a function of y for fixed π. Finally, let $g(\beta \mid y)$ denote the probability function for the posterior distribution of β after we observe the data. In these symbols, from *Bayes' Theorem*,

$$g(\beta \mid y) \text{ is proportional to } p(y \mid \beta) f(\beta).$$

Now, after we observe the data, $p(y \mid \beta)$ is the likelihood function $\ell(\beta)$ when we view it as a function of the parameter. Therefore, the posterior distribution of the parameter is determined by the product of the likelihood function with the probability function for the prior distribution. When the prior distribution is relatively flat, as data analysts often choose in practice, the posterior distribution for the parameter has a similar shape to the likelihood function.

Except in a few simple cases, such as presented next for the binomial parameter, the posterior distribution cannot be easily calculated and software uses simulation methods to approximate it. The primary method for doing this is called *Markov chain Monte Carlo* (MCMC). It is beyond our scope to discuss the technical details of how an MCMC algorithm works. In a nutshell, software generates a very long sequence of values taken from an approximation for the posterior distribution. The data analyst takes the sequence to be long enough so that the Monte Carlo error is small in approximating the posterior distribution and summary measures of it, such as the mean.

For a particular parameter, Bayesian inference methods using the posterior distribution parallel those for frequentist inference. For example, analogous to the frequentist 95% confidence interval, we can construct an interval that contains 95% of the posterior distribution. Such an interval is referred to as a *posterior interval* or *credible interval*. A simple posterior interval uses percentiles of the posterior distribution, with equal probabilities in the two tails. For example, the 95% equal-tail posterior interval for a parameter is the region between the 2.5 and 97.5 percentiles of the posterior distribution. The mean of the posterior

[10] For a continuous distribution such as the normal, this is called the *probability density function*.