MULTIVARIATE APPLICATIONS SERIES

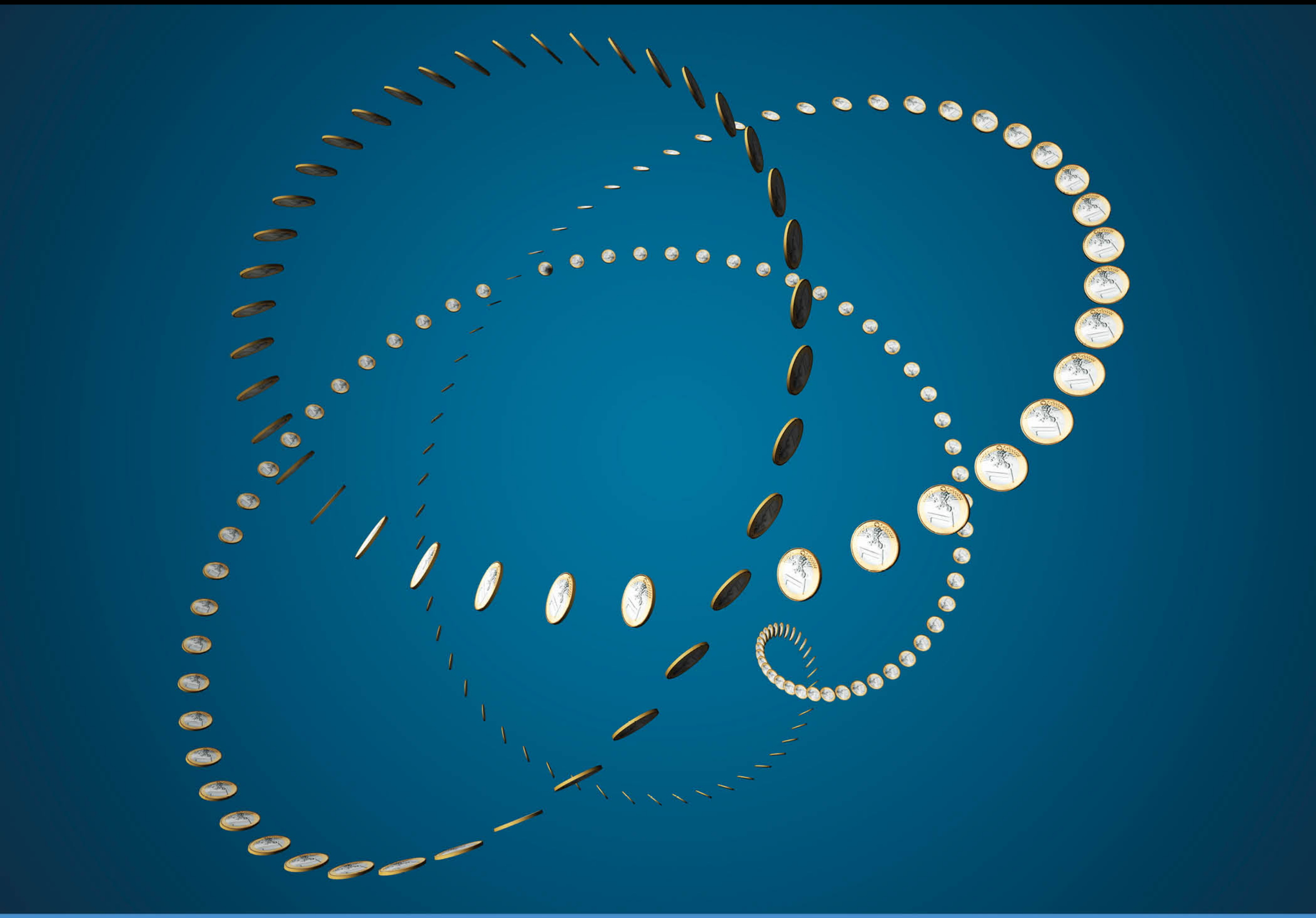
FREE INSTRUCTOR & STUDENT RESOURCES

ROUTLEDGE

CATEGORICAL AND NONPARAMETRIC DATA ANALYSIS

CHOOSING THE BEST STATISTICAL TECHNIQUE

E. MICHAEL NUSSBAUM



SECOND EDITION

“This is an excellent, hands-on, introduction to the core concepts of nonparametric statistics, with additional sections allowing for a gentle, in-depth study of advanced material. I was especially delighted when I found the nonparametric tests for factorial designs and repeated measurements. Widely disregarded for years, I am so happy to see these topics finally being covered in a textbook for psychology students, including cutting-edge results such as pseudorank-based methodology for unbalanced designs. The large number of applied examples and exercises fosters deep processing, comprehension of the concepts, and routine use of statistical software. It is obvious that Nussbaum and Brunner used their vast experience in teaching and research to provide a comprehensive overview on nonparametric statistics topics while keeping a realistic estimate with respect to academic formats and students’ skills.”

Matthias Gondan-Rochon, *University of Innsbruck, Austria*

“Highly recommended for graduate students in the social or biological sciences or education fields. Nussbaum clearly describes modern techniques for nonparametric and categorical data analysis using accessible language, numerous examples, and thought-provoking questions. The accompanying PowerPoint slides and newly added R code are invaluable.”

Jason E. King, *Baylor College of Medicine, USA*

“*Categorical and Nonparametric Data Analysis: Choosing the Best Statistical Technique* (2nd ed.) is an exceptional book about the why and how of nonparametric data analysis. It gives a comprehensive overview of the most important statistical tests, illustrating their use with careful explanations and examples. What makes this book stand out to me, however, is the amount of care the authors took to help readers appreciate the reasoning behind the statistical tests. In particular, the book provides a lot of guidance that makes it easy to understand which test is appropriate in a given scenario and why. When there are multiple tests that could be appropriate, the book provides explicit algorithms that help to decide when and why which test to use. Summaries wrap up the key points to make sure the reader always keeps sight of the main points rather than getting lost in the details.”

Nikos Bosse, *London School of Hygiene and Tropical Medicine, UK*

“This is a timely, up-to-date introduction to essential social science research tools that makes the complex accessible, and provides budding researchers with the tools they need—from the simple to the state of the art—in a consistent framework.”

Brendan Halpin, *University of Limerick, Ireland*

“The book uses real examples, step-by-step explanations and straightforward language to help the reader not only understand the statistical methods available for categorical and non-parametric data analysis, but also how to implement them in practice. . . . Also important, the book includes comprehensive explanations about computational and estimation methods often neglected in other texts.”

Irini Moustaki, *London School of Economics, UK*

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

CATEGORICAL AND NONPARAMETRIC DATA ANALYSIS

Now in its second edition, this book provides a focused, comprehensive overview of both categorical and nonparametric statistics, offering a conceptual framework for choosing the most appropriate test in various scenarios. The book's clear explanations and *Exploring the Concept* boxes help reduce reader anxiety. Problems inspired by actual studies provide meaningful illustrations of these techniques. Basic statistics and probability are reviewed for those needing a refresher with mathematical derivations placed in optional appendices.

Highlights include the following:

- Three chapters co-authored with Edgar Brunner address modern nonparametric techniques, along with accompanying R code.
- Unique coverage of both categorical and nonparametric statistics better prepares readers to select the best technique for particular research projects.
- Designed to be used with most statistical packages, clear examples of how to use the tests in SPSS, R, and Excel foster conceptual understanding.
- *Exploring the Concept* boxes integrated throughout prompt students to draw links between the concepts to deepen understanding.
- Fully developed Instructor and Student Resources featuring datasets for the book's problems and a guide to R, and for the instructor PowerPoints, author's syllabus, and answers to even-numbered problems.

Intended for graduate or advanced undergraduate courses in categorical and nonparametric statistics taught in psychology, education, human development, sociology, political science, and other social and life sciences.

E. Michael Nussbaum is a Professor of Educational Psychology at The University of Nevada, Las Vegas, USA. Dr. Nussbaum holds a PhD from Stanford University and an MPP from the University of California, Berkeley. He is the author of numerous research publications and serves on the editorial boards of the *Journal of Educational Psychology* and the *Educational Psychologist*.

MULTIVARIATE APPLICATIONS SERIES

Sponsored by the Society of Multivariate Experimental Psychology, the goal of this series is to apply statistical methods to significant social or behavioral issues, in such a way so as to be accessible to a nontechnical-oriented readership (e.g., non-methodological researchers, teachers, students, government personnel, practitioners, and other professionals). Applications from a variety of disciplines such as psychology, public health, sociology, education, and business are welcome. Books can be single- or multiple-authored or edited volumes that (1) demonstrate the application of a variety of multivariate methods to a single, major area of research; (2) describe a multivariate procedure or framework that could be applied to a number of research areas; or (3) present a variety of perspectives on a topic of interest to applied multivariate researchers.

Anyone wishing to submit a book proposal should send the following: (1) author/title; (2) timeline including completion date; (3) brief overview of the book's focus, including table of contents and, ideally, a sample chapter (or chapters); (4) a brief description of competing publications; and (5) targeted audiences.

For more information, please contact the series editor, Lisa L. Harlow, at Department of Psychology, University of Rhode Island, 10 Chafee Road, Suite 8, Kingston, RI 02881-0808; phone (401) 874-4242; fax (401) 874-5562; or e-mail LHarlow@uri.edu.

- *Higher-order Growth Curves and Mixture Modeling with Mplus: A Practical Guide Second Edition* by Kandauda A. S. Wickrama, Tae Kyoung Lee, Catherine Walker O'Neal, & Frederick O. Lorenz
- *The Essence of Multivariate Thinking: Basic Themes and Methods, Third Edition* written by Lisa L. Harlow
- *Longitudinal Structural Equation Modeling: A Comprehensive Introduction, Second Edition* by Jason T. Newsom
- *Categorical and Nonparametric Data Analysis: Choosing the Best Statistical Technique, Second Edition* by E. Michael Nussbaum

For more information about this series, please visit: www.routledge.com/Multivariate-Applications-Series/book-series/LEAMAS

CATEGORICAL AND NONPARAMETRIC DATA ANALYSIS

Choosing the Best Statistical Technique

Second Edition

E. Michael Nussbaum

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

Cover Image: Getty Images © Hiroshi Watanabe

Second edition published 2024
by Routledge
605 Third Avenue, New York, NY 10158

and by Routledge
4 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

First edition published by Routledge 2014

Library of Congress Cataloging-in-Publication Data
Names: Nussbaum, E. Michael, author.
Title: Categorical and nonparametric data analysis : choosing the best statistical technique / E. Michael Nussbaum.
Description: Second edition. | New York, NY : Routledge, 2024. | Series: Multivariate applications | Includes bibliographical references and index.
Identifiers: LCCN 2023052369 (print) | LCCN 2023052370 (ebook) | ISBN 9780367702540 (hardback) | ISBN 9780367698157 (paperback) | ISBN 9781003145264 (ebook)
Subjects: LCSH: Multivariate analysis. | Nonparametric statistics.
Classification: LCC QA278 .N87 2024 (print) | LCC QA278 (ebook) | DDC 519.5/35—dc23/eng/20240222
LC record available at https://lccn.loc.gov/2023052369
LC ebook record available at https://lccn.loc.gov/2023052370

ISBN: 978-0-367-70254-0 (hbk)
ISBN: 978-0-367-69815-7 (pbk)
ISBN: 978-1-003-14526-4 (ebk)

DOI: 10.4324/9781003145264

Typeset in Stone Serif
by Apex CoVantage, LLC

Access the Instructor and Student Resources: www.routledge.com/9780367698157

CONTENTS

DETAILED CONTENTS

ABOUT THE AUTHORS

E. Michael Nussbaum, University of Nevada, Las Vegas. Dr. Nussbaum holds a PhD from Stanford University and an MPP from the University of California, Berkeley. He is the author of numerous research publications and serves on the editorial boards of the *Journal of Educational Psychology* and the *Educational Psychologist.* https://orcid.org/0000-0002-1974-2632

Edgar Brunner (co-author of Chapters 6–8), University of Göttingen (University Medical School), Germany. Dr. Brunner holds a PhD in Mathematics from the Technical University of Aachen. He was the director of the Institute of Medical Statistics from 1976 until 2010, and served as the editor of the *Biometrical Journal* from 2004 to 2009. He has numerous publications and several textbooks on nonparametric statistics. https://orcid.org/0000-0001-7119-8622

PREFACE

When scientists perform research, they often collect data that do not come in the form suitable for traditional methods such as analysis of variance (ANOVA) or ordinary linear regression. Categorical and nonparametric data analysis are useful for data that are nominal or ordinal, or for metric/continuous data when assumptions of traditional tests have been violated, for example when the sample is small. Finally, even when all the assumptions of traditional tests have been met, alternative tests are sometimes more statistically powerful, meaning that those tests are more likely to find a statistically significant result when there is a real effect to be found. There is nothing more frustrating than spending a year or two on a research project, only to have your results come in as not quite significant (e.g., $p = .07$).

This book was developed both out of my efforts to advise students on dissertations, as well as my own research in the field of argumentation. I found that while much of what I was taught about statistics in graduate school was useful at times, it was also somewhat limited. It was as if I was only taught "half the story." In fact, one of my statistics teachers at Stanford University was upset that coverage of "categorical data analysis" (CDA) was no longer required on the grounds that there was too much material already crammed into the statistics core. It was only later, after several years conducting research with actual messy data sets, that I discovered the immense value of categorical and nonparametric data analysis.

It is important to know when to use these techniques and when to use traditional parametric analysis. One major goal of the book is to provide a conceptual framework for choosing the most appropriate type of test in a given situation. One has to consider the underlying assumptions of each test and the factors that impact each test's statistical power. One also has to be able to explain these assumptions (and how the test works) conceptually to both oneself and others, including the audiences of a journal article or dissertation. This allows one to make statistically based arguments that can serve as a rationale for using that test. Therefore another major goal of this book is to provide readers with a conceptual framework that underlies the statistical methods examined, and to some extent traditional parametric methods as well.

Intended Audience

The primary intended audience is researchers and students in the social sciences—particularly psychology, sociology, and political science—and in related professional domains such as education. Readers should have some prior knowledge of descriptive statistics, *t*-tests, and ANOVA. It is preferable for readers to additionally have some prior knowledge of linear regression, although students can also pick up the basics of regression from reading Chapter 9.

This book is fairly unique in covering both nonparametric statistics and CDA in one volume. With the exception of contingency tables, most textbooks address either one or the other. In the one-semester course that I teach on this topic, I successfully cover both and—as noted above—provide a framework for choosing the best statistical technique. Using the framework requires knowledge of both CDA and nonparametrics. Although it is a challenging course, given students' financial and time constraints it is often not practical for graduate students who are not specializing in statistics to take separate courses in nonparametric statistics and CDA. Instructors who can provide a two-semester course can of course cover more topics and provide students with more extensive practice; however, a one-semester course can still provide students with some familiarity with the topics and a framework for choosing a statistical methodology for particular research projects. (Students can then obtain more practice and mastery of that particular technique, as the best way to learn a technique is to use it in the context of applied research.) This book could also be used for a course in only CDA or nonparametrics by just focusing on the applicable chapters, leaving the rest of the material for optional, supplementary reading. Finally, the book is also suitable for researchers and graduate students who are not necessarily enrolled in a course but who desire some knowledge of these alternative techniques and approaches for enhancing statistical power.

Unique Features of the Book

The book is closely tied to those techniques currently available in both IBM SPSS and R, and it includes examples of R code and reference to R-packages that were available when this book was written. Some of the examples and problems in the book also use Excel as a pedagogical tool for building conceptual understanding. Also available is a website for the book that contains selected data sets, a brief guide on using R, and, for instructors, PowerPoint slides for each chapter.

A distinguishing feature of the book is that three of the chapters on nonparametrics (co-authored with Edgar Brunner of the University of Göttingen) present modern techniques related to relative effects, factorial designs, multiple contrast tests, and other advances in nonparametric statistics. R-packages that can conduct these tests are identified.

An important feature of the text is its conceptual focus. Simple computer simulations and the inclusion of *Exploring the Concept* boxes are used to help attach meaning to statistical formulas. Most homework problems were inspired by actual research studies; these problems therefore provide authentic and meaningful illustrations of the techniques presented. Mathematical derivations have been kept to a minimum in the main text but are available in appendices at the end of each chapter.

Content

The book is structured as followed. Chapters 1–3 cover basic concepts in probability—especially the binomial formula—that are foundational to the rest of the book. Chapters 4–5 address the analysis of contingency tables (i.e., analyzing the relationship between two or more nominal variables). Chapters 6–8 address nonparametric tests involving at least one ordinal variable, including contemporary techniques for testing nonparametric interaction effects, a topic omitted from many other texts. The book then turns to situations that involve at least one metric variable. Chapter 9 reviews some concepts from linear regression, such as exponential growth and dummy variables, as well as the concept of generalized linear models. All of these concepts are foundational to CDA, which is the focus of the remaining portion of the book. Chapters 10–11 cover various types of logistic, ordinal, and Poisson regression. Chapter 12 overviews loglinear models, and Chapter 13 presents the General Estimating Equations (GEE) methodology for measuring outcomes measured at multiple time points. Chapter 14 covers estimation methods, such as Newton-Raphson and Fisher scoring, for readers desiring a deeper understanding of how the various CDA techniques work. The chapter provides preparation for reading more advanced statistical texts and articles. Finally, Chapter 15 summarizes the various factors that need to be taken into consideration when choosing the best statistical technique.

Overall, the book's organization is intended to take the reader on a step-by-step journey from basic statistical concepts into more advanced terrain.

FOREWORD

Dr. Nussbaum offers a readable and informative second edition of his book on *Categorical and Nonparametric Data Analysis*. With the plethora of data that is available for current researchers, it is important to consider whether the data would meet the assumptions of traditional parametric analyses. In particular, when data just include nominal or named categories (e.g., country, college major, sport type), or ordinal-level data (e.g., birth order, Likert scale responses, contest or competition rank), conventional analyses (e.g., ANOVA, linear regression) may not be appropriate to use. Dr. Nussbaum's new edition is very helpful in describing and demonstrating the use of analyses that are more warranted in these commonly found scenarios.

A distinct plus of the new edition is that it shows examples of conducting categorical and nonparametric analyses with the open-source R program, in addition to IBM SPSS and Excel. The inclusion of R computing, along with specific input on using R, goes a long way in reaching an even wider audience. The book also includes problems, and answers to odd-numbered problems, for Chapters 1 to 14, making the book even more readable and understandable.

Finally, the second edition continues to cover a range of topics, including measurement, estimation and hypothesis testing, random variables, chi-square test of independence, contingency tables in special situations; nonparametric tests for ordinal data, independent samples, and related samples, where these last three topics are co-authored with Edgar Brunner; linear regression and generalized linear models, binary and multinomial logistic regression, loglinear analysis, general estimating equations, estimation procedures, and an excellent summary on choosing the best statistical technique to use. Readers will find the book accessible and illuminating for analyzing data that doesn't meet restrictive assumptions.

By Lisa L. Harlow

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

ACKNOWLEDGMENTS

I wish to thank Arne Bathke, Matthias Gondan-Rochon, Nikos Bosse, and Georg Zimmermann for reviewing selected chapters of the manuscript. Special thanks also to my graduate assistant, Kaylie Kurland, for her help with the book's figures and cover design. I am also extremely grateful to Edgar Brunner, the co-author of Chapters 6–8, for his invaluable insights and feedback. Finally, the book would not be possible without all the encouragement and feedback provided by the graduate students who completed my course in categorical/nonparametric data analysis.

CHAPTER 1

LEVELS OF MEASUREMENT, PROBABILITY, AND THE BINOMIAL FORMULA

Categorical and nonparametric data analysis is designed for use with nominal or ordinal data, or for metric data in some situations. This chapter reviews these different levels of measurement before turning to the topic of probability.

Levels of Measurement

In statistics, there are four basic types of variables: (a) nominal, (b) ordinal, (c) interval, and (d) ratio.

A *nominal* variable relates to the presence or absence of some characteristic. For example, an individual will be biologically either male or female. Gender is a dichotomous nominal variable. In contrast, with a multinomial nominal variable, cases are classified in one of several categories. For example, ethnicity is multinomial: Individuals can be classified as Caucasian, African American, Asian/Pacific Islander, Latinx, or Other. There is not any particular ordering to these categories.

With an *ordinal* variable, there is an ordering. For example, an art teacher might look at student drawings and rank them from the most to the least creative. These rankings comprise an ordinal variable. Ordinal variables often take the form of ordered categories, for example: "highly creative," "somewhat creative," and "uncreative." A number of individuals may fall into these categories, so that all the drawings classified as highly creative would technically be tied with one another (the same for the moderately creative and uncreative categories). With ranks, on the other hand, there may be few if any ties.

With an *interval* or *ratio* variable, a characteristic is measured on a scale with equal intervals. A good example is height. The scale is provided by a ruler, which may be marked off in inches. Each inch on the ruler represents the same distance; as a result, the difference between eight and ten inches is the same as between one and three inches. This is not the case with an ordinal variable. If Drawing A

DOI: 10.4324/9781003145264-1

is ranked as more creative than Drawing B, we do not know if Drawing A is just slightly more creative or significantly more creative; in fact, the distances are technically undefined. As a result, we need to use different statistics and mathematical manipulations for ordinal variables than for interval/ratio variables. (Much of this course will be devoted to this topic.)

As for the difference between an interval and ratio variable, the defining difference is that in a ratio variable, a score of zero indicates the complete absence of something. Height is a ratio variable because zero height indicates that an object has no height and is completely flat (existing in only two dimensions). Counts of objects are also ratio variables. The number of people in a classroom can range from zero on up, but there cannot be a negative number of people. With an interval variable, on the other there can be negative values. Temperature is a good example of something measured by an interval scale, since 0° Celsius is just the freezing point of water, and negative temperatures are possible. However, for the tests discussed in this book, it will usually not be necessary to differentiate between ratio and interval variables, so we will lump them together into one level of measurement. The distinction will only become important when we consider the analysis of count data with Poisson regression. For ease of exposition, in this book I will use the term *metric variable* to refer to those at the interval or ratio levels of measurement.

Metric variables are often also referred to as *continuous*, but this usage fails to recognize that some metric variables are discrete. For example, a count cannot have fractional values; for example, it would be incorrect to say that there are 30.5 people enrolled in a class.

Figure 1.1 shows the three levels of measurement. The metric level is shown on top because it is the most informative. Metric data can always be reduced to ordinal

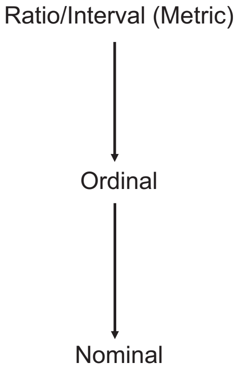


Figure 1.1 Three levels of measurement. The figure shows that metric data can be reduced to ordinal data, which can in turn be reduced to nominal data. Metric data are the most informative because they carry information on how different the cases are on a variable in quantitative terms. Nominal data are the least informative because they contain no information regarding order or ranks.

data by using the numerical values to rank the data (for example, ranking people from the tallest to the shortest based on their heights). Likewise, ordinal data can be reduced to nominal data by performing a median split and classifying cases as "above" or "below" the median. Transforming data from a higher level to a lower level is known as *data reduction*. Data reduction throws away information; for example, knowing that Marie is taller than Jennifer does not tell one how much taller Marie is. Nevertheless, data reduction is sometimes performed if the assumptions of a statistical test designed for metric data are not met. Then one might reduce the data to ordinal and perform a statistical test that is designed for ordinal data. One can also reduce ordinal (or metric) data to nominal. One cannot move from a lower level to a higher level in the figure because that requires information that is missing.

Categorical and nonparametric statistics is concerned with statistical methods designed for ordinal and nominal level data. Nonparametric methods are often used with metric data when sample sizes are small (and therefore some of the assumptions of *t*-tests, ANOVA, and linear regression are not met), and both categorical data analysis (CDA) and nonparametrics are useful when the data are skewed or otherwise highly abnormal. In the latter cases, standard methods may not be as statistically powerful as categorical and nonparametric ones. In reading this book, it is very important to remember the definition of statistical power.

> ***Statistical power*** refers to the ability to reject the null hypothesis and find a "result." (To be more technically precise, it is the probability that one will reject the null hypothesis when the alternative hypothesis is true.)

Because conducting a study is labor intensive, one usually wants to use the most powerful statistical methods. (Obtaining a *p*-value of .06 or .07 is not sufficient to reject the null hypothesis and therefore can be very disappointing to researchers.) That is why, in planning a study, one should use the most valid and statistically powerful methods one can.

Probability

All statistical methods—including categorical/nonparametric ones—require an understanding of probability. In the remainder of this chapter, I review the basic axioms of probability and use them to derive the binomial formula, which is the foundation of many of the tests discussed in this book.

In this section, I use a canonical example of tossing coins up in the air and asking questions about the probability of a certain number coming up heads. Note that whether a coin comes up heads or tails is a nominal outcome (it either happens or it doesn't)—which is why understanding probability is essential to CDA.

The Meaning of Probability

If one flips a coin, what is the probability that the coin will come up heads?

You may reply "one-half," but what exactly does this statement mean?

One definition of probability is given by the following equation:

$$\textit{Probability} = \frac{\textit{Number of favorable possibilities}}{\textit{Number of total possibilities}}, \qquad \text{(Eq. 1.1)}$$

assuming that all possibilities are equally likely and mutually exclusive.

So, a probability of one-half means that there is one favorable possibility (heads) out of two (heads or tails). However, this example assumes that the coin is fair and not biased, meaning that the possibilities are equally likely. A biased coin might have a little more metal on one side, so that heads result 60% of the time. Such coins have been created to cheat at games of chance.

Eq. 1.1 is often useful but because of the restrictive equal-probability assumption, a more general definition of probability is needed. The probability of some event *A* occurring is the proportion of time that *A* will occur (as opposed to not-*A*) in the limit as *n* approaches infinity, that is:

$$\text{Prob}(A) = \lim\ (n \rightarrow \infty)\ \frac{f(A)}{n}, \qquad \text{(Eq. 1.2)}$$

where $f(A)$ is the frequency of *A*. Thus, the meaning of the statement "the probability that the coin will come up heads is one-half" is that over a large number of flips, about half the time the coin will come up heads. The amount of error decreases as *n* (the number of flips) increases, so that the proportion will approach Prob(*A*) as *n* approaches infinity. Now to assess the probability, one might flip the coin 1,000 times and gage the relative proportion that the coin comes up heads as opposed to tails. This procedure will only give one an estimate of the true probability, but it will give one a pretty good idea as to whether the coin is fair or biased. Basically, what we are doing is taking a random sample of all the possible flips that could occur.

The definition in Eq. 1.2 reflects a *frequentist* view of probability. Technically, Eq. 1.2 only assigns probabilities to general statements, not particular facts or events. For example, the statement that 80% of Swedes are Lutherans is a meaningful probability statement because it is a generalization; but the statement that "the probability that John is Lutheran, given that he is Swedish, is 80%" is not meaningful under this definition of probability, because probability applies to relative frequencies, not to unique events. This position, however, is extreme, given that we apply probability to unique events all the time in ordinary discourse (maybe not about Lutherans, but certainly about the weather, or horse racing, etc.). In my view, probability statements can be meaningfully applied to particular cases if one makes the appropriate background assumptions. For example, if one randomly selects one individual out of the population (of Swedes), then one can meaningfully say that there is an 80% chance that she or he is Lutheran. The background assumption here is that the selection process is truly random (it may or may not be). The existence of background assumptions means that probability values cannot be truly objective because they depend on whether one believes the background assumptions. Nevertheless, these assumptions are often rational to make in many situations, so in this book I will assign probability values to unique events. (For further discussion of subjectivist, Bayesian notions of probability, which view probability statements as measures of certainty in beliefs, see Nussbaum, 2011.)

Probability Rules

Probability of Joint Events

What is the probability that if one flips two coins, both will come up heads? Using Eq. 1.1, it is ¼, because there is one favorable possibility out of four, as shown below.

H **H**
H T
T H
T T

Another way of calculating the joint probability is to use the following formula:

Prob(A & B) = Prob(A) $*$ Prob(B) [if Prob(A) and Prob(B) are statistically independent]. (Eq. 1.3)

Here, A represents the first coin coming up heads and B represents the second coin doing so; the joint probability is ½ $*$ ½ = ¼. The formula works because A represents one-half of all possibilities and of these, one-half represent favorable possibilities, where B also comes up heads. One-half of one-half is, mathematically, the same as ½ $*$ ½.

An important background assumption, and one that we shall return to repeatedly, is that A and B are statistically independent. What that means is that A occurring in no way influences the probability that B will occur. This assumption is typically a reasonable one, but we could imagine a scenario where it is violated. For example, suppose someone designs a coin with an electrical transmitter, so that if the first coin comes up heads, this information will be transmitted to the second coin. There is a device in the second coin that will tilt it so that it will always come up on heads if the first coin does. In other words: $\text{Prob}(B \mid A) = 1$. This statement means that the probability of *B* occurring, if *A* occurs, is certain. We will also assume the converse: $\text{Prob}(\text{not-}B \mid \text{not-}A) = 1$. There are therefore only two total possibilities:

H H
T T

The joint probability of two heads is therefore one-half. The more general rule for joint probabilities (regardless of whether or not two events are statistically independent) is:

$$\text{Prob}(A \,\&\, B) = \text{Prob}(A) * \text{Prob}(B \mid A). \tag{Eq. 1.4}$$

The joint probability in the previous example is $\frac{1}{2} * 1 = \frac{1}{2}$.

EXPLORING THE CONCEPT

If two events (*A* and *B*) are statistically independent, then: $\text{Prob}(A) = \text{Prob}(A \mid B)$ and $\text{Prob}(B) = \text{Prob}(B \mid A)$. Can you use this fact to derive Eq. 1.3 from Eq. 1.4?

Note that in the rigged coin example, the joint probability has increased from one-fourth (under statistical independence) to one-half (under complete dependence). This outcome reflects a more general principle that the probability of more extreme events increases if cases are not statistically independent (and positively correlated). For example, if I throw ten coins up in the air, the probability that they will all come up heads is:

$$\frac{1}{2} * \frac{1}{2} * \frac{1}{2} * \frac{1}{2} * \frac{1}{2} * \frac{1}{2} * \frac{1}{2} * \frac{1}{2} * \frac{1}{2} * \frac{1}{2} = (1/2)^{10} = 0.0001,$$

or one in ten thousand. But if the coins are programmed to all come up heads if the first coin comes up heads, then the joint probability is again just one-half. If the first coin coming up heads only creates a general tendency for the other coins to come up heads (say $\text{Prob}(B \mid A) = 0.8$), then the joint probability of two coins coming up heads would be $0.5 * 0.8 = 0.4$ (which is greater than one-fourth, the result assuming independence). If ten coins are flipped, the probability is $0.5 * (0.8)^9 = .067$. This probability is still far greater than the one under statistical independence.

Violations of statistical independence are serious, as they contravene the first axiom of statistical theory. For example, my own area of research addresses how students construct and critique arguments during small-group discussions. In small-group settings, students influence one another, so, for example, if one student makes a counterargument, it becomes more likely that other students will do so as well, due to modeling effects and other factors. The probability that Student (A) makes a counterargument is therefore not statistically independent from the probability that Student (B) will, if they are in the same discussion group. Analyzing whether some intervention, such as the use of a graphic organizer, increases the number of counterarguments, without adjusting for the lack of statistical independence, will make the occurrence of Type I errors more likely. That is because the probability of extreme events (i.e., many students making counterarguments) goes way up when there is statistical dependence, so *p*-*va*lues can be seriously inflated. There are statistical techniques such as multilevel modeling that address the problem but the technique's appropriateness in argumentation research is still being debated (given that certain sample sizes may be required). Further discussion of multilevel modeling is beyond the scope of this book, but the example illustrates why an understanding of basic probability theory is important.

EXPLORING THE CONCEPT

(a) Suppose one randomly selects 500 households, and sends a survey on political attitudes towards female politicians to all the adult males in each household and the same survey to all the adult females (e.g., one to a husband and one to a wife). Would the individual responses be statistically dependent? How might this fact affect the results? (b) How might sampling without replacement involve a violation of statistical independence? For example, what is the probability that if I shuffle a complete deck of 52 cards, the first two cards will be spades? Use Eq. 1.4 to calculate this probability.

Probabilities of Alternative Events

What is the probability that if I select one card from a full deck of 52 cards, that the card will be spades or clubs? This question relates to alternative events occurring (*A or B*) rather than joint events (*A and B*). The probability of alternative events is given by:

$$\text{The Prob}(A \text{ or } B) = \text{Prob}(A) + \text{Prob}(B) \text{ [if Events A and B are mutually exclusive].} \qquad \text{(Eq. 1.5)}$$

Applying Eq. 1.5 to our question, we find that the Prob(spades *or* clubs) = Prob(spades) + Prob(clubs) = $\frac{1}{4}+\frac{1}{4}=\frac{1}{2}$. This result makes sense because half the cards will be black. The reason Eq. 1.5 works is because Event *A* (card being spades) represents one-fourth of the favorable possibilities, and Event *B* (card being clubs) represents another one-fourth of the favorable possibilities, so together, the probability that the selected card will be black is the sum of all the favorable possibilities (which is why we add). The background assumption is that the card cannot be both spades and clubs, that is, A and B cannot both occur—the events are mutually exclusive. If the events are not mutually exclusive, one should use Eq. 1.6:

$$\text{Prob}(A \text{ or } B) = \text{Prob}(A) + \text{Prob}(B) - \text{Prob}(A \,\&\, B). \qquad \text{(Eq. 1.6)}$$

For example, if one flips two coins, the probability that one or the other (or both) will come up heads is $\frac{1}{2}+\frac{1}{2}-\frac{1}{4}=\frac{3}{4}$, which is the correct answer. If one were to erroneously use Eq. 1.5 to calculate the probability, one would double count the possibility where both coins come up heads.

In summary, remember that if joint events are involved (involving *and's*), one should multiply (assuming statistical independence) and where alternative events are involved (involving *or'* s), one should add (assuming the events are mutually exclusive).

Conditional Probabilities

The last type of probability to be considered is conditional probability. These take an IF . . . THEN form, for example, if a student is male, then what is the probability that they will complete high school? Conditional probabilities will be discussed in Chapter 4.

Some Probability Thought Experiments

We shall now progressively consider cases of increasing complexity.

Probability Distribution of the Two-Coin Example

In tossing two coins there are four possible outcomes.

		Number of Heads (H)
1. H	H	2
2. H	T	1
3. T	H	1
4. T	T	0

Each of the four outcomes is called a *permutation*. If one counts the number of hits (i.e., heads), outcomes numbers two and three both reflect one head. Although these two outcomes reflect different permutations, they reflect the same *combination,* that is, a combination of one head and one tail. Combinations reflect the number of heads, but whether the head comes up on the first trial or the second doesn't matter. The order does matter with permutations.

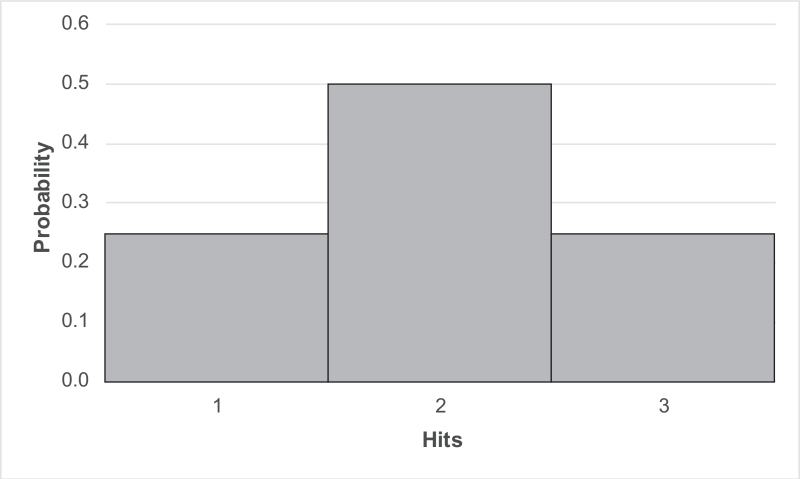


Figure 1.2 Probability histogram for the two-coin problem.

Figure 1.2 is a frequency graph of the different combinations.

This figure is shaped like an inverted U, except that it is not smooth. Many probability distributions—like the normal curve—have this upward concave characteristic. This is because there are more permutations corresponding to combinations in the middle than at the extremes, making the more central possibilities "more likely."

One Hundred-Coin Example

Let us take a more extreme case. Suppose one throws up in the air 100 coins. This resulting histogram is shown in Figure 1.3. (We have not yet covered the formula that was used in making this figure.)

The most extreme positive case is that all the coins come up heads. This situation corresponds to just one permutation. The probability of each head is one-half, and these are independent events, so the probability of their joint occurrence, given by the multiplication rule, is 0.5^{100}. Not surprisingly, this number is extremely small:

0.000000000000000000000000000788861.

On the other hand, the probability of obtaining 50 heads out of 100 is much larger (8%), because there are many ways of obtaining a combination of 50 coins out of 100 (i.e., there are a lot of corresponding permutations). The first 50 coins could come up heads and the second 50 could come up tails, the first 50 coins could come up tails and the second 50 could come up heads, every other coin could come up heads, etc. There are also a lot more "random looking" permutations, such as:

H H H T T T H T T H . . .,

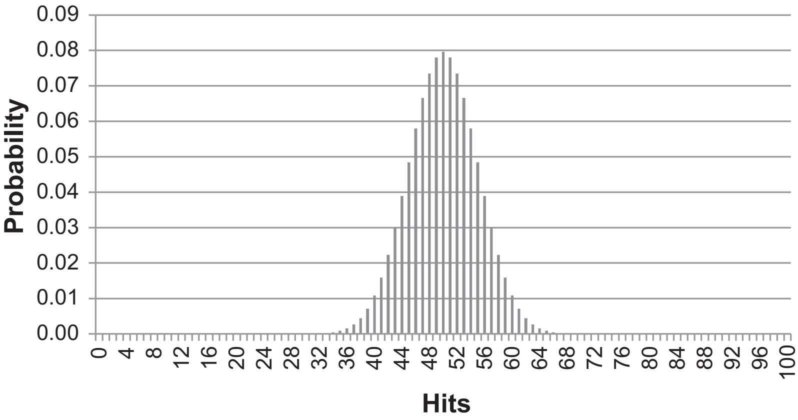


Figure 1.3 Probability histogram for the 100-coin problem.

which is a random sequence that I generated in Excel. In fact, the number of permutations comes out to be $1.00891 * 10^{29}$. How this number is derived will be explained later in the chapter. Notice that in comparing Figures 1.2 and 1.3, the one with 100 coins is skinnier, if one thinks of the extremes in percentage terms (100 heads is 100% hits, tails is 0% hits). This result occurs because the more extreme outcomes become less likely as the number of trials increases. For example, the probability of 100% hits in the two-coin case is 25%, not 0.5^{100} (as calculated above). Later we will see that this is the reason why larger samples provide more precise estimates.

The Binomial Formula

The binomial formula is used to measure the probability of a certain number of hits in a series of yes/no trials. It is a complicated formula, so I will introduce the formula bit by bit. In my example, I will use a biased rather than a fair coin to make the mathematics easier to follow. The Greek letter pi (π) denotes the probability of a hit. (Notice that the words “pi” and “probability” both start with the letter “p.” Statisticians often choose Greek letter for concepts that, in English, start with the same letter.) Let us suppose that the probability of this particular biased coin coming up heads is 70% (π = 70%), and ask:

> What is the probability of tossing ten coins (with $\pi = 0.7$) and obtaining eight heads?

This question can be broken down into two subparts:

1. What is the probability of a *permutation* involving eight heads out of ten coins?

2. How many permutations make up a *combination* of eight heads out of ten coins?

Subquestion 1: What Is the Probability of One "Permutation"?

Let us consider one permutation, where the first eight coins come up heads: H H H H H H H H T T. Assuming the tosses are independent, the joint probability is 70% * 70% * 70% * 70% * 70% * 70% * 70% * 70% * 30% * 30% = $0.7^8 * 0.3^2 = \pi^8 * (1–\pi)^2$. If we define *k* as the number of hits (in this case *k* = 8), and *n* as the number of trials (in this case *n* = 10), then the calculation becomes:

$$\pi^k * (1-\pi)^{n-k} \tag{Eq. 1.7}$$

For the permutation considered above, the probability is $0.7^8 * 0.3^2 = 0.5\%$. Note that all the permutations associated with eight coins out of ten coming up heads are equally likely. For example, the permutation "H H H T H H H H H T" has the probability 70% * 70% * 70% * 30% * 70% * 70% * 70% * 70% * 30% = $0.7^8 * 0.3^2 = 0.5\%$. The only thing that has changed is the order in which the coins come up heads or tails.

Remember that the overall goal is to find the probability of a *combination* of eight hits out of ten. We shall see that this combination is associated with 45 different permutations, including the two shown above. The probability of one or another of these 45 different permutations occurring can be calculated by adding the individual probabilities together. Because the permutations are mutually exclusive, we can use the addition formula (Eq. 1.5). Because the permutations are equally likely, the calculation reduces to:

$$45 * \underline{\pi^k * (1-\pi)^{n-k}} \tag{Eq. 1.8}$$

The underlined portion is the probability of one of the *permutations* occurring (from Eq. 1.7). What I have yet to address is how the 45 is calculated.

Subquestion 2: How Many Permutations Make Up a Combination?

How many possible permutations are associated with a combination of eight hits out of ten? The applicable formula, which is derived conceptually in Appendix 1.1, is:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}. \tag{Eq. 1.9}$$

Here, $\binom{n}{k}$ is the notation for k hits out of n trials. On the right-hand side of the equation, the exclamation mark signifies a factorial product (e.g., 4! = 4 ∗ 3 ∗ 2 ∗1).

Substituting the values from our example yields:

$$\frac{10!}{8!(10-8)!} = \frac{40{,}320}{3{,}628{,}800 * 2} = 45.$$

Putting Eqs. 1.9 and 1.7 together yields the *binomial formula,* which gives the probability of k hits out of n trials:

$$P(k\,\text{hits out of}\,n\,\text{trials}) = \frac{n!}{k!(n-k)!} * \pi^{k} * (1-\pi)^{n-k} \qquad \text{(Eq. 1.10)}$$

In the example, the terms are 45 ∗ 0.5% = 23%. The first term represents the number of permutations associated with a combination of eight heads out of ten coin tosses, and the second term reflects the probability of each permutation.

The calculation can also be performed in Excel with the following command:

= BINOMDIST(k, n, π, 0),

with the 0 indicating a noncumulative calculation (explained below). So =BINOMDIST(8, 10, 0.7, 0) ➔ 23%. There is almost a one-quarter chance of obtaining exactly eight heads with this particular biased coin.

EXPLORING THE CONCEPT

Concepts can often be better understood when applied to very simple examples. Use the binomial formula to find the probability of obtaining exactly one head in a toss of two fair coins.

Cumulative vs. Noncumulative Probabilities

Now suppose we wanted to calculate the chance of obtaining, out a toss of ten coins, eight heads *or less.* This probability could be derived by calculating the chance of obtaining *exactly* eight heads or seven heads or six heads or . . . one heads or zero heads. These are mutually exclusive outcomes, so according to Eq. 1.5 we can just add the probabilities: $23\% + 27\% + 20\% + 10\% + 4\% + 1\% + 0\% + 0\% + 0\% = 85\%$. (The 0% values are not perfectly equal to zero but are so small that these round to zero.) I calculated the individual probabilities using the Excel command "=BINOMDIST(k, n, π, 0)," but an easier way is to calculate the cumulative probability by changing the zero in this expression to one. (A one in the last place of the Excel command tells Excel to use the cumulative probabilities.) So, we have:

=BINOMDIST(8, 10, 0.7, 1) ➔ 85%.

Now I am going to change the problem a little. The question is "What is the probability of obtaining nine heads or more?" (again, out of a toss of ten coins, with π = 70%). This is calculated by subtracting the 85% from 100% (= 15%). Note that the probabilities of obtaining (a) eight coins or less and (b) nine coins or more must sum to 100%. This is because the two possibilities are exhaustive of all possibilities and mutually exclusive; one or the other must happen.

Now, instead of asking, "What is the probability of obtaining nine hits or more?" we ask "What is the probability of obtaining eight hits or more?" This value is 100% minus the probability of obtaining seven hits or less. The Excel syntax is:

1–BINOMDIST(7, 10, 0.7, 1) ➔ 38%.

Some students err by putting an eight rather than a seven in the formula, because the question asks about the probability of getting eight heads or more, yielding an incorrect answer of 15%. This approach is incorrect because it is tantamount to calculating the probability of eight heads or more plus eight heads or less. These are not mutually exclusive possibilities because if one obtains exactly eight heads, both scenarios occur. According to the formula for alternative probabilities, one has to subtract out the probability of getting exactly eight heads so as to not double count it (see Eq. 1.6). We previously calculated the probability of obtaining eight heads or less at 85% and that for eight heads or more at 38%. The two possibilities sum to 123%, which is not a legal probability. We have double counted the probability of obtaining exactly eight heads (which we calculated previously at 23%).

This fact is so important that one should spend a minute now making a mental rule to remember that when you are asked cumulative probability problems that involve *X or more* (rather than *X or less),* you need to use *X* –1 in the Excel formula. Excel will only give you cumulative probabilities for *X or less* problems.

The Bernoulli and Binomial Distributions

A probability distribution gives the probability of each and every mutually exclusive outcome of an event. The simplest probability distribution is the Bernoulli distribution, named after the eighteenth-century mathematician, Daniel Bernoulli. It is the probability distribution associated with a single, discrete event occurring. Table 1.1 presents the Bernoulli distribution for the situation where the probability of a biased coin coming up heads is 70%.

Probability distributions always sum to 100%. Because the different outcomes listed are exhaustive, one of them must happen, so the probability that one or another will happen is found by just adding all the probabilities up (the possibilities are mutually exclusive).

Charting the above Bernoulli distribution yields Figure 1.4.

With a Bernoulli distribution, we just toss one coin. With the binomial distribution, we toss more than one coin. So, my previous examples using two, ten, or 100 coins are all associated with binomial distributions. With the binomial distribution, the tosses of the coins must be independent. The formal definition of the binomial distribution is "the distribution resulting from an independent series of Bernoulli trials." We shall refer back to this definition in later chapters but for now, it is just a fancy way of saying that there is more than one nominal event (e.g., we toss more than one coin).

The shape of a binomial distribution will depend on the value of π (probability of a hit). A binomial distribution can be generated in Excel using the binomial formula, with the Excel syntax =BINOMDIST(k, n, π, 0).

Figure 1.5 (Panel A) shows a binomial distribution when using a fair coin (π = 50%) for different values of k with n = 10 (in other words, when we toss ten fair coins up in the air). The distribution is symmetric. However, with our biased coin (π = 70%), the distribution is skewed.

Now we are going to make something important happen. As the number of trials increase, the distribution becomes more symmetric. Figure 1.6 shows that with

Table 1.1 A Bernoulli Distribution

Heads or Tails	*Probability*
Heads (X = 1)	70%
Tails (X = 0)	30%
Sum	100%

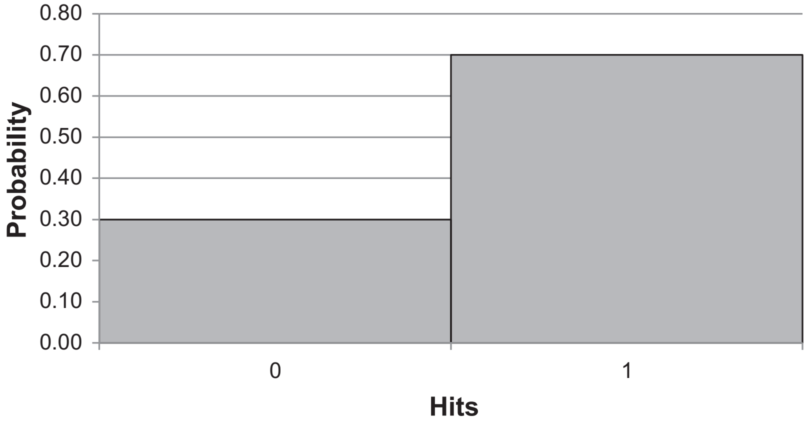


Figure 1.4 Bernoulli distribution for π = 0.7.

n = 100 (Panel B), the distribution is extremely symmetric. (The distribution is in fact normal except for the fact that the variable, "number of hits," is discrete rather than continuous.) Even at n = 20 (Panel A) the distribution is somewhat symmetric. Remember that as n increases, the probability of extreme outcomes decrease. So does the probability of somewhat extreme values, although not at the same rate. So in Figure 1.6 (with n = 20), consider the probability of moderately extreme values (such as k = 8, that is 40% heads), a value which contributes to the skew in Figure 1.5 (specifically, see Panel B at Prob(k = 4) = 40%). The probability of obtaining 40% hits is becoming less likely with the larger sample, making the distribution more symmetric.

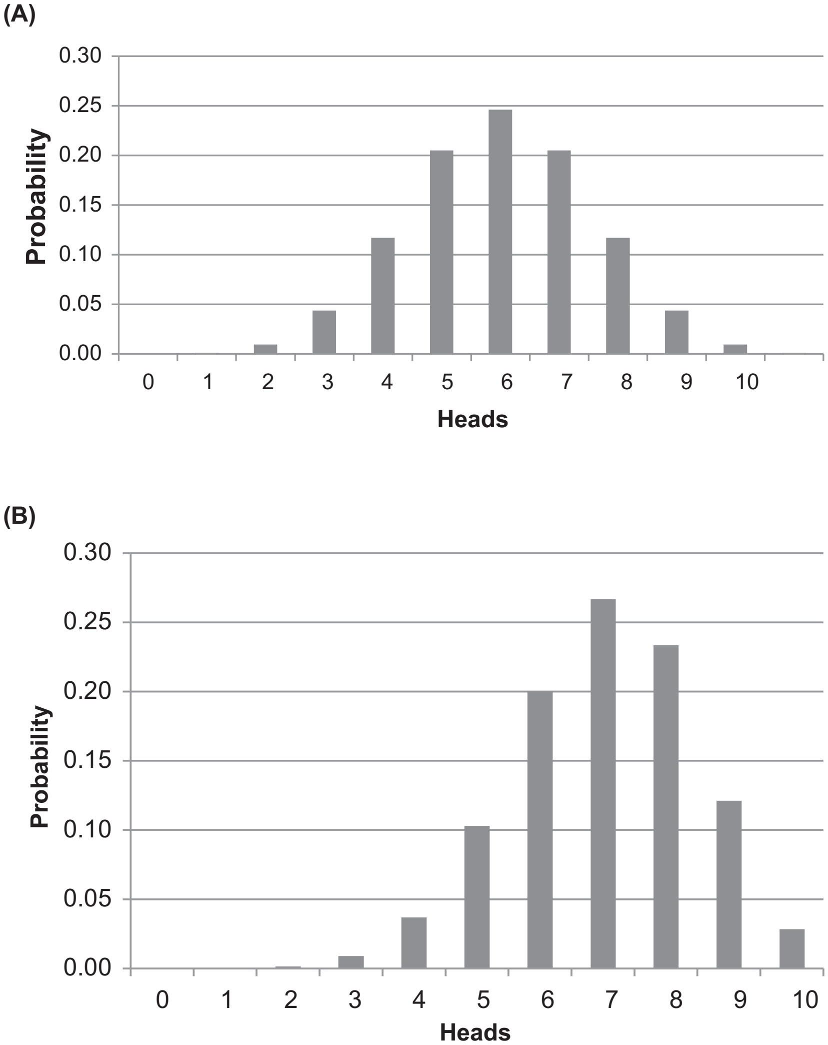


Figure 1.5 Binomial distribution (n = 10) for different probabilities of a hit. (A) π = 50%. (B) π = 70%.

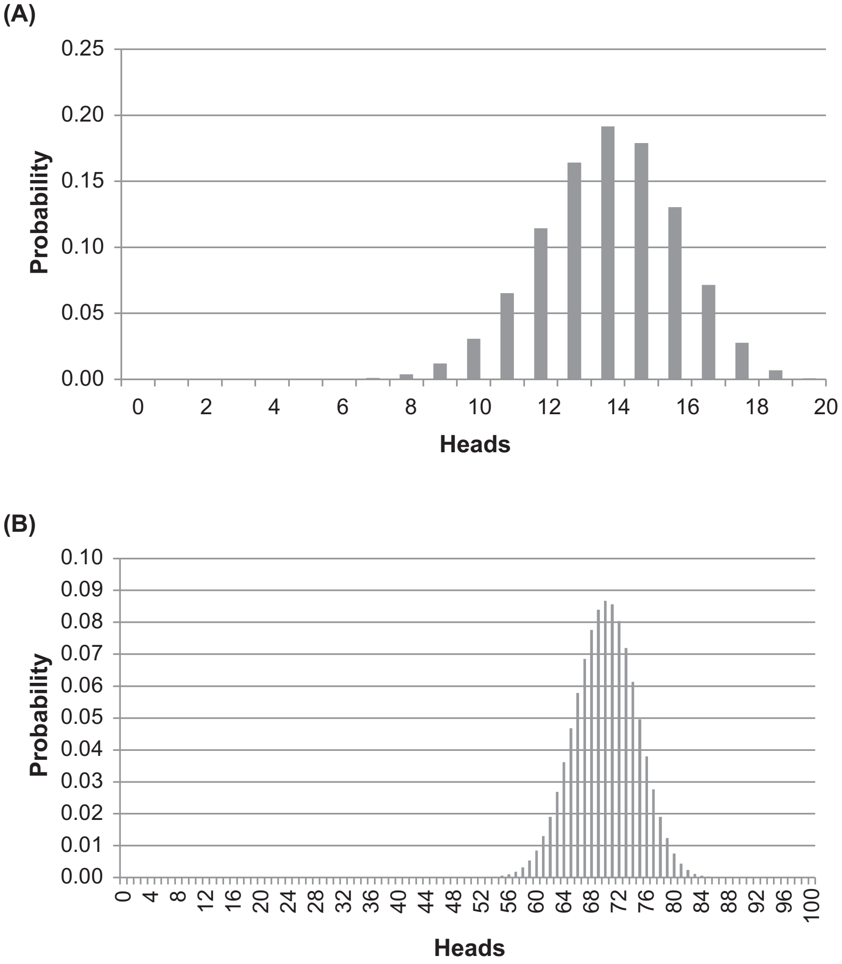


Figure 1.6 Binomial distribution for π = 70% for different values of n. (A) n = 20. (B) n = 100.

The Normal Approximation of the Binomial Distribution

Normal Distribution

A normal distribution is also known as a Gaussian distribution, and when standardized a z-distribution. The normal distribution has the property that about two-thirds of the distribution falls within one standard deviation (SD) of the mean, and most of the distribution falls within two SDs of the mean.

To *standardize* a score means to express the score in terms of the number of SDs from the mean. When a variable is normally distributed, use the following formula to compute z-scores:

$$z = \frac{X - mean}{SD}. \qquad \text{(Eq. 1.11)}$$

We shall use this equation repeatedly in this book.

One can compute the cumulative probability of different z-scores using the Excel command: =NORMSDIST(z). So, for example, if a variable is normally distributed, the cumulative probability of obtaining a value of, say, 0.23, is 59%. With Excel syntax, =NORMSDIST(0.23) ➔ 59%.

It is frequently the case in statistics that we need to build a 95% confidence interval with 2.5% of the z-distribution in each of the tails of the distribution. A 95% confidence interval around the mean can be built if we find the points that are about two SDs from the mean (1.96 SDs to be precise). These would correspond to z-scores of 1.96 and –1.96. Using Excel, one can verify that: =NORMSDIST(–1.96) ➔ 0.025, and =NORMSDIST(+1.96) ➔ 0.975.

EXPLORING THE CONCEPT

Some students wrongly expect that there should be a 95% chance of obtaining a z-score of +1.96 because we are building a 95% confidence interval. Why is the probability of obtaining a z-score of +1.96 or less 97.5% rather than 95%?

Suppose that X is normally distributed, has a mean of three and a SD of two. What is the probability of obtaining a raw score of six? One could first compute the z-score (1.5) and then consult Excel for the probability: =NORMSDIST(1.5) ➔93.3%. One could also use the unstandardized normal distribution and just enter the mean and SD into Excel using the command =NORMDIST (this one does not have an "S" after "NORM"). Specifically enter "=NORMDIST(X, mean, SD, 1); the one requests a cumulative distribution. In the example, =NORMDIST(6, 3, 2, 1) ➔ 93.3%. Which command to use is one of personal preference.

The Normal Distribution Is Continuous, Not Discrete

A discrete distribution is one where X cannot take on fractional values, whereas a continuous variable can take on any fractional value. A nominal (0, 1) variable is always discrete, whereas metric variables can be either continuous or discrete. For example, the probability distribution for the following variable is discrete:

X here only takes on the discrete values of 1, 2, 3, 4, and 5. On the other hand, if X could take on values such as 1.3, 2.67, or 4.3332, it would be continuous.

Figure 1.7 displays graphically the probability distribution for the discrete variable shown in Table 1.2. Note that the area under the curve sums to 1.0. The area for the value X = 2.0 is given by the width of the bar (one unit) times the height (which is 0.33), so the area is 0.33. The individual subareas sum to one because the probabilities sum to one.

Table 1.2 A Discrete Probability Distribution

X	*Probability*
1	15%
2	33%
3	7%
4	12%
5	33%
Total	100%

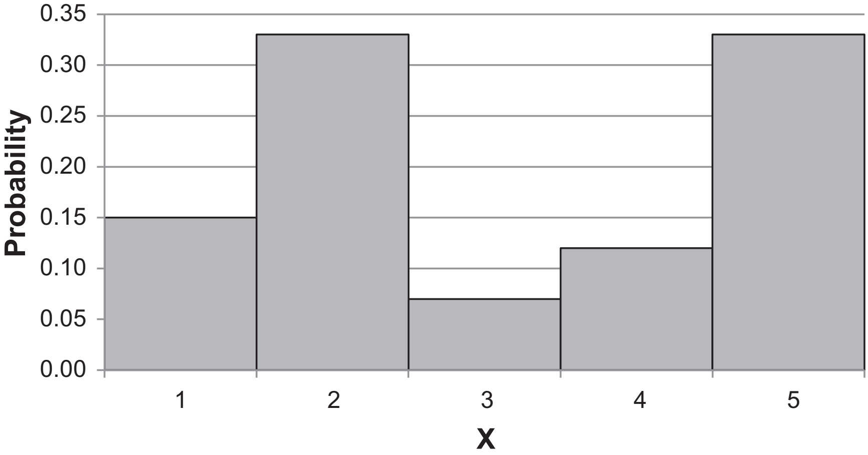


Figure 1.7 Example probability distribution for a discrete variable.

If *X* were continuous, the area for a value (such as 2.67, a number I arbitrarily chose) is given by the height (2.67) times the width, except that the width would be zero. This produces a paradox that the probability of any specific value of a continuous variable is zero! Obviously this situation cannot occur, as the variable will take on specific values in particular cases. To resolve this paradox, we stipulate that for continuous variables, one can only meaningfully talk about the values of cumulative probabilities, for example, the probability of obtaining a value of 2.67 or less, or the probability of obtaining a value of 2.67 or more. The probability of obtaining exactly 2.67 is undefined.

The normal distribution is a continuous distribution. *Z*-scores are therefore also continuous and the probability of obtaining a particular *z*-score cumulative. Thus while it is meaningful to write such expressions as Prob($z \leq 0.23$), it is not meaningful to write expressions such as Prob($z = 0.23$).

EXPLORING THE CONCEPT

In thinking about the normal curve and calculating the probability of *z*-scores, why do we typically compute the area of the curve that is to the right (or left) of a certain point?

The Normal Approximation (the 100-Coin Example)

Suppose the problem is to compute the probability that if one tosses 100 fair coins, that 60 *or less* will come up heads. I present in this section two different methods for approaching this problem: (a) With the binomial formula, and (b) with the normal curve. The latter is more complicated with this particular problem, but the approach will prove useful on problems considered in other chapters.

The Binomial Formula

To use the binomial formula, simply use the Excel syntax:

= BINOMDIST(*k, n*, π, 1) (use a 1 because we want cumulative probabilities).
= BINOMDIST(60, 100, 0.5, 1) ➔ 98.2%.

The Normal Approximation of the Binomial

It can be shown that as *n* approaches infinity, the binomial distribution approaches the normal distribution. When *n* is large, the differences between the distributions are negligible, so we speak about the normal *approximation* of the binomial. However, the binomial distribution is discrete, whereas the normal distribution is continuous, and this fact causes the approximation to be off a little, requiring what is known as a *continuity correction*. Let us set the continuity correction aside for a moment, as it will be the last step in our calculations.

We will first need to calculate a *z*-score, and to do that we need to know the mean and SD of the binomial distribution. Table 1.3 shows the mean and variances of both the Bernoulli and binomial distributions. For the binomial distribution, there are different means and variances depending on whether the outcome is a count (e.g., 50 hits) or a proportion of the trials that involve hits (e.g., 50% hits). We are presently concerned with the former situation.

In our example, the mean is $n\,\pi$ ($100 * 0.50 = 50$). If we toss 100 coins in the air, the most likely outcome is 50 heads (see Figure 1.3). Although the likelihood of this

Table 1.3 Means and Variances of the Bernoulli and Binomial Distributions

Parameter	*Distribution*		
	Bernoulli	*Binomial Count*	*Binomial Proportion*
Mean	π	$n\pi$	π
Variance	$\pi(1-\pi)$	$n\pi(1-\pi)$	$\pi(1-\pi)/n$

occurring is only 8%, it is still the most likely outcome. It is certainly possible that we might get 49 or 53 coming up heads rather than exactly 50, but the number of heads is still likely to be around 50. The mean is also known as the *expected value*.

In our example, the variance is $n\,\pi(1-\pi)$. This is sometimes written as *nPQ*, with *P* representing the probability of a hit, and Q the probability of a miss. In statistics, by convention Greek letters refer to population parameters and English letters to sample estimates of these parameters. Because we are not dealing with samples here, I use the Greek letters in writing the formulas.

To continue with the example, the variance is $n\,\pi(1-\pi) = 100 * 50\% * 50\% = 25$. The standard deviation is the square root of the variance, or 5. The *z*-score is then: $z = \frac{X-mean}{SD} = \frac{60-50}{5} = 2.00$, Prob($z \leq 2.00$) = 97.7%. This result does not jive with the calculation using the binomial formula (98.2%) because we have not yet applied the continuity correction. When using *z*-scores, it is assumed that the values are continuous. But we are here using discrete raw score values, for example, 62, 61, 60, 59, 58, etc. We want to find the probability of getting a value of $X \leq 60$. The area of the probability distribution corresponding to 60 is assumed with the normal curve method to have a width of zero (because continuous variables are assumed), when in fact the bar for 60 has a width of one. The area of the bar to the right of the midpoint corresponds to the probability excluded from the calculation with the normal curve method. This method leaves out "half a bar" associated with the binomial formula method, which gives the exact value. The point of the continuity correction when using the normal approximation is basically to "add half a bar" to the probability estimate. We can do this if we use the value of $X = 60.5$ rather than 60 in the *z*-score calculation.

$$z = \frac{X - mean}{SD} = \frac{60.5 - 50}{5} = 2.10, \text{ Prob}(z \leq 2.10) = 98.2\%.$$

This result is equal to the one from using the binomial method.

If we were to use proportions, rather than counts, the variance would be $\frac{(50\% * 50\%)}{100} = 0.25\%$ and so $z = \frac{60.5\% - 50\%}{\sqrt{0.25\%}} = 2.10$, which is the same result as if counts were used, but the formula for the variance is different (see Table 1.3 and Technical Note 1.1).

The normal approximation is an important procedure for the material discussed in this book. It may seem unnecessary here to go through all the gyrations when we can calculate the exact probability with the binomial formula, but in many other situations we will not have recourse to the binomial formula or some other "exact" formula and therefore will need to use the normal approximation.

Continuity Corrections With X or More Problems

One last problem: What is the probability of obtaining 60 *or more* heads when tossing 100 fair coins. Once again, we should apply a continuity correction. However, there is a second adjustment we will need to make, namely the one described a few pages back for "X or more" problems (specifically using $X - 1$ in the Excel formula). Therefore, instead of using $X = 60$ in the formula, we should use $X = 59$. But for the continuity correction, we also need to add a half a bar. Therefore, the correct X to use is 59.5:

$$z = \frac{X - mean}{SD} = \frac{59.5 - 50}{5} = 1.90, \text{ Prob}(z \leq 1.90) = 97.1\%.$$

Subtracting from 100%, the probability of obtaining 60 or more heads is 2.9%.

EXPLORING THE CONCEPT

The probability of obtaining 60 or more heads is 2.9%. The probability of obtaining 60 or less heads is 98.2%. These probabilities sum to more than 100%; they sum to 101.1%. Can you explain why?

Problems

1. What is the scale of measurement for the following random variables?
   a) Someone's weight (in pounds).
   b) The state in which a person resides (Nevada, California, Wisconsin, etc.).
   c) Someone's IQ score.
2. What is the scale of measurement for the following random variables?
   a) A student's percentile ranking on a test.
   b) Self-report of the number of close friends an adult has.
   c) Political party affiliation.
   d) Categorizing individuals into the highest college degree obtained: Doctorate, master, bachelor, associate, or none.

3. One throws two dice. What is the probability of:
   a) Obtaining a seven (combining the two numbers)? (HINT: Try to think of the different permutations that would produce a seven, and use the multiplication rule for joint events.)
   b) Obtaining a nine?
4. In a single throw of two dice, what is the probability that:
   a) Two of the same kind will appear (a "doublet")?
   b) A doublet or a six will appear? (HINT: These are not mutually exclusive events.)
5. In three tosses of a fair coin, what is the probability of obtaining at least one head?
6. Three cards are drawn at random (*with replacement*) from a card deck with 52 cards. What is the probability that all three will be spades?
7. Three cards are drawn at random *(without replacement)* from a card deck with 52 cards. What is the probability that all three will be clubs? (HINT: when two events are not statistically independent, and A occurs first, Prob(A & B) = Prob(A) $*$ Prob($B \mid A$).)
8. Evaluate the following expression $\binom{4}{1}$. (HINT: Remember that $\binom{n}{k} = \frac{n!}{k!(n-k)!}$.)
9. Evaluate the following expression: $\binom{4}{3}$.
10. You toss four coins. Using the binomial coefficient formula, compute how many permutations are associated with a combination of two heads and two tails.
11. You toss four coins. Using the binomial formula, compute the probability of obtaining exactly two heads if the coin is fair. (HINT: The probability is given by $\binom{n}{k} * \pi^k (1-\pi)^{(n-k)}$.)
12. In a toss of four coins, use the binomial formula to compute the probability of obtaining exactly two heads if the coin is biased and the probability of a head is 75%.
13. Repeat the previous problem using Excel.
14. Using Excel, find the probability of obtaining two heads or *less* (again, with $n = 4$ and $\pi = 75\%$).
15. Using Excel, find the probability of obtaining two heads or *more* (again, with $n = 4$ and $\pi = 75\%$).
16. Chart the frequency distribution when $n = 4$ and $\pi = 75\%$.
17. You construct a science achievement test consisting of 100 multiple-choice questions. Each question has four alternatives, so the probability of obtaining a correct answer, based on guessing alone, is 25%. If a student randomly guesses on each item, what is the probability of obtaining a score of 30 or more correct?
   a) Use the binomial function in Excel. (HINT: Find the probability of 29 or less, and then subtract from one.)
   b) Use the normal approximation method, without a continuity correction. You will need to calculate a z-score, which means you will need to calculate

the mean expected correct ($n\pi$) and the standard deviation of this $\sqrt{n\pi(1-\pi}$. (HINT: Find the probability of 29 or less, and then subtract from one.)

c) Use the normal approximation with a continuity correction. (HINT: Find the probability of 29.5 or less, then subtract from one.)

18. If Y is a binomial random variable with parameters $n = 60$ and $\pi = 0.5$, estimate the probability that Y will equal or exceed 45, using the normal approximation with a continuity correction.
19. Let X be the number of people who respond to an online survey of attitudes toward social media. Assume the survey is sent to 500 people, and each has a probability of 0.40 of responding. You would like at least 250 people to respond to the survey. Estimate Prob($X \geq 250$), using the normal approximation with a continuity correction.

Technical Note

1.1 The variance of a proportion is $\frac{\pi(1-\pi)}{n}$. The rationale is as follows. If one were to double all the values of a random variable, so that $x^* = 2x$, then the SD would double and the variance, which is the square of the SD, would quadruple. In general, $Var(kx) = k^2 Var(x)$. In calculating a proportion, the count is divided by n, which is the same as multiplying by $1/n$. The variance of a count is therefore multiplied by $\frac{1}{n^2}$ to obtain the variance of a proportion. Specifically, $\frac{1}{n^2}\left[n\pi(1-\pi)\right] = \frac{\pi(1-\pi)}{n}$

APPENDIX 1.1

LOGIC BEHIND COMBINATION FORMULA (EQ. 1.9)

$\binom{n}{k}$ is the number of ways of combining n things so that there are k hits, for example, combining 100 coins so that there are 60 heads showing. The formula is: $\binom{n}{k} = \frac{n!}{k!(n-k)!}$. This appendix explains the logic behind this formula.

Three Object Example

First consider a simpler scenario. Suppose three people (Harry, Dick, and Tom) go to the theatre. There are a number of possible seating arrangements:

HDT
DHT
HTD
DTH
THD
TDH

There are six permutations. Any one of the three people can fill the first seat. Once the first seat is filled, there are two people left, so either one could fill the second seat. Once the second seat is filled, there is just one person left, and so there is just one possibility for filling the third seat. The total number of possibilities is $3 * 2 * 1 = 6$. This is the factorial of three (3!). More generally, there are n ! distinct ways of arranging n objects.

Suppose now that Harry and Dick are identical twins. Some of the six permutations will no longer appear distinct. If H = D and substituting H for D, the six permutations become:

HDT ➔ HHT
DHT ➔ HHT
HTD ➔ HTH
DTH ➔ HTH
THD ➔ THH
TDH ➔ THH

The number of distinct arrangements has been cut in half. We could just modify the formula by dividing six by two, or more specifically, 2!). In general, when k of the objects (or people) are identical, the number of distinct permutations is given by $\frac{n!}{k!}$.

We need to divide by *k* !, and not just *k,* because there are *k* ! ways of arranging the *k* identical objects, and each of these correspond to a possibility that is duplicative of another possibility. There are therefore only three possible outcomes: HHT, HTH, and THH. If instead of people, we let H represent a coin toss coming up heads, and T tails, we can see that there are three permutations corresponding to a combination of two heads and one tail. In other words, $\binom{3}{2} = \frac{3!}{2!(3-2)!} = 3$. More generally, $\binom{n}{k} = \frac{n!}{k!(n-k)!}$, where $(n - k)$ is the number of misses. Although in the example, $(n - k)! = 1$ in more complex examples $(n - k)!$ may reflect duplicative possibilities, so we need to divide by this term as well as *k*!.

Reference

Nussbaum, E. M. (2011). Argumentation, dialogue theory, and probability modeling: Alternative frameworks for argumentation research in education. *Educational Psychologist*, *46*, 84–106. https://doi.org/10.1080/00461520.2011.558816

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

CHAPTER 2

ESTIMATION AND HYPOTHESIS TESTING

In this chapter, we examine how to estimate the value of a population parameter such as a probability (π) or mean (μ). We will also examine the binomial test, which tests hypotheses about π.

Estimation

The Estimation Problem(s)

In teaching this topic in my course on categorical data analysis, I begin the lesson by asking students to answer on a piece of paper, "Have you ever been married?" This question asks about a nominal variable and therefore it is appropriate to calculate or estimate proportions. The goal of the exercise is to infer what percentage of students at the university as a whole have ever been married (N = 30,000). In other words, we need to estimate the population proportion. Doing so requires taking a sample and using sample statistics to estimate the population parameter. (A population parameter is a numerical feature of a population that we are trying to estimate.) The sample should be random, but in this exercise, we suppose that the class represents a random sample. The population is all the students at the university.

By convention, population parameters are represented by Greek letters: In this case, pi (π) for a proportion. (For a metric variable, we would estimate μ, the population mean.) By convention, English letters (e.g., P or $\bar{X}$) represent sample statistics. Note that π, the population proportion, is also a probability. Therefore, if 60% of the students at the university have been married, then if one *randomly* draws an individual out of the population, there will be a 60% probability that she or he has been married. Suppose our sample consists of 20 students (n = 20). If X is the number of students in our sample who have been married, the expected value ($n\pi$) is the most likely value of X in the sample. Likewise, the expected value of P is π. This example shows why it is important to use random samples: The

DOI: 10.4324/9781003145264-2

laws of probability make it likely that the characteristics of the sample will be representative (i.e., similar) to the population. In this case, if the sample contained 12 students who had been married, the *P* will be 60%, just like the population, where $\pi = 60\%$.

However, recall that this expected value is only the most likely value of our sample estimate *P*; it is possible that we could draw a sample with 11 or 14 married students. We should expect that we will obtain a sample of *about* 12 students who have been married, but that our estimate will be off a little. The "off a little" reflects *sampling error*, the fact that even with a random sampling, the sample might not represent the population perfectly. Because of sampling error, we always represent statistical estimates in the form of confidence intervals, for example .58 ± .02 (with 95% confidence).

Forming a Confidence Interval for a Proportion

In the previous example, we stipulated that 60% of the population had been married. However, we typically will not know the population parameters, which is why we must estimate them. Suppose we now take a random sample of 30 individuals, and 19 of them indicate that they have been married. (This category would include people who are presently married as well as those who have been divorced or widowed.) Then $P = \frac{19}{30} = 63.3\%$. *P* is our sample estimate of π. How do we form a confidence interval?

To address this question, we need to understand the concept of a *sampling distribution*. This concept is, next to probability, the most foundational one in all of inferential statistics. A sampling distribution for a sample statistic, such as *P*, given a particular sample size (e.g., $n = 30$), is the distribution of the statistic associated with all the different possible samples of size *n* that could be drawn from the population. For each such sample, there is a P. The sampling distribution of *P* is the distribution of all the possible *P*'s (one for each sample).

In the previous example, we drew a sample with a *P* of 63.3%. If we were to draw another sample, we might obtain a *P* of 70%. If we were to draw a third sample, we might obtain a *P* of 55%. Theoretically, we could go on indefinitely drawing samples and plotting the values of the sampling distribution. In practice, we typically draw only one sample, at most two. We therefore only observe one or two points of the sampling distribution.

Although we never observe most of the points of the sampling distribution, we can still make theoretical inferences about it. To construct a confidence interval, we need to make inferences about the sampling distribution's: (a) Mean, (b) standard deviation (standard error), and (c) shape.

It can be shown that the mean of the sampling distribution is π, the population proportion. (This claim is demonstrated in Chapter 3 and in Appendix 2.1.)

EXPLORING THE CONCEPT

The mean of a sampling distribution is by definition the expected value. If the population mean (π) were 50%, why is the most likely value of *P*, from all the samples that could be drawn, 50%?

The population distribution is a Bernoulli distribution (see Figure 1.4 in Chapter 1), and so the graph of the population distribution will have two bars: One for $X = 1$ (with $\text{Prob}(X = 1) = \pi$) and one for $X = 0$ (with $\text{Prob}(X = 0) = 1–\pi$). Recall that the variance of a Bernoulli distribution is $\pi(1–\pi)$. Because we are using *P* to estimate π, we can use *PQ* to estimate the population variance, where *Q* is the probability of a miss ($Q = 1–P$).

However, our interest is in finding the standard deviation of the sampling distribution, not of the population distribution. (The former is known as the *standard error*, or *SE*.) The applicable formula is:

$$\textit{Estimated SE of a Proportion} = \sqrt{\frac{PQ}{n}}. \qquad \text{(Eq. 2.1)}$$

Appendix 2.1 presents the proof (see also Chapter 3).

EXPLORING THE CONCEPT

Conduct a thought experiment by answering the following questions: (a) If your sample consisted of just one observation ($n = 1$) from a population of 30,000, how many different samples could you draw? (b) Would the shape of the sampling distribution be just like the population distribution (i.e., Bernoulli)? (c) Would the SDs of the two distributions be the same?

Note from Eq. 2.1 that as *n* increases, the SE decreases, meaning that the sampling distribution becomes "thinner." Eq. 2.1 implies that the SE approaches zero as *n* approaches infinity. In fact, the SE would approach zero as *n* approaches *N*¸ which is the size of the population. Technically, therefore, we should include this additional constraint in the formula by writing:

$$\textit{Estimated SE of a Proportion} = \sqrt{\frac{PQ}{n} - \frac{PQ}{N}}. \qquad \text{(Eq. 2.2)}$$

When the population size is large, the second term is negligible, so it is typically left out of the formula for the SE. One should use Eq. 2.2 when the population size is small.

EXPLORING THE CONCEPT

Suppose you sample the entire population where $n = N = 10$, and $P = 50\%$. To what value would Eq. 2.2 reduce? How many different samples could one draw out of the population? Does it make sense that the SE would be zero?

In summary, the mean of the sampling distribution of P is π and the estimated SE is $\sqrt{\frac{PQ}{n}}$. In regard to shape, although the shape of the *population* distribution is Bernoulli, the shape of the *sampling* distribution will be binomial when $n > 1$. This fact is illustrated in Figure 2.1. The sampling distribution for $n = 1$ is Bernoulli (same as the population distribution). For larger n ($n = 2$ and $n = 3$ are illustrated), the shape is binomial. Here is the rationale. The shape of the sampling distribution for each observation will be Bernoulli (identical to the population distribution), and a series of independent Bernoulli trials yields a binomial distribution. Thus, in a sample of size $n = 2$, there are four possible outcomes, as shown in Table 2.1. Combining the middle two permutations where the number of hits is 1 yields the sampling distribution of P, shown in Table 2.2 and Figure 2.1. Just as the binomial distribution characterizes the probabilities resulting from flipping two coins, it also characterizes the distribution associated with selecting a sample of two.

In Chapter 1, we noted that the variance of a binomial distribution could be estimated with the expression nPQ. However, the astute reader may have noticed that I indicated previously that the variance of the sampling distribution is $\frac{PQ}{n}$,

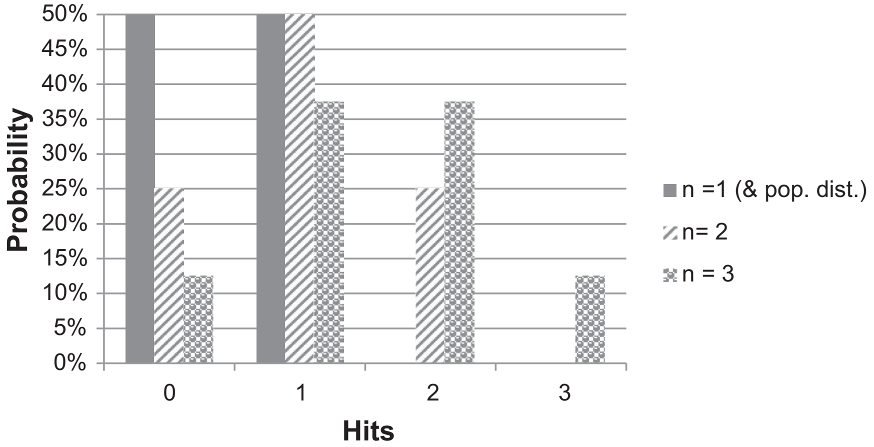


Figure 2.1 Bernoulli and binomial sampling distributions for varying sample sizes. For n = 1, the distribution is Bernoulli (and equal to the population distribution). For n = 2 and n = 3, the distributions are binomial.