Categorical Data Analysis for the Behavioral and Social Sciences

Second Edition

Razia Azen and Cindy M. Walker

ROUTLEDGE

Praise for a Previous Edition:

"An accessible walkthrough of the world of Bernoulli, Poisson, log-linear, multinomial, logistic, and all things non-interval. . . . I enjoyed reading this book and I will come back to it both as a reference and to digest some of the weightier chapters."

—*Chris Beeley, Institute of Mental Health, Nottingham, UK, in* The Psychologist

"This book fills an important need for a practitioner-oriented book on categorical data analyses. It not only could serve as an excellent resource for researchers working with categorical data, but would also make an excellent text for a graduate course in categorical data analysis."

—*Terry Ackerman, University of North Carolina–Greensboro, USA*

"A much needed book . . . it fills a significant gap in the market for a user-friendly categorical data analysis book. . . . Anyone wishing to learn categorical data analysis can read this book . . . The integration of both SPSS and SAS . . . increases the usability of this book."

—*Sara Templin, University of Alabama, USA*

"An accessible treatment of an important topic. . . . Through practical examples, data, and . . . SPSS and SAS code, the Azen and Walker text promises to put these topics within reach of a much wider range of students. . . . The applied nature of the book promises to be quite attractive for classroom use."

—*Scott L. Thomas, Claremont Graduate University, USA*

"Many social science students do not have the mathematical background to tackle the material covered in categorical data analysis books. What these students need is an understanding of what method to apply to what data, how to use software to analyze the data, and most importantly, how to interpret the results. . . . The book would be very useful for a graduate level course . . . for Sociologists and Psychologists. It might also be appropriate for Survey Methodologists."

—*Timothy D. Johnson, University of Michigan, USA*

"This book does an admirable job of combining an intuitive approach to categorical data analysis with statistical rigor. It has definitely set a new standard for textbooks on the topic for the behavioral and social sciences."

—*Wim J. van der Linden, Chief Research Scientist, CTB/McGraw-Hill*

Categorical Data Analysis for the Behavioral and Social Sciences

Featuring a practical approach with numerous examples, the second edition of *Categorical Data Analysis for the Behavioral and Social Sciences* focuses on helping the reader develop a conceptual understanding of categorical methods, making it a much more accessible text than others on the market. The authors cover common categorical analysis methods and emphasize specific research questions that can be addressed by each analytic procedure, including how to obtain results using SPSS, SAS, and R, so that readers are able to address the research questions they wish to answer.

Each chapter begins with a "Look Ahead" section to highlight key content. This is followed by an in-depth focus and explanation of the relationship between the initial research question, the use of software to perform the analyses, and how to interpret the output substantively. Included at the end of each chapter are a range of software examples and questions to test knowledge.

New to the second edition:

- The addition of R syntax for all analyses and an update of SPSS and SAS syntax.
- The addition of a new chapter on GLMMs.
- Clarification of concepts and ideas that graduate students found confusing, including revised problems at the end of the chapters.

Written for those without an extensive mathematical background, this book is ideal for a graduate course in categorical data analysis taught in departments of psychology, educational psychology, human development and family studies, sociology, public health, and business. Researchers in these disciplines interested in applying these procedures will also appreciate this book's accessible approach.

Razia Azen is Professor at the University of Wisconsin–Milwaukee, USA, where she teaches basic and advanced statistics courses. Her research focuses on methods that compare the relative importance of predictors in linear models. She received her MS in statistics and PhD in quantitative psychology from the University of Illinois, USA.

Cindy M. Walker is President and CEO of Research Analytics Consulting, LLC. Previously, she was a professor at the University of Wisconsin–Milwaukee, where she taught basic and advanced measurement and statistics courses. Her research focuses on applied issues in psychometrics. She received her PhD in quantitative research methodologies from the University of Illinois at Urbana–Champaign, USA.

Categorical Data Analysis for the Behavioral and Social Sciences

Razia Azen and Cindy M. Walker

Second Edition

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

Second edition published 2021
by Routledge
52 Vanderbilt Avenue, New York, NY 10017

and by Routledge
2 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 2011

Library of Congress Cataloging-in-Publication Data
Names: Azen, Razia, 1969– author. | Walker, Cindy M., 1965– author.
Title: Categorical data analysis for the behavioral and social sciences / Razia Azen and Cindy M. Walker.
Description: Second edition. | New York, NY : Routledge, 2021. | Includes bibliographical references and index. |
Identifiers: LCCN 2020051576 (print) | LCCN 2020051577 (ebook) | ISBN 9780367352745 (hardback) | ISBN 9780367352769 (paperback) | ISBN 9780429330308 (ebook)
Subjects: LCSH: Social sciences—Statistical methods.
Classification: LCC HA29 .A94 2021 (print) | LCC HA29 (ebook) | DDC 300.72/7—dc23
LC record available at https://lccn.loc.gov/2020051576
LC ebook record available at https://lccn.loc.gov/2020051577

ISBN: 978-0-367-35274-5 (hbk)
ISBN: 978-0-367-35276-9 (pbk)
ISBN: 978-0-429-33030-8 (ebk)

Typeset in Bembo
by Apex CoVantage, LLC

Access the Support Material: www.routledge.com/9780367352769

To our students: past, present, and future

Contents

Preface

While teaching categorical data analysis in an educational psychology department, we found that the textbooks currently available for categorical data analysis courses, although very good, are frequently too technical and require a more extensive understanding of mathematics than is typically the case for many students and researchers in the social sciences. Therefore, our approach in writing this book was to focus on the concepts and applications, rather than the technical details, of common categorical data analysis methods. We do present theory along with some technical details, as we believe these to be extremely important components of understanding statistical methods, but the main focus in this book is on conceptual understanding and application. Thus, students and researchers should be able to understand and apply these statistical procedures even without extensive prior training in mathematics.

The main goal of writing this book was to provide an accessible resource on categorical or cross-classified data analysis to students and researchers in the social and behavioral sciences. This book is intended to serve students and researchers in fields such as education, psychology, sociology, business, and health, as well as many others that require the analysis of categorical variables. The book is primarily intended for a course in categorical data analysis, which is usually taught at the graduate level in the behavioral and social sciences, and thus probably requires one or two prerequisite statistics classes that cover conventional statistical methods (e.g., factorial analysis of variance, multiple regression). Some experience with statistical software is also expected, although we provide extensive examples and instructions for using SAS (version 9.4), IBM SPSS 26 (which we refer to as SPSS in this book), and R (newly added to the second edition).

This book covers the most commonly used categorical data analysis techniques, and emphasis is placed on techniques appropriate for analyzing variables measured by a nominal scale. Although the procedures presented can also be applied with ordinal variables, more advanced methods that are specific to ordinal variables are not covered extensively. The order of the chapters reflects the complexity of the material presented. We start with the building blocks of scales of measurement (Chapter 1), probability distributions (Chapter 2), and inferential methods for a single categorical variable (Chapter 3), and then move on to methods involving contingency tables with two (Chapter 4) and three (Chapter 5) categorical variables. The transition to modeling is made by introducing the overarching ideas of generalized linear models (Chapter 6) and then demonstrating these specifically as they apply to loglinear models (Chapter 7), logistic regression with continuous (Chapter 8) and categorical (Chapter 9) predictors, and concluding with logistic models for multicategory response variables (Chapter 10). A new chapter (Chapter 11) on generalized linear mixed models, namely logistic regression for clustered data (where observations are clustered in groups), is available in the second edition.

We intend the treatment to be both conceptual and practical, in that we present the concepts in general terms and demonstrate them with practical examples. In general, each chapter begins with a "Look Ahead" section that gives an overview of the material covered in the chapter, and concludes with a "Summary" section that summarizes the main topics covered. Each chapter typically contains a general conceptual development of the ideas and methods, illustrated with examples, along with a discussion of the relationships between and among the new procedure being introduced and existing schema (i.e., analytic procedures covered in prerequisite courses or previous chapters). The conceptual development is then followed by more extensive examples that use a data set to answer research questions, and explanations of how to obtain and interpret relevant output from statistical software packages. We demonstrate the examples using SAS (version 9.4), SPSS (version 26), and R, and provide computing instructions as well as selected output from these packages in the chapters. We also include both conceptual and practical problems at the end of each chapter so that students can practice the concepts and methods covered in the chapter. The data sets used in the book can be accessed at the author's web site, https://sites.uwm.edu/azen/categoricaldata/. We wish to thank Michael Ioffe for providing valuable feedback and assisting with the addition of R code to this book.

Ultimately, our goal in writing this book was to create a resource that can be used extensively to train and assist students and researchers in appropriately conducting analyses that address the sorts of research questions they wish to answer. We were inspired by our students to write a book that would appeal to social scientists, and so the examples we use are primarily from the social sciences. We hope that we accomplished this goal and that you will find this book useful and informative.

1 Introduction and Overview

A Look Ahead

There are many statistical procedures that can be used to address research questions in the social sciences, yet the required courses in most social science programs often pay little to no attention to procedures that can be used with categorical data. There are good reasons for this, both pedagogical and conceptual. Statistical procedures for categorical data require a new set of "tools", or a new way of thinking, because of the different distributional assumptions that must be made for these types of variables. Therefore, these procedures differ conceptually from those that may already be familiar to readers of this book, such as multiple regression and analysis of variance (ANOVA). Nonetheless, acquiring a conceptual understanding of the procedures that can be used to analyze categorical data opens a whole new world of possibilities to social science researchers. In this chapter, we introduce the reader to these possibilities, explain categorical variables, and give a brief overview of the history of the methods for their analysis.

1.1 What Is Categorical Data Analysis?

Categorical data arises whenever a variable is measured on a scale that simply classifies respondents into a limited number of groups or categories. For example, respondents' race, gender, marital status, and political affiliation are categorical variables that are often of interest to researchers in the social sciences. In addition to distinguishing a variable as either categorical (qualitative) or continuous (quantitative), variables can also be classified as either independent or dependent. The term **independent** refers to a variable that is experimentally manipulated (e.g., the treatment group each person is assigned to) but is also often applied to a variable that is used to predict another variable even if it cannot be externally manipulated (e.g., socioeconomic status). The term **dependent** refers to a variable that is of primary interest as an outcome or response variable; for example, the outcome of a treatment (based on treatment group) or the educational achievement level (predicted from socioeconomic status) can be considered dependent variables. Introductory statistics courses may give the impression that categorical variables can only be used as independent variables; this is likely because the analytic procedures typically learned in these courses assume that the dependent variable follows a normal distribution in the population, and this is not the case for categorical variables. Nonetheless, treating categorical variables exclusively as independent variables can ultimately restrict the types of research questions posed by social science researchers.

For example, suppose you wanted to determine whether charter schools differed in any substantial way from non-charter schools based on the demographics of the school (e.g., location: urban, suburban, or rural; type: public or private; predominant socioeconomic status of students: low, medium, or high; and so on). You would be unable to study this phenomenon without knowledge of categorical data analytic techniques because all variables involved are categorical. As another example, suppose that a researcher wanted to predict whether a student will graduate from high school based on information such as the student's attendance record (e.g., number of days in attendance), grade point average (GPA), income of parents, and so on. In this case, a categorical analysis approach would be more appropriate because the research question requires treating graduation status (yes or no) as the dependent variable. Indeed, a naive researcher might decide to use the graduation status as an independent variable, but this approach would not directly address the research question and would ultimately limit the results obtained from the analysis. The purpose of this book is to describe and illustrate analytic procedures that are applicable when the variables of interest are categorical.

1.2 Scales of Measurement

In general, measurement can be thought of as applying a specific rule to assign numbers to objects or persons for the sole purpose of differentiating between objects or persons on a particular attribute. For example, one might administer an aptitude test to a sample of college students to differentiate the students in terms of their ability. In this case, the specific rule being applied is the administration of the same aptitude test to all students. If one were to use different aptitude tests for different respondents, then the scores could not be compared across students in a meaningful way. For some variables, measurement can often be as precise as we want it to be; for example, we can measure the length of an object to the nearest centimeter, millimeter, or micromillimeter. For the variables typically measured in the social sciences, this is often not the case, and thus the only way to ensure quality measurement is to use instruments with good psychometric properties such as validity and reliability.

Measurement precision is typically defined by the presence or absence of the following four characteristics, presented in order in terms of the level of the information or precision they provide: (1) distinctiveness, (2) magnitude, (3) equal intervals, and (4) absolute zero. A measurement scale has the characteristic of distinctiveness if the numbers assigned to persons or objects simply differ on the property being measured. For example, if one were to assign a 0 to female respondents and a 1 to male respondents, then gender would be measured in a manner that had the characteristic of distinctiveness. A measurement scale has the characteristic of magnitude if the different numbers that are assigned to persons or objects can be ordered in a meaningful way based on their magnitude. For example, if one were to assign a score of 1 to a respondent who was very liberal, 2 to a respondent who was somewhat liberal, 3 to a respondent who was somewhat conservative, and 4 to a respondent who was very conservative, then political affiliation would be measured in a manner that had the characteristic of magnitude or rank ordering. A measurement scale has the characteristic of equal intervals if **equivalent differences** between two numbers assigned to persons or objects have an equivalent meaning. For example, if one were to consider examinees' scores from a particular reading test as indicative of reading proficiency, then, assuming that examinees' scores were created by summing the number of items answered correctly on the test, reading proficiency would be measured in a manner that had the characteristic of magnitude. Note that a score of 0 on the reading test does not necessarily represent an examinee who has no reading ability. Rather, a score of 0 may simply imply that the test was too difficult, which might be the case if a second-grade student was to be given an eighth-grade reading test. This is an important distinction between

Table 1.1 Properties of the four levels of measurement

Level of measurement	*Characteristic*				*Examples*
	Distinctiveness	*Magnitude*	*Equal intervals*	*Absolute zero*	
Nominal	✓				Race, religious affiliation, sex, eye color, personality type
Ordinal	✓	✓			Proficiency classification, level of agreement to survey item, class rank
Interval	✓	✓	✓		Achievement, aptitude, temperature
Ratio	✓	✓	✓	✓	Time, age, length, height, weight, number of spelling errors

measurement scales that have the property of equal intervals and those that have the property of having an absolute zero.

A measurement scale has the characteristic of having an absolute zero if assigning a score of 0 to persons or objects indicates an **absence** of the attribute being measured. For example, if a score of 0 represents no spelling errors on a spelling exam, then number of spelling errors would be measured in a manner that had the characteristic of having an absolute zero.

Table 1.1 indicates the four levels of measurement in terms of the four characteristics just described. Nominal measurement possesses only the characteristic of distinctiveness and can be thought of as the least precise form of measurement in the social sciences. Ordinal measurement possesses the characteristics of distinctiveness and magnitude and is a more precise form of measurement than nominal measurement. Interval measurement possesses the characteristics of distinctiveness, magnitude, and equal intervals and is a more precise form of measurement than ordinal measurement. Ratio measurement, which is rarely attained in the social sciences, possesses all four characteristics of distinctiveness, magnitude, equal intervals, and having an absolute zero and is the most precise form of measurement.

In categorical data analysis, the dependent or response variable, which represents the characteristic or phenomenon that we are trying to explain or predict in the population, is measured using either a nominal scale or an ordinal scale. Methods designed for ordinal variables make use of the natural ordering of the measurement categories, although the way in which we order the categories (i.e., from highest to lowest or from lowest to highest) is usually irrelevant. Methods designed for ordinal variables cannot be used for nominal variables. Methods designed for nominal variables will give the same results regardless of the order in which the categories are listed. While these methods can also be used for ordinal variables, doing so will result in a loss of information (and usually loss of statistical power) because the information about the ordering is lost.

In this book we will focus on methods designed for nominal variables. The analyses we will discuss can be used when all variables are categorical or when just the dependent variable is categorical. The independent or predictor variables (used to predict the dependent variable) can usually be measured using any of the four scales of measurement.

1.3 A Brief History of Categorical Methods

The early development of analytical methods for categorical data took place at the beginning of the 20th century and was spearheaded by the work of Karl Pearson and G. Udney Yule. As is typically the case when something new is introduced, the development of these procedures was not without controversy. While Pearson argued that categorical variables were simply

proxies of continuous variables, Yule argued that categorical variables were inherently discrete (Agresti, 1996). This in turn led the two statisticians to approach the problem of how to summarize the relationship between two categorical variables in vastly different ways. Pearson maintained that the relationship between two categorical variables could be approximated by the underlying continuum, and given his prestige in the statistical community, he was rarely challenged by his peers. Yule, however, challenged Pearson's approach to the problem and developed a measure to describe the relationship between two categorical variables that did not rely on trying to approximate the underlying continuum (Yule, 1912). Needless to say, Pearson did not take kindly to Yule's criticism and publicly denounced Yule's approach, going so far as to say that Yule would have to withdraw his ideas to maintain any credibility as a statistician (Pearson & Heron, 1913). One hundred years later, we have come to realize that both statisticians were partially correct. While some categorical variables, especially those that are measured in an ordinal manner, can be thought of as proxies to variables that are truly continuous, others cannot.

Pearson's work was also critiqued by R. A. Fisher, who maintained that one of Pearson's formulas was incorrect (Fisher, 1922). Even though statisticians eventually realized that Fisher was correct, it was difficult for Fisher to get his work published due to Pearson's reputation in the field (Agresti, 1996). Moreover, while Pearson's criticisms of Fisher's work were published (Pearson, 1922), Fisher was unable to get his rebuttals to these criticisms published, ultimately leading him to resign from the Royal Statistical Society (Cowles, 2001). Although Fisher's scholarly reputation among statisticians today is primarily due to other theoretical work, particularly in the area of ANOVA, he did make several contributions to the field of categorical data analysis, not the least of which is his approach to small sample techniques for analyzing categorical data.

1.4 Organization of This Book

Given the fact that most of the groundwork for categorical data analysis was developed in the early part of the 20th century, the procedures presented in this book are relatively new. Indeed, it was not until the middle of the 20th century that strong theoretical advances were made in the field, and clearly there is still more work to be done. In this book, we chose to present a few of the more widely used analytic procedures for categorical data in great detail, rather than inundating the reader with all of the models that can be used for categorical data and their associated nuances. The primary goal of this book is to help social scientists develop a conceptual understanding of the categorical data analytic techniques presented. Therefore, while extensive training in mathematics will certainly be of benefit to the reader, a lack of it should not prevent students and researchers from understanding and applying these statistical procedures. This is accomplished by utilizing examples that are reflective of realistic applications of data analytic techniques in the social sciences, and by emphasizing specific research questions that can be addressed by each analytic procedure.

This book begins by introducing the reader to the different types of distributions that most often underpin categorical variables. This is followed by a discussion of the estimation procedures and goodness-of-fit tests that are used with the subsequent categorical data analytical procedures. Procedures designed to analyze the relationship between two categorical variables are then presented, followed by a discussion of procedures designed to analyze the relationships among three categorical variables. This second half of the book presents models for categorical data, beginning with an overview of the generalized linear model. Specific applications of the generalized linear model are then presented in chapters on log-linear models, binomial logistic regression models, and multinomial logistic regression models.

1.5 Summary

In this chapter, we introduced the reader to the types of research questions that can be addressed with statistical procedures designed to analyze categorical data. We gave a brief history on the development of these procedures, discussed scales of measurement, and provided the readers with the organizational structure of this book. In the next chapter, we turn to the different distributions that are assumed to underlie categorical data.

Problems

1.1 Indicate the scale of measurement used for each of the following variables and explain your answer by describing the probable scale:

a. Sense of belongingness, as measured by a 20-item scale.
b. Satisfaction with life, as measured by a 1-item scale.
c. Level of education, as measured by a demographic question with five categories.

1.2 Indicate the scale of measurement used for each of the following variables and explain your answer by describing the probable scale:

a. Self-efficacy, as measured by a 10-item scale.
b. Race, as measured by a demographic question with six categories.
c. Income, as measured by yearly gross income.

1.3 For each of the following research scenarios, identify the dependent and independent variables (or indicate if not applicable) as well as the scale of measurement used for each variable. Explain your answers by describing the scale that might have been used to measure each variable.

a. A researcher would like to determine if boys are more likely than girls to be proficient in mathematics.
b. A researcher would like to determine if people in a committed relationship are more likely to be satisfied with life than those who are not in a committed relationship.
c. A researcher is interested in whether females tend to have lower self-esteem, in terms of body image, than males.
d. A researcher is interested in predicting religious affiliation from level of education.

1.4 For each of the following research scenarios, identify the dependent and independent variables (or indicate if not applicable) as well as the scale of measurement used for each variable. Explain your answers by describing the scale that might have been used to measure each variable.

a. A researcher would like to determine if people living in the United States are more likely to be obese than people living in France.
b. A researcher would like to determine if the cholesterol levels of men who suffered a heart attack are higher than the cholesterol levels of women who suffered a heart attack.
c. A researcher is interested in whether gender is related to political party affiliation.
d. A researcher is interested in the relationship between amount of sleep and grade point average for high school students.

1.5 Determine whether procedures for analyzing categorical data are needed to address each of the following research questions. Provide a rationale for each of your answers by

identifying the dependent and independent variables as well as the scale that might have been used to measure each variable.

a. A researcher would like to determine whether a respondent will vote for the Republican or Democratic candidate in the US presidential election based on the respondent's annual income.
b. A researcher would like to determine whether respondents who vote for the Republican candidate in the US presidential election have a different annual income than those who vote for the Democratic candidate.
c. A researcher would like to determine whether males who have suffered a heart attack have higher fat content in their diets than males who have not suffered a heart attack in the past six months.
d. A researcher would like to predict whether a man will suffer a heart attack in the next six months based on the fat content in his diet.

1.6 Determine whether procedures for analyzing categorical data are needed to address each of the following research questions. Provide a rationale for each of your answers by identifying the dependent and independent variables as well as the scale that might have been used to measure each variable.

a. A researcher would like to determine whether a student will complete high school based on the student's grade point average.
b. A researcher would like to determine whether students who complete high school have a different grade point average than students who do not complete high school.
c. A researcher would like to determine whether the families of students who attend college have a higher annual income than the families of students who do not attend college.
d. A researcher would like to determine whether a student will attend college based on his or her family's annual income.

1.7 Determine whether procedures for analyzing categorical data are needed to address each of the following research questions. Indicate what analytic procedure (e.g., ANOVA, regression) you would use for those cases that do **not** require categorical methods, and provide a rationale for each of your answers.

a. A researcher would like to determine if scores on the verbal section of the SAT can be used to predict whether students are proficient in reading on a state-mandated test administered in 12th grade.
b. A researcher is interested in whether income differs by gender.
c. A researcher is interested in whether level of education can be used to predict income.
d. A researcher is interested in the relationship between political party affiliation and gender.

1.8 Provide a substantive research question that would need to be addressed using procedures for categorical data analysis. Be sure to specify how the dependent and independent variables would be measured, and identify the scales of measurement used for these variables.

2 Probability Distributions

A Look Ahead

The ultimate goal of most inferential statistics endeavors is to determine how likely it is to have obtained a particular **sample** given certain assumptions about the **population** from which the sample was drawn. For example, suppose that in a random sample of students, obtained from a large school district, 40% of the students are classified as minority students. Based on this result, what can we infer about the proportion of students who are minority students in the school district as a whole? This type of inference is accomplished by determining the **probability** of randomly selecting a particular sample from a specific population, and this probability is obtained from a sampling distribution. In this chapter, we introduce several probability and sampling distributions that are appropriate for categorical variables.

The properties of a sampling distribution typically depend on the properties of the underlying distribution of the random variable of interest, a distribution that also provides the probability of randomly selecting a particular observation, or value of the variable, from a specific population. The most familiar example of this concept involves the sampling distribution of the mean, often used to determine the probability of obtaining a sample with a particular sample mean value. The properties of the sampling distribution depend on the properties of the probability distribution of the random variable in the population (e.g., its mean and variance). In the case of continuous variables, for sufficient sample sizes the sampling distribution of the mean and the probabilities obtained from it are based on the normal distribution. In general, probability distributions form the basis for inferential procedures. In this chapter, we discuss these distributions as they apply to categorical variables.

2.1 Probability Distributions for Categorical Variables

We begin with a simple example, where we suppose that the population of interest is a fifth-grade class at a middle school consisting of 50 students: 10 females and 40 males. In this case, if a teacher randomly selected one student from the class, the teacher would be more likely to choose a male student than a female student. In fact, because the exact number of male and female students in the population is known, the exact probability of randomly selecting a male or female student can be determined. Specifically, the probability of any particular outcome is defined as the number of ways a particular outcome can occur out of the total number of possible outcomes; therefore, the probability of randomly selecting a

male student in this example is 40 / 50 = 0.8, and the probability of randomly selecting a female student is 10 / 50 = 0.2.

However, contrary to the example with the population of fifth-grade students, it is atypical to know the exact specifications (i.e., distribution) of the population. The goal of inferential statistical procedures is to make inferences about the population from observed sample data, not the other way around. This is accomplished by considering a value that is obtained as the result of some experiment or data collection activity to be only one possible outcome out of many different outcomes that may have occurred; that is, this value is a variable because it can vary across different studies or experiments. For example, if 10 students were randomly selected from the fifth grade discussed earlier, the proportion of males in that sample of 10 students could be used to infer the proportion of males in the larger group or population (of all 50 students). If another random sample of 10 students was obtained, the proportion of males may not be equal to the proportion in the first sample; in that sense, the proportion of males is a variable. **Random variable** is a term that is used to describe the possible outcomes that a particular variable may take on. It does not describe the actual outcome itself and cannot be assigned a value, but is rather used to convey the fact that the outcome obtained was the result of some underlying random process. A **probability distribution** is a mathematical function that links the actual outcome obtained from the result of an experiment or data collection activity (e.g., a random sample) to the probability of its occurrence.

Most methods that deal with continuous dependent variables make the assumption that the values obtained are random observations that come from a normal distribution. In other words, when the dependent variable is continuous, it is assumed that a normal distribution is the underlying random process in the population from which the variable was obtained. However, there are many other probability distributions and, when the dependent variable is categorical, it can no longer be assumed to have been obtained from a population that is normally distributed. The purpose of this chapter is to describe several probability distributions that are assumed to underlie the population from which categorical data are obtained.

2.2 Frequency Distribution Tables for Discrete Variables

A discrete variable is a variable that can only take on a finite number of values. Categorical data almost always consist of discrete variables. One way to summarize data of this type is to construct a frequency distribution table, which depicts the number of responses in each category of the measurement scale as well as the probability of occurrence of a particular response category and the percentage of responses in each category. In fact, a frequency distribution table is a specific example of a probability distribution. For example, suppose a random sample of individuals in the United States were asked to identify their political affiliation using a 7-point response scale that ranged from extremely liberal to extremely conservative, with higher values reflecting a more liberal political affiliation. Table 2.1 is a frequency distribution table summarizing the (hypothetical) responses.

Note that the **probabilities** depicted in the table are also the **proportions**, computed by simply dividing the **frequency** of responses in a particular category by the total number of respondents (e.g., the proportion of those who are extremely liberal is 30 / 1443 ≈ .021), and the **percentages** depicted in the table can be obtained by multiplying these values by 100 (e.g., the percentage of those who are extremely liberal is (100)(.021) = 2.1%). More formally, if the frequency is denoted by f and the total number of respondents is denoted by N, then the probability or proportion is $\frac{f}{N}$ and the percentage is $100\left(\frac{f}{N}\right)\%$. Note also that the frequency can be obtained from the proportion (or probability) by $f = N$(proportion).

Table 2.1 Frequency distribution table depicting political affiliation of respondents

Political affiliation (X)	*Frequency (f)*	*Percentage*	*Probability*
Extremely liberal (7)	30	2.1	0.021
Liberal (6)	163	11.3	0.113
Slightly liberal (5)	193	13.4	0.134
Moderate (4)	527	36.5	0.365
Slightly conservative (3)	248	17.2	0.172
Conservative (2)	241	16.7	0.167
Extremely conservative (1)	41	2.8	0.028
Total	1443	100.0	1.000

A frequency distribution table such as the one depicted in Table 2.1 summarizes the data obtained so that a researcher can easily determine, for example, that respondents were most likely to consider their political affiliation to be moderate and least likely to consider themselves to be extremely liberal. Yet, how might these data be summarized more succinctly?

Descriptive statistics, such as the mean and standard deviation, can be used to summarize discrete variables just as they can for continuous variables, although the manner in which these descriptive statistics are computed differs with categorical data. In addition, because it is no longer appropriate to assume that a normal distribution is the underlying random mechanism that produces the categorical responses in a population, distributions appropriate to categorical data must be used for inferential statistics with these data (just as the normal distribution is commonly the appropriate distribution used for inferential statistics with continuous data). The two most common probability distributions assumed to underlie responses in the population when data are categorical are the binomial distribution and the Poisson distribution, although there are also other distributions that are appropriate for categorical data. The remainder of this chapter is devoted to introducing and describing common probability distributions appropriate for categorical data.

2.3 The Hypergeometric Distribution

The hypergeometric distribution can be used with discrete variables to determine the number of "successes" in a sequence of n draws from a finite population where sampling is conducted without replacement. It should be noted that "success" is simply a label for the occurrence of a particular event of interest. For example, suppose the admissions committee at a medical college has 15 qualified applicants, 3 of which are minority applicants, and can only admit 2 new students. In this case, we might define success as the admission of a minority applicant, and the hypergeometric distribution could then be used to determine the probability of admitting at least one minority student if the admissions committee were to randomly select two new students from the 15 qualified applicants.

In general, if Y denotes the total number of successes desired (e.g., minority students selected for admission), m denotes the number of possible successes (e.g., minority applicants), N denotes the total sample size (e.g., number of qualified applicants), and n denotes the total number of draws (e.g., number of applicants to be admitted), then the probability that $Y = k$ (where k is a specific number of successes) can be expressed by

$$P\left(Y = k\right) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}, \tag{2.1}$$

where the notation $\binom{m}{k}$ is read as "m choose k" and refers to the number of ways that k individuals (or objects) can be selected from a total of m individuals (or objects). This is computed as

$$\binom{m}{k} = \frac{m!}{k!(m-k)!} = \frac{m(m-1)(m-2)\cdots(1)}{[k(k-1)(k-2)\cdots(1)][(m-k)(m-k-1)(m-k-2)\cdots(1)]}.$$

> Note that 1! and 0! (1 factorial and 0 factorial, respectively) are both defined to be equal to 1, and any other factorial is the product of all integers from the number in the factorial to 1. For example, $m! = m(m-1)(m-2)\ldots 1$.

The general formulation in Equation 2.1 can be conceptualized using the theory of combinatorics. The number of ways to select or choose n objects (e.g., applicants) from the total number of objects, N, is $\binom{N}{n}$. Similarly, the number of ways to choose k minority applicants from the total number of minority applicants, m, is $\binom{m}{k}$. For each possible manner of choosing k minority applicants from the total of m minority applicants, there are $\binom{N-m}{n-k}$ possible ways to select the remaining number of nonminority applicants to fulfill the requirement of admitting n applicants. Therefore, the number of ways to form a sample consisting of exactly k minority applicants is

$$\binom{m}{k}\binom{N-m}{n-k}.$$

Further, because each of the samples is equally likely, the probability of selecting exactly k minority applicants is

$$\frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}.$$

Using the numerical example provided earlier, with $m = 3$ minority applicants, $k = 1$ minority admission, $N = 15$ qualified applicants, and $n = 2$ total admissions, the probability of admitting exactly one minority student if applicants are selected at random can be computed as follows:

$$P(Y=1) = \frac{\binom{3}{1}\binom{15-3}{2-1}}{\binom{15}{2}} = \frac{\binom{3}{1}\binom{12}{1}}{\binom{15}{2}} = \frac{\left(\frac{3!}{1!(2!)}\right)\left(\frac{12!}{1!(11!)}\right)}{\left(\frac{15!}{2!13!}\right)} = \frac{3(12)}{\frac{15(14)}{2}} = \frac{36}{105} = 0.34.$$

Similarly, the probability of admitting exactly two minority students if applicants are selected at random can be computed as follows:

$$P\left(Y=2\right)=\frac{\binom{3}{2}\binom{15-3}{2-2}}{\binom{15}{2}}=\frac{\binom{3}{2}\binom{12}{0}}{\binom{15}{2}}=\frac{\left(\frac{3!}{2!\left(1!\right)}\right)\left(\frac{12!}{0!\left(12!\right)}\right)}{\left(\frac{15!}{2!13!}\right)}=\frac{3\left(1\right)}{\frac{15\left(14\right)}{2}}=\frac{3}{105}=0.03.$$

Therefore, the probability of admitting at least one minority student is

$$P\left(Y=1\right)+P\left(Y=2\right)=0.34+0.03=0.37.$$

We could also use this procedure to compute the probability of admitting no minority students, but it is easier to make use of the laws of probability to determine this; that is, because only two applicants are to be admitted, either none, one, or both of those admitted could be minority students. Therefore, these three probabilities must sum to 1.0, and, because the probability of one or two minority student admissions was shown to be 0.37, the probability that no minority students are admitted can be obtained by 1.0 − 0.37 = 0.63. In this case, if applicants are randomly selected from the qualified applicant pool, it is most likely (with 63% probability) that none of the applicants selected by the admissions committee will be minority students, so to increase diversity at this particular medical college, the committee should select applicants in a nonrandom manner.

Most any distribution can be described by its mean and variance. In the case of the hypergeometric distribution, the mean (μ) and variance (σ^2) can be computed using the following formulas:

$$\mu=\frac{nm}{N}$$

and

$$\sigma^2=\frac{n\left(m/N\right)\left(1-m/N\right)\left(N-n\right)}{N-1}.$$

For our example, the mean of the distribution is

$$\mu=\frac{3\left(2\right)}{15}=\frac{6}{15}=0.4$$

and the variance is

$$\begin{aligned}\sigma^2&=\frac{n\left(m/N\right)\left(1-m/N\right)\left(N-n\right)}{N-1}=\frac{3\left(2/15\right)\left(1-2/15\right)\left(15-3\right)}{15-1}\\&=\frac{3\left(.13\right)\left(.87\right)\left(12\right)}{14}=0.297,\end{aligned}$$

which implies that the standard deviation is $\sigma=\sqrt{0.297}=0.525$.

What do these values represent? The mean can be interpreted as the expected or most likely number of minority applicants chosen, if they are randomly selected from the applicant pool. Because this value is less than 1, the most likely number of minority applicants selected is less than 1, which mirrors the information obtained when the probabilities were calculated directly. The standard deviation can be interpreted in the usual manner; that is, it can be thought of as the average difference between the observed number of minority applicants selected in any given sample and the mean of minority applicants selected (which in this case is 0.4), assuming the hypergeometric distribution underlies the selection process.

2.4 The Bernoulli Distribution

The Bernoulli distribution is perhaps the simplest probability distribution for discrete variables and can be thought of as the building block for more complicated probability distributions used with categorical data. A discrete variable that comes from a Bernoulli distribution can only take on one of two values, such as pass/fail or proficient/not proficient. This distribution can be used to determine the probability of success (i.e., the outcome of interest) if only one draw is made from a finite population. Therefore, this distribution is a special case of the hypergeometric distribution with $n = 1$.

Using the example presented earlier, if $n = 1$ then only one applicant is to be selected for admission. Because this candidate can either be a minority applicant or not, the number of successes, k, is also equal to 1. In this case, to compute the probability of a success ($k = 1$), Equation 2.1 can be simplified as follows:

$$P\left(Y = k = 1\right) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} = \frac{\binom{m}{1}\binom{N-m}{0}}{\binom{N}{1}} = \frac{\left(\frac{m!}{(m-1)!(1!)}\right)\left(\frac{(N-m)!}{(N-m)!}\right)}{\left(\frac{N!}{(N-1)!(1!)}\right)} = \frac{m}{N}. \quad (2.2)$$

Therefore, the probability of selecting a minority applicant from the pool, which can be thought of as the probability of a single success, is simply the number of minority candidates divided by the total number of candidates. In other words, it simply represents the proportion of minority candidates in the candidate pool. This probability is typically denoted by π in the population (or p in the sample), and is equal to $\pi = 3 / 15 = 0.2$ in this example. Using the laws of probability, the probability of not selecting a minority candidate from the candidate pool can be obtained by $1 - \pi$, which in our example is $1 - 0.2 = 0.8$. The mean, or expected value, for this distribution is simply $\mu = \pi$ and the variance for this distribution can be calculated using the formula $\sigma^2 = \pi(1 - \pi)$, which is $0.2(0.8) = 0.16$ for the example presented.

2.5 The Binomial Distribution

Categorical variables that follow the binomial distribution can only take on one of two possible outcomes, as is the case with variables that follow the Bernoulli distribution. However, whereas the Bernoulli distribution only deals with one trial, outcome, or event, the binomial distribution deals with multiple trials (denoted by n). Therefore, the binomial distribution can be thought of as an extension of the Bernoulli distribution and can also be considered to be akin to the hypergeometric distribution.

Like the hypergeometric distribution, the binomial distribution can be used with discrete variables when trying to determine the number of successes in a sequence of n trials that are drawn from a finite population. Unlike the hypergeometric distribution, where sampling is conducted without replacement, the events or trials are independent when they follow a binomial distribution. With the hypergeometric distribution there is a dependency among the events considered because sampling is done without replacement; that is, the probability of success may become larger in subsequent trials, and this change can be dramatic if the population is small. With the binomial distribution, the probability of success does not change in subsequent trials because the probability of success for each trial is independent of previous or subsequent trials; in other words, sampling is conducted *with* replacement. For example, to draw 2 applicants from a total of 15, sampling without replacement proceeds such that once the first applicant has been selected for admission there are only 14 remaining applicants from which to select the second admission. Therefore, the probability of success for the second selection differs from the probability of success for the first selection. On the other hand, sampling with replacement proceeds such that each selection is considered to be drawn from the pool of all 15 individuals. In other words, the first selection is not removed from the total sample before the second selection is made, and each selection has the same probability, thereby making each selection independent of all other selections. In this case, it is possible that the same applicant will be selected on both draws, which is conceptually nonsensical. However, this is very unlikely with large samples, which is why the binomial distribution is often used with large samples even though conceptually the hypergeometric distribution may be more appropriate. In fact, the binomial distribution is an excellent approximation to the hypergeometric distribution if the size of the population is relatively large compared to the number of cases that are to be sampled from the population (Sheskin, 2007).

In addition, the binomial distribution is more appropriate when selections are truly independent of each other. For example, suppose the probability that a female applicant will be admitted to an engineering program (a success) is 0.8 across all such programs. In this case, to determine the probability that 3 out of 15 (or, more generally, k out of n) engineering programs would admit a female applicant, the number of trials would be $n = 15$ and the trials would be considered independent given that each program's admission decision is not influenced by any other program's decision.

In general, for a series of n independent trials, each resulting in only one of two particular outcomes, where π is the probability of "success" and $1 - \pi$ is the probability of "failure", the probability of observing k "successes" can be expressed by:

$$P\left(Y = k\right) = \binom{n}{k} \pi^{k} \left(1 - \pi\right)^{(n-k)}. \tag{2.3}$$

For example, suppose that the probability of being proficient in mathematics (a success) is 0.7 and three students are chosen at random from a particular school. The binomial distribution can be used, for example, to determine the probability that of the three randomly selected students ($n = 3$), all three are proficient in mathematics ($k = 3$):

$$P\left(Y = 3\right) = \binom{3}{3} 0.7^{3} \left(0.3\right)^{(3-3)} = 0.7^{3} = 0.343.$$

Further, the probabilities of all possible outcomes can be computed to determine the most likely number of students that will be found to be proficient. Specifically:

- The probability that none of the three students is proficient

$$= P\left(Y = 0\right) = \binom{3}{0} 0.7^{0} \left(0.3\right)^{(3-0)} = 0.3^{3} = 0.027.$$

- The probability that one of the three students is proficient

$$= P\left(Y = 1\right) = \binom{3}{1} 0.7^{1} \left(0.3\right)^{(3-1)} = 3\left(0.7\right)^{1} \left(0.3\right)^{2} = 0.189.$$

- The probability that two of the three students are proficient

$$= P\left(Y = 2\right) = \binom{3}{2} 0.7^{2} \left(0.3\right)^{(3-2)} = 3\left(0.7\right)^{2} \left(0.3\right)^{1} = 0.441.$$

Therefore, it is most likely that the number of mathematically proficient students will be two. Note that, since the preceding computations exhausted all possible outcomes, their probabilities should and do sum to one.

The mean and variance of the binomial distribution are expre.sed by $\mu = n\pi$ and $\sigma^2 = n\pi(1 - \pi)$, respectively. For our example, the mean of the distribution is $\mu = 3(0.7) = 2.1$, so the expected number of students found to be proficient in mathematics is 2.1, which is comparable to the information that was obtained when the probabilities were calculated directly. Moreover, the variance of the distribution is $\sigma^2 = 3(0.7)(0.3) = 0.63$, and the standard deviation is $\sigma = \sqrt{0.63} = 0.794$.

2.6 The Multinomial Distribution

The multinomial distribution is a multivariate extension of the binomial distribution when there are more than two possible outcomes. The binomial distribution can be used to represent the number of successes in *n* independent Bernoulli trials when the probability of success is the same for each trial. In a multinomial distribution, each trial results in one of *I* outcomes, where *I* is some fixed finite number and the probability of each possible outcome can be expressed by $\pi_1, \pi_2, \ldots, \pi_I$ such that the sum of all probabilities is $\sum_{i=1}^{I} \pi_i = 1$. Therefore, while the binomial distribution depicts the probability of obtaining a specific outcome pattern across two categories in *n* trials, the multinomial distribution depicts the probability of obtaining a specific outcome pattern across *I* categories in *n* trials. In this sense, the binomial distribution is a special case of the multinomial distribution with $I = 2$.

In the illustrative example provided for the binomial distribution, students were classified as either proficient or not proficient, and the probability of observing a particular number of proficient students was of interest, so the binomial distribution was appropriate. However, if students were instead categorized into one of four proficiency classifications, such as minimal, basic, proficient, and advanced, and the probability of observing a particular number of students in each of the four proficiency classifications was of interest, then the multinomial distribution would be appropriate.

In general, for n trials with I possible outcomes, where π_i = the probability of the i^{th} outcome, the (joint) probability that $Y_1 = k_1$, $Y_2 = k_2$, . . . , and $Y_I = k_I$ can be expressed by

$$P\left(Y_1 = k_1, Y_2 = k_2, \cdots, \text{and } Y_I = k_I\right) = \frac{n!}{k_1!k_2!\cdots k_I!}\pi_1^{k_1}\pi_2^{k_2}\cdots\pi_I^{k_I}. \tag{2.4}$$

Note that the sum of probabilities is

$$\sum_{i=1}^{I}\pi_i = 1$$

and the sum of outcome frequencies is

$$\sum_{i=1}^{I}k_i = n.$$

For example, suppose that the probability of being a minimal reader is 0.12, the probability of being a basic reader is 0.23, the probability of being a proficient reader is 0.47, and the probability of being an advanced reader is 0.18. If five students are chosen at random from a particular classroom, the multinomial distribution can be used to determine the probability that one student selected at random is a minimal reader, one student is a basic reader, two students are proficient readers, and one student is an advanced reader:

$$\begin{aligned}
&P\left(Y_1 = 1, Y_2 = 1, Y_3 = 2, \text{and } Y_4 = 1\right) \\
&\quad = \frac{5!}{(1!)(1!)(2!)(1!)}(0.12)^1(0.23)^1(0.47)^2(0.18)^1 \\
&= \frac{5(4)(3)(2)(1)}{2}(0.12)(0.23)(0.47)(0.47)(0.18) = 0.066.
\end{aligned}$$

There are 120 different permutations in which the proficiency classification of students can be randomly selected in this example (e.g., the probability that all five students were advanced, the probability that one student is advanced and the other four students are proficient, and so on), thus we do not enumerate all possible outcomes here. Nonetheless, the multinomial distribution could be used to determine the probability of any of the 120 different permutations (i.e., combinations of outcomes) in a similar manner.

There are I means and variances for the multinomial distribution, each dealing with a particular outcome, i. Specifically, for the i^{th} outcome, the mean can be expressed by $\mu_i = n\pi_i$ and the variance by $\sigma_i^2 = n\pi_i(1 - \pi_i)$. Table 2.2 depicts these descriptive statistics for each of the four proficiency classifications in the previous example.

Table 2.2 Descriptive statistics for proficiency classifications following a multinomial distribution

Proficiency classification (i)	π_i	μ_i	σ_i^2	σ_i
Minimal	0.12	5(0.12) = 0.60	5(0.12)(0.88) = 0.53	0.73
Basic	0.23	5(0.23) = 1.15	5(0.23)(0.77) = 0.89	0.94
Proficient	0.47	5(0.47) = 2.35	5(0.47)(0.53) = 1.25	1.12
Advanced	0.18	5(0.18) = 0.90	5(0.18)(0.82) = 0.74	0.86

2.7 The Poisson Distribution

The Poisson distribution is similar to the binomial distribution in that both distributions are used to model count data that varies randomly over time. In fact, as the number of trials gets larger (i.e., $n \to \infty$), the binomial distribution and the Poisson distribution tend to converge when the probability of success remains fixed. The major difference between the binomial and the Poisson distributions is that for the binomial distribution the number of observations (trials) is fixed, whereas for the Poisson distribution the number of observations is not fixed but rather the period of time in which the observations occur must be fixed. In other words, all eligible cases are studied when categorical data arise from the binomial distribution, whereas only those cases with a particular outcome in a fixed time interval are studied when data arise from the Poisson distribution.

For example, suppose that a researcher was interested in studying the number of accidents at a particular highway interchange in a 24-hour period. To study this phenomenon using the binomial distribution, the researcher would have to know the total number of cars (i.e., n) that had traveled through the particular interchange in a 24-hour period as well as the number of cars that had been involved in an accident at the particular interchange in this 24-hour period (which, when divided by n, provides π or p). This is because the binomial distribution assumes that there are two possible outcomes to the phenomenon: success or failure. On the other hand, to study this phenomenon using the Poisson distribution, the researcher would only need to know the mean number of cars that had been involved in an accident at the particular interchange in a 24-hour period, which is arguably much easier to obtain. For this reason, the Poisson distribution is often used when the probability of success is very small.

In general, if λ equals the number of successes expected to occur in a fixed interval of time, then the probability of observing k successes in that time interval can be expressed by

$$P\left(Y = k\right) = \frac{e^{-\lambda}\lambda^{k}}{k!}. \tag{2.5}$$

For example, suppose that over the last 50 years the average number of suicide attempts that occurred at a particular correctional facility each year is approximately 2.5. The Poisson distribution can be used to determine the probability that five suicide attempts will occur at this particular correctional facility in the next year. Specifically,

$$P\left(Y = 5\right)\frac{e^{-\lambda}\lambda^{5}}{5!} = \frac{e^{-2.5}2.5^{5}}{5!} = \frac{\left(0.082\right)\left(2.5\right)^{5}}{5\left(4\right)\left(3\right)\left(2\right)\left(1\right)} = 0.067.$$

Similarly, the probabilities for a variety of outcomes can be examined to determine the most likely number of suicides that will occur at this particular correctional facility in the next year. For example,

$$P\left(1 \text{ suicide attempt will occur}\right) = \frac{e^{-2.5}2.5^{1}}{1!} = \frac{\left(0.082\right)\left(2.5\right)}{1} = 0.205\,,$$

$$P\left(2 \text{ suicide attempts will occur}\right) = \frac{e^{-2.5}2.5^{2}}{2!} = \frac{\left(0.082\right)\left(2.5\right)^{2}}{2\left(1\right)} = 0.256,$$

$$P\left(3 \text{ suicide attempts will occur}\right) = \frac{e^{-2.5}2.5^{3}}{3!} = \frac{\left(0.082\right)\left(2.5\right)^{3}}{3\left(2\right)\left(1\right)} = 0.214,$$

$$P\left(4 \text{ suicide attempts will occur}\right) = \frac{e^{-2.5}2.5^4}{4!} = \frac{\left(0.082\right)\left(2.5\right)^4}{4\left(3\right)\left(2\right)\left(1\right)} = 0.138,$$

$$P\left(5 \text{ suicide attempts will occur}\right) = \frac{e^{-2.5}2.5^5}{5!} = \frac{\left(0.082\right)\left(2.5\right)^5}{5\left(4\right)\left(3\right)\left(2\right)\left(1\right)} = 0.067,$$

$$P\left(6 \text{ suicide attempts will occur}\right) = \frac{e^{-2.5}2.5^6}{6!} = \frac{\left(0.082\right)\left(2.5\right)^6}{6\left(5\right)\left(4\right)\left(3\right)\left(2\right)\left(1\right)} = 0.028,$$

$$P\left(7 \text{ suicide attempts will occur}\right) = \frac{e^{-2.5}2.5^7}{7!} = \frac{\left(0.082\right)\left(2.5\right)^7}{7\left(6\right)\left(5\right)\left(4\right)\left(3\right)\left(2\right)\left(1\right)} = 0.010,$$

$$P\left(8 \text{ suicide attempts will occur}\right) = \frac{e^{-2.5}2.5^8}{8!} = \frac{\left(0.082\right)\left(2.5\right)^8}{8\left(7\right)\left(6\right)\left(5\right)\left(4\right)\left(3\right)\left(2\right)\left(1\right)} = 0.003, \text{ and}$$

$$P\left(9 \text{ suicide attempts will occur}\right) = \frac{e^{-2.5}2.5^9}{9!} = \frac{\left(0.082\right)\left(2.5\right)^8}{9\left(8\right)\left(7\right)\left(6\right)\left(5\right)\left(4\right)\left(3\right)\left(2\right)\left(1\right)} = < 0.001.$$

Figure 2.1 depicts this distribution graphically, and Figure 2.2 depicts a comparable distribution if the average number of suicide attempts in a year had only been equal to 1. Comparing the two figures, note that the expected (i.e., average or most likely) number of suicides in any given year at this particular correctional facility, assuming that the number of suicides follows the Poisson distribution, is greater in Figure 2.1 than in Figure 2.2, as would be expected. Moreover, the likelihood of having a high number of suicides (e.g., four or more) is much lower in Figure 2.2 than in Figure 2.1, as would be expected.

Figure 2.3 illustrates a comparable distribution where the average number of suicide attempts in a year is equal to 5. Notice that Figure 2.3 is somewhat reminiscent of a normal distribution. In fact, as λ gets larger, the Poisson distribution tends to more closely resemble the normal distribution and, for sufficiently large λ, the normal distribution is an excellent approximation to the Poisson distribution (Cheng, 1949).

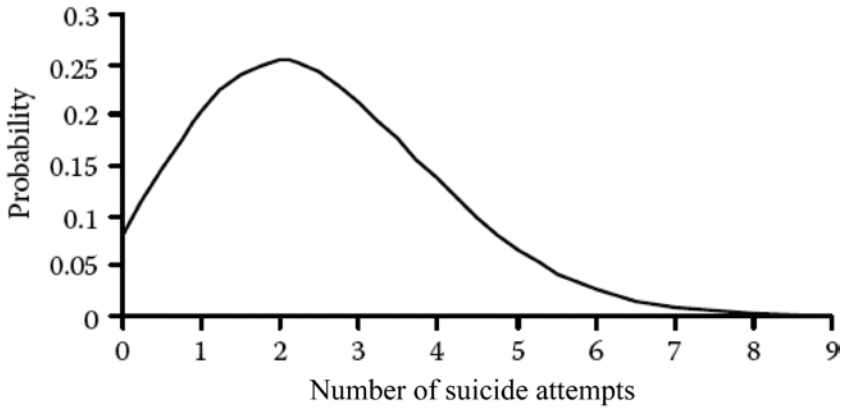

Figure 2.1 Illustration of Poisson distribution when $\lambda = 2.5$

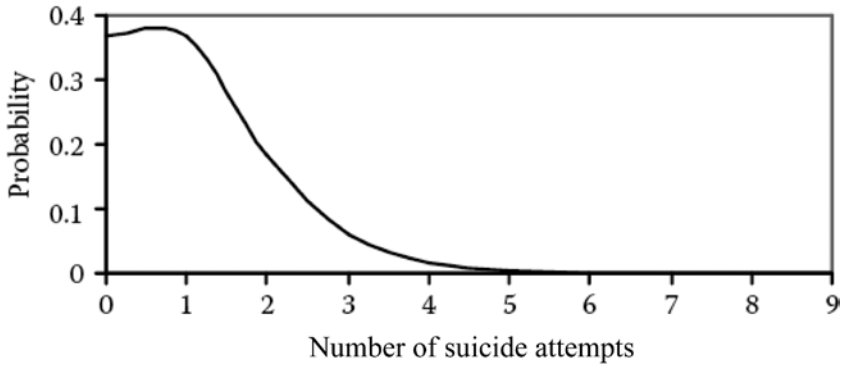

Figure 2.2 Illustration of Poisson distribution when $\lambda = 1.0$

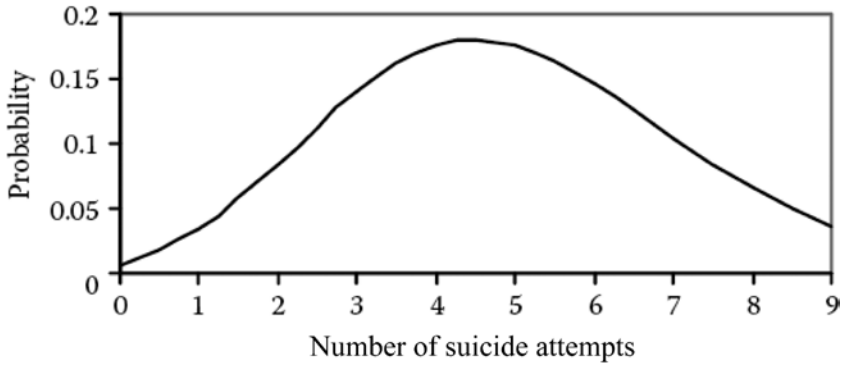

Figure 2.3 Illustration of Poisson distribution when $\lambda = 5.0$

The mean and variance of the Poisson distribution are both expressed by $\mu = \sigma^2 = \lambda$. This implies that as the average number of times an event is expected to occur gets larger, so does the expected variability of that event occurring, which also makes sense intuitively. However, with real count data the variance often exceeds the mean, a phenomenon known as **overdispersion.**

2.8 Summary

In this chapter, we discussed various probability distributions that are often used to model discrete random variables. The distribution that is most appropriate for a given situation depends on the random process that is modeled and the parameters that are needed or available in that situation. A summary of the distributions we discussed is provided in Table 2.3.

In the next chapter, we make use of some of these distributions for inferential procedures; specifically, we will discuss how to estimate and test hypotheses about a population proportion based on the information obtained from a random sample.

Table 2.3 Summary of common probability distributions for categorical variables

Distribution	*Process modeled*	*Parameters*
Hypergeometric	Probability of k successes in n dependent trials	m (total number of successes available) n (total number of selections) N (total number available)
Bernoulli	Probability of a success in one trial	$\pi = m / N$ (probability of success)
Binomial	Probability of k successes in n independent trials	n (total number of trials) π (probability of success)
Multinomial	Probability of k_i successes in each of I categories in n independent trials	n (total number of trials) $\pi_1, \pi_2, \ldots \pi_I$ (probability of success in each of I categories)
Poisson	Probability of k successes in a fixed time interval	λ (number of successes expected to occur in a fixed interval)

Problems

2.1 A divorce lawyer must choose 5 out of 25 people to sit on the jury that is to help decide how much alimony should be paid to his client, the ex-husband of a wealthy business woman. As luck would have it, 12 of the possible candidates are very bitter about having to pay alimony to their ex-spouses. If the lawyer were to choose jury members at random, what is the probability that none of the five jury members he chooses are bitter about having to pay alimony?

2.2 At Learn More School, 15 of the 20 students in second grade are proficient in reading.

a. If the principal of the school were to randomly select two second-grade students to represent the school in a poetry reading contest, what is the probability that both of the students chosen will be proficient in reading?
b. What is the probability that only one of the two students selected will be proficient in reading?
c. If two students are selected, what is the expected number of students that are proficient in reading?

2.3 Suppose there are 48 Republican senators and 52 Democrat senators in the United States Senate and the president of the United States must appoint a special committee of 6 senators to study the issues related to poverty in the United States. If the special committee is appointed by randomly selecting senators, what is the probability that half of the committee consists of Republican senators and half of the committee consists of Democrat senators?

2.4 The CEO of a toy company would like to hire a vice president of sales and marketing. Only 2 of the 5 qualified applicants are female, and the CEO would really like to hire a female VP if at all possible to increase the diversity of his administrative cabinet. If he randomly chooses an applicant from the pool, what is the probability that the applicant chosen will be a female?

2.5 Suppose that the principal of Learn More School from Problem 2.2 is only able to choose one second-grade student to represent the school in a poetry contest. If she randomly selects a student, what is the probability that the student will be proficient in reading?

2.6 Researchers at the Food Institute have determined that 67% of women tend to crave sweets over other alternatives. If 10 women are randomly sampled from across the

country, what is the probability that only 3 of the women sampled will report craving sweets over other alternatives?

2.7 For a multiple-choice test item with four response options, the probability of obtaining the correct answer by simply guessing is 0.25. If a student simply guessed on all 20 items in a multiple-choice test:

a. What is the probability that the student would obtain the correct answers to 15 of the 20 items?
b. What is the expected number of items the student would answer correctly?

2.8 The probability that an entering college freshman will obtain his or her degree in four years is 0.4. What is the probability that at least one out of five admitted freshmen will graduate in four years?

2.9 An owner of a boutique store knows that 45% of the customers who enter her store will make purchases that total less than $200, 15% of the customers will make purchases that total more than $200, and 40% of the customers will simply be browsing. If five customers enter her store on a particular afternoon, what is the probability that exactly two customers will make a purchase that totals less than $200 and exactly one customer will make a purchase that totals more than $200?

2.10 On average, 10 people enter a particular bookstore every 5 minutes.

a. What is the probability that only four people enter the bookstore in a 5-minute interval?
b. What is the probability that eight people enter the bookstore in a 5-minute interval?

2.11 Telephone calls are received by a college switchboard at the rate of four calls every 3 minutes. What is the probability of obtaining five calls in a 3-minute interval?

2.12 Provide a substantive illustration of a situation that would require the use of each of the five probability distributions described in this chapter.