STATISTICS FOR NURSING

THIRD EDITION

A Practical Approach

ELIZABETH HEAVEY

STATISTICS FOR NURSING

THIRD EDITION

A Practical Approach

ELIZABETH HEAVEY, PhD, RN, CNM
Professor of Nursing
SUNY College at Brockport
Brockport, New York

JONES & BARTLETT
LEARNING

World Headquarters
Jones & Bartlett Learning
5 Wall Street
Burlington, MA 01803
978-443-5000
info@jblearning.com
www.jblearning.com

Jones & Bartlett Learning books and products are available through most bookstores and online booksellers. To contact Jones & Bartlett Learning directly, call 800-832-0034, fax 978-443-8000, or visit our website, www.jblearning.com.

Substantial discounts on bulk quantities of Jones & Bartlett Learning publications are available to corporations, professional associations, and other qualified organizations. For details and specific discount information, contact the special sales department at Jones & Bartlett Learning via the above contact information or send an email to specialsales@jblearning.com.

14399-7

Production Credits
VP, Product Management: David D. Cella
Director of Product Management: Amanda Martin
Product Manager: Rebecca Stephenson
Product Assistant: Christina Freitas
Senior Vendor Manager: Sara Kelly
Senior Marketing Manager: Jennifer Scherzay
Product Fulfillment Manager: Wendy Kilborn
Composition and Project Management: S4Carlisle Publishing Services
Cover Design: Kristin E. Parker
Rights & Media Specialist: Wes DeShano
Media Development Editor: Troy Liston
Cover Image (Title Page, Part Opener, Chapter Opener): © sinemaslow/iStock/Getty Images Plus/Getty
Printing and Binding: Edwards Brothers Malloy
Cover Printing: Edwards Brothers Malloy

Library of Congress Cataloging-in-Publication Data
Names: Heavey, Elizabeth, author.
Title: Statistics for nursing : a practical approach / Elizabeth Heavey, Ph.D., R.N., C.N.M., Professor of Nursing, SUNY College at Brockport.
Description: Third edition. | Burlington, MA : Jones & Bartlett Learning, [2019] | Includes index.
Identifiers: LCCN 2017054187 | ISBN 9781284142013 (paperback)
Subjects: LCSH: Nursing--Statistical methods.
Classification: LCC RT68 .H43 2019 | DDC 610.73072/7--dc23 LC record available at https://lccn.loc.gov/2017054187

6048

Printed in the United States of America
22 21 20 19 18 10 9 8 7 6 5 4 3 2 1

Dedication

This book is dedicated to my RN to BSN students, who remind me every day how much effort, persistence, and determination it takes to return to school while balancing work, family, and professional responsibilities. Your feedback and willingness to challenge yourselves, despite the many obligations and responsibilities you have in your lives, inspires me. I have watched so many of you arrive in class tired, very unsure, anxious, and stressed about taking statistics, and yet you persevere. You put your best foot forward, tentatively realizing that, though it isn't easy, you have the capacity to master this content and reach your goals. Sometimes you stumble, but you get back up and try again and because of that, I have watched you accomplish so much in all avenues of your lives. It makes me so proud to watch you figure out difficult content, develop understanding of how this really will have an impact on your patient care, and grow as nurses and individuals. Nothing makes me happier than to be part of your success, to watch you walk across the stage at graduation, ready for the next challenge in your educational and professional path. I watch your families, so proud of all that you are and all that you do. I watch your children realize they too can dream big and succeed because they have watched you do just that. It is a joy and privilege to work with each of you, and just like you, I keep trying to grow and improve. So here is my third try at this book, and I hope you find that listening to my students has helped make this edition even better—because you are the reason it exists in the first place.

Beth

CONTENTS

INTRODUCTION ix
ACKNOWLEDGMENTS xi

CHAPTER 1: INTRODUCTION TO STATISTICS AND LEVELS OF MEASUREMENT 1

Introduction 2
Population Versus Sample 4
Quantitative Versus Qualitative 4
Independent Versus Dependent Variables 5
Continuous Versus Categorical Variables 6
Levels of Measurement 6
Summary 11
Review Questions 12

CHAPTER 2: PRESENTING DATA 19

Frequency Distributions 21
Percentages 22
Bar Charts 23
Histograms 28
Line Graphs 29
Scatterplots 29
Box and Whiskers Plot 30
Summary 31
Review Questions 31

CHAPTER 3: DESCRIPTIVE STATISTICS, PROBABILITY, AND MEASURES OF CENTRAL TENDENCY 41

Descriptive Statistics: Properties of Variables 43
Measures of Central Tendency 43
Range and Sample Standard Deviation 44
Calculating the Standard Deviation 45
Using a Box and Whiskers Plot to Display Central Tendency and Range 47

Moving Forward: Inferential Statistics 51
Frequency Distributions Versus Probability Distributions 51
The Normal Distribution 54
Skewed Distributions 56
Summary 57
Review Questions 58

CHAPTER 4: MEASURING DATA 63

Feasibility 65
Validity 65
Reliability 66
Screening Tests 68
Sensitivity 69
Specificity 69
Positive Predictive Value of a Screen 71
Negative Predictive Value 73
Efficiency 74
Summary 74
Review Questions 75

CHAPTER 5: SAMPLING METHODS 83

Sampling Methods 84
Probability Sampling 85
Sampling Error Versus Sampling Bias 86
Sampling Distributions 87
Nonprobability Sampling 90
Inclusion and Exclusion Criteria 92
Sample Size 92
Summary 93
Review Questions 93

CHAPTER 6: GENERATING THE RESEARCH IDEA 99

Hypothesis Testing 100
Statistical Significance 102
Statistical Significance Versus Clinical Significance 104
How Does the Test Statistic Compare to the Null Hypothesis? 106
Applying the Decision Rule 107
Test Statistics and Corresponding *p*-Values 107
Summary 107
Review Questions 108

Chapter 7: Sample Size, Effect Size, and Power 113

Effect Size 114
Type Two Error 117
A Quick Review of Type One and Type Two Errors 118
Sample Size 119
Summary 121
Review Questions 121

Chapter 8: Chi-Square 127

Chi-Square (X^2) Test 128
The Null and Alternative Hypotheses 128
2 × 2 Table 128
Degrees of Freedom 129
Statistical Significance 129
Direction of the Relationship 129
When Not to Use Chi-Square: Assumptions and Special Cases 132
Summary 135
Review Questions 136

Chapter 9: Student *t*-Test 145

The Student *t*-Test 146
The Null and Alternative Hypotheses 152
Statistical Significance 152
Degrees of Freedom for Student *t*-Tests 152
Summary 160
Review Questions 161

Chapter 10: Analysis of Variance (ANOVA) 171

Comparing More Than Two Samples 172
The Null and Alternative Hypotheses 172
Degrees of Freedom 174
Statistical Significance 174
Appropriate Use of ANOVA 175
Repeat-Measures ANOVA 176
Summary 181
Review Questions 181

Chapter 11: Correlation Coefficients 191

Looking for a Relationship in One Sample 192
The Null and Alternative Hypotheses 192

Selecting the Best Correlation Test to Use 193
Direction of the Relationship 193
Sample Size 194
Strength of the Relationship 194
Statistical Significance 194
Appropriate Use of Correlation Coefficients 195
More Uses for Pearson's *r* 195
Summary 198
Review Questions 200

CHAPTER 12: REGRESSION ANALYSIS 207

Quantifying an Association 208
Summary 239
Review Questions 239

CHAPTER 13: RELATIVE RISK, ODDS RATIO, AND ATTRIBUTABLE RISK 249

Epidemiology 251
Study Designs Used in Epidemiology 251
Attributable Risk 265
Summary 266
Review Questions 267

APPENDIX A: TABLES FOR REFERENCE 277

APPENDIX B: WORKING WITH SMALL SAMPLES 287

REFERENCES 291

EPILOGUE 295

INDEX 297

Introduction

When the first edition of *Statistics for Nursing: A Practical Approach* came out, I was very happy to hear from many nurses about how useful it was in making statistics accessible for anyone just beginning to work with these concepts. They also had some very helpful suggestions, much like my own students, who provided the motivation and feedback that helped create the first edition of the book. I also heard from quite a few of you in DNP programs that the book was helpful to get you started as well. Wonderful! I am thrilled to be a part of nurses going on into advanced practice!

In this third edition, I have again acted on the feedback from students, and I have included even more practice questions at the end of each chapter, giving you more opportunities to practice, practice, practice. A whole new test bank provides instructors the opportunity to use the older questions as practice quizzes, which also provides students with additional feedback and practice. I have also provided new research article reviews with practice questions within the analysis chapters of the text. Teaching from the text myself has given me the opportunity to identify areas where students needed additional support, so I have added content like decision trees and tables showing the different tests and how to differentiate which one is appropriate. However, I have stayed true to the original premise of the text, which is that all of this is at an introductory level, without a lot of ancillary information to confuse you.

If you are teaching from this text at an undergraduate level, it is perfectly appropriate to skip the regression chapter; the rest of the content from the book will still work fine. You can also include it if it is appropriate for your students or course. As with the previous edition, the

"From the Statistician" features examine some of the chapter concepts in greater detail. A new "From the Statistician" feature in Chapter 13 provides additional background about confidence intervals. These features are set apart from the rest of the text and are available for students who prefer a more mathematical approach or want to have a better understanding of "why." Students who want to stick with the clinically applied information can skip these sections without experiencing problems in understanding the essential content.

The third edition of *Statistics for Nursing* also includes updated recorded lectures with closed captioning available and computer application updates. I have started using skeletal notes when I teach this class, and I provide these for my students to use in each chapter; these notes are included for all students as well. They provide an outline for note taking and helpful graphics to fill in, thus promoting student engagement with the content rather than just their passive printing of PowerPoint slides.

I would love to hear from any of you who use the new content and supports in the third edition of this text. What are your thoughts? Did these new resources help you? Do you have any other ideas for useful learning tools? Send me a quick email and help me make this material even better.

I hope you find the third edition of *Statistics for Nursing* helpful and that you continue on your quest to becoming a nurse who understands statistics! You never know where it may take you someday. I certainly didn't!

All the best,
Beth

Acknowledgments

This book is the product of the combined effort of many individuals who were gracious enough to contribute their time, knowledge, and effort.

Brendan Heavey is the contributing author for all of the “From the Statistician” features in the text and has developed the computer application content for Excel. Brendan has a statistical knowledge significantly beyond my own and spent many hours writing, rewriting, and explaining concepts to make sure my simplified explanations were technically correct. I am ever grateful not only for his statistical contributions to the text but also for his interest in and support of the project from the early proposal days. He is an incredibly gifted human being. I am proud to call him my brother.

Dr. Renee Biedlingmaier, a former star student and now valued colleague, shares her experience working with small samples in the new Appendix B for this book. Several of the DNP instructors asked for a brief overview of this challenging topic. Her expertise and willingness to share it with our students enhance the book, and her contribution is greatly appreciated.

The RN to BSN faculty members at SUNY Brockport believe this content is essential to the knowledge and future success of nurses, and they have supported the inclusion of the course for all RN to BSN students in our program. Thank you! My colleagues, Dr. Biedlingmaier and Professor Bingham, have been graciously willing to teach the class with me and have helped solidify student support content and data collection for evaluation of new material.

Ms. Shelby Brown, one of our traditional nursing students, dedicated many hours to helping me input content and complete computer work efficiently. Her effort and skill are greatly appreciated!

Many other undergraduate and graduate students emailed me from afar with feedback and thoughts about the book. I always enjoy hearing how this book has had an impact on your understanding and your career, as well as your suggestions for improvement. Thank you for taking the time to share these ideas with me.

Thank you also to all the instructors who are using the book and letting me know how well it is working in your classrooms. You inspire and encourage me with all of your great ideas and dedication to student learning.

Thank you to the publishing team at Jones & Bartlett Learning, who saw the potential in the first edition before I did and helped make it happen, and then came back for more!

As always, my heartfelt gratitude goes to my family and friends, who loved and supported me throughout this project. I would not be where I am today without all of you. Thank you for all the hours spent watching swim meets, hosting sleepovers, and climbing rock walls and kayaking on camping trips; the homecooked meals, quiet hugs, and heartfelt phone calls; and the belief in me no matter what crazy plan I come up with next. I will always be grateful for each of you.

And to my children, Gabrielle and Nathaniel, you are the reason behind it all, why every day matters and making the most of it counts. Watching you grow into the young people that you are has been a journey that has involved very little sleep but has brought me tremendous joy. You have made me more humble, reflective, forgiving, and determined to have an impact on the world you will inherit from our generation. Being your mother puts the meaning in everything that I do. I love you to the moon and the stars, to infinity and beyond and back again, forever and ever.

Thank you all,
Beth

CHAPTER 1

Introduction to Statistics and Levels of Measurement

How to figure things out.

OBJECTIVES

By the end of this chapter students will be able to:

- State the question that statistics is always trying to answer.
- Define the empirical method.
- Compare quantitative and qualitative variables.
- Differentiate a population from a sample and a statistic from a parameter, giving an example of each.
- Explain the difference between an independent and a dependent variable, citing examples of each.
- Identify continuous and categorical variables accurately.
- Distinguish the four levels of measurement, and describe each.
- Apply several beginning-level statistical techniques to further develop understanding of the concepts discussed in this chapter.

KEY TERMS

Categorical variable
A variable that has a finite number of classification groups or categories, which are usually qualitative in nature.

Continuous variable
A variable that has an infinite number of potential values, with the value being measured falling somewhere on a continuum containing in-between values.

Dependent variable
The outcome variable or final result.

Empirical method
A way of gathering information through systematic observation and experimentation.

Estimate
A preliminary approximation.

Independent variable
A variable measured or controlled by the experimenter; the variable that is thought to affect the outcome.

Interval data
Data whose categories are exhaustive, exclusive, and rank-ordered, with equally spaced intervals.

Nominal data
Data that indicates a difference only, with categories that are exhaustive and exclusive but not rank-ordered.

Ordinal data
Data whose categories are exhaustive, exclusive, and rank-ordered.

Parameter
Descriptive result for the whole group.

Population
The whole group.

Probability
The likelihood that an outcome will occur.

Qualitative measure
A measure that describes or characterizes an attribute.

Quantitative measure
A measure that reflects a numeric amount.

Ratio data
Data whose categories are exhaustive, exclusive, and rank-ordered with equally spaced intervals and a point at which the variable does not exist.

Sample
A group selected from the population.

Statistic
An estimate derived from a sample.

Variable
The changing characteristic being measured.

INTRODUCTION

So here you are. You've worked hard, you are in nursing school, and you are ready to begin. But wait! Why do you have to take statistics? Why do you need to understand all those numbers and equations when you are a nurse and want to help people?

Most nursing students experience a mild sense of panic when they discover they have to take statistics—or any other kind of math, for that matter. That reaction is common. Here is a calming thought to remember: You already practice statistics, but you just may not know it.

Statistics boils down to doing two things:

- Looking at data
- Applying tests to find out either (1) that what you observe is what you expected or (2) that your observation differs enough from what you expected that you need to change your expectations

You might be convinced that you don't use statistics in your life, so let me give you an example. New York State, where I live, has four seasons. The summer is usually June, July, and August. Fall is September, October, and November. Winter is December, January, and February. And that leaves March, April, and May for the spring. If you walk outside in July and find it to be 80° and humid, you would draw an unspoken conclusion that what you just observed is what you were expecting, and you would put on your sunglasses. However, what if you walk outside in January and find it to be 80° and humid? You would probably be startled, take off your overcoat and boots, and read about global warming. The difference between the weather you expect in winter and what you actually encounter is so different that you might need to change your expectations. You are already practicing statistics without knowing it!

Of course, that day in January might just be a fluke occurrence (a random event), and the temperature could be below freezing again the next day. That is why we need to use the **empirical method**, otherwise known as systematic observation and experimentation. The empirical method allows you to determine whether the temperature observed is *consistently* different from what you expect. To use the empirical method, you need to check the daily temperature on more than one day. So you might decide to monitor the daily temperature for a whole month of winter to see whether readings are consistently different from what you expect in the winter months. In this scenario, you would be using the empirical method to practice statistics (see **Figure 1-1**).

FIGURE 1-1 **Long-Range Winter Forecast for 2017.**

Courtesy of Farmers' Almanac.

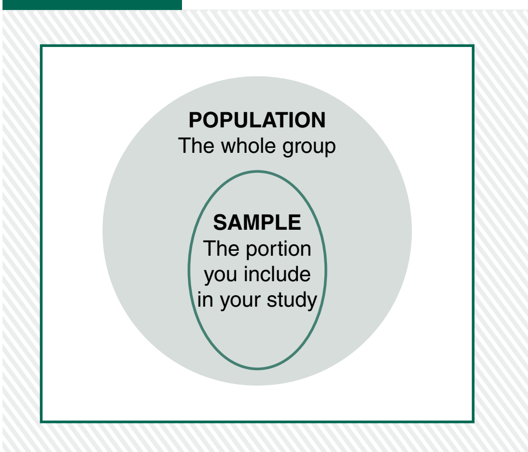
POPULATION VERSUS SAMPLE

To answer questions in research, we need to set up a study of the concepts we're interested in and define multiple **variables**, that is, the changing characteristics being measured. In our example, the temperature is a variable, a measured characteristic. Each variable has an associated **probability** for each of its possible outcomes, that is, how likely it is the outcome will occur. For example, how likely is it that the temperature will be below freezing as opposed to being in the eighties in winter? In your study, you recorded the daily temperature for a winter month, and those readings make up a **sample** of all the daily temperatures in the months of winter. The manner in which you collect your sample is dependent on the purpose of your study (**Figure 1-2**).

A sample is always a subset of a **population**, or an overall group (sometimes referred to as the reference population). In this case, our population includes all the daily temperatures in the winter months, and the subset, or sample, is all the daily temperatures recorded during your month of data collection. If you calculate the average temperature based on this sample data, you create what is called a **statistic**, which is an **estimate** generated from a sample.

A measured characteristic of a population is called a **parameter**. In our example, if you measured the daily temperature for December, January, and February and then calculated the average temperature, you would be determining a parameter. A really good way to remember the relationships among these four terms is with the following analogy: Statistic is to sample as parameter is to population.

FIGURE 1-2 **Population vs. Sample.**



QUANTITATIVE VERSUS QUALITATIVE

While you are collecting the weather data, you may realize that the data can be recorded in several ways. You could write down the actual temperature on that day, which would be a **quantitative measurement**, or you could describe the day as "warm" or "cold," which would be a **qualitative measurement**. A numeric amount or measure is associated with quantitative measurement (such as 80°F), and qualitative measures describe or characterize things (such as, "So darn cold I can't feel my toes").

Be careful with this difference: You can easily get confused. Qualitative variables do not contain quantity information, even if numbers are assigned. The assigned numbers have no quantitative information, rank, or distance. For example, a survey question asks, "What color scrubs are you wearing?" and lists choices numbered 1 to 3. Even if you selected choice 2, neon orange, you do not necessarily have any more scrubs than someone who chooses 1, lime green (although both respondents may want to purchase new scrubs). Even though these qualitative variables have numbers assigned to them, the numbers simply help with coding. The variables are still qualitative.

Independent versus Dependent Variables

Being as inquisitive as you are, you have probably asked yourself a number of times about a relationship you observe in your patients. For example, you notice that many supportive family members visit Sally Smith after her hip replacement recovery and that she is discharged 3 days after her surgery. Joanne Jones, on the other hand, has no visitors during her hip replacement recovery and is not discharged until day 6. As an observant nurse researcher, you have been wondering how variable *x* (the **independent variable**, which is measured or controlled by the experimenter) affects variable *y* (the **dependent variable**, or outcome variable) (**Figure 1-3**). You wonder, does having family support (the independent variable) affect the duration of a hospital stay (the dependent, or outcome, variable)?

To answer this question, you create a study. Obviously, other factors might be involved as well, but in your experiment, you are interested in how family support, the independent variable, affects hospital stay, the dependent variable. If you are correct, then the duration of the hospital stay *depends* on family support. The independent variable can be a suspected causative agent, and the dependent variable is the measured outcome or effect (**Figure 1-4**).

FIGURE 1-3 **Relationship of Independent and Dependent Variables.**

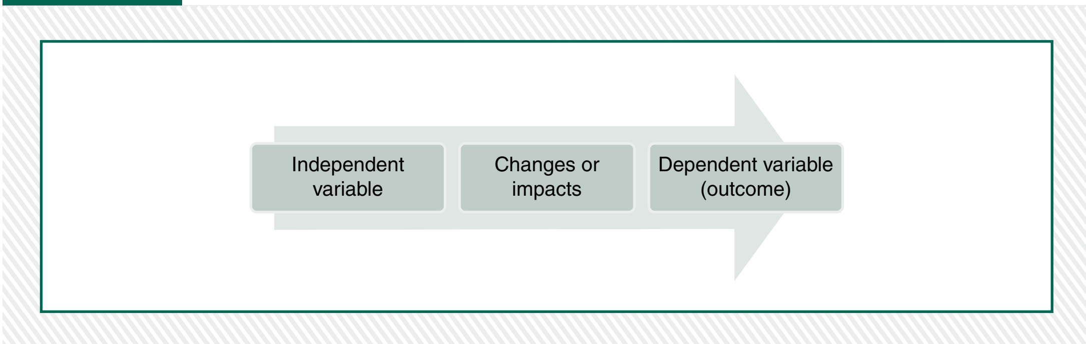
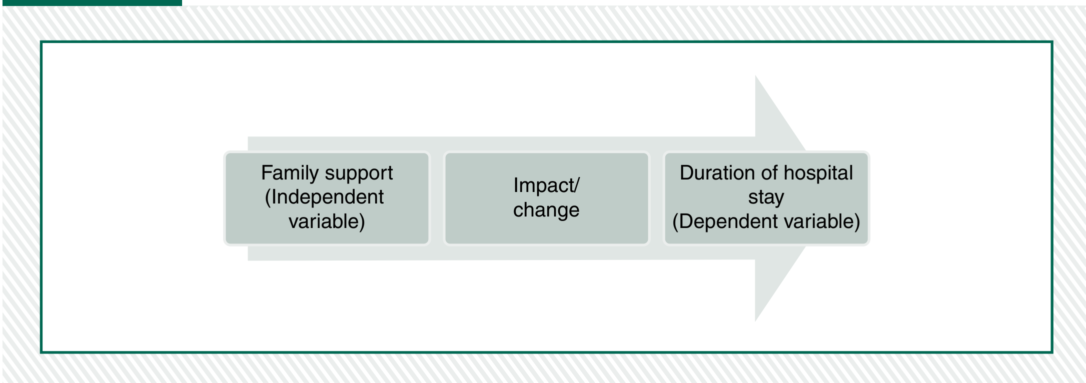


FIGURE 1-4 **Does Family Support Affect the Duration of a Subject's Hospital Stay?**



Note: Additional criteria must be met to say that a variable is causative, so I refer here only to the "suspected" causative agent.

Continuous versus Categorical Variables

Some data have an infinite number of potential values, and the value you measure falls somewhere on a continuum containing in-between values. These values are called **continuous variables**. As a nurse, when you measure your patient's temperature, you are measuring a continuous variable. The reading could be 98° or 98.6° or 98.66666°. The infinite possibilities are all quantitative in nature. Actually, the only limit to the measurement is the accuracy of the measuring device. For example, if you have a thermometer that measures only in whole degrees, you will not have as much information as you would using a thermometer that measures to the one-thousandth of a degree.

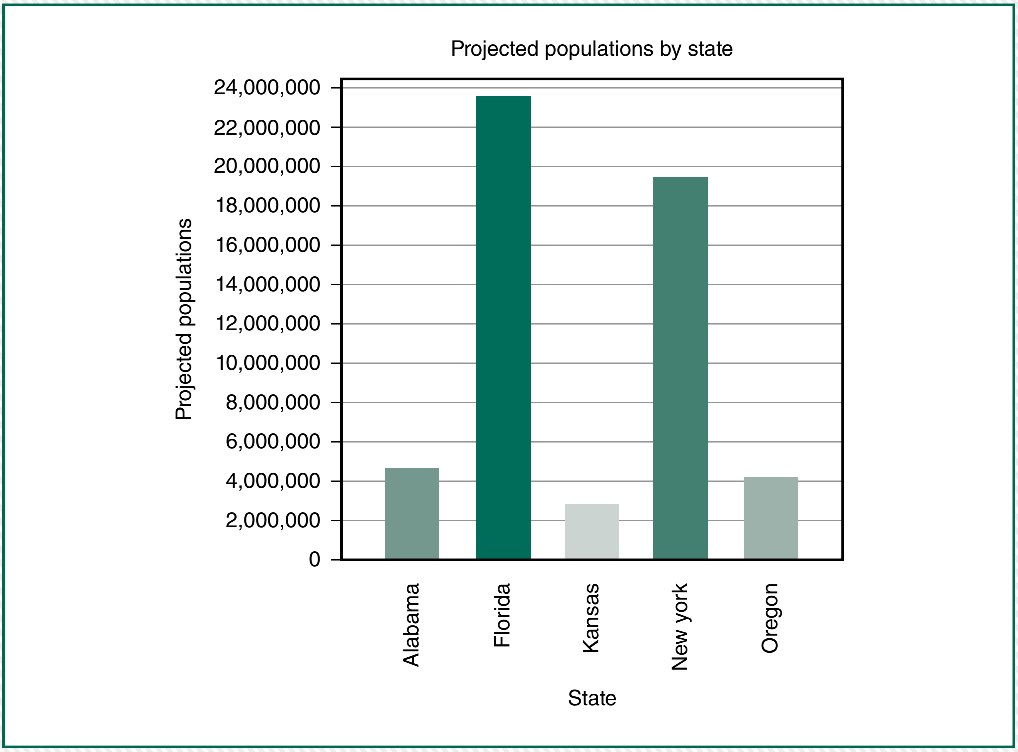
Continuous variables can be contrasted with **categorical variables**, sometimes called discrete variables, which have a finite number of classification groups, or categories, that are usually qualitative in nature. For example, as part of your research you may need to collect information about your patients' racial background. The choices available are African American, Native American, Caucasian, Asian, Latino, mixed race, and other. Race is an example of a categorical variable, a measurement that is restricted to a specific value and does not have any fractional or in-between values. When you read a study, the demographic information about the sample involved usually contains quite a few categorical variables including marital status, gender, race, geographic region, educational level, language spoken, smoking status, and so on.

Let's look at an example where we can see both types of variables in a study. If you were reading a public health study examining statewide variation in population estimates you might have the information in **Figure 1-5** available. Your sample was collected and reported about five states, so the state becomes one of the demographic variables you will want to report. Note that "state" is a categorical/qualitative variable: It just tells you the location of the sample subject and does not include any quantitative information. You also record the state population, which is a continuous/quantitative variable where the value can fall anywhere within the range of population values.

Levels of Measurement

Let's say that your interest in the relationship between family support (the independent variable) and duration of stay (the dependent variable) is extensive enough that you apply for a program at your hospital that includes a small research fellowship. You win the fellowship and proceed to collect data about each patient admitted to your orthopedic unit for hip replacement over a 3-month period. The study protocol calls for you to complete the usual admission forms and then for patients to complete a short survey about perceived family support. After your institutional review board approves your

FIGURE 1-5 **Projected Populations by State.**



Reproduced from CDC. (n.d). Population Projections, United States, 2004–2030, by state, age and sex. CDC WONDER Online Database, September 2005. Accessed at http://wonder.cdc.gov/population-projections.html on Apr 4, 2017.

study, you begin. The level of measurement of your data determines what type of analysis you are able to perform in your study, so let's look at the different types and what makes each level unique.

Your first survey question asks the patient's gender (male, female, other). The data you gather for this question is an example of **nominal data**; it simply indicates a difference between the three answers. One is neither greater than nor less than the other, and they are not in any particular order. Also, the categories are exclusive and exhaustive; that is, the patient cannot answer "both" or "neither." Asking about

FROM THE STATISTICIAN *Brendan Heavey*

What Is a Statistic?

As a student of statistics, you will run into questions regarding parameters and statistics all the time. Determining the difference between the two can be difficult. To get a concrete idea of the difference, let's look at an example. According to the Bureau of Labor Statistics, registered nurses constitute the largest healthcare occupation, with 2.7 million jobs nationwide. Because this text is primarily designed for nursing students, let's use this number for our example.

Let's say that you are a consultant working for a fledgling company that is planning to make scrubs for nurses. Let's call this company Carol's Nursing Scrubs, Inc. Scrubs at Carol's come in small, medium, and large. The company offers all kinds of styles and prints, but the underlying sizes are intended to remain the same. Carol just received her first bit of seed money to mass-produce 20,000 pairs of scrubs. Carol, an overly demanding boss, wants the medium-size scrubs to fit as many nurses nationwide as possible. To make that happen, she needs to know the average height and weight of nurses nationwide, so she has instructed you to conduct a nationwide poll. She thinks you should ask every nurse in the country his or her height and weight and then calculate the average of all the numbers you get.

Now, you are an intelligent, well-grounded employee who's in demand everywhere and working for Carol only because her health plan comes with a sweet gym membership and you get a company car. You realize it would be pretty difficult to set up a nationwide poll and ask all the nurses in the country for their height and weight. Even if you tried a mass mailing, the data returned to you would be filled with so many incompletes and errors that it wouldn't be trustworthy.

So what are you to do? Your first instinct might be to respond to your boss by saying, "Geez, Carol, that's so absurd and impossible I don't even know where I'd start," and then finish your day on the golf range. After this course, however, you'll be not only a nurse but a nurse with some training in statistics. You'll be able to deal with this situation more effectively.

Jenna the Statistical Nursing Guru (you):	Carol, I recommend we take a few *samples* of nurses nationwide and *survey* them rather than attempting to contact every nurse in the country. Then we could *estimate* the true average height and weight based on our samples.
Carol:	How would that work, Jenna?
Jenna:	Well, I'd go down to the University Hospital and poll 30 RNs on their height and weight. Then I'd go to the next state and do the same. My third and final sample would contain 30 RNs from a hospital in Springfield. I'd calculate the average from my total *sample* (90 RNs), which is a *statistic*, and use that to estimate the overall average in the United States, which is a *parameter* of the total *population*. You see, Carol, anytime you calculate an estimate with data from a sample or list the data from the sample itself, you calculate a statistic. If you calculate an estimate from data in an entire population, you're calculating a parameter.

the patient's marital status (married, divorced, separated, living together, and other) is another example of nominal data.

Your next survey question asks the patient to rate his or her family support level as low, medium, or high. This question is an example of **ordinal data**. Ordinal data must be exhaustive and exclusive, just like nominal data, but the answers are also rank-ordered. With rank-ordered data, each observation/category is higher or lower, or better or worse, than another, but you do not know the level of difference between the observations/categories. In this example, a high level of family support indicates a greater quantity of the variable in question than does a moderate or low level of family support

A routine part of admitting each patient also includes a baseline set of vital signs, which you want to include in your survey data. One of the vital signs you check is each subject's temperature. Temperature is an example of **interval data**, which is exhaustive, exclusive, and rank-ordered, and has numerically equal intervals. In this example, the interval is a degree of Fahrenheit. You may also decide you want to look at your survey data by age group and develop the table shown in **Table 1-1**. In this example, age group is interval data; it is exhaustive, exclusive, and rank-ordered; and each interval is 4 years, so the intervals are all equal.

TABLE 1-1 Number of Patients Surveyed in Each Age Group

Age Group	Number of Patients
40–44 years	4
45–49 years	22
50–54 years	48
55–59 years	84

After recording each patient's temperature, you go on to examine each patient's blood pressure. Blood pressure is an example of **ratio data**, which is exhaustive, exclusive, and rank-ordered, with equal intervals *and* a point at which the variable is absent. (If the blood pressure reading is "absent" in any of your patients, you need to begin CPR!)

If you look at the diagram in **Figure 1-6**, you will see the relationship between the levels of measurement. Each increase in level includes the factors of the previous level, plus it adds another qualifier. Thus, if a variable is at the ratio level, it meets all the criteria for the nominal, ordinal, and interval levels, plus there is a point where it does not exist.

Ratio data is the highest level of measurement you can collect and gives you the greatest number of options for data analysis, but not all variables can be measured at this level. As a general rule of thumb, always collect the highest-level data you can for all your variables, especially your dependent variable. In your study of how family support (the independent variable) affects the duration of hospital stay (the dependent variable), you could have measured the length of hospital stay as short, medium, or long (ordinal) or in the number of actual days (the interval/ratio level). Obviously, the actual number of days gives you a higher level of measurement. A dependent variable with a higher level of measurement allows for a more robust data analysis. So collect the highest level you can! (See **Figure 1-7**.)

Note: Ordinal data may be quantitative (age group) or qualitative (mild/moderate/severe).

FIGURE 1-6 **Relationship between the Levels of Measurement.**

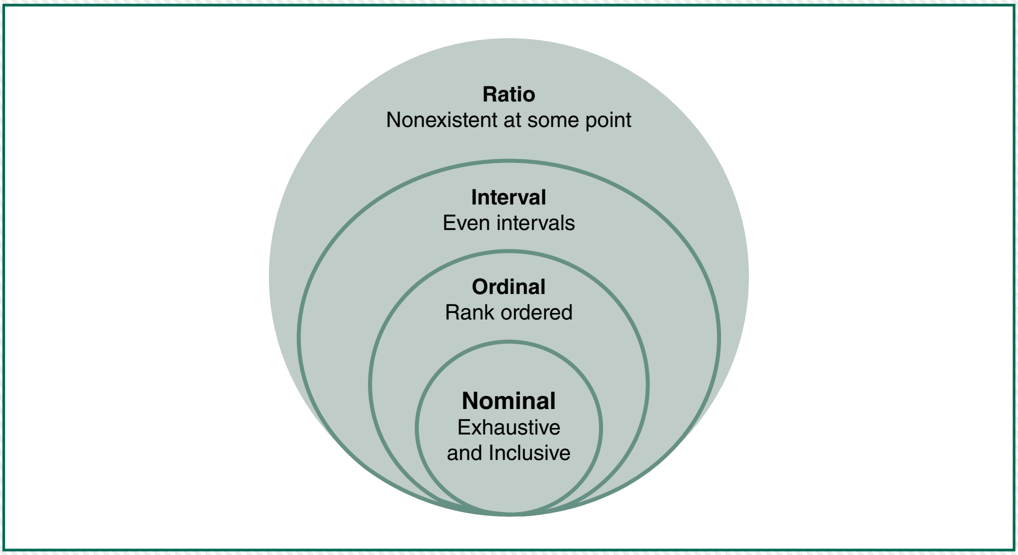
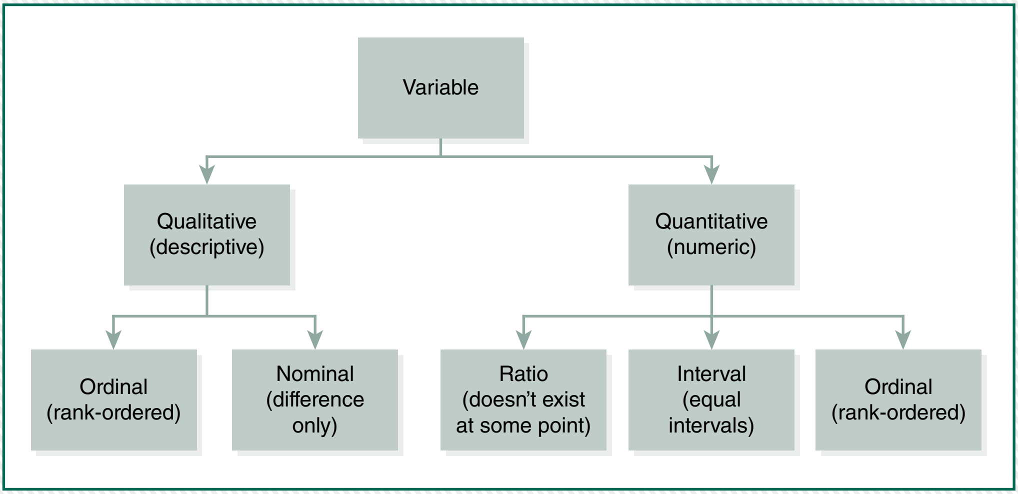


FIGURE 1-7 **The Relationship between Variable Descriptions.**



THINK IT THROUGH

How Can I Determine the Level of Measurement of a Variable?

Your study examines placental weight, which is measured in grams. What level of measurement is this variable?

Nominal Level: Ask yourself, does the variable show a difference? Yes, it does, different scores indicate different placental weight. This variable is at least at the nominal level.

If your answer is yes, then go to the next step because the variable may be at a higher level.

Ordinal Level: Is the difference rank-ordered? Yes, a lower score means less placental weight, and a higher score means more placental weight. Your variable is at least at an ordinal level.

If your answer to this question is no, then you do not have the criteria for this level and should identify your variable as the level before, in this case, nominal. If your answer is yes, then go to the next step.

Interval: Does this variable have equal intervals? Yes, each gram is an equal interval. Then your variable is at least interval level.

If your answer to this question is no, then you do not have the criteria for this level and should identify your variable as the level before, in this case, ordinal. If your answer is yes, then go to the next step.

Ratio: Is there a point where the variable can be equal to zero? No, every placenta will have at least some level of mass to it, so the variable will never be equal to zero.

Because your answer to this question is no, you do not have the criteria for this level and should identify your variable as the level before, in this case, interval. If your answer is yes, then you have satisfied the criteria for the highest level of measurement, which is ratio.

Let's look at another example, using the same steps. Your study examines amniotic fluid volume (AFV) measured as minimal, adequate, and excessive. What level measurement is this variable?

Nominal Level: Ask yourself, does the variable show a difference? Yes, it does; different categories indicate different amounts of AFV. This variable is at least at the nominal level.

If your answer is yes, then go to the next step because the variable may be at a higher level.

Ordinal Level: Is the difference rank-ordered? Yes, a person with minimal AFV has less AFV than someone with excessive AFV. Then your variable is at least at the ordinal level.

If your answer to this question is no, then you do not have the criteria for this level and should identify your variable as the level before, in this case, nominal. If your answer is yes, then go to the next step.

Interval: Does this variable have equal intervals? No, we don't know the intervals for these categories, so we can't say they are even.

Because your answer to this question is no, you do not have the criteria for this level and should identify your variable as the level before, in this case, ordinal.

SUMMARY

Talk about exhausting, but you survived! So let's wrap it up here. Statistics really boils down to asking:

- Is what you observe what you expect?
- Or, using the empirical method, have you determined that what you observe is different enough from what you would expect that you need to change your expectations?

Using qualitative (descriptive) and quantitative (numeric) variables, you can assess the impact of independent variables on dependent (outcome) variables. Always collect the highest level of measurement possible, especially for your dependent variable. Doing so gives you the widest range of analysis options when you are ready to "crunch the numbers."

If you understand these concepts, you are ready to move on to the review exercises. If you are still struggling, don't despair. These concepts sometimes take a while to absorb. Read the review questions and then the chapter again, and slowly start to look at the review questions. You will get the hang of statistics; sometimes you just need practice. My students frequently look at me as though I am an alien when I tell them that by the end of the course this chapter will seem really simple. You may not believe it either. As you develop your understanding and apply these concepts, however, they will become clearer, and you too will look back in amazement. You are a statistical genius in the making!

CHAPTER 1 REVIEW QUESTIONS

1. A researcher asks hospitalized individuals about their comfort in a new type of hospital gown. This is an example of what type of data?
   a. ratio
   b. independent
   c. quantitative
   d. qualitative

2. If a researcher is examining how exposure to cigarette ads affects smoking behavior, the cigarette ads are what type of variable?
   a. qualitative
   b. quantitative
   c. dependent
   d. independent

3. A nurse practitioner measures how many times per minute a heart beats when an individual is at rest versus when running. She is measuring the heartbeat at what level of measurement?
   a. interval/ratio
   b. nominal
   c. independent
   d. ordinal

4. If a researcher is examining how exposure to cigarette ads affects smoking behavior, smoking behavior is what type of variable?
   a. ratio
   b. independent
   c. dependent
   d. nominal

5. The research nurse is coding adults according to size. A person with a below-average body mass index (BMI) is coded as 1, average is 2, and above average is 3. What level of measurement is this?
   a. nominal
   b. ratio
   c. ordinal
   d. interval

6. You are asked to design a study measuring how nutritional status is related to serum lead levels in children. You assess calcium and fat intake, as well as serum lead levels in a sample of 30 children who are 2 years old. Lead levels are measured in micrograms per deciliter (mcg/dL). One child had a lead level of 17 mcg/dL. This is an example of what type of variable?
   a. quantitative
   b. qualitative
   c. independent
   d. nominal

Questions 7–9: You are asked to design a study to examine the relationship between preoperative blood pressure and postoperative hematocrit.

7. What is your independent variable?

8. What is your dependent variable?

9. How will you measure each, and what level of measurement is this?

Questions 10–13: You are later asked to do a follow-up study to see whether requiring an intraoperative blood transfusion had an impact on postoperative rates of poor mental health, specifically depression.

10. What is your independent variable?

11. What is your dependent variable?

12. How will you measure these variables, and why?

13. Is your dependent variable measured at the highest level? If not, why not?

Questions 14–18: You decide to measure depression on the following scale: 1 = low, 2 = moderate, 3 = high.

14. What level of measurement is this?

15. How could this measure be improved?

16. Why might you want to improve it?

17. You decide to measure postoperative hematocrit by serum levels. Is this a quantitative or qualitative measurement?

18. You discover that all but those with the lowest hematocrits had higher levels of depression after their surgery and transfusion. Why might the group that had the most critical need for the transfusions not have the subsequent depression associated with this result in the rest of your sample?

Questions 19–25: Elevated serum lead levels in childhood are associated with lower IQ, hyperactivity, aggression, poor growth, diminished academic performance, increased delinquency, seizures, and even death. The neurological damage that occurs cannot be reversed, even once exposure is stopped.

19. You have been asked to follow up in your community and determine what outcomes are associated with lead exposure in children. List three dependent variables for your study and how you will measure them.

20. What level of measurement are your dependent variables? Are they continuous or categorical?

21. Can you increase the level of measurement for any of them?

22. If you are looking at what outcomes are associated with lead exposure in children, what is your independent variable?

23. Why might this independent variable be difficult to measure?

24. Describe how this independent variable could be measured quantitatively or qualitatively.

25. Which way do you prefer to measure the independent variable? Why?

Questions 26–34: A nurse researcher is assessing how well patients respond to two different dosing regimens of a new drug approved to treat diabetic neuropathy. Two different dosing regimens are administered, and side effects are monitored. Results are shown in **Table 1-2**.

TABLE 1-2 Self-Reported Side Effects of Two Randomized Groups of 100 Individuals Treated for Diabetic Neuropathy

Side Effect Reported	Low Dosage	High Dosage
Nausea	8	21
Headache	3	5
Weight gain	1	0
Weight loss	0	6
Lethargy	3	11
Skin rash	13	13

26. What is the independent variable?

27. What is the dependent variable?

28. In this study, the nurse researcher measures the side effects as present or not present. This variable is what level of measurement?

29. If instead the nurse researcher decided to measure weight gain in pounds, what level of measurement would it be? Would it be a continuous or categorical variable?

30. If the nurse researcher decided to measure nausea as present, limiting, or debilitating, what level of measurement would it be? Would nausea be a continuous or categorical variable?

31. If the nurse researcher measured nausea as the number of hours of nausea experienced in a day, what level of measurement would it be?

32. If the nurse researcher asked the subjects to describe their headache, would this be a quantitative or a qualitative variable?
    In the second phase of this study, the nurse researcher asks the study participants to report changes in signs and symptoms of their neuropathy. She determines that those on the low-dose regimen had a similar level of pain relief and improvement in mobility as those who took the high-dose drug regimen.

33. What is the dependent variable in the second phase of the study?

34. Considering the information you now know about the side effects and relief of neuropathy symptoms, what might you prefer as a patient? Why? What else might you want to know before making the decision?

Questions 35–38: Relate to the following content.

35. You complete a study in which you categorize the subject's blood pressure as normal, prehypertensive, high blood pressure stage 1, or high blood pressure stage 2 using the following criteria. What level of measurement is the stage of high blood pressure?
36. You are now interested in examining compliance with a DASH diet. You ask your subjects if they have or have not complied with the diet this week. Your dietary compliance variable is what level of measurement?
37. After meeting with your statistician, you measure compliance with the DASH diet on a scale of 1 to 7. For analysis purposes, the dietary compliance variable is now what level of measurement? Why might the statistician have recommended this change?
38. You conclude your study by examining how compliance with the DASH diet affects the stage of high blood pressure. What is your independent variable?
39. What is your dependent variable? Is it continuous or categorical?

Statistics: Is what you observe what you expected?

ANSWERS TO ODD-NUMBERED CHAPTER 1 REVIEW QUESTIONS

1. d
3. a
5. c
7. Preoperative blood pressure
9. Answers will vary: actual blood pressure ratio, lab-reported hematocrit ratio, and so on.
11. Depression
13. Answers will vary.
15. Use of interval data, such as Beck's depression scale
17. Quantitative
19. Answers will vary: including IQ, school enrollment, crime, pregnancy, hematocrit, learning disabilities, growth, hearing, and behavior.
21. Answers will vary.
23. Answers will vary: including "It requires a blood draw," "There are different testing mechanisms," "The level may change depending on when the exposure occurred and the time that has lapsed since then," "Levels may differ from fingersticks versus serum draws."
25. Answers will vary.
27. Side effects, nausea, headache, weight gain, weight loss, lethargy, skin rash
29. Ratio, continuous
31. Ratio
33. Signs and symptoms of neuropathy
35. Ordinal
37. Interval; because collecting the data at a higher level of measurement gives you more analysis options
39. Stage of high blood pressure, categorical

CHAPTER 2

Presenting Data

Will my audience be able to see what the data is saying?

OBJECTIVES

By the end of this chapter students will be able to:

- Describe a frequency distribution.
- Calculate the cumulative frequency and the cumulative percentages for a group of data.
- Identify situations in which a grouped frequency distribution is helpful.
- Develop a frequency distribution.
- Calculate a percentage.
- Identify the best visual representation for various types of data.
- Determine the percentile rank of an observation.

KEY TERMS

Bar chart
A chart that has the nominal variable on the horizontal axis and the frequency of the response on the vertical axis, with spaces between the bars on the horizontal axis.

Box and whisker plot
A diagram of the central value and variability seen in a data set, with a box containing the median and middle two quartiles of a data set and lines extending out to form "whiskers," which represent the first and fourth quartiles of a data set.

Cumulative frequency
The number of observations with a value less than the maximum value of the variable interval.

Cumulative percentage
The percentage of observations with a value less than the maximum value of the variable interval.

Cumulative relative frequency
A number calculated by adding together all the relative frequencies less than or equal to the selected upper limit point.

Frequency distribution
A summary of the numerical counts of the values or categories of a measurement.

Grouped frequency
A frequency distribution with distinct intervals or groups created to simplify the information.

Histogram
A chart that usually has an ordinal variable on the horizontal axis and the frequency of the response on the vertical axis, with no spaces between the columns on the horizontal axis.

Line graph
A chart in which the horizontal axis shows the passage of time and the vertical axis marks the value of the variable at that particular time.

Outlier
An extreme value of a variable, one that is outside the expected range.

Percentage
A portion of the whole.

Percentile rank
The percentage of observations below a particular value.

Percentiles
One hundred equal portions of a data set.

Quartiles
Four equal portions of a data set, with the first quartile being the 25th percentile, the second quartile being the 50th percentile, and the third quartile being the 75th percentile.

Relative frequency
The number of times a particular observation occurs divided by the total number of observations.

Scatterplot
A chart in which each point represents the measurement of one subject in terms of two variables.

Frequency Distributions

Once you have designed a study and collected data, the next step is to decide how to present the assembled data. You have several options for doing so. The first and most common choice is a **frequency distribution**, which shows the frequency of each measure of a variable. A frequency distribution is created by gathering all the responses collected from a sample of variables into a table like the one in **Table 2-1**. The first column of the frequency distribution in the table shows the number of days spent postoperatively in the hospital (the dependent variable), sorted from the shortest stay to the longest. The second column shows how frequently that length of stay was needed, that is, the number of patients who spent each number of postoperative days in the hospital. These two columns display the total numeric value of the variable of interest (in this case, the dependent variable, days spent in the hospital), usually ordered from the lowest to the highest. You can see the frequency of each level of the variable and its spread (distribution).

Sometimes it is also helpful to include the **relative frequency** of an observation, which is just the number of times a particular observation occurs divided by the total number of observations. For example, according to Table 2-1, seven patients stayed in the hospital for 3 days. If you want to report the relative frequency of staying 3 days, divide the number of patients who stayed for 3 days by the total number of patients included in the study; in this example that would be 7 divided by 50, or a relative frequency of 14%. Relative frequency is a helpful concept to illustrate what proportion of the observations this particular observation is. The frequency distribution table presents a big-picture view of your data.

To augment a frequency distribution table and really impress your colleagues, you can add a **cumulative frequency** column, which simply lists the number of observations with a value less than the maximum value of the variable interval. For example, in the third column of Table 2-1, in the second row, the number 9 means that nine patients have stayed either 0, 1, 2, or 3 days postop in the hospital. The number 9 is the cumulative frequency for the first four intervals (0, 1, 2, and 3 days) of the variable. Let's say you are putting together an in-service presentation and decide to collect data from a set of patients on how many postop days they spent in the hospital. You find that nine patients were discharged on day 3 or earlier; that total includes all the patients who stayed for 0, 1, 2, or 3 days. It is a cumulative frequency. You may also want to report a **cumulative relative frequency**, which totals the relative frequencies up to the upper limit point you have selected.

TABLE 2-1 Frequency Distribution Table for the Length of the Hospital Stay

Days Spent in the Hospital Postop	Number of Patients Who Stayed This Long (Frequency)	Number of Patients Who Stayed This Long or Less (Cumulative Frequency)
0	0	0
1	0	0
2	2	2
3	7	9
4	23	32
5	14	46
6	4	50

In this example, if you wanted to report the cumulative relative frequency for staying 3 days or less, you would simply add the relative frequencies for those staying 0, 1, 2, and 3 days (0/50 + 0/50 + 2/50 + 7/50 = 9/50, or 0.18, or 18%). This can be helpful when presenting your report because most individuals can understand percentages fairly well.

Now your nurse manager approaches you in a panic because the accreditation agency is coming next week, and she needs to know how many patients were discharged after more than 4 days of recovery. The best way to answer that question visually is to create a new table that includes **grouped frequencies**, which are frequency distributions with distinct intervals or groups created to simplify the information. In **Table 2-2**, the values of the frequency distribution in Table 2-1 have been collected into two groups: (1) patients who spent 4 days or fewer in postop and (2) those who spent 5 days or more. Grouped frequencies are typically used when working with a lot of data and an entire frequency distribution is simply too large to be meaningful.

Unfortunately, when data is grouped, some information is lost. For example, how many patients in Table 2-2 stayed for only 2 days? The answer is not discernible from this table. This is the first drawback to be aware of when using grouped frequencies; you can lose a lot of information when you convert your data into groups, especially if you use large intervals. You can even make the intervals so large that they are meaningless. In our example, if one interval were more than 7 days and the other were less than 7 days, Table 2-2 would not be very useful anymore because all the patients in the study were discharged by day 7. On the other hand, make sure not to make the intervals too small, or your grouped frequency won't have any benefit over a standard frequency distribution.

Let's return to our example of the poor nurse manager who needed to get ready for the accreditation agency visit. After retabulating the data as shown in Table 2-2, you can calmly tell her that during the period of time in your study, 18 patients were discharged after more than 4 days of recovery.

PERCENTAGES

A **percentage** is a part of the whole. To calculate a percentage, divide the partial number of items by the total number of items, and then multiply that quantity by 100. For example, what if that same nurse manager asked you, "What percentage of our patients do those 18 represent?" You could do the simple calculation shown in **Figure 2-1**. In

TABLE 2-2 **Frequency Distribution Table for the Length of Hospital Stay Using Grouped Frequencies**

Days Spent in the Hospital Postop	Number of Patients Who Stayed This Long (Frequency)
$\leq$ 4 days	32
5 or more days	18

FIGURE 2-1 **Calculating a Percentage.**

$$\frac{18}{50} = 0.36 \times 100 = 36\%$$

TABLE 2-3 Cumulative Percentage Table for Length of Hospital Stay

Days Spent in the Hospital Postop	Number of Patients Who Stayed This Long	Cumulative Frequency	Cumulative Percentage
0	0	0	$0/50 = 0 \times 100 = 0\%$
1	0	0	$0/50 = 0 \times 100 = 0\%$
2	2	2	$2/50 = 0.04 \times 100 = 4\%$
3	7	9	$9/50 = 0.18 \times 100 = 18\%$
4	23	32	$32/50 = 0.64 \times 100 = 64\%$
5	14	46	$46/50 = 0.92 \times 100 = 92\%$
6	4	50	$50/50 = 1.0 \times 100 = 100\%$

this example, the number of patients of interest (those who were discharged after day 4) is 18. The total number of patients studied is 50. (See the last line of the third column in Table 2-1.) The first step in our calculation is 18 divided by 50. This division results in 0.36, which is then multiplied by 100 to get a percentage of 36%.

Exam scores are a classic example of percentages. If you take an exam with 30 questions and get 27 correct, what is your overall score? In this case, divide 27 by 30, and then multiply by 100 to get 90%.

A statistics concept commonly associated with percentages is **cumulative percentage**, which is the percentage of observations with a value less than the maximum value of the variable interval. The idea is the same as cumulative frequency, but it is expressed as a percentage. The last column in **Table 2-3** shows the conversion of each cumulative frequency (from Table 2-1) into a cumulative percentage: the percentage of patients who had a hospital stay of less than or equal to the number of days listed in that row. For example, 18% of patients had a hospital stay of less than or equal to 3 days, and all of the patients (100%) were discharged on or before day 6 (see the last line of the last column in the table).

Percentages are also closely related to percentiles, which are explained in the next "From the Statistician."

Percentages are also related to the concept of the **percentile rank** of a score, which is the percentage of observations lower than that score in a frequency distribution. For example, if your test score is greater than 75% of all the scores for the class, it is at the 75th percentile.

Bar Charts

Remember nominal categorical data (the categorical data that shows only a difference and is not rank-ordered)? A **bar chart** is one way to display this type of data. A common way to set up the bar chart is to line up the responses for the nominal variable along the horizontal axis and place the frequencies of the responses on the vertical axis. Bar charts are typically used for nominal categorical data with spaces between the bars because

FROM THE STATISTICIAN *Brendan Heavey*

Quantiles, Quartiles, and Percentiles—Oh My!

The terms *quantiles*, *quartiles*, and *percentiles* cause a lot of grief because they are so closely related. So let's break them down a little. Using quantiles is just like dividing a data set into different portions or bins. Two special cases of quantiles are percentiles and quartiles.

- **Percentiles** divide a data set into 100 equal portions. You see this concept used with body mass index (BMI). If a patient's BMI is in the 90th percentile, then 90% of the BMIs in the reference population used to develop the distribution were at or below this patient's BMI. Put another way, this patient's BMI is in the top 10% of the reference population.
- **Quartiles** divide a data set into four equal parts. For example, suppose your nursing manager wants to hire only students who finished in the top quarter of the class on a particular exam. She would calculate the third quartile and select all the scores above it. Because your score would clearly be near the top, she would then rush to your school and attempt to woo you with new scrubs and tuition benefits!

How would the nursing manager compute a percentile? Let's say a sample of 331 nurses at Massachusetts General Hospital is asked how many patients they see on average each shift. The results of this survey are shown in **Table 2-4**. A nice formula to find percentiles in the ordered data set is shown in **Figure 2-2**. For instance, based on the ordered data set in Table 2-4, we apply the formula as shown in Figure 2-2 to find what is called the median, or the middle observation when the observations are lined up in rank order (least to greatest). The median is also the 50th percentile. Therefore, our median is our 166th observation. Using **Figure 2-3**, we can see that the nurse who was observation number 166 saw 16 patients.

Because the total sample is 331, the middle observation is the 166th observation. If you add all the nurses who saw fewer than 16 patients, you find that 155 of them reported seeing fewer than 16 patients, and 22 others saw 16 patients. After lining up the observations in rank order, the 166th observation falls into the group who reported seeing 16 patients; that group is the median. Using the formula takes less time than the old-fashioned way: adding them up.

You can also check yourself by looking at the cumulative relative frequency column. The median should be where the midpoint is. In Table 2-4, look at the row for 15 patients, and see the corresponding cumulative frequency percentage; 47% of the nurses reported seeing 15 or fewer patients. Look at the next line, for 16 patients; 53% of the nurses reported seeing 16 or fewer patients. Therefore, the 50th percentile is above 15 and less than or equal to 16. Because we cannot split a patient into parts (at least for statistical purposes!), the median number of patients is 16.

FROM THE STATISTICIAN *Brendan Heavey*

TABLE 2-4 **Frequency Table for a Sample of 331 Nurses at Massachusetts General Hospital**

Number of Patients	Number of Nurses	Relative Frequency (%)	Cumulative Relative Frequency (%)	Number of Patients	Number of Nurses	Relative Frequency (%)	Cumulative Relative Frequency (%)
1	0	0	0	16	22	7	53
2	0	0	0	17	20	6	60
3	2	1	1	18	19	6	65
4	3	1	2	19	18	5	71
5	5	2	3	20	19	6	76
6	8	2	5	21	17	5	82
7	8	2	8	22	15	5	86
8	9	3	11	23	12	4	90
9	9	3	13	24	8	2	92
10	12	4	17	25	9	3	95
11	14	4	21	26	5	2	96
12	18	5	27	27	5	2	98
13	22	7	33	28	3	1	99
14	23	7	40	29	3	1	100
15	22	7	47	30	1	0	100

Three hundred and thirty-one nurses at Massachusetts General Hospital were asked how many patients they saw on their shifts that night. Relative frequencies are computed by dividing the number of nurses who saw each number of patients by the total number of nurses (331). Each cumulative relative frequency is the result of adding the previous row's relative frequency to its cumulative relative frequency, so in essence, cumulative relative frequency is an accumulation of relative frequency.

(continues)

FROM THE STATISTICIAN *Brendan Heavey*

FIGURE 2-2 **Formula for Calculating Percentiles in an Ordered Data Set.**

$$P^{th}_{obs} = (n+1) \times \frac{y}{100}$$

Where

P^{th}_{obs} = The number of the observations at the percentile for which you are looking

n = The number of observations in your data set

y = The percentile you're looking for

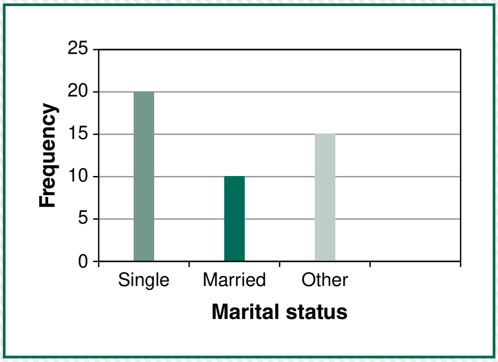
FIGURE 2-3 **Equation for the Massachusetts General Sample.**

$$P^{th}_{obs} = (331+1) \times \frac{50}{100} = 166$$

each answer is distinct and in no particular order. For example, if you collected data about the marital status of fellow nurses on your unit, you might find the data shown in **Figure 2-4**. A quick look at the bar chart makes it apparent that more of the nurses working on this unit are single than either married or other. The bar chart gives you a good visual representation of nominal categorical data.

Figure 2-5 shows another example of a bar chart that represents the data we first examined

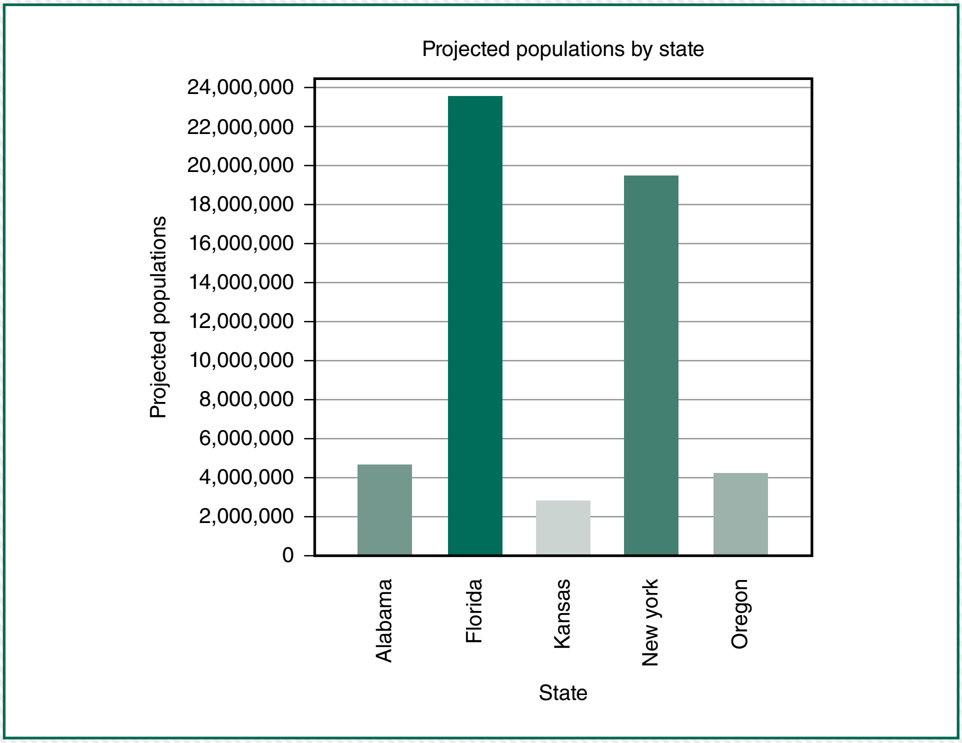
FIGURE 2-4 **Bar Chart for Marital Status.**



in Chapter 1 about population projections in various states for our public health study. By glancing at the bar chart, we can quickly determine that the population in the state of Florida is substantially larger than the population in Kansas.

Bar charts can be used for ordinal data as well, but then the bars should follow the rank order of the variable categories. For example, if you are presenting the data in **Figure 2-6** about pain levels post hysterectomy at a conference and a listener reports that the policy at his institution is to prescribe pain medication for

FIGURE 2-5 **Projected Populations by State, 2004–2030.**



Data from CDC. Population Projections, United States, 2004–2030, by state, age, sex, on CDC WONDER Online Database, September 2005. Retrieved from http://wonder.cdc.gov/population-projections.html.

FIGURE 2-6 **Pain Level 2 Weeks Post Hysterectomy.**

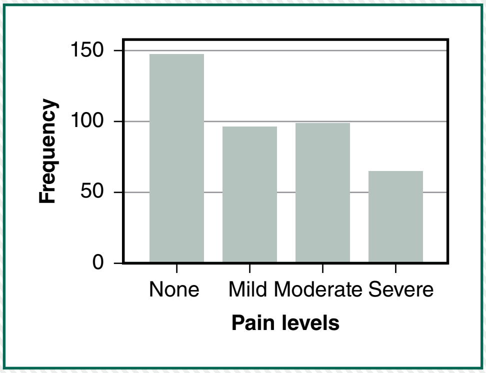
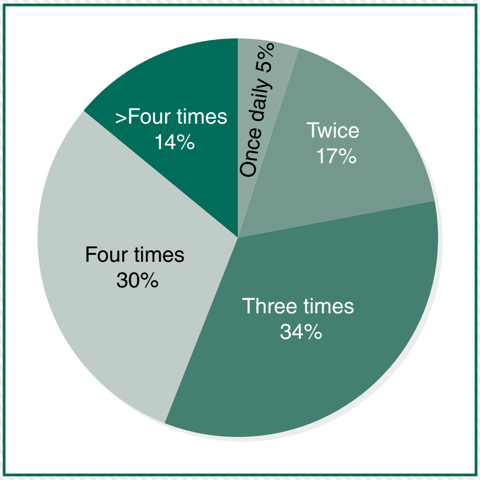


FIGURE 2-7 **How Often Do Subjects Eat Each Day.**



2 weeks upon discharge because patients are fully recovered at that point, your bar chart might be very helpful. You would be able to show quickly that, in your study, the majority of patients are pain-free or close to it (mild pain) by 2 weeks post hysterectomy, but there are still a substantial number who have moderate to severe pain 2 weeks after their procedure. Because the pain variable is measured at an ordinal level, it is useful to have it in rank order on the *x* axis to illustrate the subsequent escalation of the pain level reported.

Another option for displaying nominal data is a pie chart. In this type of graph, a large circle (the pie) is divided into smaller pieces, and each piece illustrates a percentage of the whole. If you are conducting a study about dietary habits and had a survey question that asks about the frequency of eating each day, you might choose to illustrate the data with the pie chart in **Figure 2-7**. Again, it is easy to see that the majority of your subjects eat 3 to 4 times a day.

HISTOGRAMS

A **histogram** is a type of bar chart. Histograms often have no spaces between the bars because these charts are most frequently used to display either ordinal data or continuous data. (Remember, ordinal data has categories that show a ranked difference; continuous data has an infinite number of possible, in-between measures.) In our previous example, pain may be rated as mild, moderate, or severe. Instead of developing a bar chart and putting the ordinal pain data in rank order on the *x* axis, we might want to put those bars right next to each other in a histogram. The lack of spaces between the bars in the histogram reinforces the idea that these responses are on a continuum and that the order is illustrated on that continuum.

Presenting these types of data in a histogram shows how frequently each response is

FIGURE 2-8 **Histogram for Pain Level.**

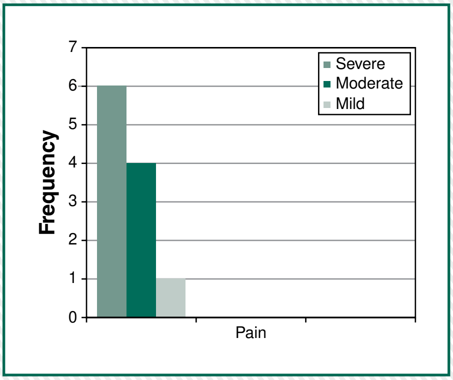
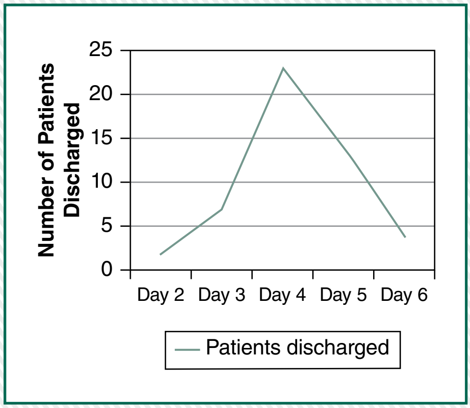


FIGURE 2-9 **Line Graph for Length of Hospitalization.**



selected and allows for visual comparison of the different levels. In **Figure 2-8**, 11 patients were interviewed 12 hours after abdominal surgery. Six rated their pain as severe, four stated it was moderate, and one felt it was only mild (she had just had her pain meds).

This histogram gives you a big-picture idea of the pain these patients experienced. In this case, the histogram seems to indicate that many postop patients report the first 24 hours as being very painful. So the next time you orient a new nurse, you might remember to point out how important it is to make sure patients have their pain medicine ordered and administered immediately after surgery. The chart also visually displays that many patients may not be getting adequate pain medication because so many report severe pain. After collecting this data, you may decide to review the unit protocols for pain management.

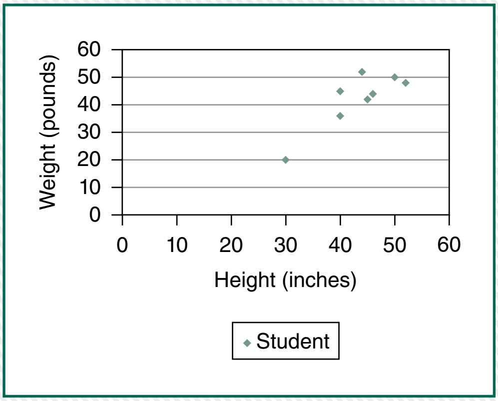
Line Graphs

Continuous variables that change over time are frequently best illustrated in a **line graph**. The horizontal axis shows the passage of time, and the vertical axis marks the value of the variable over time. For example, the data from the cumulative frequency example about days of hospitalization is illustrated in **Figure 2-9**. The chart shows that most of the patients needed to stay 4 days postop before going home. You might want to compare this line graph to another after you institute an early mobilization plan with your surgical patients to see whether the length of hospitalization has changed.

Scatterplots

Scatterplots are a little different from the previously discussed graphs in that each point represents how one subject relates to two variables. For example, **Figure 2-10** shows a scatterplot of height in inches and weight in pounds for a group of eight kindergartners. Each square on the scatterplot represents one student. The horizontal axis displays that student's height,

FIGURE 2-10 **Scatterplot for Student Height (*x* axis) and Weight (*y* axis).**



and the vertical axis displays his or her weight. You can see from the direction of the plotted points that as students get taller, they usually get heavier as well; that's the relationship between the two variables. When points are close together or seem to follow a line closely, the relationship between the variables on the horizontal and vertical axes are relatively strong.

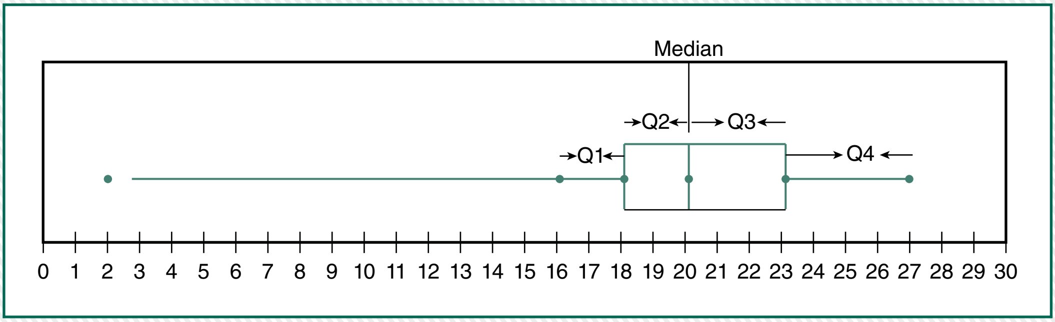
When you look at a scatterplot, note the general trend. In this example, the plotted points start low on the left side and move up as they progress toward the right side. This pattern indicates a *positive relationship* between height and weight (in other words, they usually move in the same direction—when height increases, weight usually does too). If the plotted points were to start in the upper left corner and slope down to the right, the pattern would indicate a *negative relationship* between the two variables (such as exercise and weight—when exercise is increased, weight usually decreases).

Scatterplots also give nurses a chance to look for **outliers**, or data that is outside the expected relationship. In Figure 2-10, the student who is only 30 inches and 20 pounds clearly stands out from the rest of the group. Perhaps the child is just extremely small, or there may have been an error in measurement, recording, or data entry. If there are a many outliers, a nurse may decide either that further investigation is needed to ensure accuracy or that the outliers may actually represent the children whom the study is designed to identify. One example of this technique is the use of growth charts. When children make pediatric visits, nurses almost always plot the children's heights and weights on a growth chart. This is one example of using a scatterplot to look for outliers. If a child isn't growing properly, recognizing the growth pattern as an outlier is one way to identify a child who needs intervention.

BOX AND WHISKERS PLOT

Sometimes when we want to find outliers, or we want to illustrate the central value and variability seen in a data set, we may opt to use a diagram called a **box and whiskers plot** (see **Figure 2-11**). In this type of diagram, the observations are lined up from least to most, and the quartiles (first 25% of observations, second 25% of observations, third 25% of observations, and fourth 25% of observations) are identified. The middle 50% (quartile 2 and 3) of the data set is contained in the box of the diagram, and the whiskers of the diagram show the range of the data. Notice that there is an observation at a value of 2, which is not contained in the range of the whiskers. Individual dots at the far end of a box and whiskers plot are considered outliers. We will talk about this in more detail in the next chapter when we start to look at measures of central tendency.

FIGURE 2-11 **Box and Whiskers Plot.**



Summary

This chapter contained quite a bit of information, so let's review and make sure you are really comfortable with it. Frequently, researchers put a great deal of time into collecting data and very little time into thinking about how to present it; however, how you present your data often determines whether your intended audience understands your work or is even interested in it. (Teachers are all aware of this point!) The most common choice for presentations is a frequency distribution, which shows the frequency of each measure of a variable. You can add these frequencies and create either cumulative frequency columns or grouped frequencies, depending on the question you are trying to answer.

You can also calculate percentages, which are parts of the whole. Because many nurses are familiar with them, percentages are sometimes a useful way to convey information. You can then add the percentages and present a cumulative percentage, which is simply the percentage of observations with a value less than the maximum value of the interval.

Another way to convey information is with a visual graph, such as a bar chart for nominal data, a histogram for ordinal or continuous data, a line graph for continuous variables that change over time, or a scatterplot in which one subject's values for two variables are graphed. You need to decide which type of chart will work the best for the audience you are trying to reach. Just remember, use color, make it bigger, and avoid using lots and lots of numbers in a row—and bring coffee because most nurses are considerably sleep deprived!

Chapter 2 Review Questions

1. Thirty fathers were asked about the highest level of education they had completed. Ten completed only elementary school, 10 completed elementary and high school, 7 completed all levels through college, and 3 completed all levels through graduate school. What was the cumulative percentage of fathers who completed only elementary school? Round to the nearest whole number.

2. In your study of 40 people, 8 had no cold symptoms, 12 had mild cold symptoms, 9 had moderate cold symptoms, and 10 had severe cold symptoms. One patient was lost to follow-up, and no data could be collected. What percentage of patients reported cold symptoms?

3. Given the information in review question 2, what percentage of patients reported no cold symptoms?

4. Use the frequency distribution in **Table 2-5** to construct a bar chart for influenza cases in your hospital during 8 months in 2017-2018. Would your chart look different if it were a histogram? Discuss at least one rationale for selecting either a bar chart or a histogram to present this data.

Questions 5–7: Your community begins a large-scale influenza vaccine effort, and the following year, the number of cases drops (see **Table 2-6**).

5. Construct a line graph showing the data from 2017 and the data from 2018. Compare the two.

6. Why didn't the numbers change significantly for August and September?

7. Do you consider the vaccine effort to be successful? Why?

TABLE 2-5 Influenza Cases for 2017-2018

Month	Number of Cases
August	18
September	29
October	68
November	107
December	158
January	166
February	160
March	111

TABLE 2-6 Influenza Cases for 2017-2018

Month	Number of Cases
August	19
September	27
October	31
November	34
December	48
January	59
February	51
March	45

Research Application

Questions 8–13: See the data in **Table 2-7**.

8. Construct a bar chart for mother's marital status.

TABLE 2-7 Demographic Characteristics of 92 Adolescents Completing a Family Planning Survey

	n **(%)**
Pregnancy status	
Pregnant	78 (84.7)
Not pregnant	14 (15.2)
Age	
$\leq$ 14 years old	2 (2.2)
15–17 years old	46 (50)
18–19 years old	43 (46.7)
*Number in household**	
$<$ 6 people	70 (79.5)
$\geq$ 6 people	18 (20.5)
Student status†	
Not in school	46 (50.5)
In school	45 (49.4)
Mother's marital status	
Single	38 (41.3)
Married	27 (29.3)
Divorced	13 (14.1)
Other	14 (15.2)
Employment status	
Employed	25 (27.2)
Not employed	67 (72.8)

*Missing $n = 4$
†Missing $n = 1$
Reproduced from Heavey, E. Moysich, K., Hyland, A., Druschel, C., & Sill, M. (2008) Female adolescents' perceptions of male partners' pregnancy desire. *Journal of Midwifery & Women's Health*, *53*(4), 338-344. Reproduced with permission of Elsevier Inc.

9. What percentage of the adolescents are employed? What percentage of the adolescents are in school? Are these variables quantitative or qualitative? Round to the nearest tenth of a percent.

10. Identify the level of measurement of each variable.

11. Could any of these variables have been measured as continuous quantitative variables?

12. Construct a histogram for the ages of the adolescents. Describe the histogram and what it tells you about this sample population. Why would a histogram be an appropriate choice for presenting this data?

13. Is the pregnancy status of this group of adolescents typical? Why might that be?

14. You've been recruited by the head of the Federal Emergency Management Agency (FEMA) to act as the head triage nurse for a large city's hurricane response team. One of your main duties is to decide which nurses will cover which facilities in the overall relief effort. Because most nursing duties will have to change during this shift in personnel, you decide to divide the group based on years of experience. Your nurse's aide carries out a brief survey of all the personnel available (100 nurses) and gives you a list of the number of years of experience for each (see **Table 2-8**). Find the quartiles of this distribution, and assign a role to each nurse. For example, the nurses with the least amount of experience should be assigned to the rescue team, and the ones with the most experience should be assigned to the intensive care unit (ICU).

Questions 15–18: A diabetes educator is working with a group of 15 patients who have been newly diagnosed with type 2 diabetes. She administers a brief pretest, reviews carbohydrate counting with them, and then asks them to complete a posttest assessing their knowledge of the total grams of carbohydrate found in one serving of four sample items brought to the class. The results are shown in **Table 2-9**.

15. Complete the Cumulative Frequency column of the table.

16. What percentage of the patients answered all four questions correctly?

TABLE 2-8 Nurses Available for the Hurricane Response Team

Number of Years' Experience	Number of Nurses	Number of Years' Experience	Number of Nurses
2	15	12	3
3	10	13	3
4	9	15	3
5	8	18	3
6	8	20	2
7	9	25	2
8	6	32	1
9	5	35	1
10	5	36	1
11	5	40	1

Coordinated Hurricane Response

Rescue Team	Stadium	Hospital	ICU

TABLE 2-9 Posttest Assessing Carbohydrate Knowledge

Number Correct on the Posttest	Number of Patients Who Answered This Number of Questions Correctly (Frequency)	Number Who Answered This Many or Fewer Questions Correctly (Cumulative Frequency)
0	1	
1	1	
2	2	
3	6	
4	5	

17. Add another column to the table, and calculate the cumulative percentage.

18. The diabetes educator would like you to present her results in a grouped frequency table showing the frequency and percentage of those who passed the posttest with >70% and those who didn't for the report she has to make to her funding agency. Show your product.

Questions 19–20: A researcher examining patients diagnosed with sarcoidosis would like to look at trends in their inflammatory markers. The researcher would like to use a graph that can illustrate the erythrocyte sedimentation rate and the C-reactive protein level for each individual in the study and thus show the relationship between the two markers.

19. What graphic representation of the data would you suggest?

20. After creating this graph, the researcher notices one of the subjects is significantly below the trend line that the other subjects seem to follow. What might be a logical explanation for this outlier?

21. An emergency room nurse working in a hospital in the wine region of the Finger Lakes in New York has noticed a seasonal trend for ocular injuries from bottle corks. She would like to illustrate this graphically and develops a histogram. Why might she prefer a histogram to a bar chart?

---

22. When the nurse looks at her data and the histogram, it is apparent that more ocular injuries from bottle corks occur in October (which is wine fermentation season) and January. She also notices that more of the injuries involve the right eye. Provide a reasonable explanation for both the January peak and the prevalence of the right-eye injuries.

---

23. The nurse would like to determine if more ocular injuries are associated with wine bottles from Sharespeak Winery or with wine bottles from Francesco's winery. What would be the independent variable?

---

24. What would be the dependent variable in review question 23?

---

25. You are tracking melanoma in your county. Calculate the race- and gender-specific mortality rates from the data provided in **Table 2-10.**

TABLE 2-10 Melanoma Data

	New Cases	Deaths	Population	Mortality Rate
Men	5	2	25,000	
Women	3	1	25,000	
White	7	3	40,000	
African American	1	0	10,000	

26. Offer a potential explanation for why the race-specific rates may be different for whites and African Americans.

---

27. Complete the following table:

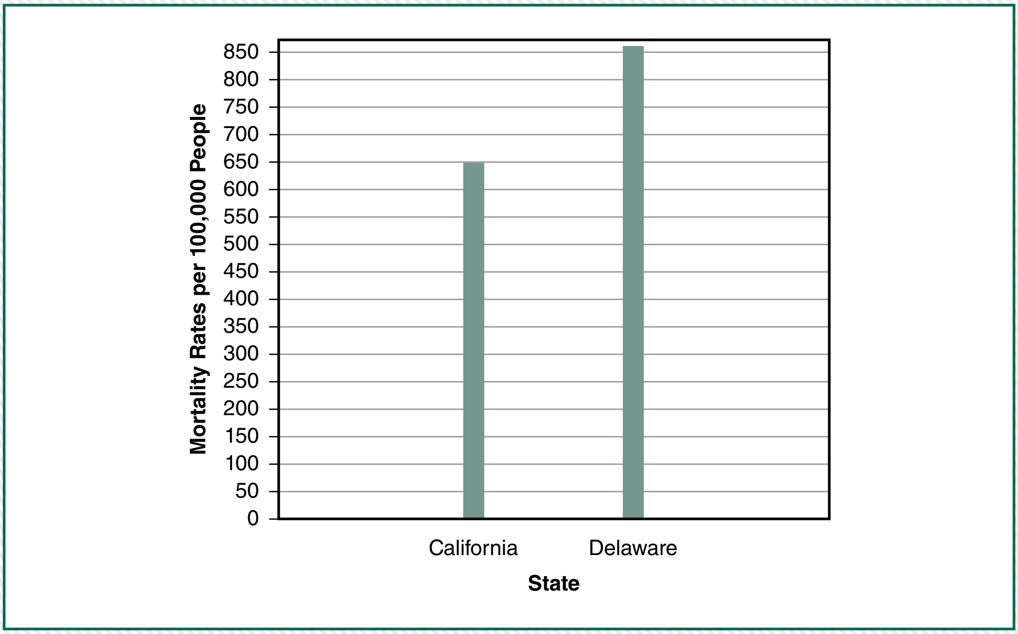
Age Range	Number of Subjects in Sample	Cumulative Frequency
0–5	1	
6–10	2	3
11–15		7
16–20	1	8
21–25		11
26–30	4	
31–35	6	

Questions 28–29: The Centers for Disease Control and Prevention (CDC) provided **Table 2-11** and **Figure 2-12** showing mortality rate data from Delaware and California in 2014.

TABLE 2-11

State	Deaths	Population	Rate per 100,000
California	3,539,688	540,469,946	654.9
Delaware	110,279	12,809,756	860.9
Total	3,649,967	553,279,702	659.7

FIGURE 2-12 **Mortality Rates per 100,000 People in California and Delaware in 2014.**



Reproduced from Centers for Disease Control and Prevention, National Center for Health Statistics. Compressed Mortality File 1999-2013. Archive on CDC WONDER Online Database, released October 2014. Data are from the Compressed Mortality File 1999-2013 Series 20 No. 2S, 2014, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program. Accessed at http://wonder.cdc.gov/cmf-icd10-archive2013.html on Apr 4, 2017.

28. Your colleague has a pending job offer and decides he will accept the offer in Delaware because too many people die in California every year. Do you agree with his decision? Why or why not?

29. Which is more effective for presenting the data about mortality rates: the table or the bar chart? Why?

30. **Table 2-12** presents data from the CDC that tracks Ebola cases in Sierra Leone in June 2014. Complete the Cumulative Frequency columns.

TABLE 2-12

Date	Number of Cases	Cumulative Frequency, Cases	Number of Deaths	Cumulative Frequency, Deaths
6/2/2014	50		6	
6/5/2014	81		7	
6/10/2014	89		7	
6/11/2014	117		19	
6/18/2014	97		49	
6/24/2014	158		34	

Data from CDC. (2016). *2014 Ebola outbreak in West Africa* — Reported case graphs. Retrieved from https://www.cdc.gov/vhf/ebola/outbreaks/2014-west-africa/cumulative-cases-graphs.html

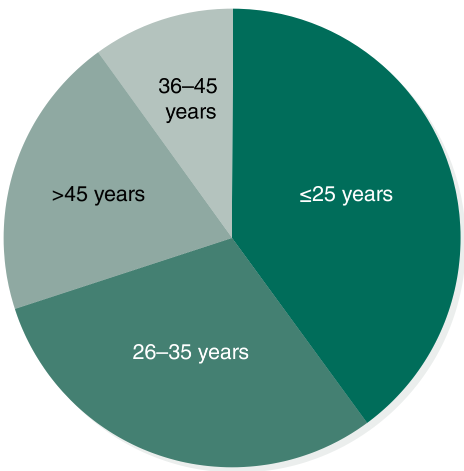
Questions 31–33: Refer to the graph in **Figure 2-13.**

31. Which age group has the highest incidence of chlamydia? What explanations might the researchers offer about this data?

32. Which group has the lowest incidence of chlamydia? What explanations might the researchers offer about this data?

33. If you were going to design a chlamydia prevention campaign, what age group would you suggest targeting? Why?

FIGURE 2-13 **Incidence of Chlamydia in Florida by Age Group.**



ANSWERS TO ODD-NUMBERED CHAPTER 2 REVIEW QUESTIONS

1. 33%
3. 20%
5. Answer includes line graph; beginning in October, there is a substantial decrease in cases.
7. Yes, there was a substantial decrease in cases beginning in October after the vaccine was administered.
9. 27.2%, 49.4%, qualitative
11. Yes, age, number in household, years in school, years employed
13. No, all the adolescents were waiting for pregnancy or family planning–related services.

15.

Number Correct on the Posttest	Number of Patients Who Answered This Number of Questions Correctly (Frequency)	Number Who Answered This Many or Fewer Questions Correctly (Cumulative Frequency)
0	1	1
1	1	2
2	2	4
3	6	10
4	5	15

17.

% Who Answered This Many or Fewer Questions Correctly (Cumulative %)
1/15 = 6.7%
2/15 = 13.3%
4/15 = 26.7%
10/15 = 66.7%
15/15 = 100%

19. A scatterplot
21. A histogram better illustrates the continuum of time from one month to the next with an order involved.
23. The winery

25.

New Cases	Deaths	Population	Mortality Rate	
Men	5	2	25,000	**0.008%**
Women	3	1	25,000	**0.004%**
White	7	3	40,000	**0.0075%**
African American	1	0	10,000	**0**

27.

Age Range	Frequency	Cumulative Frequency
0–5	1	1
6–10	2	3
11–15	4	7
16–20	1	8
21–25	3	11
26–30	4	15
31–35	6	21

29. Putting the rates rather than the absolute numbers in the bar chart helps to illustrate more clearly the different mortality rates between states that are so vastly different in size.

31. The highest incidence is in those who are $\leq$25 years old. Explanations may vary but may include having more sexual partners, practicing riskier sex, and engaging in more frequent sexual activity for people in this age bracket.

33. Answers will vary but may include ages $\leq$25 because of higher incidence and more long-term risks from early infection. May also include all under age 35 because that captures the two highest-incidence age groups.

CHAPTER 3

Descriptive Statistics, Probability, and Measures of Central Tendency

What does the data tell me?

OBJECTIVES

By the end of this chapter students will be able to:

- Compare and contrast descriptive statistics and inferential statistics.
- Define, distinguish between, and interpret the mean, median, mode, and standard deviation.
- Identify unimodal, bimodal, and multimodal distributions.
- Determine which measure of central tendency is appropriate in a given data set.
- Calculate and interpret a standard deviation and range for a given data set.
- Explain descriptive results from a given data set using a Statistical Package for the Social Sciences (SPSS) printout.
- Define probability and explain the range of possible probabilities.
- Compare and contrast frequency and probability distributions.
- Contrast positive and negative distribution skews and describe where the outliers are present.

KEY TERMS

Bimodal
Having two values or categories that have the highest occurrence and that are equal frequencies.

Central tendency
An indicator of the center of the data.

Extreme outlier
An outlier in a data set that is more than 3 times the interquartile range either above or below the interquartile range ($\pm 3 \times$ the length of the box above or below the box on a box and whiskers plot).

Frequency distribution
Lists all the possible outcomes of an experiment and tallies the number of times each outcome occurs.

Mean
The sum of the values divided by the total number of observations. It is the most commonly known measure of central tendency but requires interval or ratio data.

Median
For ordinal, interval, and ratio data, the value in the middle when all the measured values are lined up in order from least to most; the 50th percentile value.

Mode
The most frequently occurring value or category in the distribution. When a distribution has only one mode, it is called unimodal.

Multimodal
Having more than two modes.

Normal distribution
A probability distribution in which the mean, median, and mode are equal with a bell-shaped distribution curve.

Probability
The chance that a particular outcome will occur after an event.

***p*-value**
The probability of finding the reported results if the null hypothesis is true.

Probability distribution
The probability of all the possible outcomes of the variable.

Range
The difference between the maximum and minimum values in a distribution.

Sampling distribution
Plots realized frequencies of a statistic versus the range of possible values that statistic can take.

Skewed distribution
An asymmetrical distribution of the values of the variable around the mean, making one tail longer than the other.

Standard deviation
The average distance that the values in a distribution are from the center.

Tukey fences
A cutoff value indicating that an observation is an outlier in a data set because it is more than 1.5 times the interquartile range either above or below the interquartile range (± 1.5 times the length of the box above or below the box on a box and whiskers plot).

***Z*-score**
A measure that indicates how many standard deviations a value is from the mean value.