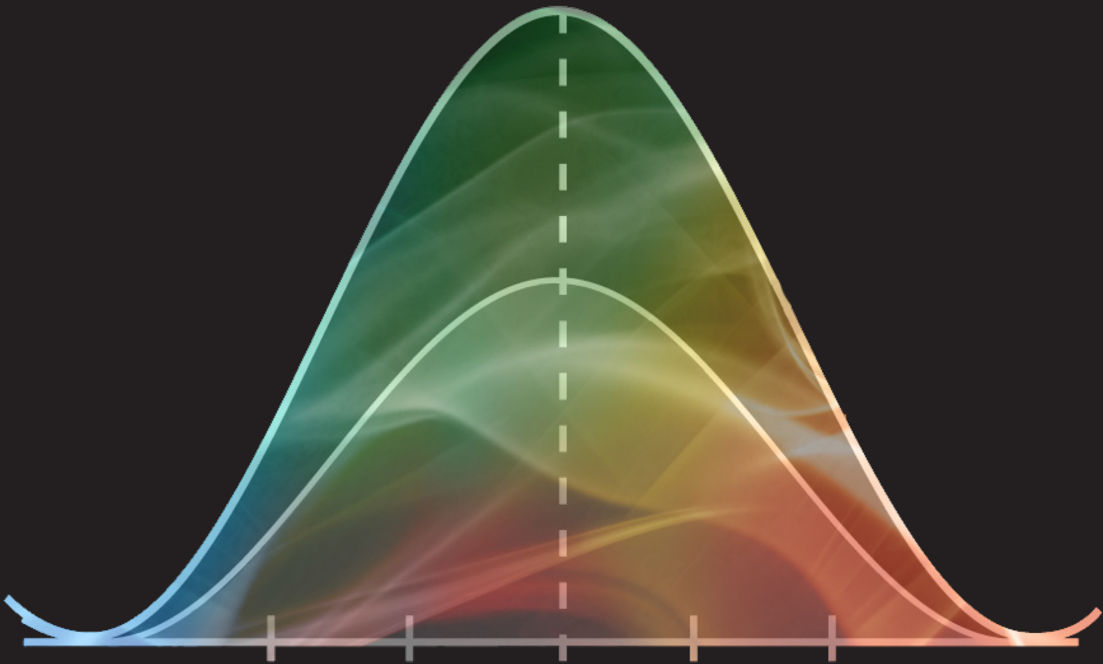


Biostatistics for Clinical and Public Health Research

Melody S. Goodman

Biostatistics for Clinical and Public Health Research

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

Biostatistics for Clinical and Public Health Research

Melody S. Goodman

Routledge
Taylor & Francis Group
London and New York

First published 2018
by Routledge
711 Third 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

Library of Congress Cataloging-in-Publication Data

Names: Goodman, Melody S., author.
Title: Biostatistics for clinical and public health research / Melody S. Goodman.
Description: Milton Park, Abingdon, Oxon ; New York, NY : Routledge, 2018. |
Includes bibliographical references and index.
Identifiers: LCCN 2017025994| ISBN 9781498784801 (hbk) | ISBN 9781138196353 (pbk) |
ISBN 9781315155661 (ebk)
Subjects: LCSH: Public health--Research--Statistical methods. | Biometry.
Classification: LCC RA440.85 .G66 2018 | DDC 362.1072/7--dc23
LC record available at https://lccn.loc.gov/2017025994

ISBN: 978-1-4987-8480-1 (hbk)
ISBN: 978-1-138-19635-3 (pbk)
ISBN: 978-1-315-15566-1 (ebk)

Please visit the eResources at: www.routledge.com/9781138196353

This book is dedicated to all of my students, past, present, and future. To all of the introduction to biostatistics instructors who believe in the "I hear, I forget, I do, I understand" approach to teaching this material. I hope this "old-school" workbook helps with learning and teaching.

It is also dedicated in loving memory of my mother, Linder White-Goodman. You were my first teacher, best friend, loudest cheerleader, and greatest critic. Thanks for always telling me I can do, be, and achieve anything I want. For believing in me when I didn't believe in myself and reassuring me that I was great but could always do better. For showing me that education was a viable path to success and could be used as a tool to change my life circumstances. You touched many lives but none more profoundly than mine, your only daughter. As hard as I tried not to follow in your footsteps, I too am an educator—just in a different context. This book is one of my greatest teaching accomplishments.

Contents

Acknowledgments

It takes a village to write a book even if there is only one name on the cover. I want to thank the Goodman Lab team at the Washington University School of Medicine (Nicole Ackermann, Goldie Komaie, Sarah Lyons, and Laurel Milam). Although I often get the credit and accolades, you make it possible for me to look good. You are an amazing group of women and have been an enormous support to me, for which I am forever grateful. I know that working for me can be challenging. This was especially the case when I decided to write two books in addition to everything else that we already had going on as a team, but you smiled through it all.

A special thanks to the talented young female statisticians who worked in the Goodman Lab and served as teaching assistants for my course. Each of you made significant contributions that helped turn my course notes into an actual book: Nicole Ackermann (Chapters 1 through 5, 15, and Lab A), Sarah Lyons (Chapters 11 through 13), and Laurel Milam (Chapters 6 through 10, 14, and Labs B through E). Thank you for creating examples, datasets, code books, figures, tables, and callout boxes; thank you for truly making this a team effort.

Thanks to Sharese T. Wills for being brave enough to take on the challenge of editing a biostatistics book. You took what I thought were solid drafts and completely reworked them, making them much better.

To my family, friends, mentors, and colleagues who have supported me through this journey—there are too many of you to name (not really; I am just too private to do so), but I hope that you know who you are. It is rare for someone to have such strong personal and professional support systems, and I am lucky to have both. It is great to feel wanted in your department and supported by your institution, but I think it is rare to feel supported by your profession. Thank you to all the academics who have helped me to get to the point of publishing this workbook on biostatistics. It was not an easy journey, and there is still more on the road ahead. However, I never would have made it this far without you.

Author

Dr. Melody S. Goodman, is an associate professor in the Department of Biostatistics at New York University College of Global Public Health. She is a biostatistician with experience in study design, developing survey instruments, data collection, data management, and data analysis for public health and clinical research projects. She has taught introductory biostatistics for masters of public health and medical students for over 10 years at multiple institutions (Stony Brook University School of Medicine, Washington University in St. Louis School of Medicine, New York University College of Global Public Health).

List of abbreviations

ACS	American Community Survey
BMI	Body mass index
BRFSS	Behavioral Risk Factor Surveillance System
CBT	Cognitive behavioral therapy
CDC	Centers for Disease Control and Prevention
CDF	Cumulative distribution function
CL	Confidence limit
CLM	Confidence limit for mean
CLT	Central limit theorem
CRP	C-reactive protein
df	Degrees of freedom
DRE	Digital rectal examination
ED	Emergency department
GED	General education diploma
GIS	Graphical information system
HIV	Human immunodeficiency virus
LCLM	Lower confidence limit
MRCI	Medication Regimen Complexity Index
mRFEI	Modified retail food environment index
MSA	Metropolitan Statistical Area
NATA	National Air Toxics Assessment
NHANES	National Health and Nutrition Examination Survey
NIH	National Institutes of Health
PDF	Probability density function
PK	Pyruvate kinase
PSA	Prostate-specific antigen
ROC	Receiver operating characteristic
RR	Relative risk
RT-PCR	Reverse transcription polymerase chain reaction
SEM	Standard error of the mean
SES	Socioeconomic status
SSA	Social Security Administration
TRUS	Transrectal ultrasound
UCLM	Upper confidence limit
VPA	Vigorous physical activity
YRBSS	Youth Risk Behavior Surveillance System

Introduction

This workbook started as a set of course notes and handouts that I used while teaching Introduction to Biostatistics. Although I love technology, I just think that "old school" is better for some things. I remember learning math in elementary school with workbooks. I am almost sure that none of my current students have even seen a workbook. Nonetheless, there is something about working through problems that helps people to grasp the concepts. We try to use examples in this workbook that are easy to understand, and we walk through problems step by step.

This introductory workbook is designed like a good set of notes from the best student in the class—the professor. Its outline format at the beginning of each chapter points to key concepts in a concise way, and chapters include highlights, bold text, and italics to point out other areas of focus. In addition, tables provide important information. Labs that include real-world clinical and public health examples walk readers through exercises, ensuring that students learn essential concepts and know how to apply them to data. The workbook provides the reader with the statistical foundation needed to pass medical boards and certification exams in public health. Also, those enrolled in online courses may find this workbook to be a great resource to supplement course textbooks. Researchers in the field, particularly those new to quantitative methods and statistical software (e.g., SAS or Stata), will find that this book starts at an appropriate level and covers a breadth of needed material with proper depth for a beginner. The workbook will also serve as a great reference to consult after the initial reading.

The workbook provides a solid foundation of statistical methods, allowing math-phobic readers to do basic statistical analyses, know when to consult a biostatistician, understand how to communicate with a biostatistician, and interpret quantitative study findings in the contexts of the hypotheses addressed. Many introductory biostatistics books spend considerable time explaining statistical theory, but what students and researchers really need to know is how to apply these theories in practice. This workbook walks readers through just that, becoming a lifelong reference.

This workbook covers the basics—from descriptive statistics to regression analysis—providing a survey of topics, including probability, diagnostic testing, probability distributions, estimation, hypothesis testing (one-sample, two-sample, means proportions, nonparametric, and categorical), correlation, regression (linear and logistic), and survival analysis. Examples are used to teach readers how to conduct analyses and interpret the results. There is no fluff or extra verbiage. The workbook provides readers with exactly what they need to know and shows them how to apply their knowledge to a problem.

The workbook not only provides the reader with an introduction to statistical methods but also a step-by-step how-to guide for using SAS and Stata statistical software packages to apply these methods to data, using lots of practical hands-on examples.

Statistical package: A collection of statistical programs that describe data and perform various statistical tests on the data.

Some of the most widely used statistical packages include the following:

- SAS—used in this book
- R
- Stata—used in this book
- SPSS
- MATLAB®
- Mathematica
- Minitab
- Excel

In addition, this workbook provides a solid foundation with concisely written text and minimal reading required. Although it is designed for an academic course, the workbook can be used as a self-help book that allows the user to learn by doing. The real-world practical examples show the user how to place results in context and that outcomes of analysis do not always go the way that the researcher predicts.

The workbook walks the readers through problems, both by hand and with statistical software. Readers can learn how the software performs the calculations, and they can gain the ability to read and interpret SAS and Stata output. The SAS and Stata code provided in the workbook provide readers with a solid foundation from which to start other analyses and apply to their own datasets.

General overview

What is statistics?

Statistics

1. "The science whereby inferences are made about specific, random phenomena on the basis of relatively limited sample material."[1]
2. "The art of learning from data. It is concerned with the collection of data, their subsequent description, and their analysis, which often lead to the drawing of conclusions."[2]

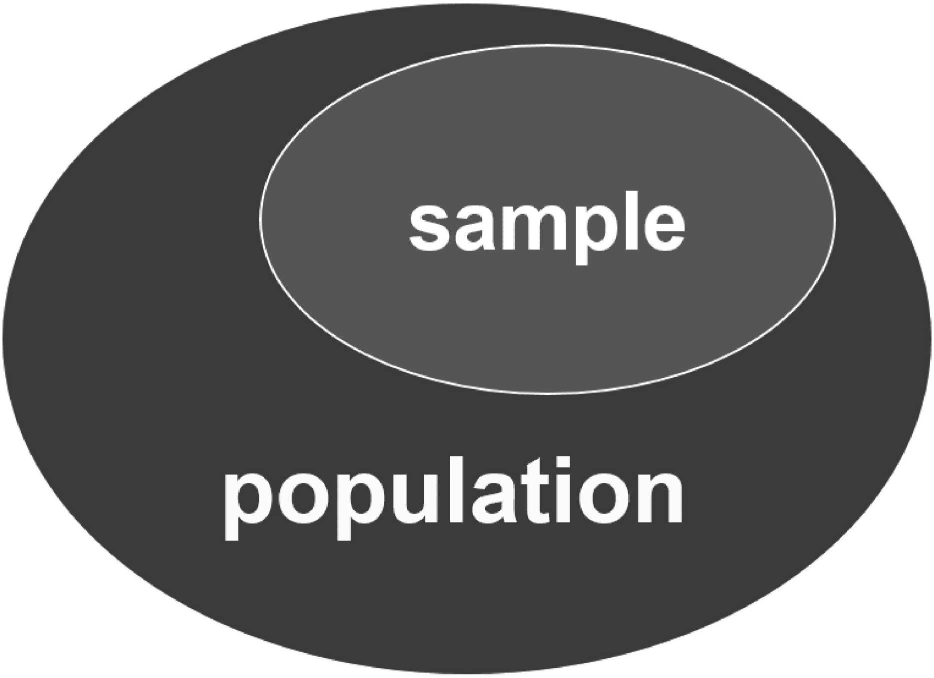
The two main branches of statistics

1. ***Mathematical statistics:*** The branch of statistics concerned with the development of new methods of statistical inference and requires detailed knowledge of abstract mathematics for its implementation.
2. ***Applied statistics:*** The branch of statistics involved with applying the methods of mathematical statistics to specific subject areas such as medicine, economics, and public health.

A ***Biostatistics:*** The branch of applied statistics that applies statistical methods to medical, biological, and public health problems. The study of biostatistics explores the collection, organization, analysis, and interpretation of numerical data.

Basic problem of statistics

Consider a sample of data $x_1, x_2, \ldots, x_n$ where x_1 corresponds to the first sample point and x_n corresponds to the nth sample point. Presuming that the sample is drawn from some population P, what inferences or conclusions can be made about P from the sample? (See figure below.)



Data

Data are often used to make key decisions. Such decisions are said to be *data driven*. With the use of technology, we are able to collect, merge, and store large amounts of data from multiple sources. Data are often the core of any public health or clinical research study, which often starts with a research question and the collection and analysis of data to answer that question. It is important for researchers and practitioners to understand how to collect (or extract from other sources), describe, and analyze data. Furthermore, the ability to understand and critique data and methods used by others is increasing in importance as the volume of research studies increases while the quality remains inconsistent.

Units and variables

In most instances, a dataset is structured as a data matrix with the unit of analysis (e.g., research participants, schools, research papers) in the rows and the variables in the columns.

- ***Units (cases):*** The research participants or objects for which information is collected.
- ***Variables:*** Systematically collected information on each unit or research participant.

Types of variables

- ***Nominal:*** Unordered categories (e.g., male, female).
- ***Ordinal:*** Ordered categories (e.g., mild, moderate, severe).
- ***Ranked:*** Data transformation where numerical or ordinal values are replaced by the rank when the data are sorted (e.g., top five causes of death, top three favorite movies).
- ***Discrete:*** Has a finite number of values where both ordering and magnitude are important (e.g., number of accidents, number of new AIDS cases in a one-year period).
- ***Continuous:*** Has an infinite number of possible values between its minimum and maximum values (e.g., volume of tumor, cholesterol level, time).

In this book, we will discuss ways to describe and analyze a single variable and the relationship between two or more variables. We demonstrate and walk through key principles by hand and supplement this with the use of a statistical package. A statistical package does what the user tells it to do. It is important to understand key concepts so that you can arrive at accurate outcomes from a software package, read statistical output, and properly interpret the results. In Chapter 1, we discuss methods for describing sample data, and, in the rest of the chapters, we discuss ways of analyzing data to test hypotheses.

References

1. Rosner B. *Fundamentals of Biostatistics.* 8th ed. Boston, MA: Cengage Learning. 2016.
2. Ross SM. *Introductory Statistics.* 2nd ed. Burlington, MA: Elsevier Academic Press. 2005.

1 Descriptive statistics

This chapter will focus on descriptive statistics and will include the following:

- Measures of central tendency (measures of location)
- Measures of spread (measures of dispersion)
- Measures of variability
- Graphic methods
- Outliers and standard distribution rules

Terms

- arithmetic mean (average)
- bar graph
- box plot
- Chebyshev Inequality
- decile
- descriptive statistics
- empirical rule
- geometric mean
- GIS map
- histogram
- interquartile range
- median
- mode
- outlier
- percentiles
- quartile
- quintile
- range
- scatterplot
- standard deviation
- stem-and-leaf plot
- tertile
- variance

Introduction

A complete statistical analysis of data has several components. A good first step in data analysis is to describe the data in some concise way, which allows the data analyst a chance to learn about the data being considered. ***Descriptive statistics*** is the part of statistics that is concerned with the description and summarization of data. Initial descriptive analysis quickly provides the researcher an idea of the principal trends and suggests where a more detailed look is necessary. The measures used in describing data include measures of central tendency, spread, and the variability of the sample. All of these measures can be represented in both tabular and graphic displays. We will go over different types of graphs and displays in this chapter.

Measures of central tendency (Measures of location)

One type of measure that is useful for summarizing data defines the center, or middle, of the sample. Thus, this type of measure is called a measure of central tendency (also "measure of location"). Several measures of central tendency exist, but four measures of central tendency will be discussed in this section:

1 Arithmetic mean (average)
2 Median
3 Mode
4 Geometric mean

Arithmetic Mean: The sum of all observations divided by the number of observations.

- The arithmetic mean is what is commonly referred to as an *average*.
- This is the most widely used measure of central tendency. However, the arithmetic mean, or average, is oversensitive to extreme values, meaning that the mean can be influenced by a value that is much higher or much lower as compared to other values in the dataset.
- We use the notation μ to denote the mean of a population and $\bar{x}$ to denote the mean of a sample.

Equation 1.1 shows the calculation of the arithmetic mean.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \ldots + x_n}{n} \tag{1.1}$$

Because the mean is based on summation, knowing several properties of summation is often useful as you begin to analyze data, specifically as the data relate to the mean.

BOX 1.1 PROPERTIES OF THE SAMPLE MEAN

Equation 1.2 shows the multiplicative property of summations:

$$\sum_{i=1}^{n} cx_i = c\sum_{i=1}^{n} x_i \tag{1.2}$$

Three important properties of the arithmetic mean:

1. If $y_i = x_i + c$ where i = 1,…,n then $\bar{y} = \bar{x} + c$
2. If $y_i = cx_i$ where i = 1,…,n then $\bar{y} = c\bar{x}$
3. If $y_i = c_1x_i + c_2$ where i = 1,…,n then $\bar{y} = c_1\bar{x} + c_2$

Median: The median is the value in the middle of the sample variable such that 50% of the observations are greater than or equal to the median and 50% of the observations are less than or equal to the median.

- The median is an alternate measure of central tendency (measure of location) and is second to the arithmetic mean in familiarity.
- The median is useful in data that have outliers and extreme values; the median is insensitive to these values.
- Calculation of the median uses only the middle points in a sample and is less sensitive to the actual numeric values of the remaining data points.

Calculation: Suppose that there are n observations in a sample and that these observations are ordered from smallest (1) to largest (n). The sample median is defined as follows:

If n is odd,

$$\text{Median} = \text{the}\left(\frac{n+1}{2}\right)^{th} \text{observation}$$

If n is even,

$$\text{Median} = \text{the average of the}\left(\frac{n}{2}\right)^{th} \text{plus the}\left(\frac{n}{2}+1\right)^{th} \text{observations}$$

EXAMPLE PROBLEM 1.1

Calculate the arithmetic mean and median of Sample 1.
Sample 1: 2.15, 2.25, 2.30

To find the mean, we add all the values and divide the sum by n, which equals 3.

$$\text{Mean} = \frac{2.15+2.25+2.30}{3} = \frac{6.7}{3} = 2.23$$

To find the median, we put all values in numerical order and find the $\left(\frac{3+1}{2}\right)^{th}$ = 2nd value.

Median = 2nd value of Sample 1 (2.15, 2.25, 2.30) = 2.25.

PRACTICE PROBLEM 1.1

Calculate the arithmetic mean and median of Sample 2.
Sample 2: 2.15, 2.25, 2.30, 2.60

Mode: The mode is the observation that occurs most frequently.

- This measure of central tendency (measure of location) is not a useful measure if there are a large number of possible values, each of which occurs infrequently.
- Some distributions can have more than one mode. We can classify a distribution by the number of modes in the data.

- If there is one mode, the distribution is unimodal. For example, in the following sequence of numbers, there is one mode because 7 appears the most out of any data value:
  - 1 2 3 5 7 7 7 8 8 9
- If there are two modes, the distribution is bimodal. For example, the following sequence of numbers has two modes because 5 and 6 both appear the most out of any data value in the sequence:
  - 2 3 4 5 5 6 6 7 10
- If there are three modes, the distribution is trimodal. For example, the following sequence of numbers has three modes because 1, 2, and 6 appear the most out of any data value in the sequence:
  - 1 1 2 2 5 6 6 8 9

Arithmetic mean versus median

Because the mean is sensitive to outliers and extreme values, it is important to determine when to use the arithmetic mean versus the median. The distribution of the data is a key factor in making this decision.

Arithmetic mean

For a symmetric distribution, the arithmetic mean is approximately the same as the median.
For a positively skewed distribution, the arithmetic mean tends to be larger than the median.
For a negatively skewed distribution, the arithmetic mean tends to be smaller than the median.

Median

If the distribution is symmetric, the relative position of the points on each side of the sample median is the same. The mean or median can be used to describe this sample.
If the distribution is positively skewed (skewed to the right), the points above the median tend to be farther from the median in absolute value than points below the median. This is sometimes referred to as "having a heavy right tail."
If a distribution is negatively skewed (skewed to the left), points below the median tend to be farther from the median in absolute value than points above the median. This is sometimes referred to as "having a heavy left tail."

See Figure 1.1 for a demonstration of the relationship between the arithmetic mean and the median and the skewed versus nonskewed distributions. The mode is also represented in the symmetric distribution on the figure.

Geometric Mean: The geometric mean is the antilogarithm of $\overline{\log x}$ (see Equation 1.3).

This measure of central tendency is not often used in practice but can be useful when dealing with biological or environmental data that are based on concentrations (e.g., biomarkers, blood lead levels, C-reactive protein [CRP], cortisol).

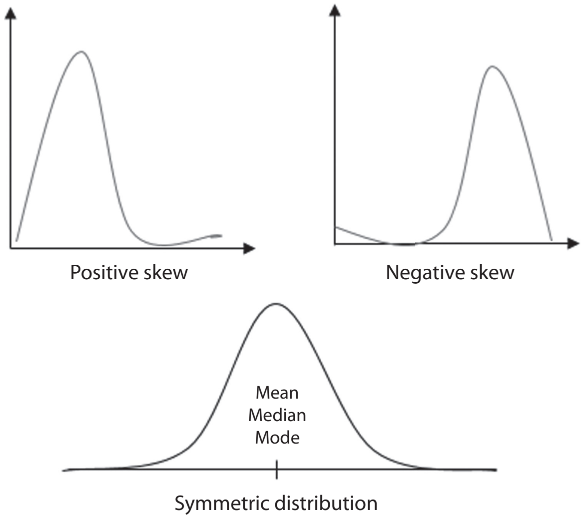


Figure 1.1 **Positively and negatively skewed distribution.** In the distribution with a positive skew (top left) and the distribution with a negative skew (top right), the median is a better measure of central tendency than the mean. The mean more accurately captures the central tendency of the data when the distribution is symmetric (center bottom) with thin tails versus when the data is skewed. The mean follows the tail of a skewed distribution. The relationship between the sample mean and the sample median can be used to assess the symmetry of a distribution.

Equation 1.3 shows the $\overline{\log x}$ calculation.

$$\overline{\log x} = \frac{1}{n}\sum_{i=1}^{n} \log x_i \qquad (1.3)$$

Example of geometric mean using data from the 2014 National Health and Nutrition Examination Survey

The 2014 National Health and Nutrition Examination Survey (NHANES) measures participants' blood lead levels.[1] Using these data, we computed the geometric mean for the participants. The geometric mean for blood lead levels for the entire group of 2014 participants with nonmissing data was 0.83 ug/mL compared to the arithmetic mean of 1.1 ug/mL. We also categorized participants into age groups and computed the geometric mean of blood lead levels by age category (see Figure 1.2).

EXAMPLE PROBLEM 1.2

Calculate the geometric mean for the dataset in Sample 1 from Example Problem 1.1.

Log (2.15) = 0.33
Log (2.25) = 0.35
Log (2.30) = 0.36

$$\overline{\log x} = \frac{0.33+0.35+0.36}{3} = \frac{1.04}{3} = 0.3467$$

Geometric mean = antilog(0.3467) = $10^{0.3467}$ = 2.22

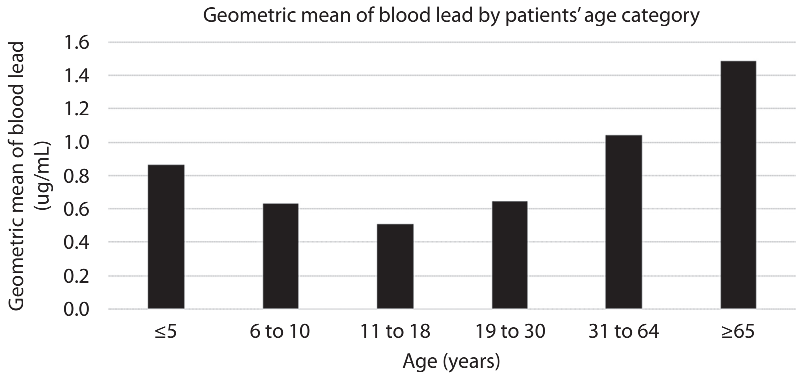


Figure 1.2 **Geometric mean of blood lead levels by patients' age category for participants of the 2014 National Health and Nutrition Examination Survey (NHANES).** (Data from the National Health and Nutrition Examination Survey, Centers for Disease Control and Prevention website, https://wwwn.cdc.gov/Nchs/Nhanes/Search/nhanes13_14.aspx, accessed February 23, 2016.)

PRACTICE PROBLEM 1.2

Calculate the geometric mean of Sample 2.

Measures of spread

Many variables can be well described by a combination of a measure of central tendency (measure of location) and a measure of spread. Measures of spread tell us how far or how close together the data points are in a sample. Six measures of spread will be discussed in this section:

1 Range
2 Quantiles
3 Percentiles
4 Interquartile range
5 Variance
6 Standard deviation

Range: The range is the difference between the largest and smallest observations of a variable.

The range measures the spread of a variable as the distance from the minimum to the maximum value. Although the range is very easy to compute, it is sensitive to extreme observations. The range depends on the sample size (n). The larger the n, the larger the range tends to be. This makes it hard to compare ranges from datasets of different sizes.

Quantiles and Percentiles: Quantiles and percentiles are measures of spread that are determined by the numerical ordering of the data. They are cut points that divide the frequency distribution into equal groups, each containing the same fraction of the total population.

Quantiles and percentiles can also be used to describe the spread of a variable. They are less sensitive to outliers and are not greatly affected by the sample size, which is an advantage over the range.

- The most useful percentiles are often determined by the sample size and subject matter.
- The p^{th} percentile is the value V_p such that p percent of the sample points is less than or equal to V_p.
- The median, which we learned about previously in this chapter, is the 50th percentile.
- Other frequently used percentiles include the following:
  - ***Tertiles:*** Two points that divide and order a sample variable into three categories, each containing a third of the population (e.g., high, medium, low). The 33rd and 66th percentiles of a sample variable are used to categorize it into tertiles.
  - ***Quartiles:*** Three points that divide and order a sample variable into four categories, each containing a fourth of the population. The 25th, 50th, and 75th percentiles of a variable are used to categorize it into quartiles.
  - ***Quintiles:*** Four points that divide and order a sample variable into five categories, each containing a fifth of the population. The 20th, 40th, 60th, and 80th percentiles of a variable are used to categorize it into quintiles.
  - ***Deciles:*** Nine points that divide and order a sample variable into ten categories, each containing a tenth of the population. The 10th, 20th, 30th, 40th, 50th, 60th, 70th, 80th, and 90th percentiles of a variable are used to categorize it into deciles.

The p^{th} percentile is defined by the following:

1. The $(k+1)th$ largest sample point if $np/100$ is not an integer. (k is the largest integer less than $np/100$).
2. The average of the $\left(\frac{np}{100}\right)^{th}$ and the $\left(\frac{np}{100}+1\right)^{th}$ largest observations if $np/100$ is an integer.

Interquartile Range: The interquartile range is between the 75th and 25th percentiles. This measure of spread encompasses the middle 50% of observations and is not easily influenced by extreme values.

EXAMPLE PROBLEM 1.3

Calculate the range and interquartile range using the dataset in Sample 1 in Example Problem 1.1.

$$\text{Range} = 2.30 - 2.15 = 0.15$$

$$75^{th}\ p = \frac{np}{100} = \frac{3\times 75}{100} = 2.25 \rightarrow V_{75} = 2.30$$

$$25^{th}\ p = \frac{np}{100} = \frac{3\times 25}{100} = 0.75 \rightarrow V_{25} = 2.15$$

$$V_{75} - V_{25} = 2.30 - 2.15 = 0.15$$

The range is equal to 0.15, and the interquartile range is also equal to 0.15.

PRACTICE PROBLEM 1.3

Calculate the range and interquartile range for the dataset in Sample 2 from Practice Problem 1.1.

Measures of variability

- Measures of variability are measures of spread that tell us how varied our data points are from the average of the sample.
- Measures of variability include the variance and the standard deviation (S).

Variance: The variance is a measure that will inform us how values in our dataset differ from the mean.

- A high variance indicates a wide range of values; a low variance indicates that values are closer to the mean.
- Equation 1.4 shows us how to calculate the variance.

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \tag{1.4}$$

where x_i is the sample observation for variable X, $\bar{x}$ is the mean of X, and n is the number of units (observations) in the sample. Variance is also noted as σ^2 in some instances; S^2 refers to a sample variance, and σ^2 refers to a population variance.

Standard Deviation: The standard deviation is the square root of the variance.

- The standard deviation is similar to the variance in that it also tells us how much variation around the mean there is in the dataset. The standard deviation is more commonly used than the variance because, in many cases, the standard deviation has the same units as the mean, whereas the variance has units squared, which can complicate interpretation.
- Equation 1.5 shows us how to calculate the standard deviation. Similar to the notation of variance, in some instances, the standard deviation will be noted as σ, which refers to the population standard deviation. When the standard deviation for a sample is being referred to, S is typically used.

$$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n=1}} = \sqrt{Variance} \tag{1.5}$$

BOX 1.2 PROPERTIES OF THE SAMPLE VARIANCE

1. Suppose there are two samples, $x_1,\ldots,x_n$ and $y_1,\ldots,y_n$, where $y_i = x_i + c$, $i = 1,\ldots,n$, and S_x^2 and S_y^2 are the respective variances, then $S_x^2 = S_y^2$.
2. Suppose there are two samples, $x_1,\ldots,x_n$ and $y_1,\ldots,y_n$, where $y_i = cx_i$, $i = 1,\ldots,n$, and S_x^2 and S_y^2 are the respective variances, then $S_y^2 = c^2S_x^2$ and $S_y = cS_x$ if $c > 0$ or $S_y = -cS_x$ if $c < 0$.

EXAMPLE PROBLEM 1.4

Calculate the variance and standard deviation of the dataset in Sample 1 in Example Problem 1.1.

$$S^2 = \frac{(2.15-2.23)^2 + (2.25-2.23)^2 + (2.30-2.23)^2}{3-1}$$

$$= \frac{0.0064+0.0004+0.0049}{2} = \frac{0.0117}{2} = 0.00585$$

$$S = \sqrt{0.00585} = 0.076$$

PRACTICE PROBLEM 1.4

Calculate the variance and standard deviation of Sample 2 from Practice Problem 1.1.

Grouped Data: Grouped data refers to data that are given as a count of observations within an interval of a continuous variable.

Suppose that data were given to us in a frequency table, such as the data in Table 1.1, but we wanted to know the mean and variance of those data. Table 1.1 contains a pair of frequency distributions of weight (lbs.) for a group of people who reported asthma and a group of people who reported no asthma. These data are from the 2014 Behavioral Risk Factor Surveillance System (BRFSS)[2] data and are limited to participants residing in the state of Missouri.

Note that the continuous measure of weight for each subject is not given; instead, the weight of each subject is within one of three ranges of weight. Suppose that you want to calculate the mean and standard deviation for these subjects. You can do this by using the formulas for the grouped mean and grouped variance.

Grouped mean

Equation 1.6 shows us how to calculate the grouped mean.

$$\bar{x} = \frac{\sum_{i=1}^{k} m_i f_i}{\sum_{i=1}^{k} f_i} \tag{1.6}$$

where k is the number of intervals, m_i is the midpoint of the i^{th} interval, and f_i is the frequency associated with the i^{th} interval.

Table 1.1 Asthma status by weight group

Weight group, lbs	*Weight group midpoint*	*Asthma (n = 890)*	*No asthma (n = 5834)*
70–150	110	206	1786
151–205	178	394	2699
206–730	468	290	1349

Source: Behavioral Risk Factor Surveillance System (BRFSS), Centers for Disease Control and Prevention, http://www.cdc.gov/brfss/annual_data/annual_2014.html, 2014. Accessed February 23, 2016.

Note: The 2014 Behavioral Risk Factor Surveillance System (BRFSS) survey asked participants if they were "(ever told) you had asthma?" The response values were *yes*, *no*, *don't know/not sure*, or *refused*. Those who answered the latter two responses were excluded from the data presented in Table 1.1. Survey participants were also asked, "About how much do you weigh without shoes?" Those who did not give weight were excluded from the data presented in Table 1.1. Also, weights given in kilograms were converted to pounds. The data presented in Table 1.1 are limited to residents of the state of Missouri.[2]

BOX 1.3 SIDE NOTE ON GROUPED MEAN

Side Note: $\sum_{i=1}^{k} f_i = n,$ where n is the number of observations in the sample.

Side Note: Zero is an integer; it is not a positive integer.

Grouped variance

Equation 1.7 shows us how to calculate grouped variance.

$$S^2 = \frac{\sum_{i=1}^{k} (m_i - \bar{x})^2 f_i}{\left(\sum_{i=1}^{k} f_i\right) - 1} \tag{1.7}$$

To get the grouped standard deviation, we could simply take the square root of the grouped variance.

EXAMPLE PROBLEM 1.5

Calculate the grouped mean and grouped standard deviation for the group with asthma (Table 1.1).

$$\bar{x} = \frac{110(206) + 178(394) + 468(290)}{890} = \frac{22660 + 70132 + 135720}{890}$$

$$= \frac{228512}{890} = 256.8\ lbs$$

$$S^2 = \frac{(110 - 256.8)^2 \times 206 + (178 - 256.8)^2 \times 394 + (468 - 256.8)^2 \times 290}{890 - 1}$$

$$= \frac{(-146.8)^2 \times 206 + (-78.8)^2 \times 394 + (211.2)^2 \times 290}{889}$$

$$= \frac{21550.24 \times 206 + 6209.44 \times 394 + 44605.44 \times 290}{889}$$

$$= \frac{4439349.44 + 2446519.36 + 1295577.6}{889}$$

$$= 22296.3\ lbs^2$$

$$S = \sqrt{22296.3}$$

$$= 149.3\ lbs$$

PRACTICE PROBLEM 1.5

Calculate the grouped mean and grouped standard deviation for the group without asthma (Table 1.1).

Types of graphs

- Graphic displays provide a quick overall impression of the data, which is sometimes difficult to obtain with numeric measures.
- Several types of graphs are shown in this section, including bar graphs, histograms, stem-and-leaf plots, scatterplots, box plots, and GIS maps.

What makes a good graphic or tabular display?

To make a good graphic or tabular display, the material should be as self-contained as possible and should be understandable without the need for additional text. These attributes require clear labeling, including the title, units, and axes on graphs or figures. The statistical terms used in tables and figures should be well defined. Keep in mind these important attributes of good displays.

Bar Graph: A widely used method for displaying grouped data, a bar graph is a pictorial representation of a frequency distribution for either nominal or ordinal data.

- Nominal data are represented by categories that have no order. An example of nominal data is type of food. We could have categories of spicy, bland, or flavorful.
- Ordinal data are also represented in categories; however, the categories are ordered. An example of ordinal data is price category of food. The dishes can be cheap, affordable, moderately expensive, or pricey.

With nominal or ordinal data, the bar graph is constructed by dividing the data into groups. For each group, a rectangle is constructed with a base of a constant width and a height proportional to the frequency within that group. The main problem with using a bar graph to present nominal or ordinal data is that the sense of the actual sample points in the respective groups is lost.

Figure 1.3 shows an example of a bar graph with nominal data. In this graph, we can see that family practitioners perform the majority of care for most of the respondents in the sample, and general practitioners/internists are in a close second.

Histogram: The most commonly used type of graph, the histograms depicts the symmetry and spread of a single variable. It shows the frequency distribution for discrete or continuous data. It allows for identification of intervals with high levels of frequency, gaps in the data, and values that are far from others.

- *Discrete* or *continuous* are ways to classify numerical data (see also Chapter 4).
- Whereas discrete variables can have only whole number values, continuous variables can have decimals.

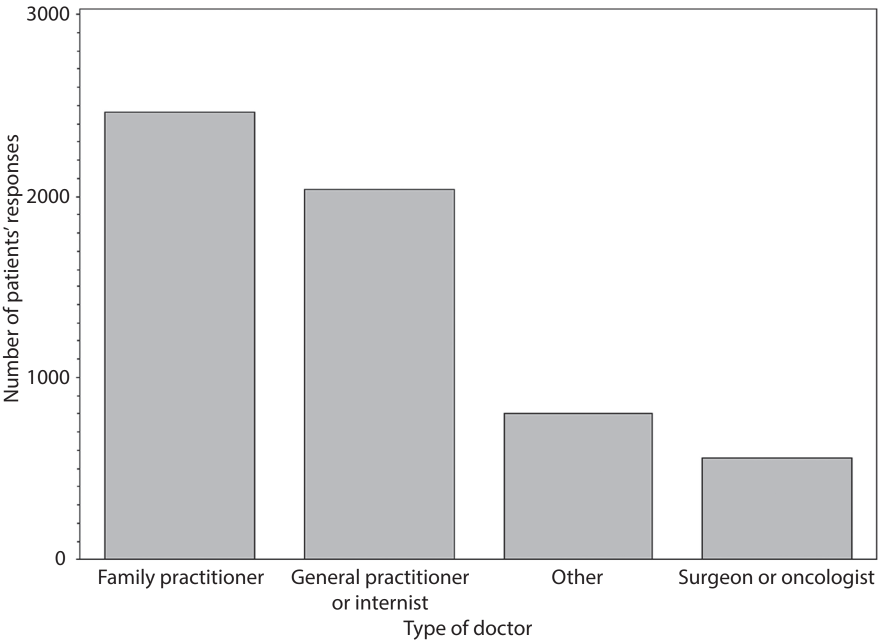


Figure 1.3 **Type of doctor providing the majority of care per patients' responses.** Data used to create the figure are from participants' responses to the question, "What type of doctor provides the majority of your healthcare?" The responses were *cancer surgeon, family practitioner, general surgeon, gynecologic oncologist, general practitioner/internist, plastic surgeon/reconstructive surgeon, medical oncologist, radiation oncologist, urologist, other,* or *don't know/not sure*. Some did not answer the question. Note that *urologist* and *other* were combined to make the *other* category in the figure. Don't know/not sure and abstentions were excluded from the figure. (Data from the Behavioral Risk Factor Surveillance System (BRFSS), Centers for Disease Control and Prevention website, http://www.cdc.gov/brfss/annual_data/annual_2014.html, 2014, accessed February 23, 2016.)

- Examples of discrete data include the following:
  - The number of Zika virus cases in a state. These would be discrete numerical data because we can have only a whole number of Zika cases (e.g., 12 Zika cases). It would not be possible to have 11.6 Zika cases in a state.
  - The number of live births at a hospital in a given year.
  - The temperature of the water in a river on a given day.
- Examples of continuous data include the following:
  - Concentration of lead in the blood of a patient.
  - Parts per million of water contamination.
  - Length of time a cancer patient survives after diagnosis.
- When there are a large number of discrete possibilities, the line between discrete and continuous data can be blurry. Rounding of measurement can also make continuous data appear discrete.

On the basis of the histogram in Figure 1.4, would you say the distribution of age in this sample is skewed or symmetric?[3–6]

This is a random sample that is based on a modest sample size (n = 93) from a real dataset; nevertheless, the distribution is fairly symmetric. However, the left tail is heavier than the right tail, suggesting a left skew.

Stem-and-Leaf Plot: In this type of plot, the actual sample values are preserved, and a grouped display of the data is presented. The collection of leaves takes on the general shape of the distribution of sample points. The stem-and-leaf plot can be used to overcome problems with bar graphs. From the stem-and-leaf plot, it is easy to compute the median and other quantiles.

To construct a stem-and-leaf plot, follow these steps:

1. Separate each data point into a stem component and a leaf component.
   a. The stem component consists of the number formed by all but the rightmost digit.
   b. The leaf component consists of the rightmost digit.
2. Write the smallest stem in the dataset in the upper left-hand corner of the plot.
3. Write the second stem, first stem + 1, below the first stem.
4. Continue with Step 3 until you reach the largest stem in the dataset.
5. Draw a vertical bar to the right of the column of stems.
6. For each number in the dataset, find the appropriate stem, and write the leaf to the right of the vertical bar.

The data[3–6] on age of the sample of patients from Figure 1.4 were used to create the stem-and-leaf plot in Figure 1.5.

EXAMPLE PROBLEM 1.6

Create a stem-and-leaf plot for Sample 3, which consists of the following observed data points: 234, 235, 243, 246, 247, 248, 250, 263, 274, 275.

To create the stem-and-leaf plot pictured in Figure 1.6, we first start by separating all of the data points into stem-and-leaf components. Here, we have stems of 23, 24, 25, 26, and 27. Next, we place the stems in numerical order—which they already happen to be in—starting with the top of the left column labeled *Stem*. Now, we draw our

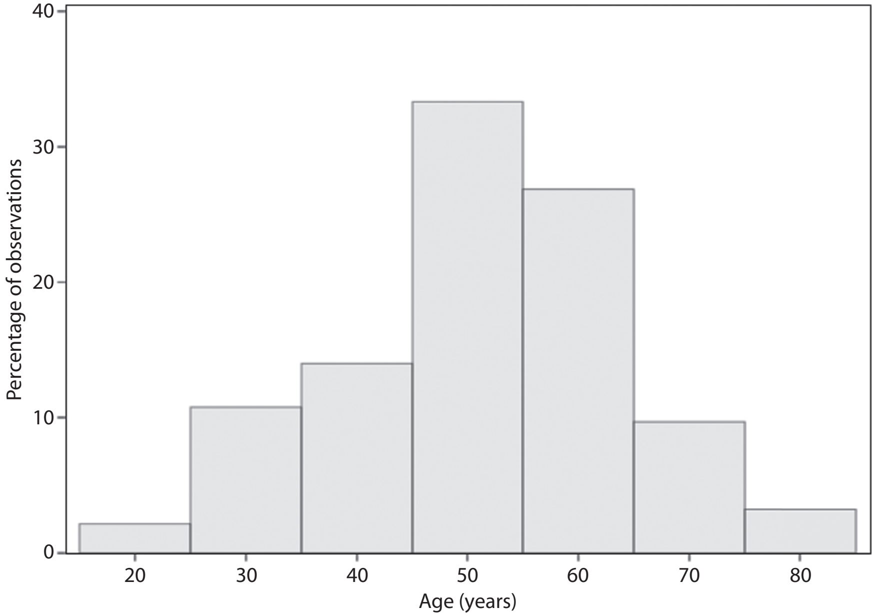


Figure 1.4 **Age of a random sample (n = 93) of primary care clinic patients seen between July 2013 and April 2014.** The data are from a random sample of 93 participants in a study conducted in a primary care clinic in St. Louis, MO. Participants were recruited from the waiting room of the clinic, and they completed the survey in English. A total of 1380 patients participated in the study. (From Goodman, M. S., Griffey, R. T., Carpenter, C. R., Blanchard, M., and Kaphingst, K. A., *J Am Board Fam Med.*, 28(5), 584–594, 2015, doi:10.3122/jabfm.2015.05.150037.)

Stem	Leaf
2	0376
3	011233445557789
4	01223456788999
5	000000111222222233344444455666666777889
6	0001222244456667
7	00127
8	23

Figure 1.5 **Stem-and-leaf plot of age of sample of primary care clinic patients.** The data are from the same random sample of 93 patients (from a total of 1380 patients) represented in Figure 1.4. (From Goodman, M. S., Griffey, R. T., Carpenter, C. R., Blanchard, M., and Kaphingst, K. A., *J Am Board Fam Med.*, 28(5), 584–594, 2015, doi:10.3122/jabfm.2015.05.150037.)

vertical bar to the right of the stem values and fill in the leaf for each data point in the column we label *Leaf*.

PRACTICE PROBLEM 1.6

Create a stem-and-leaf plot for Sample 4, which consists of the following observed data points: 348, 345, 342, 347, 374, 369, 350, 368, 364, 358, 359, 365, 373, 354, 352, 358, 372, 366, 364, 355.

Stem	Leaf
23	45
24	3678
25	0
26	3
27	45

Figure 1.6 **Stem-and-leaf plot of Sample 3.** The stem-and-leaf plot shows the distribution of the data in Sample 3 from Example Problem 1.6.

Box Plots: Box plots show the relationship among the median, upper quartile (75th percentile), and lower quartile (25th percentile) to demonstrate the skewness of a distribution. The box represents the middle 50% of observations (interquartile range) with the lower end of the box at the 25th percentile and the upper end of the box at the 75th percentile. The line in the middle of the box is the median (50th percentile). From the bottom of the box is a line with a horizontal bar at the 5th percentile, and from the top of the box is a line with a horizontal bar at the 95th percentile.

If the distribution is symmetric, the 75th and 25th percentiles should be approximately equally spaced from the median.
If the 75th percentile is farther from the median than the 25th percentile, the distribution is positively skewed.
If the 25th percentile is farther from the median than the 75th percentile, the distribution is negatively skewed.

Box plots will also show the presence of outliers. In the box plot in Figure 1.7, we can see that there is one outlier in the non-Hispanic white category. This is represented by the point above the upper horizontal bar representing the 95th percentile.

- Does the box plot in Figure 1.7 give you a sense of the mean of the BMI in the two groups?
  - Yes, the mean is represented by the small diamonds inside the box plots.
- Does the box plot in Figure 1.7 give you a sense of the median BMI in the two groups?
  - Yes, the median is represented by the bars through the middle of the boxes.
- Does the box plot in Figure 1.7 give you a sense of the mode(s) of BMI in the two groups?
  - No, we do not get a sense of the mode from a box plot.
- How would you compare the distribution of BMI among non-Hispanic white participants and non-Hispanic black participants based on Figure 1.7?
  - The distributions appear similar between the two groups. The BMI distribution for non-Hispanic blacks has a larger spread than the BMI distribution for non-Hispanic whites.

Scatterplots: Scatterplots represent data by depicting single points for each observation.

- There are two types of scatterplots: a one-way scatterplot and a two-way scatterplot.
  - A *one-way scatterplot* uses a single horizontal axis to display the relative position of each data point in a group. This plot is not often used in practice.
  - A *two-way scatterplot* is used to depict the relationship between two continuous measurements. Each point on the graph represents a pair of values.

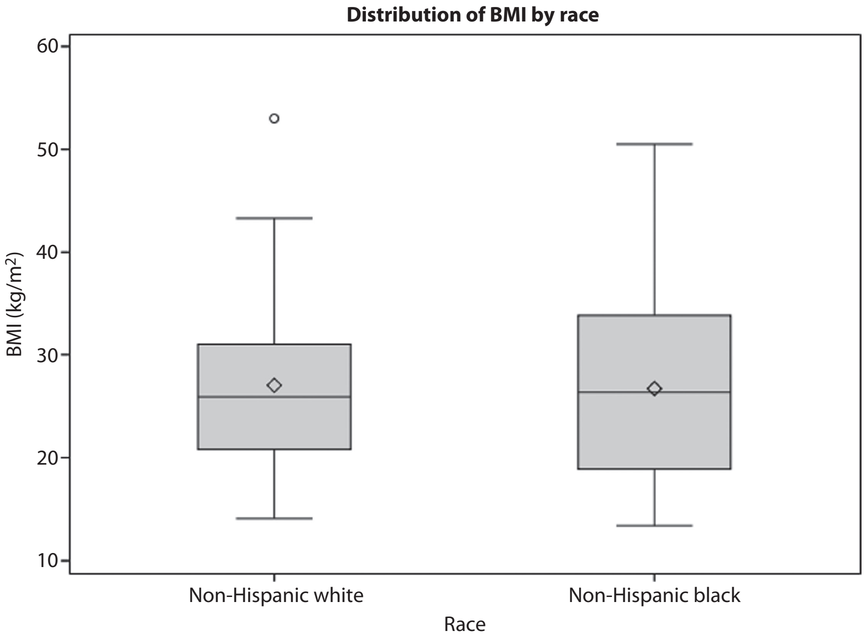


Figure 1.7 **Distribution of BMI by race.** BMI indicates body mass index. The box plot shows BMI by race for a random sample (n = 100) of non-Hispanic white and non-Hispanic black participants in the 2014 NHANES survey. (Data from the National Health and Nutrition Examination Survey, Centers for Disease Control and Prevention website, https://wwwn.cdc.gov/Nchs/Nhanes/Search/nhanes13_14.aspx, accessed February 23, 2016.)

Figure 1.8 shows an example of a two-way scatterplot. The two continuous variables being represented in the graph are waist circumference and body mass index.

Is there a relationship between the two variables?

We can see, on the basis of the scatterplot, that as body mass index goes up, waist circumference also goes up.

GIS Map: Graphical information system (GIS) maps show data that have some kind of geospatial attribute.

- There are two main types of geospatial data: Vector data and raster data.
  - Types of vector data:
    - Points, such as longitude and latitude
    - Polygons, such as city boundaries
    - Lines, such as streets
  - Types of raster data:
    - Elevations
    - Temperature
- Figure 1.9 shows an example of a GIS map from a study on playground safety in the neighborhoods of St. Louis, Missouri.[7]

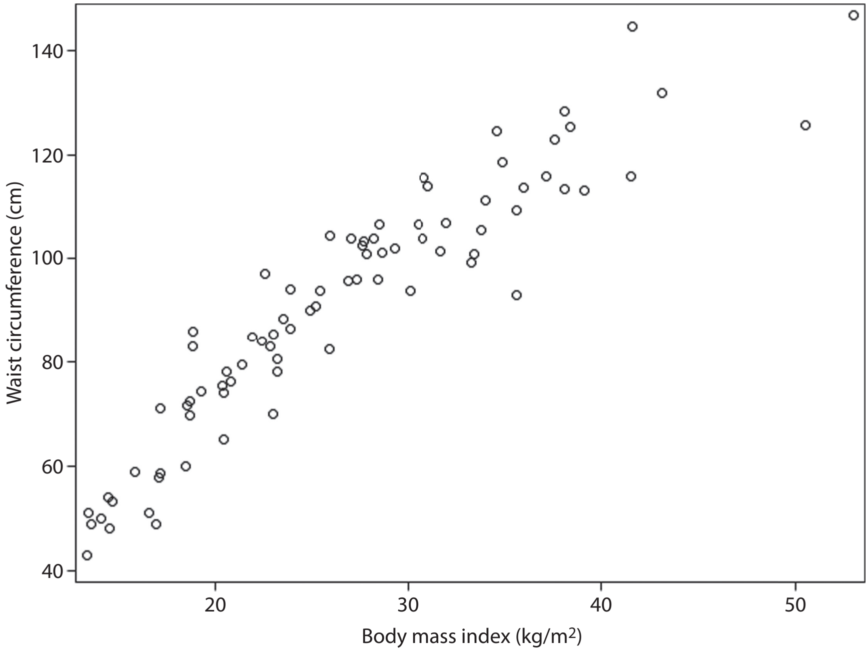


Figure 1.8 **Two-way scatterplot of BMI by waist circumference for a random sample (*n* = 100) of non-Hispanic white and non-Hispanic black participants in the 2014 NHANES Survey.** NHANES indicates National Health and Nutrition Examination Survey. The body mass index (BMI) variable was calculated using participants' measured weight and height. The waist circumference variable is the participants' measured waist circumference in centimeters. All measures were collected by trained health professionals. (Data from the National Health and Nutrition Examination Survey, Centers for Disease Control and Prevention website, https://wwwn.cdc.gov/Nchs/Nhanes/Search/nhanes13_14.aspx, accessed February 23, 2016.)

Outliers and standard distribution rules

Outlier: An outlying value is a value x such that either of the following is true:
(1) $x >$ upper quartile + 1.5*(interquartile range) or
(2) $x <$ lower quartile – 1.5*(interquartile range)

- An extreme outlying value is a value x such that either of the following is true:
  (1) $x >$ upper quartile + 3.0*(interquartile range) or
  (2) $x <$ lower quartile – 3.0*(interquartile range)
- Outliers can influence measures of data and should be considered when presenting descriptive statistics.

Empirical Rule: The empirical rule describes a unimodal and symmetric distribution using the mean and standard deviation.

When dealing with a unimodal and symmetric distribution, the empirical rule states that

- *Mean* ± 1 *standard deviation* covers approximately 67% of observations.
- *Mean* ± 2 *standard deviation* covers approximately 95% of observations.
- *Mean* ± 3 *standard deviation* covers approximately all observations.

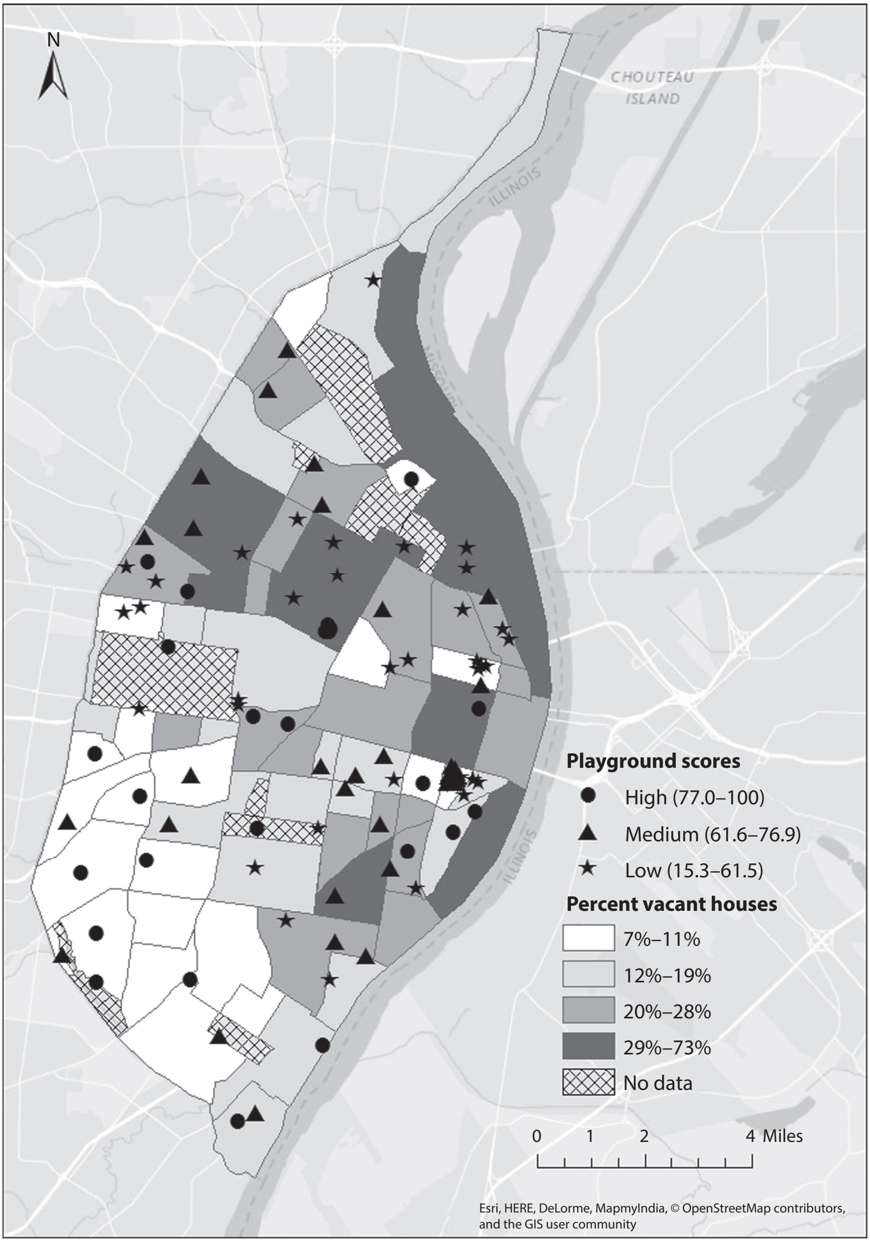


Figure 1.9 **GIS map from playground safety study in St. Louis, MO.** The graphic information system (GIS) map shows levels of playground safety in neighborhoods in St. Louis, MO, and the percent of vacant houses in each neighborhood. (Data on playground scores from Arroyo-Johnson, C., Woodward, K., Milam, L., Ackermann, N., Goodman, M. S., and Hipp, J. A., J Urban Heal., 93(4), 627-638, 2016, doi:10.1007/s11524-016-0063-8; data on percentage of vacant housing from Neighborhood Data Gateway, RISE Community Development, http://www.datagateway.org, accessed on June 15, 2015. The cartography is licensed as CC BY-SA. For more information visit http://openstreetmap.org, http://open datacommons.org, and http://creativecommons.org.)

Table 1.2 Practice Problem 1.7 blank table

k	*Left limit*	*Right limit*	*Empirical rule, %*	*Chebyshev, %*	*Actual, %*
1					
2					
3					

Note: This table should be filled in with the appropriate values based on the data provided in Practice Problem 1.7 and Figure 1.5.

Chebyshev Inequality: The Chebyshev Inequality states that at least $1-\left(\frac{1}{k}\right)^2$ of the x observations lies within k standard deviations of the mean.

- The Chebyshev Inequality makes no assumptions about the distribution of the variable. That is, when we have a dataset that does not have a unimodal and symmetric distribution, the Chebyshev Inequality still applies, whereas the empirical rule is no longer valid.
- The Chebyshev Inequality still holds for unimodal and symmetric data but is more conservative than the empirical rule.

PRACTICE PROBLEM 1.7

Fill in Table 1.2 using the information provided.

The mean age of the study participants discussed in Figures 1.4 and 1.5 is 51.3 years with a standard deviation of 12.7 years.[3–6] Use Figure 1.5 to fill in the column of Table 1.2 with the heading *Actual*.

$$\bar{x} \pm k \times (SD)$$

PRACTICE PROBLEM 1.8

Suppose that the scale for a dataset is changed by multiplying each observation by a positive constant, c.

A What is the effect on the median?

B What is the effect on the mode?

C What is the effect on the geometric mean?

D What is the effect on the range?

E What is the effect on the arithmetic mean?

F What is the effect on the variance?

G What is the effect on the standard deviation?

PRACTICE PROBLEM 1.9

State whether each of the following observations is an example of discrete or continuous data.

A The number of Zika diagnoses in Puerto Rico during May of 2016.

B The weight of a newborn infant.

C The concentration of albumin in a sample of blood.

D The number of children injured on playgrounds in 2015.

Table 1.3 Pulse rate by shortness of breath status

Pulse rate, 60-second pulse rate	
Shortness of breath on stairs/inclines	*No shortness of breath on stairs/inclines*
76	76
76	68
72	64
62	84
74	68
82	58
70	82
80	74
80	58
64	74

Source: National Health and Nutrition Examination Survey, Centers for Disease Control and Prevention, https://wwwn.cdc.gov/Nchs/Nhanes/Search/nhanes13_14.aspx. Accessed February 23, 2016.

Note: The data are from a stratified random sample of 20 participants from the 2014 National Health and Nutrition Examination Survey. Participants over age 40 were asked, "{Have you/Has SP} had shortness of breath either when hurrying on the level or walking up a slight hill?" Respondents who answered *don't know* or who refused were excluded from this sample. In a separate setting, participants also had their pulse rate measured by a trained health technician.

Table 1.4 Practice Problem 1.10 descriptive statistics blank table

Descriptive statistics	*Shortness of breath*	*No shortness of breath*
Median		
Mean		
Range		
Interquartile range		
Standard deviation		

Note: Find the descriptive statistics named in the rows for each group named in the columns. This table should be solved using the data presented in Table 1.3.

PRACTICE PROBLEM 1.10

In Table 1.3 are data from a stratified random sample of participants in the 2014 NHANES survey.[1] The first column lists pulse rates, measured as beats per 60 seconds, for those who reported a shortness of breath while on stairs or an incline. The second column lists pulse rates for those who reported no shortness of breath while on stairs or an incline.

Fill in Table 1.4 (descriptive statistics table).

A Is the typical pulse rate per minute larger for those who experience a shortness of breath when walking up stairs or an incline or for those who do not?

B Which of the two groups has a greater amount of variability in the measurement?

References

1. National Health and Nutrition Examination Survey. Centers for Disease Control and Prevention website. https://wwwn.cdc.gov/Nchs/Nhanes/Search/nhanes13_14.aspx. Accessed February 23, 2016.
2. Behavioral Risk Factor Surveillance System (BRFSS). Centers for Disease Control and Prevention website. http://www.cdc.gov/brfss/annual_data/annual_2014.html. Published 2014. Accessed February 23, 2016.
3. Fan JH, Lyons S, Goodman MS, Blanchard MS, Kaphingst KA. Relationship between health literacy and unintentional and intentional medication nonadherence in medically underserved patients with type 2 diabetes. *Diabetes Educ*. 2016;42(2):199–208. doi:10.1177/0145721715624969.
4. Gilbert KL, Elder K, Lyons S, Kaphingst K, Blanchard M, Goodman M. Racial composition over the life course: Examining separate and unequal environments and the risk for heart disease for African American men. *Ethn Dis*. 2015;25(3):295–304. doi:10.18865/ed.25.3.295.
5. Goodman MS, Griffey RT, Carpenter CR, Blanchard M, Kaphingst KA. Do subjective measures improve the ability to identify limited health literacy in a clinical setting? *J Am Board Fam Med*. 2015;28(5):584–594. doi:10.3122/jabfm.2015.05.150037.
6. Seo J, Goodman MS, Politi M, Blanchard M, Kaphingst KA. Effect of health literacy on decision-making preferences among medically underserved patients. *Med Decis Mak*. 2016;(314):1–7. doi: 10.1177/0272989X16632197.
7. Arroyo-Johnson C, Woodward K, Milam L, Ackermann N, Goodman MS, Hipp JA. Still separate, still unequal: Social determinants of playground safety and proximity. *J Urban Heal*. 2016;93(4):627–638. doi:10.1007/s11524-016-0063-8.

Descriptive statistics

National Health and Nutrition Examination Survey . Centers for Disease Control and Prevention website. https://wwwn.cdc.gov/Nchs/Nhanes/Search/nhanes13_14.aspx. Accessed February 23, 2016.

Behavioral Risk Factor Surveillance System (BRFSS) . Centers for Disease Control and Prevention website. http://www.cdc.gov/brfss/annual_data/annual_2014.html. Published 2014. Accessed February 23, 2016.

Fan JH , Lyons S , Goodman MS , Blanchard MS , Kaphingst KA . Relationship between health literacy and unintentional and intentional medication nonadherence in medically underserved patients with type 2 diabetes. Diabetes Educ. 2016;42(2):199208. doi:10.1177/0145721715624969.

Gilbert KL , Elder K , Lyons S , Kaphingst K , Blanchard M , Goodman M . Racial composition over the life course: Examining separate and unequal environments and the risk for heart disease for African American men. Ethn Dis. 2015;25(3):295304. doi:10.18865/ed.25.3.295.

Goodman MS , Griffey RT , Carpenter CR , Blanchard M , Kaphingst KA . Do subjective measures improve the ability to identify limited health literacy in a clinical setting? J Am Board Fam Med. 2015;28(5):584594. doi:10.3122/jabfm.2015.05.150037.

Seo J , Goodman MS , Politi M , Blanchard M , Kaphingst KA . Effect of health literacy on decision-making preferences among medically underserved patients. Med Decis Mak. 2016;(314):17. doi: 10.1177/0272989X16632197.

Arroyo-Johnson C , Woodward K , Milam L , Ackermann N , Goodman MS , Hipp JA . Still separate, still unequal: Social determinants of playground safety and proximity. J Urban Heal. 2016;93(4):627638. doi:10.1007/s11524-016-0063-8.

Lab A: Introduction to SAS

Emergency Room MICA. Jefferson City, MO: Bureau of Health Care Analysis & Data Dissemination, Missouri Department of Health and Senior Services. 2014. http://health.mo.gov/data/mica/EmergencyRoomMICA/. Updated August 30, 2016. Accessed June 15, 2016.

National Health and Nutrition Examination Survey. Hyattsville, MD: Centers for Disease Control and Prevention, National Center for Health Statistics. 2014. https://wwwn.cdc.gov/Nchs/Nhanes/Search/nhanes13_14.aspx. Accessed February 23, 2016.

Lab A: Introduction to Stata

Emergency Room MICA. Jefferson City, MO: Bureau of Health Care Analysis & Data Dissemination, Missouri Department of Health and Senior Services. 2014. http://health.mo.gov/data/mica/EmergencyRoomMICA/. Updated August 30, 2016. Accessed June 15, 2016.

National Health and Nutrition Examination Survey. Hyattsville, MD: Centers for Disease Control and Prevention, National Center for Health Statistics. 2014. https://wwwn.cdc.gov/Nchs/Nhanes/Search/nhanes13_14.aspx. Accessed February 23, 2016.

Probability

HIV Risk Behaviors . Centers for Disease Control and Prevention website. http://www.cdc.gov/hiv/risk/estimates/riskbehaviors.html. Published 2015.

United States Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, Division of Vital Statistics . Linked Birth/Infant Death Records 20072013, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program, on CDC WONDER Online Database. http://wonder.cdc.gov/lbd-current.html. Accessed July 7, 2016.

Arroyo-Johnson C , Mincey KD , Ackermann N , Milam L , Goodman MS , Colditz GA . Racial and ethnic heterogeneity in self-reported diabetes prevalence trends across Hispanic subgroups, National Health Interview Survey, 19972012. Prev Chronic Dis. 2016;13(E10). doi:10.5888/pcd13.150260.

Centers for Disease Control and Prevention . National Hospital Ambulatory Medical Care Survey: 2011 Emergency Department Summary Tables. 2011. https://www.cdc.gov/nchs/data/ahcd/nhamcs_emergency/2011_ed_web_tables.pdf.

Hing E , Decker SL , Jamoom E . Acceptance of New Patients with Public and Private Insurance by Office-Based Physicians: United States, 2013. Hyattsville, MD: National Center for Health Statistics. 2015. http://www.cdc.gov/nchs/data/databriefs/db195.pdf.

Diagnostic testing/screening

Lyons SA , Kaphingst KA , Goodman MS . Validating self-reported survey measures using SAS. Paper presented at SAS 2014 Global Forum; March 2326, 2014; Nashville, TN. http://support.sas.com/resources/papers/proceedings14/1675-2014.pdf. Accessed August 4, 2016.

Warusevitane A , Karunatilake D , Sim J , Smith C , Roffe C . Early diagnosis of pneumonia in severe stroke: Clinical features and the diagnostic role of C-reactive protein. PLoS One. 2016;11(3):e0150269. doi:10.1371/journal.pone.0150269.

Hidalgo B , Kaphingst KA , Stafford J , Lachance C , Goodman MS . Diagnostic accuracy of self-reported racial composition of residential neighborhood. Ann Epidemiol. 2015;25(8):597604. doi:10.1016/j.annepidem.2015.04.003.

Semper AE , Broadhurst MJ , Richards J et al. Performance of the GeneXpert Ebola assay for diagnosis of Ebola virus disease in Sierra Leone: A field evaluation study. PLoS Med. 2016;13(3):e1001980. doi:10.1371/journal.pmed.1001980.

Deschasaux M , Souberbielle J-C , Andreeva VA et al. Quick and easy screening for vitamin D insufficiency in adults: A scoring system to be implemented in daily clinical practice. Medicine (Baltimore). 2016;95(7). doi:10.1097/MD.0000000000002783.

Uppara M , Adaba F , Askari A et al. A systematic review and meta-analysis of the diagnostic accuracy of pyruvate kinase M2 isoenzymatic assay in diagnosing colorectal cancer. World J Surg Oncol. 2015;13:48. doi:10.1186/s12957-015-0446-4.

106 SEER stat fact sheets: Colon and rectum cancer . National Cancer Institute website. http://seer.cancer.gov/statfacts/html/colorect.html. Accessed August 9, 2016.

U.S. and world population clock . United States Census Bureau website. http://www.census.gov/popclock/. Accessed August 9, 2016.

Mettlin C , Murphy GP , Babaian RJ et al. The results of a five-year early prostate cancer detection intervention. Investigators of the American Cancer Society National Prostate Cancer Detection Project. Cancer. 1996;77(1):150159. doi:10.1002/(sici)1097-0142(19960101)77:1<150::aid-cncr25>3.0.co;2-3.

Discrete probability distributions

Pagano M , Gauvreau K . Principles of Biostatistics. Vol 2. Pacific Grove, CA: Duxbury. 2000.

Pratt LA , Brody DJ . Depression in the U.S. Household Population, 20092012. Hyattsville, MD: National Center for Health Statistics. 2014. NCHS Data Brief 172.

Arroyo-Johnson C , Mincey KD , Ackermann N , Milam L , Goodman MS , Colditz GA . Racial and ethnic heterogeneity in self-reported diabetes prevalence trends across Hispanic subgroups, National Health Interview Survey, 19972012. Prev Chronic Dis. 2016;13(E10). doi:10.5888/pcd13.150260.

Injury Prevention & Control: Opioid Overdose: State Data . Centers for Disease Control and Prevention, National Center for Injury Prevention and Control, Division of Unintentional Injury Prevention. http://www.cdc.gov/drugoverdose/data/statedeaths.html. Published 2016. Accessed April 2, 2017.

Quealy K , Sanger-Katz M . Compare these gun death rates: The U.S. is in a different world. New York Times. June 13, 2016. http://www.nytimes.com/2016/06/14/upshot/compare-these-gun-death-rates-the-us-is-in-a-different-world.html?_r=0. Accessed August 16, 2016.

124 Alcohol Facts and Statistics . National Institute on Alcohol Abuse and Alcoholism. https://www.niaaa.nih.gov/alcohol-health/overview-alcohol-consumption/alcohol-facts-and-statistics. Published 2016. Accessed April 2, 2017.

Epidemiology of the IBD . National Center for Chronic Disease Prevention and Health Promotion, Centers for Disease Control and Prevention. http://www.cdc.gov/ibd/ibd-epidemiology.htm. Published 2015.

Continuous probability distributions

National Health and Nutrition Examination Survey. Hyattsville, MD: Centers for Disease Control and Prevention, National Center for Health Statistics. https://wwwn.cdc.gov/Nchs/Nhanes/Search/nhanes13_14.aspx. Published 2014. Accessed February 23, 2016.

Arroyo-Johnson C , Woodward K , Milam L , Ackermann N , Goodman MS , Hipp JA . Still separate, still unequal: Social determinants of playground safety and proximity. J Urban Heal. 2016;93(4):627638. doi:10.1007/s11524-016-0063-8.

Quealy K , Sanger-Katz M . Compare these gun death rates: The U.S. is in a different world. New York Times. June 13, 2016. http://www.nytimes.com/2016/06/14/upshot/compare-these-gun-death-rates-the-us-is-in-a-different-world.html?_r=0. Accessed August 16, 2016.

Lab B: Probability distributions

World Health Organization . Global Health Observatory Data Repository: Exclusive breastfeeding under 6 months. http://apps.who.int/gho/data/node.main.1100?lang=en. Updated April 2, 2015. Accessed September 29, 2016.

Adams D , Fullerton K , Jajosky R et al. Summary of notifiable infectious diseases and conditionsUnited States, 2013. MMWR Morb Mortal Wkly Rep. 2015;62(53):1122. doi:10.15585/mmwr.mm6253a1.

National Health and Nutrition Examination Survey. Hyattsville, MD: Centers for Disease Control and Prevention, National Center for Health Statistics. 2014. https://wwwn.cdc.gov/Nchs/Nhanes/Search/nhanes13_14.aspx. Accessed February 23, 2016.

Estimation

Webb Hooper M , Antoni MH , Okuyemi K , Dietz NA , Resnicow K . Randomized controlled trial of group-based culturally specific cognitive behavioral therapy among African American smokers. Nicotine Tob Res. 2016. doi:10.1093/ntr/ntw181.

Heatherton TF , Kozlowski LT , Frecker RC , Fagerstrom KO . The Fagerstrom Test for Nicotine Dependence: A Revision of the Fagerstrom Tolerance Questionnaire. Br J Addict. 1991;86:11191127. doi:doi:10.1111/j.1360-0443.1991.tb01879.x.

National Health and Nutrition Examination Survey. Hyattsville, MD: Centers for Disease Control and Prevention, National Center for Health Statistics. https://wwwn.cdc.gov/Nchs/Nhanes/Search/nhanes13_14.aspx. Published 2014. Accessed February 23, 2017.

173 Kaphingst KA , Blanchard M , Milam L , Pokharel M , Elrick A , Goodman MS . Relationships between health literacy and genomics-related knowledge, self-efficacy, perceived importance, and communication in a medically underserved population. J Health Commun. 2016;21 Suppl 1:5868. doi:10.1080/10810730.2016.1144661.

U.S. Environmental Protection Agency . National Air Toxics Assessment. https://egis-hud.opendata.arcgis.com/datasets/c7e2c62560bd4a999f0e0b2f4cee2494_0. Published 2005. Accessed May 16, 2017.

Centers for Disease Control and Prevention . Youth Risk Behavior Survey Questionnaire. www.cdc.gov/yrbs. Published 2013. Accessed April 11, 2017.

Centers for Disease Control and Prevention . Behavioral Risk Factor Surveillance System. http://www.cdc.gov/brfss/annual_data/annual_data.htm. Published 2014. Accessed April 3, 2017.

One-sample hypothesis testing

National Health and Nutrition Examination Survey. Hyattsville, MD: Centers for Disease Control and Prevention, National Center for Health Statistics. https://wwwn.cdc.gov/Nchs/Nhanes/Search/nhanes13_14.aspx. Published 2014. Accessed February 23, 2016.

Behavioral Risk Factor Surveillance System. Atlanta, GA: Centers for Disease Control and Prevention. http://www.cdc.gov/brfss/annual_data/annual_data.htm. Published 2014. Updated September 11, 2015. Accessed February 23, 2016.

Centers for Disease Control and Prevention . Secondhand smoke facts. http://www.cdc.gov/tobacco/data_statistics/fact_sheets/secondhand_smoke/general_facts/. Updated August 20, 2015. Accessed September 22, 2016.

United States Census Bureau . American Community Survey 2014: Gini index of income inequality, 20092014. http://www.socialexplorer.com/tables/ACS2014_5yr/R11234719?ReportId=R11234719. Accessed March 18, 2016.

United States Census Bureau . American Community Survey 2014: Poverty status in 2014 for population age 65 and over. http://www.socialexplorer.com/tables/ACS2014/R11234722?ReportId=R11234722. Accessed March 18, 2016.

Lab C: One-sample hypothesis testing, power, and sample size

National Health and Nutrition Examination Survey. Hyattsville, MD: Centers for Disease Control and Prevention, National Center for Health Statistics. 2014. https://wwwn.cdc.gov/Nchs/Nhanes/Search/nhanes13_14.aspx. Accessed February 23, 2016.

Centers for Disease Control and Prevention . National Center for Health Statistics. National Health Interview Survey: Tables of summary health statistics, Table A-3. 2014. http://ftp.cdc.gov/pub/Health_Statistics/NCHS/NHIS/SHS/2014_SHS_Table_A-3.pdf.

National Health and Nutrition Examination Survey. Hyattsville, MD: Centers for Disease Control and Prevention, National Center for Health Statistics. 2012. http://wwwn.cdc.gov/nchs/nhanes/search/nhanes11_12.aspx.

Two-Sample Hypothesis Testing

United States Census Bureau . American Community Survey: Health insurance, 2009. http://www.socialexplorer.com/tables/ACS2009/R11239760?ReportId=R11239760. Accessed March 9, 2016.

United States Census Bureau . American Community Survey: Health insurance, 2014. http://www.socialexplorer.com/tables/ACS2014/R11239768?ReportId=R11239768. Accessed March 9, 2016.

Behavioral Risk Factor Surveillance System . 2014 BRFSS Survey Data and Documentation. Atlanta, GA: Centers for Disease Control and Prevention. http://www.cdc.gov/brfss/annual_data/annual_2014.html. September 11, 2015. Accessed February 23, 2016.

Fan JH , Lyons S , Goodman MS , Blanchard MS , Kaphingst KA . Relationship between health literacy and unintentional and intentional medication nonadherence in medically underserved patients with type 2 diabetes. Diabetes Educ. 2016;42(2):199208. doi:10.1177/0145721715624969.

National Health and Nutrition Examination Survey. Hyattsville, MD: Centers for Disease Control and Prevention, National Center for Health Statistics. https://wwwn.cdc.gov/Nchs/Nhanes/Search/nhanes13_14.aspx. Published 2014. Accessed February 23, 2016.

Nonparametric hypothesis testing

World Health Organization . Global Health Observatory Data Repository: Human resources, data by country, 2015. http://apps.who.int/gho/data/node.main.MHHR?lang=en. Updated August 5, 2015. Accessed March 11, 2016.

DAgostino McGowan L , Stafford JD , Thompson VL , Johnson-Javois B , Goodman MS . Quantitative evaluation of the community research fellows training program. Front Pblic Heal. 2015;3(July):112. doi:10.3389/fpubh.2015.00179.

National Health and Nutrition Examination Survey. Hyattsville, MD: Centers for Disease Control and Prevention, National Center for Health Statistics. https://wwwn.cdc.gov/Nchs/Nhanes/Search/nhanes13_14.aspx. Published 2014. Accessed February 23, 2016.

Centers for Disease Control and Prevention . mRFEI data table. http://www.cdc.gov/obesity/downloads/2_16_mrfei_data_table.xls. Accessed October 1, 2014.

Social Security Administration . Data catalog: SSA disability claim data. http://catalog.data.gov/dataset/ssa-disability-claim-data. Updated September 28, 2016. Accessed April 13, 2016.

Lab D: Two-sample (parametric and nonparametric) hypothesis testing

United States Census Bureau . American Community Survey: Hispanic or Latino by race, 20102014. http://www.socialexplorer.com/tables/ACS2014_5yr/R11256228?ReportId=R11256228. Accessed August 17, 2016.

World Health Organization . Global Health Observatory Data Repository: Measles (MCV), data by country, 2015. http://apps.who.int/gho/data/node.main.A826?lang=en. Updated August 25, 2016. Accessed August 18, 2016.

National Health and Nutrition Examination Survey, Hyattsville, MD: Centers for Disease Control and Prevention, National Center for Health Statistics. 2014. https://wwwn.cdc.gov/Nchs/Nhanes/Search/nhanes13_14.aspx. Accessed February 23, 2016.

County Health Rankings & Roadmaps . Years of Potential Life Lost, 2013. http://www.socialexplorer.com/tables/HD2013/R11256222?ReportId=R11256222. Accessed August 22, 2016.
County Health Rankings & Roadmaps . Years of Potential Life Lost, 2016. http://www.socialexplorer.com/tables/HD2016/R11256226?ReportId=R11256226. Accessed August 22, 2016.

Hypothesis testing with categorical data

Behavioral Risk Factor Surveillance System (BRFSS). Atlanta, GA: Centers for Disease Control and Prevention; 2014. http://www.cdc.gov/brfss/annual_data/annual_data.htm. September 11, 2015. Accessed February 23, 2016.
National Health and Nutrition Examination Survey. Hyattsville, MD: Centers for Disease Control and Prevention, National Center for Health Statistics. 2014. https://wwwn.cdc.gov/Nchs/Nhanes/Search/nhanes13_14.aspx. Accessed February 23, 2016.
Fan JH , Lyons S , Goodman MS , Blanchard MS , Kaphingst KA . Relationship between health literacy and unintentional and intentional medication nonadherence in medically underserved patients with type 2 diabetes. Diabetes Educ. 2016;42(2):199208. doi:10.1177/0145721715624969.
Youth Risk Behavior Surveillance System (YRBSS). Atlanta, GA: Centers for Disease Control and Prevention. 2013. www.cdc.gov/yrbs. Updated August 11, 2016. Accessed March 11, 2016.
Centers for Disease Control and Prevention . Multistate outbreak of Salmonella enteritidis infections linked to bean sprouts. http://www.cdc.gov/salmonella/enteritidis-11-14/index.html. Update January 23, 2015. Accessed October 27, 2016.

Analysis of variance (ANOVA)

National Health and Nutrition Examination Survey. Hyattsville, MD: Centers for Disease Control and Prevention, National Center for Health Statistics. 2014. https://wwwn.cdc.gov/Nchs/Nhanes/Search/nhanes13_14.aspx. Accessed August 1, 2016.
Fan JH , Lyons S , Goodman MS , Blanchard MS , Kaphingst KA . Relationship between health literacy and unintentional and intentional medication nonadherence in medically underserved patients with type 2 diabetes. Diabetes Educ. 2016;42(2):199208. doi:10.1177/0145721715624969.
Hamilton BE , Martin JA , Osterman MJK , Curtin SC , Mathews TJ . National Vital Statistics Reports: Births: Final Data for 2014. Hyattsville, MD: National Center for Health Statistics; 2015:64(12). http://www.cdc.gov/nchs/data/nvsr/nvsr64/nvsr64_12.pdf.
State and territorial data . Centers for Disease Control and Prevention website. http://www.cdc.gov/nchs/fastats/state-and-territorial-data.htm. Updated October 6, 2016. Accessed October 10, 2016.
Kochanek K , Murphy SL , Xu J , Tejada-Vera B . National Vital Statistics Reports: Deaths: Final Data for 2014. Hyattsville, MD: National Center for Health Statistics; 2016:65(4). http://www.cdc.gov/nchs/data/nvsr/nvsr65/nvsr65_04.pdf.

Correlation

Mukaka MM . Statistics corner: A guide to appropriate use of correlation coefficient in medical research. Malawi Med J. 2012;24(3):6971. doi:10.1016/j.cmpb.2016.01.020.
Fan JH , Lyons S , Goodman MS , Blanchard MS , Kaphingst KA . Relationship between health literacy and unintentional and intentional medication nonadherence in medically underserved patients with type 2 diabetes. Diabetes Educ. 2016;42(2):199208. doi:10.1177/0145721715624969.
World Health Organization . Global Health Observatory Data Repository: Total fertility rate. http://apps.who.int/gho/data/node.main.HE-1548?lang=en. Updated May 15, 2015. Accessed August 1, 2016.
World Health Organization . Global Health Observatory Data Repository: Demand for family planning satisfied. http://apps.who.int/gho/data/view.main.94330. May 15, 2015. Accessed August 1, 2016.
National Health and Nutrition Examination Survey. Hyattsville, MD: Centers for Disease Control and Prevention, National Center for Health Statistics. 2014. https://wwwn.cdc.gov/Nchs/Nhanes/Search/nhanes13_14.aspx. Accessed August 1, 2016.
World Health Organization . Global Health Observatory Data Repository: Diptheria tetanus toxoid and pertussis (DTP3). http://apps.who.int/gho/data/view.main.80200. Updated July 17, 2015. Accessed August 1,

2016.
World Health Organization . Global Health Observatory Data Repository: Number of under-five deaths. http://apps.who.int/gho/data/node.main.ChildMort-1?lang=en. Updated August 27, 2015. Accessed August 1, 2016.
Youth Risk Behavior Surveillance System. Atlanta, GA: Centers for Disease Control and Prevention. 2013. www.cdc.gov/yrbs. Updated August 11, 2016. Accessed August 1, 2016.

Linear regression

Centers for Disease Control and Prevention . Hospital actions affect breastfeeding. Vital Signs. http://www.cdc.gov/vitalsigns/breastfeeding2015/index.html. Updated October 6, 2015. Accessed September 16, 2016.
Baby-Friendly USA . The ten steps to successful breastfeeding. http://www.babyfriendlyusa.org/about-us/baby-friendly-hospital-initiative/the-ten-steps/. Accessed September 16, 2016.
Centers for Disease Control and Prevention . Maternity Practices in Infant Nutrition and Care (mPINC) Survey. http://www.cdc.gov/breastfeeding/data/mpinc/index.htm. Updated November 4, 2016. Accessed December 10, 2016.
Mathews TJ , MacDorman MF . Infant mortality statistics from the 2010 period linked birth/infant death data set. Natl Vital Stat Reports. 2013;62(8):126. http://www.cdc.gov/nchs/data/nvsr/nvsr62/nvsr62_08.pdf. Accessed December 10, 2016.
Arroyo-Johnson C , Woodward K , Milam L , Ackermann N , Goodman MS , Hipp JA . Still separate, still unequal: Social determinants of playground safety and proximity. J Urban Heal. 2016;93(4):627638. doi:10.1007/s11524-016-0063-8.

Logistic regression

National Health and Nutrition Examination Survey. Hyattsville, MD: Centers for Disease Control and Prevention, National Center for Health Statistics. 2014. https://wwwn.cdc.gov/Nchs/Nhanes/Search/nhanes13{_}14.aspx. Accessed February 23, 2016.
Behavioral Risk Factor Surveillance System. Atlanta, GA: Centers for Disease Control and Prevention. 2014. http://www.cdc.gov/brfss/annual_data/annual_2014.html. Updated September 11, 2015. Accessed February 23, 2016.

Survival analysis

National Health and Nutrition Examination Survey. Hyattsville, MD: Centers for Disease Control and Prevention, National Center for Health Statistics. 2014. https://wwwn.cdc.gov/Nchs/Nhanes/Search/nhanes13_14.aspx. Accessed February 23, 2016.
National Health and Nutrition Examination Survey, NNYFS 2012 . Centers for Disease Control and Prevention, National Center for Health Statistics. https://wwwn.cdc.gov/Nchs/Nhanes/Search/Nnyfs12.aspx. Published 2012. Accessed January 11, 2016.

Lab E: Data analysis project

Vital Statistics Data Available Online. Hyattsville, MD: Centers for Disease Control and Prevention, National Center for Health Statistics. 2016. http://www.cdc.gov/nchs/data_access/vitalstatsonline.htm. Accessed November 22, 2016.