

PRINCIPLES and APPLICATIONS of BIOSTATISTICS

RAY M. MERRILL



PRINCIPLES and APPLICATIONS of BIOSTATISTICS

PRINCIPLES and APPLICATIONS of BIOSTATISTICS

RAY M. MERRILL, PhD, MPH, MS, FAAHB, FACE

Professor of Biostatistics and Epidemiology

Director of Gerontology

Brigham Young University



JONES & BARTLETT
LEARNING



World Headquarters

Jones & Bartlett Learning
25 Mall Road, 6th Floor
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.

Copyright © 2022 by Jones & Bartlett Learning, LLC, an Ascend Learning Company

All rights reserved. No part of the material protected by this copyright may be reproduced or utilized in any form, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without written permission from the copyright owner.

The content, statements, views, and opinions herein are the sole expression of the respective authors and not that of Jones & Bartlett Learning, LLC. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not constitute or imply its endorsement or recommendation by Jones & Bartlett Learning, LLC and such reference shall not be used for advertising or product endorsement purposes. All trademarks displayed are the trademarks of the parties noted herein. *Principles and Applications of Biostatistics* is an independent publication and has not been authorized, sponsored, or otherwise approved by the owners of the trademarks or service marks referenced in this product.

There may be images in this book that feature models; these models do not necessarily endorse, represent, or participate in the activities represented in the images. Any screenshots in this product are for educational and instructive purposes only. Any individuals and scenarios featured in the case studies throughout this product may be real or fictitious but are used for instructional purposes only.

This publication is designed to provide accurate and authoritative information in regard to the Subject Matter covered. It is sold with the understanding that the publisher is not engaged in rendering legal, accounting, or other professional service. If legal advice or other expert assistance is required, the service of a competent professional person should be sought.

24316-1

Production Credits

VP, Product Development: Christine Emerton
Director of Product Management: Matt Kane
Product Manager: Sophie Teague
Content Strategist: Sara Bempkins
Project Manager: Jessica deMartin
Senior Project Specialist: Jennifer Risdén
Digital Project Specialist: Rachel DiMaggio
Senior Marketing Manager: Susanne Walker
Content Services Manager: Colleen Lamy

VP, Manufacturing and Inventory Control: Therese Connell
Composition: Straive
Project Management: Straive
Cover Design: Briana Yates
Media Development Editor: Faith Brosnan
Rights Specialist: Maria Leon Maimone
Cover Image (Title Page, Part Opener, Chapter Opener):
©MicroStockHub/Getty Images
Printing and Binding: McNaughton & Gunn

Library of Congress Cataloging-in-Publication Data

Names: Merrill, Ray M., author.

Title: Principles and applications of biostatistics / Ray M. Merrill. PhD, MPH, MS, FAAHB, FACE,
Professor, Brigham Young University, Department of Public Health.

Description: First edition. | Burlington, Massachusetts : Jones & Bartlett Learning, [2022] | Includes
bibliographical references and index.

Identifiers: LCCN 2021009317 | ISBN 9781284225976 (paperback)

Subjects: LCSH: Biometry. | BISAC: EDUCATION / General

Classification: LCC QH323.5 .M467 2022 | DDC 570.1/5195--dc23

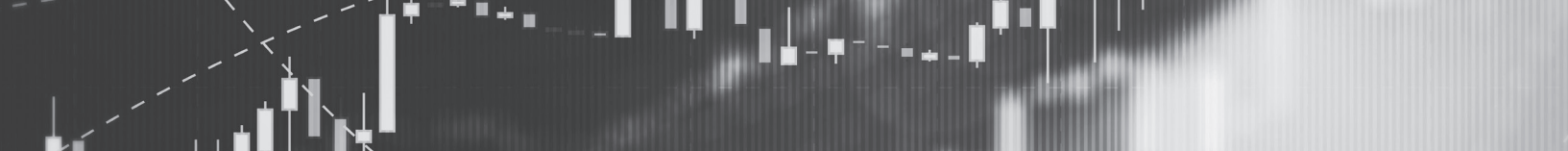
LC record available at <https://lcn.loc.gov/2021009317>

6048

Printed in the United States of America

25 24 23 22 21 10 9 8 7 6 5 4 3 2 1

This book is dedicated with love and appreciation to Marla,
Phillip, Dallin, Andrew, Ron, and Cindy.



Contents

Preface	xiii
About the Author	xv
Reviewers	xvii

CHAPTER 1 Introduction to Biostatistics..... 1

Key Concepts	1
1.1 What Is Biostatistics?	2
1.2 Questions for Biostatistics	2
1.3 Data	3
1.3.1 Data Collection.....	4
1.3.2 Data Analysis	5
1.3.3 Data Presentation	5
1.3.4 Data Interpretation	5
1.4 Methods in Biostatistics	5
1.4.1 Descriptive Biostatistics	6
1.4.2 Probability	6
1.4.3 Inferential Statistics.....	7
1.4.4 Statistical Techniques	8
1.5 Statistical Software	9
1.6 Summary	11
Exercises	12
References	13

CHAPTER 2 Descriptive Summaries of Data..... 15

Key Concepts	15
2.1 Questions for Descriptive Biostatistics	16
2.2 Empirical Frequency Distributions	19
2.2.1 Frequency Distribution.....	19
2.2.2 Relative Frequency	19
2.2.3 Proportion	19
2.2.4 Rate	19
2.2.5 Adjusting Rates.....	20
2.2.5.1 <i>Direct Method of Adjusting Rates</i>	21
2.2.5.2 <i>Indirect Method of Adjusting Rates</i>	22
2.2.6 Ratio.....	23
2.3 Notation	23
2.4 Measures of Central Tendency.....	24
2.4.1 Geometric Mean	24
2.4.2 Arithmetic Mean.....	24

2.4.3 Weighted Mean	25
2.4.4 Median	25
2.4.5 Mode	25
2.4.6 Choosing a Measure of Central Tendency	25
2.5 Measures of Variation	26
2.5.1 Variance and Standard Deviation	26
2.5.2 Degrees of Freedom	27
2.5.3 More on the Standard Deviation	28
2.5.4 Coefficient of Variation	28
2.5.5 Range and Interquartile Range	29
2.6 Describing Data with Statistics and Graphs	29
2.7 Misleading Graphs	33
2.8 Five-Number Summary	34
2.9 Statistical Software	35
2.10 Summary	36
Exercises	37
References	40
CHAPTER 3 Probability	41
Key Concepts	41
3.1 Probability Questions in Biostatistics	42
3.2 Probability Concepts	42
3.2.1 Frequentist Probability	43
3.2.2 Classical Probability	43
3.2.3 Subjective Probability	44
3.3 Properties of Probabilities	44
3.4 Calculating Probabilities	45
3.5 Bayes' Theorem	48
3.6 Evaluating Screening and Diagnostic Tests	51
3.7 Sampling	53
3.8 Summary	55
Exercises	56
References	58
CHAPTER 4 Probability and Sampling Distributions	61
Key Concepts	61
4.1 Probability Distributions for Addressing Questions in Biostatistics	62
4.2 Random Variables	63
4.3 Probability Distributions for Discrete Random Variables	64
4.3.1 Binomial Probability Distributions	66
4.3.2 Poisson Probability Distributions	68
4.4 Probability Distributions for Continuous Random Variables	70
4.4.1 Normal Probability Distribution	71
4.4.2 Normal Approximation to the Binomial	73

4.5 Sampling Distribution	74
4.5.1 Sampling Distribution of the Mean	74
4.5.2 The Central Limit Theorem	76
4.6 t Distribution	77
4.7 Other Sampling Distributions	78
4.7.1 Chi-Square Distribution	78
4.7.2 F Distribution	79
4.8 Finite Population Correction	79
4.9 Summary	79
Exercises	81
References	84

CHAPTER 5 Estimation and Hypothesis Testing . . . 85

Key Concepts	85
5.1 Questions and Hypotheses	86
5.2 Estimators	87
5.3 Interval Estimates	88
5.3.1 Confidence Intervals for Normal Means and Known Variance	88
5.3.2 Confidence Intervals for Normal Means and Unknown Variance	89
5.3.3 Confidence Intervals for the Binomial with Large Sample Sizes	90
5.4 Tests of Hypotheses	91
5.4.1 Type I and Type II Errors	94
5.4.2 p -Value	95
5.5 Hypotheses Involving a Parameter in One Group	96
5.5.1 Hypotheses Involving the Mean in One Group	97
5.5.2 Paired Design: Mean	98
5.5.3 Hypotheses Involving a Proportion in One Group	99
5.5.4 Paired Design: Proportion	101
5.6 Hypotheses Involving Means in Two Independent Groups	104
5.6.1 Testing the Hypothesis $\mu_1 = \mu_2$	104
5.6.2 Testing the Hypothesis $p_1 = p_2$	111
5.7 Chi-Square Distribution	114
5.7.1 Chi-Square for Test of Goodness of Fit	115
5.7.2 Chi-Square for Test of Independence	115
5.7.3 Chi-Square for Test of Independence with Small Numbers	117
5.7.4 Chi-Square for Test of a Single Variance	119
5.8 Evaluating Differences Using Error Bar Graphs	120
5.9 Summary	121
Exercises	124
References	130

CHAPTER 6 Analysis of Variance 133

Key Concepts	133
6.1 Analysis of Variance Concepts and Computations	134
6.2 Two-Sample t Test Following Significant F Test	138
6.3 ANOVA for Completely Randomized Design	144
6.4 Analysis of Covariance (ANCOVA)	148

6.5 ANOVA for Randomized Block Design	149
6.6 Repeated Measures ANOVA	151
6.7 Equality of Proportions and Trends	154
6.7.1 χ^2 Test of Independence for Assessing Three or More Proportions	154
6.7.2 Cochran-Armitage Trend Test	156
6.7.3 Cochran-Mantel-Haenszel Trend Test	157
6.8 Summary	158
Exercises	160
References	165

CHAPTER 7 Measures of Association 167

Key Concepts	167
7.1 Questions About Statistical Associations	168
7.2 Measurement Scales	169
7.3 Nominal Data	171
7.3.1 Risk Ratio	171
7.3.2 Rate Ratio	173
7.3.3 Attributable Fraction in the Population (AF_p)	175
7.3.4 Attributable Fraction in the Exposed Cases (AF_e)	176
7.3.5 Prevalence Ratio	176
7.3.6 Odds Ratio	177
7.3.7 Odds Ratio for Matched Case-Control Study	179
7.4 Mantel-Haenszel Method to Estimate an Adjusted Measure of Association	180
7.5 Continuous Data	185
7.5.1 Correlation Coefficient	186
7.5.2 Spearman Rank Correlation Coefficient	187
7.5.3 Coefficient of Determination	187
7.5.4 Fisher's z -Transformation for Testing the Correlation Coefficient	190
7.5.5 Confidence Interval for a Correlation Coefficient	191
7.6 Comparing Two Correlation Coefficients	191
7.7 Statistical Software	192
7.8 Summary	195
Exercises	198
References	205

CHAPTER 8 Regression 207

Key Concepts	207
8.1 Questions	208
8.2 Regression Function	208
8.3 Simple Linear Regression	209
8.3.1 Least Squares Method	211
8.3.2 Statistical Inference About the Slope	212
8.3.3 Confidence Interval for Mean Y Given a Value of X	215
8.3.4 Prediction Interval for an Individual Value Y Given a Value X	216
8.4 Logistic Regression	217

8.5 Poisson Regression	222
8.6 Log-Binomial Regression	224
8.7 Cox Proportional Hazard Model	225
8.8 Multiple Regression	226
8.8.1 Statistical Inference for Regression Coefficients	227
8.8.2 Evaluating the Model	230
8.8.3 Interaction Term	231
8.9 Multiple Logistic Regression	233
8.10 Multiple Poisson Regression	235
8.11 Multiple Log-Binomial Regression	236
8.12 Multiple Cox Proportional Hazard Model	238
8.13 Polynomial Regression	242
8.14 Summary	244
Exercises	247
References	253

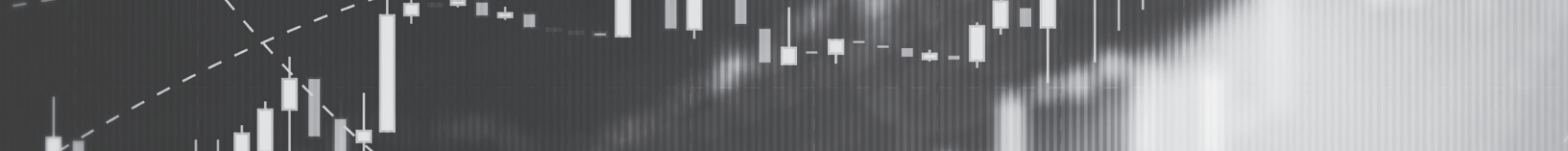
CHAPTER 9 Nonparametric Methods255

Key Concepts	255
9.1 What Is Nonparametric Statistics?	256
9.2 Reasons to Use Nonparametric Statistical Tests	256
9.3 Choosing an Appropriate Statistical Test	257
9.4 Nonparametric Statistical Tests	258
9.5 Kolmogorov-Smirnov Test	258
9.6 Sign Test	260
9.7 Wilcoxon Signed-Rank Test	262
9.8 Mood's Median Test	265
9.9 Mann-Whitney U Test (Also Called Wilcoxon Rank-Sum Test)	266
9.10 Kruskal-Wallis Test	269
9.11 Friedman Test	271
9.12 Kendall Tau-b Correlation Coefficient	272
9.13 Other Selected Tests	273
9.14 Summary	274
Exercises	274
References	277

CHAPTER 10 Survey Research279

Key Concepts	279
10.1 Write a Proposal	281
10.2 Purpose of the Survey	281
10.3 Survey Population	282
10.4 Sample Size	283
10.4.1 Sample Size Calculation	283
10.4.2 Sample Size Calculation for Strata	285
10.4.3 Sample Size Calculation for Clusters	285

10.5	Survey Method	286
10.5.1	Questionnaire or Interview	286
10.5.2	Cross-Sectional or Longitudinal	287
10.5.3	Probability or Nonprobability Sampling	287
10.5.4	Question Types	287
10.5.5	Rating Scales	288
10.6	Cronbach's Alpha	292
10.7	Factor Analysis	295
10.8	Content	298
10.9	Things to Consider When Writing Questions	298
10.9.1	Ask Questions That Can Be Answered	298
10.9.2	Social Desirability Bias	299
10.9.3	Avoid Ambiguity	299
10.9.4	Phrasing of Questions	299
10.10	Layout	300
10.11	Pretest	300
10.12	Conduct the Survey	301
10.12.1	Response Rate	301
10.12.2	Minimizing Errors	302
10.13	Analysis Plan	303
10.14	Report of the Results	303
10.15	Existing Data	303
10.15.1	Secondary Data	304
10.15.2	Ancillary Data	304
10.15.3	Systematic Reviews and Meta-Analysis	305
10.15.3.1	Fixed Effects Model	306
10.15.3.2	Random Effects Model	308
10.16	Summary	309
	Exercises	311
	References	314
Appendix A: Answers to Odd-Numbered Exercises		317
Appendix B: Tables		333
Appendix C: Formula Sheet		347
Appendix D: Detailed Formula Sheet		349
Glossary		353
Index		361



Preface

Statistics is the science of data. It involves collecting, classifying, summarizing, organizing, analyzing, and interpreting data. Data are pieces of information used as a basis for our understanding problems and influencing solutions. In the current era of SARS-Cov-2, we can see that statistics has added to our knowledge of various aspects of the resulting disease, from a description of its frequency and pattern to certain long-term health consequences. A large number of research papers are now being published, rapidly advancing our knowledge of the disease through the help of statistics.

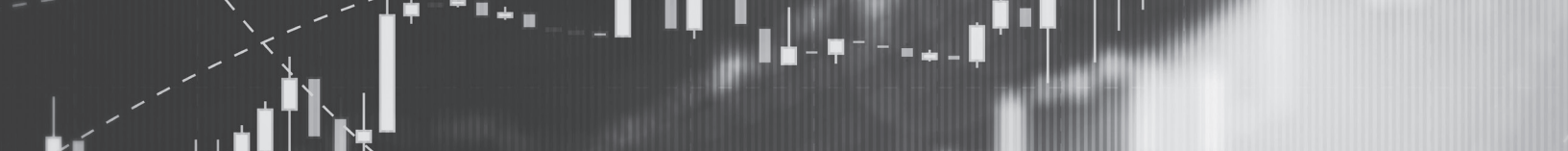
In general, a variety of disciplines rely on statistics: biology, health, medicine, epidemiology, business, psychology, engineering, agriculture, and others. When the focus of statistics is on the biological, health, and medical sciences, it is called biostatistics. Biostatistics involves statistical methods applied to biological data. Biostatistics, along with epidemiology and health services, is the foundation of public health.

All of the primary concepts and methods that are fundamental to our understanding of biostatistics are covered in this text, from data collection and presentation to probability and statistical inference to multiple regression and analysis of variance. Substantive, real-world biological, health, and medical science examples and news files are included, letting the reader see a number of ways biostatistics can help advance knowledge.

This book is written for students and practitioners in the biological, health, and medical sciences. Its primary aim is to present principles and methods of biostatistics in the context of relevant, interesting, and up-to-date applications. The principles and applications are presented with clarity, brevity, and accuracy. Specifically, statistical principles and applications are developed with understandable and focused explanations and practical examples. The aim is to make learning easier for students and pedagogy more effective for instructors.

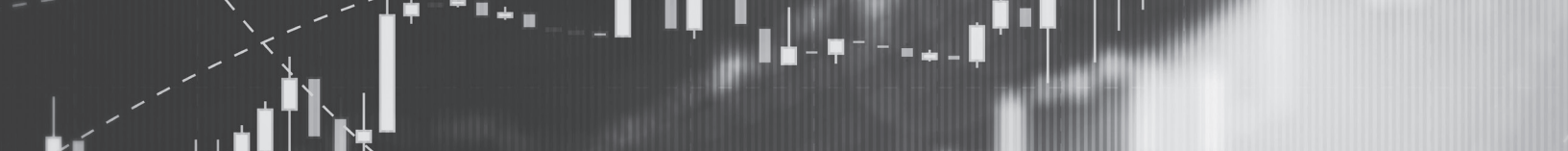
The book is concise. Key topics in biostatistics are capsulized in order to be manageable in a one-semester course. At the beginning of most chapters is a list of the type of questions that can be addressed using the principles and methods covered in that chapter. Highlighted in the various sections are bolded key words, definitions, and concepts. Concepts and formulas are developed with reasoning suitable for application in biostatistics. Focus is given to the conditions that are necessary to apply the methods correctly.

Excel and SAS computer software will be used to demonstrate data analysis and graphing throughout the book. Other computer software options are readily available.



About the Author

Ray M. Merrill is a professor of biostatistics and epidemiology and program director of gerontology at Brigham Young University. He holds adjunct positions in Family and Preventive Medicine and Health Education at the University of Utah. He is a former Cancer Prevention Fellow at the National Cancer Institute and visiting scientist at the International Agency for Research on Cancer, Lyon, France. He is currently a fellow of the American Academy of Health Behavior and the American College of Epidemiology. He teaches introductory and advanced level courses in biostatistics and epidemiology. He is the author of more than 300 peer-reviewed research articles and text books (*Environmental Epidemiology*, *Reproductive Epidemiology*, *Behavioral Epidemiology*, *Principles of Epidemiology Workbook*, *Fundamentals of Epidemiology and Biostatistics*, *Statistical Methods in Epidemiologic Research*, *Fundamental Mathematics for Epidemiology Study*, and *Introduction to Epidemiology*). In his spare time he hikes mountains and runs marathons.



Reviewers

Chakra Budhathoki, PhD

Associate Professor
Johns Hopkins University
School of Nursing
Baltimore, Maryland

Macey Buker, PhD, CPA

Assistant Professor, Program Director
Dixie State University

Susan Draine, EdD, MSN, CNS, MBA, RN**Lenis Chen-Edinboro, PhD**

Assistant Professor in Public Health
University of North Carolina Wilmington

Sarbani Ghoshal, MPH, PhD

Assistant Professor
Queensborough Community College, CUNY

Jayanta Gupta, MD, PhD

Florida Gulf Coast University

Shayesteh Jahanfar, PhD

Professor
Central Michigan University

Maureen K. Johnson, PhD

Associate Professor
Department of Applied Health Sciences
Indiana State University

Seung-Hwan Kim, PhD

Merrimack College

Bernadette McCrory, PhD, MPE, PE, CHFP

Mechanical and Industrial Engineering
Montana State University

Mary Moore, PhD

Colorado State University Global

Ramzi W. Nahhas, PhD

Professor
Department of Population and Public Health Sciences
Department of Psychiatry
Boonshoft School of Medicine,
Wright State University

Oluwatomi Oluwasanmi, MPH

Hunter College Urban School of Public Health

Erin Reynolds, PhD, MPH

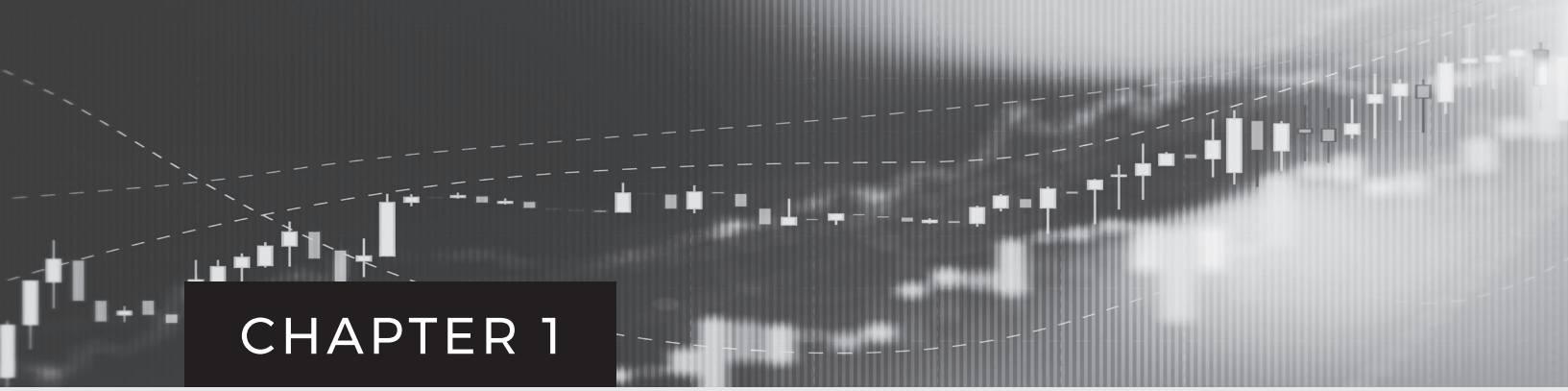
Associate Professor
University of Southern Indiana

Ramona Stone, PhD, MPH

West Chester University of Pennsylvania

Jing Xu, PhD, MHA

University of North Florida



CHAPTER 1

Introduction to Biostatistics

KEY CONCEPTS

- Biostatistics is the branch of statistics concerned with the development and application of methods for collecting, analyzing, and interpreting biological data to assist decision-making.
- Problems in the biological sciences lead to research questions (representing uncertainty to be resolved), which may be addressed using biostatistics.
- Data are information taken from variables.
- A variable is a characteristic that varies from one subject to the next and can take on a specified set of values.
- Quantitative data are information about quantities or numbers.
- Qualitative data are information about the quality, nature, or essence of something.
- Nominal scale refers to data that fall into categories with no logical order or structure.
- Ordinal scale refers to data that fall into categories with an inherent order.
- Interval scale refers to measurements where the difference between the intervals is meaningful, but there is no true definition of zero.
- Ratio scale has the same properties as interval scale, but there is a true zero.
- The four broad areas of biostatistics are descriptive statistics, probability, inference, and statistical techniques.
- Descriptive biostatistics are measures of frequency (e.g., count, percent), central tendency (e.g., mean, median), dispersion (variation) (e.g., standard deviation, range, interquartile range), and position (e.g., percentile, rank) that represent either a sample or a population.
- Probability provides a basis for evaluating the reliability of the conclusions we reach and the inferences we make under uncertainty.
- Inferential statistics draw conclusions about a population based on sample information taken from that population.
- Statistical techniques are analytic approaches that utilize statistical methods to investigate a range of questions.
- Application of statistical software is essential as we address research questions in biostatistics.

This chapter provides an introduction to biostatistics. First, the meaning of the term biostatistics is discussed. Second, the research question and its importance for motivating the application of biostatistics is developed. Third, basic concepts related to data collection, analysis, presentation, and interpretation are covered. Fourth, the four broad areas of biostatistics, upon which this book is organized, are presented: descriptive statistics, probability, inference, and statistical techniques. Fifth, a brief introduction to statistical software and coding is given.

1.1 What Is Biostatistics?

The word statistics is from the German word *statistik*, which derives from the Latin word *statisticum* (“of the state”) and the Italian word *statista* (“statesman,” “politician”). In 1749, it was used to refer to the science dealing with data about the condition of a community or state.¹ Today we refer to statistics as the area of mathematics that involves the collection, analysis, presentation, and interpretation of data. Biostatistics is the science or practice of statistics applied to data that relate to living organisms, which primarily involves human biology, health, and medicine. In other words, it is a contraction of biology and statistics. Biostatistics contributes to a number of fields, including health, medicine, nutrition, genetics, biology, epidemiology, and many more. It is important in the health sciences for a number of reasons, including helping us identify the natural course of a disease; individuals at greatest risk for disease, injury, or death; risk factors for disease, injury, or death; and the value of new drugs, medical procedures, and healthcare interventions.

Biostatistics is the branch of statistics concerned with the development and application of methods for collecting, analyzing, and interpreting biological data to assist decision-making.

Biostatistics finds application in many areas, including clinical trials to evaluate the safety and efficacy of medications or medical devices. In clinical trials, biostatistics plays an important role at all levels: designing the study; determining the sample size; and collecting, analyzing, and interpreting the data. In public health programs, biostatistics is often used to evaluate a program’s efficacy, effectiveness, and potential for replication in other areas. In literature reviews and meta-analyses, biostatistics provides the tools for quantifying patterns and trends, thus showing evidence for medical delivery and treatment practices. There are yet several other examples of application areas for biostatistics, many of which will be discussed in this book.

1.2 Questions for Biostatistics

When a problem is identified in the biological sciences, it leads to a research question or questions. A research question reflects the uncertainty that the investigator wishes to resolve by conducting a study. The application of methods in biostatistics allows us to effectively address questions arising in the biological sciences.

Consider the question, “Does physical activity increase your life expectancy?” A study showed that more leisure-time physical activity leads to longer life expectancy.² Physical activity equivalent to brisk walking for approximately 75 minutes per week resulted in 1.8 more years in life expectancy compared with no physical activity. Brisk walking for 450+ minutes per week resulted in 4.5 additional years in life expectancy compared with being inactive. Improvements in life expectancy with greater physical activity were found across weight classifications (normal, overweight, and obese). Individuals of normal weight with physical activity equivalent to 150+ minutes per week of brisk walking had 7.2 more years in life expectancy compared to those who were physically inactive and obese. Other questions about physical activity and

Table 1.1 Selected Questions Related to COVID-19

What is the frequency and pattern of this disease?
What are the clinical characteristics of this disease?
How do the symptoms vary by age group?
How does population density affect this disease?
How does comorbid health problems affect this disease?
How does nutrition and obesity affect this disease?
How does climate influence the virus?
Who is at the greatest risk?
What is the length of time after exposure before clinical symptoms appear (i.e., incubation period)?
When are infected people most contagious?
How is the disease spread?
How long is an asymptomatic carrier contagious?
What is the probability of reinfection of the disease following recovery?
What is the probability of false negatives and false positives for the test?
How effective is wearing a mask at preventing the spread of the disease?
What is the death-to-case ratio, and how does it vary by selected groups?
What are the long-term health consequences of the disease?
Is there an effective treatment for the disease?
Is there an effective vaccine for the disease?

health have also been addressed in recent years with the application of biostatistics.^{3–7}

With the current COVID-19 pandemic, several research questions are being investigated, some of which are listed in **Table 1.1**. As answers are discovered, the disease can be better managed. For example, identifying the duration of time from exposure to first symptoms (incubation period) for COVID-19 tells us about the possible source of the virus, the appropriate length of time for quarantine, and the time a person can spread the virus prior to knowing they have it. Consider the question about the incubation period for COVID-19. One study found that the median incubation period is 5 days (range 1–14 days).⁸ Incubation periods for several other infectious diseases are listed elsewhere.⁹

The Centers for Disease Control and Prevention has presented a general list of questions and answers related to COVID-19.¹⁰ In the coming months, data will become increasingly more available to address such questions.

1.3 Data

The heart of biostatistics is data. Data are information and may be thought of as observations or measurements of something of interest.

Data are information obtained through observation, experiment, or measurement of a phenomenon of interest.

Data are either quantitative or qualitative.

Quantitative data are information about quantities or numbers, measured on a numerical scale.

Identifying additional years of life associated with increased exercise is an example of quantitative data.

Qualitative data are information about the quality, nature, or essence of something, with information gathered and ordered into larger themes as the researcher works from the specific to the general.

Describing how regular exercise makes people feel is an example of qualitative data.

Qualitative data are also called categorical data when the levels of the variable fit into categories. The levels of categorical data have no logical order. For example, males versus females or disagree, neutral, agree.

Data consist of measurements taken from variables. A variable is a characteristic that varies from one entity to another and can be measured or categorized. A variable can take on a specified set of values. A random variable is any outcome of a variable that occurs by chance, such as for every 1000 COVID-19 patients, how many are hospitalized.

Data measurements are made on different scales: nominal, ordinal, interval, and ratio, listed in **Table 1.2**. The statistical techniques presented and applied in this book depend on the type of data involved.

Interval and ratio data can be divided into one of two types: discrete or continuous. Discrete data have a finite number of measurements based on counts. Continuous data have a theoretically infinite number of measurements; there are no limits in the area between measurements. While discrete data represent things that are counted, continuous data represent things that are measured. Continuous data are more precise and provide more information than discrete data.

1.3.1 Data Collection

Data collection is the systematic approach in which information is gathered and measured in order to obtain an accurate picture of a phenomenon of interest. Data collection involves various methods, depending on whether quantitative or qualitative data are desired. Quantitative data collection methods consist of scales, tests, questionnaires, and surveys, whereas qualitative data collection

Table 1.2 Measurement Scales

Scale	Description	Example
Nominal	Refers to placing data into categories, where there is no logical order or structure	Exposed ("Yes" or "No") Diseased ("Yes" or "No") Sex, race, marital status, educational status
Ordinal	Refers to placing data into categories where the gross order of the categories is informative but the relative positional distances are not quantitatively meaningful	Preference rating (e.g., agree, neutral, disagree) Rank-order scale
Interval	Refers to a measurement where the difference between the intervals is meaningful, but there is no true definition of zero	Temperature, since zero on Fahrenheit or Celsius scales does not mean no temperature Calendar year
Ratio	Has the same properties as interval scale data, but has a clear definition of zero (e.g., height, weight, and blood pressure)	Height, weight, blood pressure Length, duration

methods consist of interviews, observations, and documents. In the physical activity and life expectancy study, leisure-time physical activity was associated with life expectancy according to six large cohort studies comprising 654,827 individuals, ages 21–90 years. Physical activity was measured as metabolic equivalent hours per week (Met-h/week). An approximate 10-year follow-up period occurred in the study with the main outcome, variable life expectancy, measured in years.

1.3.2 Data Analysis

Data analysis is the process of evaluating data to discover useful information. While quantitative data can be statistically analyzed, qualitative data are generally used to build concepts, hypotheses, or theories. In the physical activity and life expectancy study, the data were quantitatively assessed using a statistical technique called proportional hazards regression.

1.3.3 Data Presentation

Data presentation is the method in which we summarize, organize, and communicate information. These methods may be textual, tabular, or graphical. Textual methods include ranking data and the stem-and-leaf plot; tabular methods include the frequency distribution table and contingency table; and graphical methods include the bar chart, histogram, frequency polygon, and **pie chart**. In general, graphs and tables are effective ways to communicate data so that they can be easily understood.

1.3.4 Data Interpretation

Once data are collected, they are analyzed and presented in various forms, such as statistics, tables, graphs, diagrams, maps, and so on. At this point, the findings need to be interpreted. Data interpretation is the process of assessing and determining the meaning and importance of the results. Data interpretation involves consideration of questions about the data as they relate to your study questions. The answer to these questions should be organized as results and conclusions. Implications and recommendations may follow.

Data may be interpreted with perspective by comparing the results among groups or with what is typically known. In clinical trials, a control group is critical for assessing new drugs and medical procedures. The control group is also an effective way to control for confounding factors. Controlling for confounding allows us to have confidence that intervention effects are not explained by extrinsic factors. In the physical activity and life expectancy study, meaning and importance of leisure-time physical activity in terms of greater life expectancy was made by comparing life expectancy among people classified into groups according to physical activity and body weight.

1.4 Methods in Biostatistics

Biostatistics has four general method areas: descriptive, probability, inferential, and statistical techniques. They do not stand alone; for example, descriptive biostatistics often requires the use of probability and statistical techniques; inferential biostatistics relies on probability; and statistical techniques often involve inference.

1.4.1 Descriptive Biostatistics

Descriptive biostatistics involves methods of organizing, summarizing, and describing biological data. Suppose we want to describe our data so that we can answer who is at greatest risk for disease and where is the disease risk greatest. In recent months, several reports have answered questions about the frequency and pattern of COVID-19. As of July 27, 2020, the observed death-to-case ratio for COVID-19 was 15.2% in the United Kingdom, 14.3% in Italy, 13.9% in France, 10.4% in Spain, 7.2% in Sweden, 6.8% in Ireland, 5.7% in Switzerland, 3.9% in Poland, 3.5% in the United States, and 2.3% in India.¹¹

Descriptive biostatistics are measures of frequency (e.g., count, percent, rate), central tendency (e.g., mean, median, mode), dispersion (variation) (e.g., standard deviation, range, interquartile range), and position (e.g., percentile, rank) representing either a sample or a population.

In a study involving 5700 hospitalized COVID-19 patients in New York, the median age was 63 years. The patients were more likely male (60.3% versus 39.7%) and had at least one comorbid condition (88% versus 6.3% with one and 6.1% with none). An estimated 56.6% had hypertension, 41.7% were obese, and 33.8% had diabetes.¹²

However, another study that controlled for underlying conditions and patient characteristics found that hypertension was no longer associated with hospitalization, indicating that underlying factors associated with hypertension (e.g., obesity and diabetes) were explaining the positive association.¹³

1.4.2 Probability

Probability provides a basis for evaluating the reliability of the conclusions we reach and the inferences we make under uncertainty. Biostatistics applies probability in several areas, such as in risk assessment, sampling and evaluating the reliability of an estimator, validity of a diagnostic test, and the chance of survival over a given time period.

Probability is a numerical description of the likelihood of an events occurrence.

A common question in biological health research is whether the probability of an event is related to the levels of a given variable, like exposure status, age, race and Hispanic origin, marital status, treatment, and so on. For example, in the United States, the probability of COVID-19 cases dying from the disease varies considerably by age (**Figure 1.1**).¹⁴

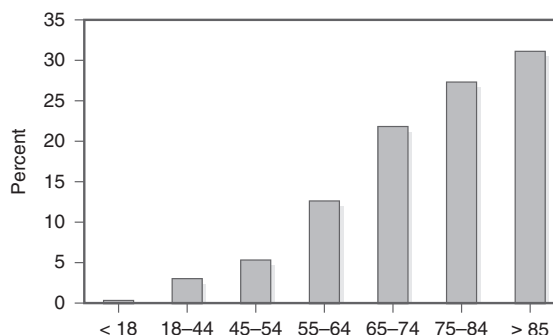


Figure 1.1 Mortality from COVID-19 in the United States by Age, February 12 through May 18, 2020

Among hospitalized cases in New York City, the probability of dying is greater among males than females in each age group (**Figure 1.2**).¹² The higher probability of death in males than females is more pronounced in the younger age groups. For example, the risk of death is 3.9 times greater in males compared with females in the age group 20–29 years and 1.4 times greater in the age group ≥ 90 years.

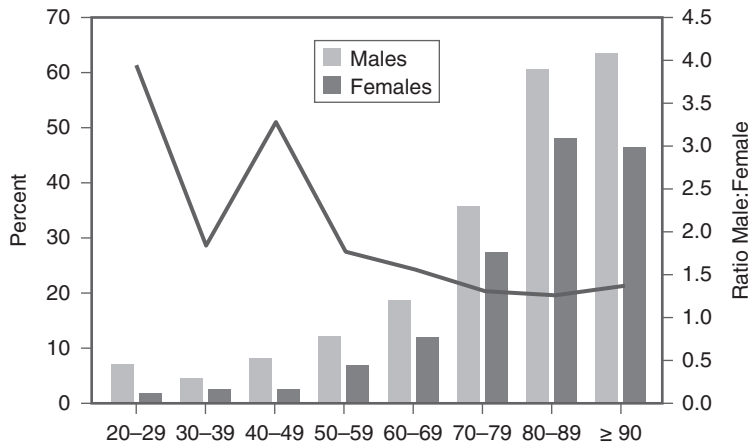


Figure 1.2 Deaths in COVID-19 Patients Hospitalized in the New York City Area by Sex and Age

1.4.3 Inferential Statistics

Inferential statistics involves a process of drawing conclusions about some aspect of a population, typically based on random sampling. For example, according to an observed subset of all customers who shop at a given store, females are more likely than males to wear a mask to prevent the spread of COVID-19.

Inferential statistics is the theory, methods, and practice of drawing conclusions about the parameters of a population based on sample information taken from the population.

The population is all customers.

A **population** is a set or collection of items of interest in a study.

The sample should be a sufficiently large, representative group from the population.

A **sample** is a subset of items that have been selected from the population.

Two methods are commonly used in inferential statistics: estimation and hypothesis testing. We may wish to infer something about a population, based on an estimated value from a sample. The estimated statistic is hopefully close to the true population parameter. Because our estimate contains some uncertainty, an accompanying statement of confidence (confidence interval) should be included.

A **statistic** is a measure from a sample that is used in statistics.

A **parameter** is a measure from a population that is often unknown. We are usually not able to calculate its value.

A **confidence interval** is a range of values wherein there is a **level of confidence** that the true parameter value lies within it.

Sampling is the process of selecting participants for the study when the number of people in the population of interest is larger than the required sample size. A number of sampling methods will be considered in this book. We will also address ways to minimize sampling bias where systematic error causes the sample of people in the study to not represent the target population.

An example of estimation, in which an estimate is derived from a sample and a statement of confidence is associated with the estimate, involves the death-to-case ratio for COVID-19. What is the death-to-case ratio for COVID-19? On March 5, 2020, researchers estimated the death-to-case ratio for China, including Hong Kong, Macau, and Taiwan, to be 3.48% (95% confidence interval [CI] 3.35–3.61%).¹⁵

In some cases, we may wish to infer that an estimated value is significantly different from another value. To assess this difference, we employ the process of hypothesis testing. For example, suppose you want to know whether the death-to-case ratio of COVID-19 is different between the United States and the United Kingdom. Your research hypothesis is that the ratios are different. On July 27, 2020, the number of confirmed cases and deaths in the United States was 4,288,012 and 148,009, respectively.¹⁶ Corresponding values in the United Kingdom were 301,706 and 45,844.¹⁶ The death-to-case ratio was 15.19% in the United Kingdom and 3.45% in the United States. The ratio of these two death-to-case estimates is 4.40 (95% CI = 4.36–4.45). In other words, the death-to-case ratio was 4.40 times greater in the United Kingdom than in the United States. Because the 95% confidence interval does not overlap 1, we conclude that this result is statistically significant at the 0.05 level.

1.4.4 Statistical Techniques

A variety of statistical techniques will be presented in this book. These statistical techniques are often applied to questions in the biological sciences. For example, rate ratios were used to address questions as to whether pregnancy associates with hospitalization and increased risk for intensive care unit admission, receipt of mechanical ventilation, and death among reproductive age women with COVID-19 infection.¹⁷ Rate ratios are commonly used to compare the risk of an outcome in one group with another group. After adjusting for age, race and Hispanic origin, and underlying health conditions, pregnant women were 5.4 (95% CI = 5.1–5.6) times more likely to be hospitalized, 1.5 (95% CI = 1.2–1.8) times more likely to be admitted to the ICU, and 1.7 (95% CI = 1.2–2.4) times more likely to receive mechanical ventilation. Each of these estimates are statistically significant because the confidence intervals do not overlap 1. On the other hand, pregnant women were 0.9 (95% CI = 0.5–1.5) times as likely to die, which is not statistically significant.

Statistical techniques are analytic approaches that utilize statistical methods to investigate a range of questions.

News File

During March 27–28, 2020, three University of Texas students showed symptoms and tested positive for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Each of these students had gone to Cabo San Lucas, Mexico for spring break (March 14–19). Symptoms for the students did not manifest until March 22–25. Other student travelers were identified by contact tracing interviews and flight manifests. Of 183 students who traveled to Cabo San Lucas and who were evaluated for SARS-CoV-2, 60 (33%) tested positive. In addition, there was 1 of 13 (8%) household contacts of the travelers and 3 of 35 community contacts of the travelers who tested positive. Of those with positive tests, 20% were asymptomatic, and no one needed hospitalization or died. Testing of all those who were at risk allowed public health officials to identify those who could spread the virus, even asymptomatic people, in an effort to control further COVID-19 outbreaks.¹⁸

1.5 Statistical Software

Throughout this text, we will give examples of programs and output from computer packages designed to assess statistical data. Common statistical software used to analyze data include Statistical Analysis System (SAS), Statistical Package for the Social Sciences (SPSS), Stata, R, and Microsoft Excel. Although we will concentrate on SAS and Excel, the supplementary content for this class will be programs and output for SPSS, STATA, and R. In this section, we will provide introductions to both SAS and Excel.

SAS has three interactive windows: the Editor, Log, and Output. Each will be briefly described here. SAS programs involve SAS statements. Each SAS statement ends with a semicolon. The semicolon completes the SAS statement. SAS statements are used to create SAS data sets and to run predefined statistical or other routines. A group of SAS statements used to define and create your SAS data set is called a Data Step. A Data Step tells SAS programs about the data. For example,

```
DATA Example;  
  INPUT Age Sex $;  
DATALINES;  
  17 Male  
  18 Female  
  20 Male  
  24 Male  
  24 Female  
  30 Male  
  32 Male  
;
```

The name of the data set is arbitrary but must begin with a letter. For categorical variables, follow the variable name on the INPUT line with “\$”.

“DATA LINES” comes prior to entering the data. Following the data set is a semicolon. It does not matter whether you choose to type words in uppercase or lowercase.

PROC (meaning procedure) begins several predefined routines. Each PROC is followed by a specific option, like PRINT, UNIVARIATE, MEAN, FREQ, REG, and so on. We end the procedure statements with “RUN”. For example, to obtain the mean age for males and females separately, we first sort the data according to sex:

```
PROC SORT DATA=Example;
  Title 'Sorted Data by Sex';
  BY Sex;
RUN;
```

Then run PROC MEANS, with the variable that the means are to be computed for preceded by VAR and, because the means are to be computed for males and females, “BY Sex” is added. The MAXDEC option allows you to specify the number of decimal values.

```
PROC MEANS DATA=Example MAXDEC=2;
  TITLE 'Mean Age by Sex';
  VAR Age;
  BY Sex;
RUN;
```

SAS programs are entered into the SAS Editor and then run by clicking “Submit” on the toolbar. When the procedure is executed, it produces a LOG and OUTPUT. The LOG is an annotated copy of the original program, with the data excluded. Any procedure coding errors and information about the data set (e.g., number of observations and variables) will be identified there. The OUTPUT provides the results requested by the PROC statement, which for our example appear as follows:

The MEANS Procedure

Sex=Female

Analysis Variable : Age				
N	Mean	Std Dev	Minimum	Maximum
2	21.00	4.24	18.00	24.00

Sex=Male

Analysis Variable : Age				
N	Mean	Std Dev	Minimum	Maximum
5	24.60	6.39	17.00	32.00

Microsoft Excel features the ability to store, organize, and manipulate data (i.e., make calculations, graphs, and more). It is used to create spreadsheets, which are special documents that allow us to store and organize data in rows (horizontal sets of boxes labeled as 1, 2, 3, etc.) and columns (vertical sets of boxes labeled as A, B, C, etc.). The data can then be read and manipulated. The interaction of each row and column is a cell wherein we enter numbers, text, or formulas.

An Excel document is a workbook. A workbook consists of one or more worksheets. Worksheets are the grid where we store and calculate data. Simple

and complex formulas can be calculated in Excel. Excel also offers a variety of charts (e.g., line graph, bar chart, scatter plot) to assess data. To introduce you to Excel, consider the partial table shown here. The data are the same as used in the SAS example. To obtain the statistics in Column B, we entered in cell B8 =COUNT(B2:B6), in cell B9 =AVERAGE(B2:B6), in cell B10 =STDEV.S(B2:B6), in cell B11 =MIN(B2:B6), and in cell B12 =MAX(B2:B6). Statistics for the data in column C are computed similarly.

	A	B	C
1		Male	Female
2		17	18
3		20	24
4		24	
5		30	
6		32	
7			
8	Number	5	2
9	Mean	24.6	21
10	Standard Deviation	6.4	4.2
11	Minimum	17	18
12	Maximum	32	24

Excel also has an add-in feature for performing data analysis. To add this feature, go to the File tab and select Options. Then select Add-ins and choose Analysis ToolPak. In the Manage box, select Excel Add-ins and click Go. Check the Analysis ToolPak and then select OK. Now, in the Excel spreadsheet under the Data tab on the far right of the toolbar, the Data Analysis option will appear.

To use the Data Analysis option, choose the Descriptive Statistics option, identify the input range, check the summary statistics option, and select OK. This will return the summary statistics shown in the previous table, and more.

1.6 Summary

This chapter provides an introduction to the meaning of biostatistics and the role it plays in addressing research questions. In biostatistics, data are the numerical values from variables. A variable is an attribute which describes something and can vary from one entity to another. Variables are quantitative if the values are quantities or numbers, and qualitative if the information is about quality. Data are measured on different scales: nominal, ordinal, interval, and ratio. The scale of measurement determines how the values should be collected, analyzed, and presented.

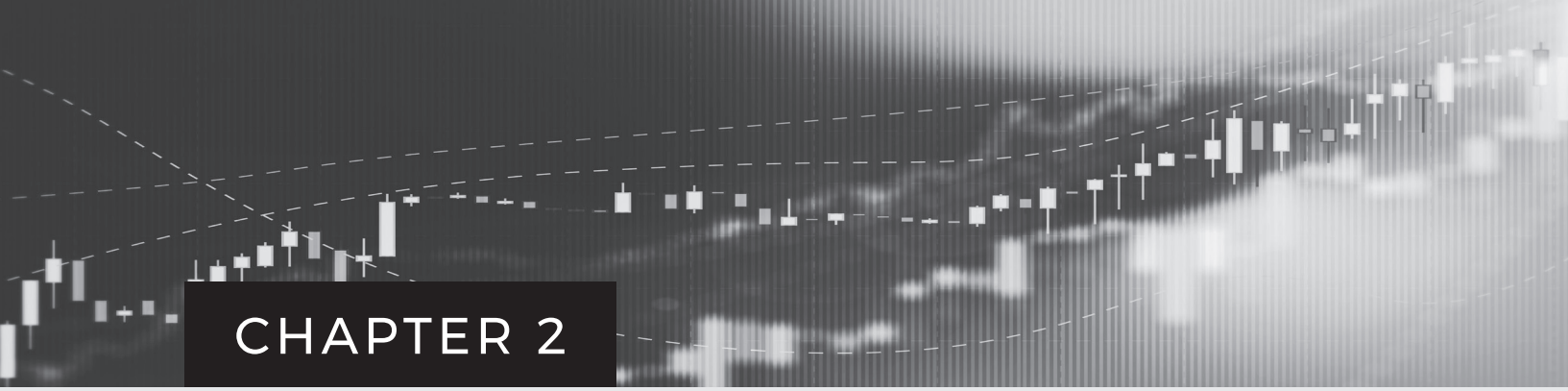
Four broad areas of biostatistics are descriptive, probability, inference, and statistical techniques. Descriptive statistics are important because they present data in a more manageable, meaningful way that leads to simpler interpretation. Probability is important because it serves as a basis for evaluating the reliability of the conclusions we reach and the inferences we make under uncertainty. Inferential statistics is important because it allows us to draw conclusions about a population based on sample information. Statistical techniques are important because they allow us to explore and answer a range of research questions. Statistical software is an efficient tool for conducting thorough analysis of each of these areas of biostatistics.

Exercises

1. List and briefly describe the four broad areas of biostatistics.
2. Classify the following as (A) nominal data, (B) ordinal data, (C) discrete data, or (D) continuous data.
 - ___ Integers of counts that differ by fixed amounts, with no intermediate values possible
 - ___ Measurable quantities not restricted to taking on integer values
 - ___ Ordered categories or classes
 - ___ Unordered categories or classes
3. Classify the following as either (A) quantitative or (B) qualitative data.
 - ___ Age
 - ___ Hometown
 - ___ Cholesterol level
 - ___ Eye color
 - ___ Number of siblings
4. Discuss whether discrete or continuous data provide more precision and information.
5. Descriptive biostatistics involves four types of statistics. List and define these.
6. Define probability and list some of its applications.
7. In what way does inferential statistics involve probability?
8. Classify the following as either (A) a statistic, (B) a parameter, or (C) a confidence interval.
 - ___ Gives a range of values for an unknown parameter
 - ___ Measure from a sample
 - ___ Measure from a population
 - ___ A numerical characteristic of a population
 - ___ A number that represents a property of a sample
9. The length of stay for a sample of eight patients admitted to the hospital were recorded: 3, 16, 12, 3, 2, 5, 4, 1. Use a statistical software package to compute the number of patients, mean, standard deviation, minimum, and maximum. For the mean and standard deviation, express your answers to the nearest hundredth. For the others, do not round.
10. In this chapter we spoke briefly about confidence intervals. Much more will be said about confidence intervals later in this book. For now, try calculating a confidence interval for the population mean based on data in the previous exercise. Express your answer to the nearest hundredth. (Hint: In SAS use the procedure MEANS and the option CLM. In Excel, go into Data Analysis, choose Descriptive Statistics, identify the input range of your data and select summary statistics and confidence level for the mean. Use the estimated mean and confidence level in the output to derive the confidence interval.)

References

1. Van der Zande J. Statistik and history in the German enlightenment. *J Hist Ideas*. 2010;71(3):411–432.
2. Moore SC, Patel AV, Matthews CE, et al. Leisure time physical activity of moderate to vigorous intensity and mortality: A large pooled cohort analysis. *PloS Med* 2012;9(11):e1001335.
3. US Department of Health and Human Services. *Physical activity and health: A report of the Surgeon General*. Atlanta, GA: CDC; 1996.
4. American Heart Association. Physical activity improves quality of life; 2018. http://www.heart.org/HEARTORG/HealthyLiving/PhysicalActivity/FitnessBasics/Physical-activity-improves-quality-of-life_UCM_307977_Article.jsp#.W2CQd9VKiUk. Accessed July 24, 2020.
5. Holme I, Anderssen SA. [Physical activity, smoking and mortality among men who participated in the Oslo studies of 1972 and 2000]. *Tidsskr Nor Laegeforen*. 2014;134(18):1743–1748.
6. US Department of Health and Human Services. *2008 Physical Activity Guidelines for Americans*. Washington, DC: US Department of Health and Human Services; 2008. <https://health.gov/paguidelines/pdf/paguide.pdf>. Accessed May 3, 2021.
7. US Department of Health and Human Services. *Physical Activity Guidelines Advisory Committee Report*, 2008. Washington, DC: US Department of Health and Human Services; 2008. https://health.gov/sites/default/files/2019-10/CommitteeReport_7.pdf instead? Accessed May 3, 2021.
8. Lauer SA, Grantz KH, Bi Q, et al. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Ann Intern Med*. 2020;172(9):577–582.
9. Merrill, RM. *Introduction to Epidemiology*, Eighth Edition. Chapter 3. Burlington, MA: Jones & Bartlett Learning; 2021.
10. Centers for Disease Control and Prevention. Clinical questions about COVID-19: Questions and answers. <https://www.cdc.gov/coronavirus/2019-ncov/hcp/faq.html>. Accessed July 24, 2020.
11. Mortality analysis. Johns Hopkins University & Medicine. <https://coronavirus.jhu.edu/data/mortality>. Accessed July 27, 2020.
12. Richardson S, Hirsch JS, Narasimhan M, et al. Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City area. *JAMA*. 2020;323(20):2052–2059.
13. Killerby ME, Link-Gelles R, Haight SC, et al. Characteristics associated with hospitalization among patients with COVID-19—metropolitan Atlanta, Georgia, March–April 2020. *MMWR*. 2020;69:790–794.
14. Wortham JM, Lee JT, Athomsons S, et al. Characteristics of persons who died with COVID-19—United States. February 12–May 19, 2020. *MMWR* 2020;69(28):923–929.
15. Wilson N, Kvalsvig A, Barnard LT, Baker MG. Case-fatality risk estimates for COVID-19 calculated using a late time for fatality. *Emerg Infect Dis*. 2020;26(6):1339–1441.
16. COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). Johns Hopkins University & Medicine. <https://coronavirus.jhu.edu/map.html>. Accessed July 27, 2020.
17. Ellington S, Strid P, Tong VT, et al. Characteristics of women of reproductive age with laboratory-confirmed SARS-CoV-2 infection by pregnancy status—United States, January 22–June 7, 2020. *MMWR*. 2020;69:769–775.
18. Lewis M, Sanchez R, Auerbach S, et al. COVID-19 outbreak among college students after a spring break trip to Mexico—Austin, Texas, March 26–April 5, 2020. *MMWR*. 2020;69(26):830–835.



CHAPTER 2

Descriptive Summaries of Data

KEY CONCEPTS

- A research design is a formal approach of scientific investigation.
- Three general types of research designs are descriptive, exploratory, and analytic.
- A descriptive design is used to provide information about patterns and frequencies in data; it is used to describe what is occurring.
- An exploratory design is the least structured of the three research designs, with a high level of uncertainty and poor understanding about the biological problem.
- An analytic design is used to measure associations and test whether those associations are statistically significant.
- A frequency distribution is a display, either tabular or graphical, of the number of observations a variable takes at each level.
- A proportion is a number compared to the whole.
- A rate is a proportion with the added dimension of time.
- A ratio is the relationship between two quantities, normally expressed as the quotient of one divided by the other.
- Greek letters are used to represent population characteristics (parameters), and Roman letters are used to represent sample characteristics (statistics).
- Measurements on a nominal or ordinal scale are evaluated using frequency, relative frequency, proportions, and rates.
- Measurements on an interval or ratio scale are evaluated using measures of central tendency and variation.
- Measures of central tendency describe what is typical for a distribution of data, such as the geometric mean, arithmetic mean, median, and mode.
- Measures of variation capture the extent of spread over the values of a data set, such as variance, standard deviation, coefficient of variation, range, and interquartile range.
- Tables and graphs are used to present information, typically involving a large number of observations, in a concise and comprehensible way.

In the previous chapter, we emphasized that studies in biostatistics begin with a research question, which represents an unknown issue to be resolved by carrying out the study. A research design reflects the purpose of the study, whether descriptive, exploratory, or analytic. It is an overall plan that describes the approach that will be taken to answer the research question—it is a formal approach of scientific investigation. We will cover descriptive research designs in this chapter and emphasize the first major topic area in biostatistics—descriptive biostatistics.

In this chapter, you will become familiar with questions for descriptive biostatistics, and selected tabular and graphical approaches will be shown for summarizing and presenting numerical data. Numerical descriptions of data obtained through arithmetic operations will be presented. These are called descriptive statistics. They measure frequency (e.g., count, percent, rate), central tendency (e.g., mean, median), dispersion or variation (e.g., standard deviation, range, interquartile range), and position (e.g., percentile, rank). Rate calculations and adjustment methods of rates will be presented.

2.1 Questions for Descriptive Biostatistics

A descriptive study design in biostatistics should provide an accurate and valid representation of the variables that are relevant to the research question. This type of research is less structured than the analytic study design, but more structured than the exploratory study design. The analytic study design investigates why and how a phenomenon occurred. The exploratory study design proceeds with a high level of uncertainty and poor understanding about a biological problem.

Descriptive measures describe and summarize numerical data. A descriptive study design aims to provide information about patterns and frequencies in the data. Descriptive statistics allow us to present the data so that simple interpretation might occur. For example, if we had the race results of all the athletes from the Boston marathon, we may be interested in how many started the race, how many finished the race, the fastest time, the average time, how spread out the times were, how the results varied according to age, sex, and race and Hispanic origin, and so on. These descriptive statistics, therefore, help us summarize the race information in a meaningful way so that patterns might emerge from the data. Descriptive statistics, however, do not allow us to draw conclusions beyond the data we have assessed or test specific hypotheses. We are simply describing what occurred in our data.

The descriptive research design, with accompanying possible questions and examples, is presented in **Table 2.1**. To help emphasize the distinct role descriptive research plays, the descriptive study design is contrasted with the exploratory and analytic study designs. Later chapters will cover the analytic study design, but briefly it involves techniques for measuring associations between variables and testing hypotheses of association.

Table 2.1 Questions and Examples Associated with Descriptive, Exploratory, and Analytic Biostatistics Study Designs

Type of Design	Question	Examples
Descriptive	What is the number of cases? What is the proportion of cases? What is the rate of cases? Are selected variables related?	How many COVID-19 cases were identified today? What proportion of COVID-19 cases resulted in death? What is the rate of COVID-19 per 100,000 in July 2020? Is the death rate for COVID-19 cases related to age?

Table 2.1 Questions and Examples Associated with Descriptive, Exploratory, and Analytic Biostatistics Study Designs (Continued)

Type of Design	Question	Examples
Exploratory	What is a case? What are the key factors involved?	What clinical characteristics tend to manifest among patients exposed to COVID-19? What reasons are there for the high percentage of students being absent from school today?
Analytic	Why? What are the causes of the health problem?	Do differences in comorbid diseases among racial/ethnic groups explain the high racial/ethnic disparity in case-fatality rates? Will a new treatment lower the risk of death among COVID-19 patients?

News File

In order to rapidly monitor changes in mental health, the Census Bureau partnered with the National Center for Health Statistics to conduct the Household Pulse Survey.¹ The survey was intended to provide relevant information about how COVID-19 affects the U.S. population. It covered April 23 through July 21, 2020.

Questions were included to assess the frequency of anxiety and depressive symptoms. The questions collected information from participants regarding the previous 7 days. It identified the percentage of adults who experienced symptoms of anxiety or depression more than half of the time.

Results from the survey are presented in **Table 2.2**. The overall increase in anxiety or depression was 14%. The greatest increase was among individuals 30–39 years, males, non-Hispanic Asians, and in those with at least a high school diploma. An increase in anxiety or depression occurred in each age group except for those 80 years and older, where it decreased by 17%. Although other reports have shown that the death-to-case ratio of COVID-19 increases with age, other factors (e.g., social isolation, unemployment) seem to be affecting younger people disproportionately hard. Although we might expect an increase in percentages of depression or anxiety in the older age groups because they are more vulnerable to dying from the disease, the reason for the decrease in percentages among ages 80 and older is unclear.

For age, sex, race and Hispanic origin, and education, the greatest increase tends to be in those groups that started with the lowest levels of anxiety or depression, and vice versa. Social isolation, employment, and economic concerns are certainly contributors to the increase, but where anxiety or depression was initially higher, some or all of these concerns may have already been greater. For example, Hispanics, non-Hispanic blacks, and non-Hispanic other or multiple race began with higher levels of anxiety or depression, but the increase during the COVID-19 pandemic was smaller than the increase among non-Hispanic whites and non-Hispanic Asians. In addition, among the education groups those with less than a high school education began with the highest level of anxiety or depression, but only increased 4% during the COVID-19 pandemic.

Table 2.2 Anxiety or Depression in the United States During April 23 through July 16

	Week Beginning April 23	Week Beginning July 16	Ratio
United States	35.9%	40.9	1.14
Age Group			
18–29 years	46.8%	53.4%	1.14
30–39 years	39.6%	47.1%	1.19
40–49 years	38.9%	44.5%	1.14
50–59 years	35.8%	40.0%	1.12
60–69 years	28.9%	32.1%	1.11
70–79 years	21.5%	24.3%	1.13
80 years and above	21.1%	17.5%	0.83
Sex			
Male	31.0%	37.0%	1.19
Female	40.7%	44.6%	1.10
Race and Hispanic Origin			
Hispanic or Latino	42.7%	46.3%	1.08
Non-Hispanic white, single race	33.6%	38.8%	1.15
Non-Hispanic black, single race	38.9%	42.6%	1.10
Non-Hispanic Asian, single race	31.9%	40.3%	1.26
Non-Hispanic, other or multiple race	43.9%	48.8%	1.11
Education			
Less than a high school diploma	45.4%	47.3%	1.04
High school diploma or GED	36.7%	42.2%	1.15
Some college/ associate's degree	38.5%	43.3%	1.12
Bachelor's degree or higher	30.7%	35.7%	1.16

2.2 Empirical Frequency Distributions

Tabular and graphical presentations of the data are generally known as empirical frequency distributions. It is often of interest to know the pattern or grouping into which data fall. In this section, we will present the frequency distribution and corresponding statistical measures.

2.2.1 Frequency Distribution

A **frequency distribution** is a display, either tabular or graphical, of the number of occurrences at each possible level of a variable. For example, the number of deaths from COVID-19 in the United States during the week ending July 25, 2020,² is distributed across selected age groups as follows:

	< 45	45–54	55–64	65–74	75–84	≥ 85
Frequency of deaths	933	834	1968	3311	4434	5925

2.2.2 Relative Frequency

Relative frequency is the number for the specific level of a variable compared with the total number across the levels of the variable. For example, the frequency distribution table is expanded to include the relative frequencies. If you sum the relative frequencies, they total to 1. The relative frequencies here tell us the proportion of cases in each age group.

	< 45	45–54	55–64	65–74	75–84	≥ 85
Frequency of deaths	933	834	1968	3311	4434	5925
Relative frequency	0.054	0.048	0.113	0.190	0.255	0.340

2.2.3 Proportion

Each of the relative frequencies just shown is a proportion. They range from 0 to 1. Proportions are often presented as a percentage (per 100) of the whole. Thus we can make statements like 0.340 (or 34.0%) of the cases of COVID-19 in the United States are in the age group 85 years and older, as of July 25, 2020.

A **proportion** is a part or number in comparative relation to a whole.

2.2.4 Rate

A related measure is a rate. A rate is the number of cases compared with the population at risk of becoming a case from which the cases derived. A rate specifies some period of time in which cases were generated. In biostatistics a rate may reflect a disease (e.g., hemophilia, meningococcal meningitis, COVID-19), an event (e.g., domestic violence, premature birth, pulmonary

embolism), or a behavior or condition (e.g., social distancing during an infectious disease outbreak, nutrition, vaccination, use of health services). In other words, rather than representing the magnitude of something, it represents the risk of something.

A **rate** is a proportion with the added dimension of time.

For example, a measure of the magnitude of COVID-19 is the number of cases. A measure of the risk of COVID-19 is the rate of the illness. Extending the distribution table to include age-specific rates per 100,000 people in the United States population is as follows:

	< 45	45–54	55–64	65–74	75–84	≥ 85
Frequency of death	933	834	1968	3311	4434	5925
Relative frequency	0.054	0.048	0.113	0.190	0.255	0.340
Rate per 100,000	0.48	2.05	4.60	10.10	26.77	88.42

In one report identifying COVID-19 hospitalization rates by race and Hispanic origin for March 1–July 18, 2020, the authors were able to make meaningful comparisons because the rates accounted for differential population sizes. The rates were further adjusted for age differences among the racial/ethnic groups, also making the comparisons more meaningful. The rate estimates (per 100,000 population) were 281 for non-Hispanic American Indians or Alaska Natives, 247 for non-Hispanic black or African American, 242 for Hispanic or Latino, 67 for non-Hispanic Asian or Pacific Islander, and 53 for non-Hispanic white.³ It would have been less informative had just the counts been reported since race/ethnicity population sizes from which cases derived vary considerably. According to population estimates for race and Hispanic origin in the United States on July 1, 2019, 76.3% were white only, 13.4% were black only, 5.9% were Asian only, 1.3% were American Indians/Alaska Natives only, 0.2% were native Hawaiians and other Pacific Islanders only, 2.8% were two or more races, and 18.5% were Hispanic or Latino.⁴

2.2.5 Adjusting Rates

If crude rates are compared between populations who have different age distributions, and if age is related to the rate variable, then the comparison of rates will be confounded. Essentially, any variable that varies across the population distributions being compared, if related to the rate variable, will make the comparison in rates potentially misleading. To account for the potential confounder and make the comparison of rates valid, the rates should be adjusted. Two methods will be presented in this section: direct and indirect. These will be illustrated for the situation where age is a potential confounder.

Sometimes age-adjusted rates are also applied to a single group where rates are being compared over time, but the age distribution is changing. For example, cancer rates tend to increase with age. In the United States, life expectancy increased from about 50 years in 1900 to about 79 years today. So some of the higher cancer rates now are because people are living to older ages where cancer is more common. To account for this, when the National Cancer Institute reports cancer rates over time, they are age-adjusted to some standard population, such as the U.S. age distribution in 2000.

Applying the direct method of age adjustment to rates over time assumes that the age-specific rates have a similar trend. If age-specific trends are not somewhat parallel, then an age-adjusted rate can mask important information. In the situation of nonparallel trends in age-specific rates, reporting trends in age-specific rates is more appropriate.

In practice, the direct method of standardizing rates is more common than the indirect method. With computer software such as Excel, it is fairly easy to compute age-adjusted rates using either the direct method or the indirect method. One rule of thumb is to use the direct method when subgroup-specific rates are available for all the groups being compared. If not, or if the subgroup rates are based on small numbers and therefore are unstable, the indirect method should be used.

2.2.5.1 Direct Method of Adjusting Rates

A comparison of the crude mortality rate of COVID-19 in the United States between males and females is appropriate if the rate is not dependent on age and the age distribution is similar between males and females. Mortality rates increase with age, and the age distribution differs between males and females (**Table 2.3**).^{2,3}

The crude mortality rate (per 100,000) of COVID-19 during the week ending in July 25, 2020, was 5.40 for males and 5.07 for females. To obtain the age-adjusted rate, for each population, we need to know the age-group specific rates. These are calculated in the table. The crude rate in each population is just a weighted average of the age-group specific rates. For example, the crude mortality rate for males is

$$\frac{\sum(\text{Rate} \times n)}{\text{Total } n} = \frac{(1.69 \times 138,890,329) + (12.49 \times 15,328,874) + (30.89 \times 7,267,187) + (96.73 \times 2,418,086)}{138,890,329 + 15,328,874 + 7,267,187 + 2,418,086} = 5.40 \text{ per } 100,000.$$

Because we are interested in obtaining rates that reflect similar age-distributions, the numbers in each category from one of the populations can be selected as the reference population and used as the weights to form weighted averages for both populations. It does not matter which of the two populations

Table 2.3 Death Rate from COVID-19 in the United States for Males and Females, During the Week Ending July 25, 2020

	Male				Female			
Age Group	Population	%	Cases	Rate per 100,000	Population	%	Cases	Rate per 100,000
< 65 years	138,890,329	84.74	2345	1.69	137,697,207	81.61	1390	1.01
65–74 years	15,328,874	9.35	1915	12.49	17,460,563	10.35	1396	8.00
75–84 years	7,267,187	4.43	2245	30.89	9,294,062	5.51	2189	23.55
≥ 85 years	2,418,086	1.48	2339	96.73	4,282,794	2.54	3586	83.73
Total	163,904,476	100.00	8844	5.40	168,734,626	100.00	8561	5.07

is selected as the reference, but it is important to note the reference population in the interpretation of the rate. Sometimes a set of age-specific frequencies from a totally separate reference population is used. What is important is that the same set of numbers (standard) is applied to both populations.

For this example, we will select the male population as our standard and apply it to the age-group specific rates in the female population, as follows:

$$\begin{aligned}\text{Age-Adjusted Rate} &= \frac{\sum (\text{Rate} \times n \text{ in standard})}{\text{Total } n \text{ in standard}} \\ &= \frac{(1.01 \times 138,890,329) + (8.00 \times 15,328,874) + (23.55 \times 7,267,187) + (83.73 \times 2,418,086)}{138,890,329 + 15,328,874 + 7,267,187 + 2,418,086} \\ &= 3.88 \text{ per } 100,000\end{aligned}$$

So, if females had the same age-distribution as males, their mortality rate (per 100,000) of COVID-19 during the week ending July 25, 2020, would be 3.88 instead of 5.07.

Therefore, the direct method of age-adjustment involves a weighted average of the age-specific rates, where the weights are the proportions of persons in the corresponding age groups of the standard population.

2.2.5.2 Indirect Method of Adjusting Rates

In situations where age-group specific rates are unstable because of small numbers or unavailable altogether in the populations being compared, the indirect method may be useful for adjusting rates.

As was the case with the direct method, a standard population is selected. Age-group specific rates are calculated according to age-group specific values from a standard population. These rates are then multiplied by the age-group specific population value in the comparison population to obtain the expected number of cases in each age group. The total number of cases observed in the comparison population is then divided by the total number of expected cases. This ratio is referred to as the standard morbidity/mortality ratio (SMR).

$$\text{SMR} = \frac{\text{Observed}}{\text{Expected}}$$

To illustrate, refer again to Table 2.3. Suppose that some or all of the female age-group specific deaths are unavailable but that the total is available. Further, suppose that the age-group specific rates for males have been calculated. Multiply the age-group specific rates in the male population by the age-group specific female population values to obtain the age-group specific number of COVID-19 deaths for females, assuming they have the same rates as males. Sum the expected counts to obtain the total number of expected deaths in the comparison population. Then divide the observed count for females by the expected count for females.

$$\begin{aligned}\text{Expected} &= (0.0000169 \times 137,607,207) + (0.0001249 \times 17,460,563) \\ &\quad + (0.0003089 \times 9,294,062) + (0.0009673 \times 4,282,794) = 11,520\end{aligned}$$

Then

$$\text{SMR} = \frac{8561}{11,520} = 0.74$$

The SMR indicates that females have a 36% lower mortality rate from COVID-19 than males in the week ending in July 25, 2020.

2.2.6 Ratio

Finally, ratios can be used for descriptive and analytic purposes. As a descriptive measure, a ratio can describe comparisons among groups. For example, the rate of hospitalization for COVID-19 for each racial/ethnic group (presented in Section 2.2.4) can be compared using ratios.³ Hospitalization rates for each of the racial/ethnic groups compared with non-Hispanic whites are 5.30, 4.66, 4.57, and 1.26, respectively.

A **ratio** is the relationship between two quantities, normally expressed as the quotient of one divided by the other.

We can now apply the following equations to ratios to calculate percent greater or less than the reference group.

If Ratio > 1 then $(\text{Ratio} - 1) \times 100\%$
 If Ratio ≤ 1 then $(1 - \text{Ratio}) \times 100\%$

Therefore, the rate of hospitalization is 430% greater for non-Hispanic American Indians or Alaska Natives, 366% greater for non-Hispanic blacks, 357% greater for Hispanic or Latino, and 26% greater for non-Hispanic Asian or Pacific Islander compared with non-Hispanic white.

Using the ratio as an analytic tool in which hypotheses are considered is also common in biostatistics. A hypothesis could be that the rate of hospitalization is greater for Hispanics than non-Hispanic whites. This hypothesis would be tested against there being no difference in rates. We would then apply a statistical test to the data in order to evaluate whether the difference in rates is statistically significant.

2.3 Notation

Recall that data are information obtained through observation, experiment, or measurement of a phenomenon of interests, such as the weekly case-fatality rate of COVID-19 for all U.S. states, the average age of COVID-19 patients for each census tract within a given state, or the proportion of COVID-19 hospital patients requiring invasive mechanical ventilation at selected hospitals. A capital letter or symbol like X is used to represent the characteristic of interest. If two characteristics are measured, we represent them by X and Y , and so on. By convention a subscript is added to the variable so that each subscript number identifies a specific item in the data set. For example, x_i represents a specific characteristic of interest for the i th person or item in the population. The number in the population is denoted by N , and the number in the sample is denoted by n . Other common statistical notation includes f_i = frequency of x_i , f = total number of observations in an interval, and Σ = sum.

Greek letters are used to represent population characteristics (parameters), and Roman letters are used to represent sample characteristics (statistics). Some examples appear as follows:

Parameter		Statistic	
Symbol	Name	Symbol	Name
μ	Population mean	$\bar{x}$	Sample mean
σ^2	Population variance	s^2	Sample variance
σ	Population standard deviation	s	Sample standard deviation

2.4 Measures of Central Tendency

For numerical data and their frequency distributions, the observed values usually tend to group toward a center. In other words, some central measure best characterizes the data. This phenomenon is called central tendency.

A **measure of central tendency** is a typical value for a distribution of data. Measures of central tendency include the geometric mean, arithmetic mean, median, and mode.

2.4.1 Geometric Mean

The geometric mean applies to positive numbers. It is used for variables measured on an exponential or logarithmic scale, such as dilution assays (1/2, 1/4, 1/8, 1/16, etc.). It is often used for environmental samples, where the levels of the variable can range over several orders of magnitude (e.g., coliforms in sample taken from a lake can range from less than 100 to greater than 100,000).

The **geometric mean** is an average that indicates the typical value of a set of numbers by taking the n th root of the product of their values.

$$\text{Geometric mean} = \sqrt[n]{(x_1)(x_2)\dots(x_N)} \quad \text{Population}$$

$$\text{Geometric mean} = \sqrt[n]{(x_1)(x_2)\dots(x_n)} \quad \text{Sample}$$

For example, suppose you are monitoring the level of *Enterococci* bacteria per 100 mL of sample over time and obtain the following data: 5, 25, 50, 100 per 100 mL. The geometric mean is $\sqrt[4]{5 \times 25 \times 50 \times 100} = 28.1$.

Now suppose two types of face masks are compared. The first has a protection score against SARS-CoV-2 of 65 and gets a 7 in reviews. The second has a protection score of 80 and gets a 5 in reviews. Comparing using the arithmetic mean gives $(65 + 7)/2 = 36$ versus $(80 + 5)/2 = 42.5$. On the other hand, the geometric means of the two face masks are: $\sqrt{65 \times 7} = 21.3$ and $\sqrt{80 \times 5} = 20$. Therefore, even though the protection is 15 points better for the second mask, the lower user rating of 5 is still important.

For data measured on an exponential or logarithmic scale, the median may be a better summary measure than the geometric mean if the data contains a negative number or a zero observation. Otherwise, when the data have a multiplicative or propagative nature, the geometric mean has some nice mathematical properties and should be used.

2.4.2 Arithmetic Mean

The **arithmetic mean** or average indicates the typical value of a set of numbers by taking their sum divided by either the population or sample size.

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i \quad \text{Population}$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{Sample}$$

For example, you observe a sample of 10 ages: 24, 24, 25, 25, 25, 25, 26, 26, 26, 27. The mean age is

$$\bar{x} = \frac{1}{10} (24 + 24 + 25 + 25 + 25 + 25 + 26 + 26 + 26 + 27) = \frac{253}{10} = 25.3.$$

2.4.3 Weighted Mean

In some situations where the numbers are not equally important, the numbers may be weighted.

$$\mu = \frac{\sum_{i=1}^N w_i X_i}{\sum_{i=1}^N w_i} \quad \text{Population}$$

$$\bar{x} = \frac{\sum_{i=1}^n w_i X_i}{\sum_{i=1}^n w_i} \quad \text{Sample}$$

For example, the weighted mean for the data in the previous example is

$$\bar{x} = \frac{2 \times 24 + 4 \times 25 + 3 \times 26 + 1 \times 27}{2 + 4 + 3 + 1} = \frac{253}{10} = 25.3.$$

2.4.4 Median

The **median** is a number in which half the data are greater than that number and half the values are less than that number. If the number of observations is an odd number, then the median is the middle value. If the number of observations is an even number, then the median is the average of the two middle values.

For example, the number of people performing COVID-19 tests at each of the nine locations is

2 3 3 5 5 9 9 10 10.

The median is 5 people. Consider if there were 10 locations in our data, such as

2 3 3 5 5 9 9 9 10 10.

The observations 5 and 9 occupy two center positions, so the median is $(5 + 9)/2$, or 7.

2.4.5 Mode

The **mode** of a set of data is the number that most frequently occurs. The mode is a desirable measure when we desire to know which number occurs most frequently. In the previous example, the mode is 9.

2.4.6 Choosing a Measure of Central Tendency

In general, the shape of the frequency distribution influences the relationship among the measures of central tendency. If the distribution consists of a symmetrical bell-shaped curve, then the mean, median, and mode will coincide. If the frequency distribution is not symmetric (skewed) then the measures of central tendency will not coincide. The mean is most sensitive to extreme values. For example, if the average age of a group of individuals was estimated in

a group of 10 or so people, and one of the people was 100, then the average age might not be characteristic of most of the people's ages. Hence, the median or mode may be preferred.

Conceptually, if we took a large number of sets of data from the same population and the mean, median, and mode were computed each time, less variation will result in the means than in the medians or the modes. Therefore, the mean is more stable. This fact, along with the mean being more amenable to mathematical treatment, makes the mean most widely chosen for statistical purposes, with the exception of some exploratory and descriptive assessment. The order of their statistical desirability is usually first the mean, second the median, and third the mode.⁵

2.5 Measures of Variation

A **measure of variation** (also called dispersion, scatter, or spread) is the extent to which a distribution is spread over the data range. Common measures of variation include the variance, standard deviation, range, interquartile range, and coefficient of variation.

2.5.1 Variance and Standard Deviation

To begin, consider the deviations from the mean as equal to $(x_1 - \mu)$, $(x_2 - \mu)$, ..., $(x_N - \mu)$. The average of the deviations from the mean is then

$$\frac{1}{N} \sum_{i=1}^N (x_i - \mu) = 0.$$

The average of the deviations from the mean will always be 0 because

$$\sum_{i=1}^N (x_i - \mu) = 0.$$

This is also true when sample data are involved:

$$\sum_{i=1}^n (x_i - \bar{x}) = 0.$$

Because the average of the deviations from the mean do not capture variation in the data, the average squared deviations from the mean is used as a dispersion measure. This is called the variance.

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 \quad \text{Population}$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{Sample}$$

The standard deviation is the square root of the variance.

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2} \quad \text{Population}$$

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{Sample}$$

For example, recall the random sample of 10 ages. The standard deviation is

$$\begin{aligned} s &= \sqrt{s^2} = \sqrt{\frac{1}{10-1} [(24-25.3)^2 + (24-25.3)^2 + \dots + (27-25.3)^2]} \\ &= \sqrt{0.9} = 0.949. \end{aligned}$$

You likely noticed the differences between the population and sample formulas for the variance. The population variance uses μ , and the sample variance uses $\bar{x}$. This is only a difference in notation. The population variance divides by N , and the sample variance divides by $n-1$. The variance is a measure of dispersion or scatter of observations around the mean. Given that a sample is a subset of the population, the range of sample data is going to be smaller than the range of population data. Hence, the mean of the deviations about the mean will be smaller for sample data than it will be for population data. To adjust the variance up to where it better estimates the population variance, we divide by $n-1$ instead of N . In addition to presenting the standard deviation for the arithmetic mean, the standard deviation for the geometric mean is shown in the following definition box.

The geometric standard deviation for a set of numbers $(x_1, x_2, \dots, x_n)$ is denoted as

$$\sigma_g = \exp \left(\sqrt{\frac{\sum_{i=1}^N \left(\ln \frac{x_i}{\mu_g} \right)^2}{N}} \right) \quad \text{Population}$$

$$s_g = \exp \left(\sqrt{\frac{\sum_{i=1}^n \left(\ln \frac{x_i}{\bar{x}_g} \right)^2}{n-1}} \right) \quad \text{Sample}$$

$$\begin{aligned} \text{For the example in 2.4.1, } s_g &= \exp \left(\sqrt{\frac{\left(\ln \frac{5}{28.1} \right)^2 + \left(\ln \frac{25}{28.1} \right)^2 + \left(\ln \frac{50}{28.1} \right)^2 + \left(\ln \frac{100}{28.1} \right)^2}{4-1}} \right) \\ &= 3.61. \end{aligned}$$

2.5.2 Degrees of Freedom

Degrees of freedom of an estimate is the number of independent pieces of information used to obtain the estimate. For the variance, several sums of squared quantities make up the estimate. In the above example, we had 10 squares of deviations from the mean. However, these 10 values are not independent because each of the sampled observations were used to obtain an estimate of the mean. Therefore, the sum of squares of deviations of the n values from their mean has $n-1$ degrees of freedom, or in our example, nine degrees of freedom. In general, the degrees of freedom for an estimate equals the number of observed values minus the number of parameters estimated in order to obtain the estimate in question.

2.5.3 More on the Standard Deviation

Because the units of the variance are squared, it is no longer in the same unit of measurement as the original data. Taking the square root of the variance, which yields the standard deviation, is a way to provide a measure of variation that is restored to the original unit of measurement. It is also important to note that when the frequency distribution is symmetrical and bell-shaped, which is called a normal distribution, the Empirical Rule in statistics applies.

The **Empirical Rule** states that for a normal distribution, almost all the data lie within three standard deviations of the mean. Under this rule, about 68% of the observations lie within the interval $\mu \pm \sigma$; about 95% of the observations lie within the interval $\mu \pm 2\sigma$; and about 99.7% of observations lie within the interval $\mu \pm 3\sigma$.

The empirical rule is one way to define an outlier when the data have an approximately symmetric distribution. Approximately 0.3% of data account for outliers. Another approach to identify outliers, useful when the data are not symmetric, will be described in section 2.7.

2.5.4 Coefficient of Variation

The **coefficient of variation** is the ratio of the standard deviation to the mean. It is a measure of relative variability that should only be derived for data measured on a ratio scale. It allows us to compare the variability between groups when the unit of measurement differs between groups.

$$CV = \frac{\sigma}{\mu} \times 100\% \quad \text{Population}$$

$$CV = \frac{s}{\bar{x}} \times 100\% \quad \text{Sample}$$

Consider two measures used to screen for depression, A and B. If we take a random sample of individuals who were assessed for depression based on these two measures we might find the following standard deviations:

$$s_A = 3 \quad s_B = 12$$

This might cause us to want to avoid the second depression measure because of its relatively high variability in scores. Before we do so, note that the range of A is 1–10 and the range of B is 1–100. Let $\bar{x}_A = 7.5$ and $\bar{x}_B = 75$.

To convert the standard deviation so that we can compare between two sets of numbers with different units of measurement, we apply the coefficient of variation. The coefficient of variation expresses the standard deviation as a percentage of the mean. It is a unitless quantity.

$$CV_A = \frac{3}{7.5} \times 100\% = 40\%$$

$$CV_B = \frac{12}{75} \times 100\% = 16\%$$

From these results, we now see that the second depression measure is preferred in terms of having smaller spread of scores scattered around its mean.

2.5.5 Range and Interquartile Range

The **range** is the difference between the largest and smallest values in the data. In a previous example (2.4.4), the range is 8 ($10 - 2$). In reporting the range, it is often informative to include the maximum and minimum values upon which it is based.

However, reasons have been given as to why this may be a less preferred measure of variation. First, it only uses a small part of the available information about variation in the data. It fails to capture the extent to which most of the observations are dispersed. Second, with larger data sets, the range tends to increase. Hence, a comparison of ranges between data sets is inappropriate unless the data sets are equal in size. Third, the range is a less stable measure than the other measures of variation considered in this section. In other words, from sample to sample, the range will vary the most.⁵ Nevertheless, when the median is used rather than the mean to represent the center of a data distribution, we tend to report with it the range, or the interquartile range.

The **interquartile range** represents the middle 50% of the data, extending from the first quartile to the third quartile. Two common approaches for calculating interquartile range are the exclusive and inclusive methods. The exclusive method does not include the median when identifying the first and third quartiles, whereas the inclusive method does include the median when identifying the quartiles. For example, for the data set referred to above that has an odd number of values (i.e., 2, 3, 3, 5, 5, 9, 9, 10, 10), to obtain the interquartile range using the exclusive method, place parentheses around the numbers above and below the median: (2, 3, 3, 5), 5, (9, 9, 10, 10). Think of the first quartile as the median of the first group of numbers and the third quartile as the median of the second group of numbers, or $Q1 = (3 + 3)/2 = 3$ and $Q3 = (9 + 10)/2 = 9.5$. Subtract $Q1$ from $Q3$ to obtain the interquartile range as $9.5 - 3 = 6.5$. Using the inclusive method, $Q1 = 3$ and $Q3 = 9$, hence the interquartile range is $9 - 3 = 6$.

2.6 Describing Data with Statistics and Graphs

As with frequency distribution tables and summary statistics, **graphs** can present a large amount of data in a concise and understandable way. Graphs often do a nice job clarifying biological health problems. They identify patterns, trends, aberrations, similarities, and differences in the data. They are important for communicating information according to selected factors like age, sex, and race and Hispanic origin, geographic location, and across time. They should be accompanied with descriptive titles, labels, and legends.

A summary of the statistical methods that we have considered in this chapter for describing data is presented in **Table 2.4**. The approach depends on the measurement scale of our data. A number of graphical approaches useful for describing data are described in **Table 2.5**.

Table 2.4 Scales of Measurement and Corresponding Statistics and Graphs

Scale	Statistics	Graphs
Nominal	Frequency Relative frequency Proportion Rate	Line graphs Bar chart/pie chart Spot map Area map
Ordinal	Frequency Relative frequency Proportion	Bar chart/pie chart
Interval or Ratio	Geometric mean Arithmetic mean Median Mode Range Variance Standard deviation Coefficient of variation	Bar chart/pie chart Histogram or frequency polygon Box plot Scatter plot Stem-and-leaf plot

Reference 6.

Table 2.5 Graphs for Describing Data

Type of Graph	Description
Arithmetic-scale line graph	Line graphs are used mostly for data plotted against time. An arithmetic graph has equal quantities along the y-axis. An arithmetic graph shows actual changes in magnitude of the number or rate of something across time.
Logarithmic-scale line graph	The y-axis is changed to a logarithmic scale. In other words, the axis is divided into cycles, with each being 10 times greater than the previous cycle. Focus is on the rate of change. A straight line reflects a constant rate of change.
Simple bar chart	A visual display of the magnitude of the different categories of a single variable is shown, with each category or value of the variable represented by a bar.
Grouped bar chart	Multiple sets of data are displayed as side-by-side bars.
Stacked bar chart	Although similar to a grouped bar chart, each of the segments in which the bar or column is divided belongs to a different data series. It shows how a total entity is subdivided into parts.
Deviation bar chart	Differences, both positive and negative, from the baseline are illustrated.
100% component bar chart	The bar is divided into proportions that are the same as the proportions of each category of the variable; it compares how components contribute to the whole in different groups.
Pie chart	Components of a whole are shown.
Population pyramid	This graphical illustration shows the distribution of age groups in a population for males and females.