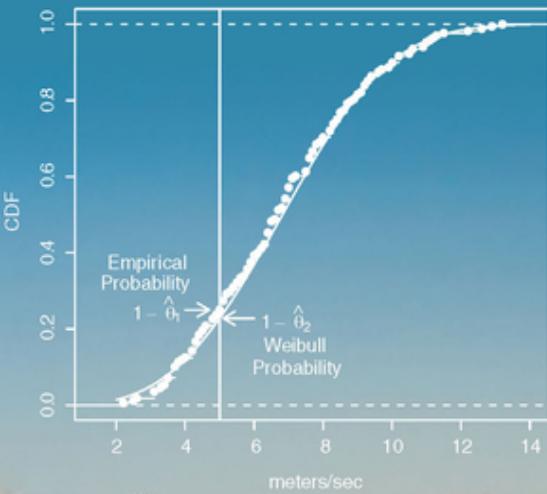


Third Edition

# Mathematical Statistics with Resampling and R

Laura M. Chihara | Tim C. Hesterberg



WILEY



## **Mathematical Statistics with Resampling and R**



# **Mathematical Statistics with Resampling and R**

Third Edition

*Laura M. Chihara*  
Carleton College, Northfield, MN, US

*Tim C. Hesterberg*  
Instacart, Seattle, WA, US

**WILEY**

This third edition first published 2022  
© 2022 John Wiley & Sons, Inc.

*Edition History*

2e: 2018, Wiley; 1e: 2011, Wiley

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by law. Advice on how to obtain permission to reuse material from this title is available at <http://www.wiley.com/go/permissions>.

The right of Laura M. Chihara and Tim C. Hesterberg to be identified as the authors of this work has been asserted in accordance with law.

*Registered Office*

John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

*Editorial Office*

John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at [www.wiley.com](http://www.wiley.com).

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

*Limit of Liability/Disclaimer of Warranty*

In view of ongoing research, equipment modifications, changes in governmental regulations, and the constant flow of information relating to the use of experimental reagents, equipment, and devices, the reader is urged to review and evaluate the information provided in the package insert or instructions for each chemical, piece of equipment, reagent, or device for, among other things, any changes in the instructions or indication of usage and for added warnings and precautions. While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

*Library of Congress Cataloging-in-Publication Data applied for*  
Hardback ISBN: 9781119874034

Cover image: Courtesy of Carleton College

Cover design by Wiley

Set in 10/12pt WarnockPro by Straive, Chennai, India

*The world seldom notices who teachers are;  
but civilization depends on what they do.*  
– Lindley Stiles

*To:  
Theodore S. Chihara*

*To:  
Bev Hesterberg*



## Contents

### Preface *xiii*

<b>1</b>	<b>Data and Case Studies</b>	<b>1</b>
1.1	Case Study: Flight Delays	1
1.2	Case Study: Birth Weights of Babies	2
1.3	Case Study: Verizon Repair Times	3
1.4	Case Study: Iowa Recidivism	4
1.5	Sampling	5
1.6	Parameters and Statistics	6
1.7	Case Study: General Social Survey	7
1.8	Sample Surveys	8
1.9	Case Study: Beer and Hot Wings	9
1.10	Case Study: Black Spruce Seedlings	10
1.11	Studies	11
1.12	Google Interview Question: Mobile Ads Optimization	13
	Exercises	16
<b>2</b>	<b>Exploratory Data Analysis</b>	<b>21</b>
2.1	Basic Plots	21
2.2	Numeric Summaries	25
2.2.1	Center	25
2.2.2	Spread	26
2.2.3	Shape	26
2.3	Boxplots	27
2.4	Quantiles and Normal Quantile Plots	29
2.5	Empirical Cumulative Distribution Functions	34
2.6	Scatter Plots	36
2.7	Skewness and Kurtosis	38
	Exercises	39

<b>3</b>	<b>Introduction to Hypothesis Testing: Permutation Tests</b>	<b>45</b>
3.1	Introduction to Hypothesis Testing	45
3.2	Hypotheses	46
3.3	Permutation Tests	50
3.3.1	Implementation Issues	54
3.3.2	One-Sided and Two-Sided Tests	58
3.3.3	Other Statistics	59
3.3.4	Conditions	62
3.3.5	Remark on Terminology	65
3.4	Matched Pairs	66
3.5	Cause and Effect	67
	Exercises	71
<b>4</b>	<b>Sampling Distributions</b>	<b>77</b>
4.1	Sampling Distributions	77
4.2	Calculating Sampling Distributions	82
4.3	The Central Limit Theorem	85
4.3.1	CLT for Binomial Data	88
4.3.2	Continuity Correction for Discrete Random Variables	90
4.3.3	Accuracy of the Central Limit Theorem*	92
4.3.4	CLT for Sampling Without Replacement	93
	Exercises	93
<b>5</b>	<b>Introduction to Confidence Intervals: The Bootstrap</b>	<b>103</b>
5.1	Introduction to the Bootstrap	103
5.2	The Plug-in Principle	109
5.2.1	Estimating the Population Distribution	110
5.2.2	How Useful Is the Bootstrap Distribution?	112
5.3	Bootstrap Percentile Intervals	115
5.4	Two Sample Bootstrap	116
5.4.1	Matched Pairs	122
5.5	Other Statistics	123
5.6	Bias	126
5.7	Monte Carlo Sampling	130
5.8	Accuracy of Bootstrap Distributions	131
5.8.1	Sample Mean, Large Sample Size	131
5.8.2	Sample Mean: Small Sample Size	132
5.8.3	Sample Median	134
5.8.4	Mean–Variance Relationship	135
5.9	How Many Bootstrap Samples Are Needed?	136
	Exercises	137

<b>6</b>	<b>Estimation</b>	<b>147</b>
6.1	Maximum Likelihood Estimation	147
6.1.1	Maximum Likelihood for Discrete Distributions	148
6.1.2	Maximum Likelihood for Continuous Distributions	150
6.1.3	Maximum Likelihood for Multiple Parameters	155
6.2	Method of Moments	158
6.3	Properties of Estimators	160
6.3.1	Unbiasedness	161
6.3.2	Efficiency	164
6.3.3	Mean Square Error	167
6.3.4	Consistency	169
6.3.5	Transformation Invariance*	171
6.3.6	Asymptotic Normality of MLE*	173
6.4	Statistical Practice	174
6.4.1	Are You Asking the Right Question?	175
6.4.2	Weights	175
	Exercises	176
<b>7</b>	<b>More Confidence Intervals</b>	<b>183</b>
7.1	Confidence Intervals for Means	183
7.1.1	Confidence Intervals for a Mean, Variance Known	183
7.1.2	Confidence Intervals for a Mean, Variance Unknown	188
7.1.3	Confidence Intervals for a Difference in Means	195
7.1.4	Matched Pairs, Revisited	201
7.2	Confidence Intervals Using Pivots	201
7.2.1	Location and Scale Parameters*	205
7.3	One-Sided Confidence Intervals	209
7.4	Confidence Intervals for Proportions	211
7.4.1	Agresti–Coull Intervals for a Proportion	214
7.4.2	Confidence Interval for a Difference of Proportions	215
7.5	Bootstrap Confidence Intervals	216
7.5.1	T Confidence Intervals Using Bootstrap Standard Errors	216
7.5.2	Bootstrap $t$ Confidence Intervals	217
7.5.3	Comparing Bootstrap $t$ and Formula $t$ Confidence Intervals	223
7.6	Confidence Interval Properties	224
7.6.1	Confidence Interval Accuracy	224
7.6.2	Confidence Interval Length	225
7.6.3	Transformation Invariance	225
7.6.4	Ease of Use and Interpretation	225
7.6.5	Research Needed	226

7.7	The Delta Method*	226
	Exercises	230
<b>8</b>	<b>More Hypothesis Testing</b>	<b>245</b>
8.1	Hypothesis Tests for Means and Proportions: One Population	245
8.1.1	A Single Mean	245
8.1.2	One Proportion	248
8.2	Bootstrap <i>t</i> Tests	250
8.3	Hypothesis Tests for Means and Proportions: Two Populations	252
8.3.1	Comparing Two Means	252
8.3.2	Comparing Two Proportions	255
8.3.3	Matched Pairs for Proportions	259
8.4	Type I and Type II Errors	261
8.4.1	Type I Errors	262
8.4.2	Type II Errors and Power	267
8.4.3	<i>P</i> -Values Versus Critical Regions	272
8.4.4	Relationship Between Confidence Intervals and Hypothesis Tests	273
8.5	Interpreting Test Results	276
8.5.1	Terminology	277
8.5.2	Arbitrary Thresholds	277
8.5.3	Statistical Discernibility Versus Practical Importance	277
8.5.4	Negative Results	278
8.5.5	Inflated False Positive Rate	279
8.6	Likelihood Ratio Tests	281
8.6.1	Simple Hypotheses and the Neyman–Pearson Lemma	281
8.6.2	Likelihood Ratio Tests for Composite Hypotheses	285
8.7	Statistical Practice	289
8.7.1	More Campaigns with No Clicks and No Conversions	293
	Exercises	294
<b>9</b>	<b>Regression</b>	<b>309</b>
9.1	Covariance	309
9.2	Correlation	313
9.3	Least Squares Regression	316
9.3.1	Regression toward the Mean	320
9.3.2	Variation	321
9.3.3	Diagnostics	323
9.3.4	Multiple Regression	328
9.4	The Simple Linear Model	329
9.4.1	Inference for $\alpha$ and $\beta$	333
9.4.2	Inference for the Response	336
9.4.3	Comments About Conditions for the Linear Model	340
9.5	Resampling Correlation and Regression	342

9.5.1	Permutation Tests	345
9.5.2	Bootstrap Case Study: Bushmeat	346
9.6	Logistic Regression	350
9.6.1	Inference for Logistic Regression	355
	Exercises	357
<b>10</b>	<b>Categorical Data</b>	<b>367</b>
10.1	Independence in Contingency Tables	367
10.2	Permutation Test of Independence	369
10.3	Chi-Square Test of Independence	371
10.3.1	Model for Chi-Square Test of Independence	373
10.3.2	$2 \times 2$ Tables	376
10.3.3	Fisher's Exact Test	378
10.3.4	Conditioning	379
10.4	Chi-Square Test of Homogeneity	380
10.5	Goodness-of-Fit Tests	382
10.5.1	All Parameters Known	382
10.5.2	Some Parameters Estimated	385
10.6	Chi-Square and the Likelihood Ratio*	388
	Exercises	389
<b>11</b>	<b>Bayesian Methods</b>	<b>399</b>
11.1	Bayes Theorem	400
11.2	Binomial Data: Discrete Prior Distributions	400
11.3	Binomial Data: Continuous Prior Distributions	408
11.4	Continuous Data	414
11.5	Sequential Data	417
	Exercises	421
<b>12</b>	<b>One-Way ANOVA</b>	<b>429</b>
12.1	Comparing Three or More Populations	429
12.1.1	The ANOVA $F$ Test	429
12.1.2	A Permutation Test Approach	438
	Exercises	439
<b>13</b>	<b>Additional Topics</b>	<b>443</b>
13.1	Smoothed Bootstrap	444
13.1.1	Kernel Density Estimate	444
13.2	Parametric Bootstrap	449
13.3	Stratified Sampling	452
13.3.1	Post-stratification	453
13.3.2	Optimal Stratified Sampling	454
13.4	Control Variates and Casual Modeling	455
13.4.1	Control Variates in Experiments	457

13.4.2	Potential Outcomes Framework	460
13.4.3	Observational Data – Causal Modeling	461
13.5	Computational Issues in Bayesian Analysis	462
13.6	Monte Carlo Integration	464
13.7	Importance Sampling	468
13.7.1	Ratio Estimate for Importance Sampling	475
13.7.2	Importance Sampling in Bayesian Applications	478
13.8	The EM Algorithm	483
13.8.1	EM in General	485
	Exercises	488

**Appendix A Review of Probability** 493

A.1	Basic Probability	493
A.2	Mean and Variance	494
A.3	Marginal and Conditional Distributions	496
A.4	The Normal Distribution	497
A.5	The Mean of a Sample of Random Variables	498
A.6	Sums of Normal Random Variables	499
A.7	The Law of Averages	500
A.8	Higher Moments and the Moment Generating Function	501

**Appendix B Probability Distributions** 505

B.1	The Bernoulli and Binomial Distributions	505
B.2	The Multinomial Distribution	506
B.3	The Geometric Distribution	508
B.4	The Negative Binomial Distribution	509
B.5	The Hypergeometric Distribution	510
B.6	The Poisson Distribution	511
B.7	The Uniform Distribution	513
B.8	The Exponential Distribution	513
B.9	The Gamma Distribution	514
B.10	The Chi-Square Distribution	517
B.11	The Student's <i>t</i> Distribution	520
B.12	The Beta Distribution	522
B.13	The <i>F</i> Distribution	523
	Exercises	525

**Appendix C Distributions Quick Reference** 527**Problem Solutions** 531**Bibliography** 545**Index** 553

## Preface

Mathematical Statistics with Resampling and R is a one term undergraduate statistics textbook aimed at sophomores or juniors who have taken a course in probability (at the level of, for instance, Ross (2009), Ghahramani (2004), or Scheaffer and Young (2010)) but may not have had any previous exposure to statistics.

What sets this book apart from other mathematical statistics texts is the use of modern resampling techniques – permutation tests and bootstrapping. We begin with permutation tests and bootstrap methods before introducing classical inference methods. Resampling helps students understand the meaning of sampling distributions, sampling variability,  $P$ -values, hypothesis tests, and confidence intervals. We are inspired by the textbooks of Wardrop (1995) and Chance and Rossman (2005), two innovative introductory statistics books which also take a non-traditional approach in the sequencing of topics.

We believe the time is ripe for this book. Many faculty have learned resampling and simulation-based methods in graduate school and/or use them in their own work, and are eager to incorporate these ideas into a mathematical statistics course. Students and faculty today have access to computers that are powerful enough to perform resampling quickly.

A major topic of debate about the Mathematical Statistics course is how much theory to introduce. We want mathematically talented students to get excited about statistics, so we try to strike a balance between theory, computing, and applications. We feel that it is important to demonstrate some rigor in developing some of the statistical ideas presented here, but that mathematical theory should not dominate the text. To keep the size of the text reasonable, we omit some topics such as sufficiency and Fisher information (though we plan to make some omitted topics available as supplements on the text web page <https://github.com/lchihara/MathStatsResamplingR>).

We have compiled the definitions and theorems of the important probability distributions into an appendix (see Appendix B). Instructors who want to prove results on distributional theory can refer to that chapter. Instructors who wish to skip the theory can continue without interrupting the flow of the statistical discussion.

Incorporating resampling and bootstrapping methods requires that students use statistical software. We use R or RStudio because they are freely available ([www.r-project.org](http://www.r-project.org) or <http://rstudio.com>), powerful, flexible, and a valuable tool in future careers. One of us worked at Google where there was an explosion in the use of R, with more and more non-statisticians learning R (the statisticians already know it). We realize that the learning curve for R is high, but believe that the time invested in mastering R is worth the effort. We have written some basic materials on R that are available on the website for this text. We recommend that instructors work through the introductory worksheet with the students on the first or second day of the term, in a computer lab if possible.

For the third edition, we decided to incorporate the packages in Hadley Wickham's **tidyverse**, including **ggplot2**. And though some R packages exist that implement some of the bootstrap and permutation algorithms that we teach, we felt that students understand and internalize the concepts better if they are required to write the code themselves. We do provide R scripts or R Markdown files with code on our website, and we may include alternate coding using some of the many R packages available.

Statistical computing is necessary in statistical practice and for people working with data in a wide variety of fields. There is an explosion of data – more and more data – and new computational methods are continuously being developed to handle this explosion. Statistics is an exciting field, dare we even say sexy?<sup>1</sup>

**Third Edition:** The issue of *P*-values has generated much discussion in the statistics community (Wasserstein and Lazar, 2016; Wasserstein et al., 2019), and has even made it into the popular press (Resnick, 2019). In light of this controversy, one of the major changes we have made in this edition is to move away from the term "statistically significant" in favor of "statistically discernible." We discuss this in Section 8.5 (Interpreting Test Results).

As noted above, we also updated the R code, using the **ggplot2** package for plots instead of base R. We've added sections on cause-and-effect, control variates, expanded on stratification (Chapter 13), moved the section on the delta method from Chapter 13 to Chapter 7, updated the General Social Survey data set, included more examples, and developed more exercises. Through-out the text, we have updated, clarified or made small changes to the exposition.

---

<sup>1</sup> Try googling "statistics sexy profession."

**Pathways:** This textbook contains more than enough material for a one term undergraduate course. We have written the textbook in such a way that instructors can choose to skip or gloss over some sections if they wish to emphasize others. In some instances, we have labeled a section or subsection with an asterisk (\*) to denote it as optional. For classes comprised primarily of students who have no statistics background, a possible sequence includes Chapters 1–10. For courses focused more on applications, instructors could omit, for example, Sections 7.2 (Confidence Intervals Using Pivots) and 8.6 (Likelihood Ratio Tests). For classes in which students come in with an introductory statistics course background, instructors could have students read the first two chapters on their own, beginning the course at Chapter 3. In this case, instructors may wish to spend more time on theory, Bayesian methods, the delta method in Chapter 7, or the topics in Chapter 13, including the parametric bootstrap, stratified sampling and importance sampling. These topics could also be assigned as final projects for an undergraduate course or a senior capstone thesis.

**Acknowledgments:** This textbook could not have been completed without the assistance of many colleagues and students. In particular, for the first edition, we would like to thank Professor Katherine St. Clair of Carleton College who bravely class-tested an early (very!) rough draft in her *Introduction to Statistical Inference* class during Winter 2010. In addition, Professor Julie Legler of St. Olaf College adopted the manuscript in her *Statistical Theory* class for Fall 2010. Both instructors and students provided valuable feedback that improved the exposition and content of this textbook. For the second edition, we would like to thank Professors Elaine Newman (Sonoma State University), Eric Nordmoe (Kalamazoo College), Nick Horton (Amherst College), and Carleton College faculty Katie St. Clair, Andy Poppick and Adam Loy for their helpful comments. We thank Ed Lee of Google for the Mobile Ads data and explanation. For the third edition, we would like to thank Jeff Witmer (Oberlin) and Miles Ott (Johnson & Johnson) for insightful comments; Ruhan Zhang of Instacart for the ad evaluation example in Exercise 13.14; and Rachel Zhang, Nick Cooley, and Jeff Moulton of Instacart for the human evaluation framework Exercise 13.15. Finally, we thank Professor Albert Y. Kim (Smith College) who compiled the data sets into an R package (**resampled****data**).

We would also like to thank Siyuan (Ernest) Liu and Chen (Daisy) Sun, two Carleton College students, for solving many of the exercises in the first edition and writing up the solutions with L<sup>A</sup>T<sub>E</sub>X.

Finally, we are also grateful for the valuable assistance provided by the staff at Wiley, including Skyler Van Valkenburgh, Judit Anbu Hena, Isabella Proietti, and Kalli Schutea. For the first and second editions, we were helped by Jon Gurstelle, Amudhapriya Sivamurthy, Kshitija Iyer, Vishnu Narayanan, Kathleen Pagliaro, Steve Quigley, Sanchari Sill, Dean Gonzalez, and Jackie Palmieri.

**Additional Resources:** We will place additional materials including R scripts, data sets, tutorials, and errata at our github site

<https://github.com/lchihara/MathStatsResamplingR>.

Northfield MN

Seattle WA

April 2022

*Laura M. Chihara*

*Tim C. Hesterberg*

# 1

## Data and Case Studies

Statistics is the art and science of collecting and analyzing data and understanding the nature of variability. Mathematics, especially probability, governs the underlying theory, but statistics is driven by applications to real problems.

In this chapter, we introduce several data sets that we will encounter throughout the text in the examples and exercises. These data sets are available in the R package `resampleddata3` or at the textbook website <https://github.com/lchihara/MathStatsResamplingR>.

### 1.1 Case Study: Flight Delays

If you have ever traveled by air, you probably have experienced the frustration of flight delays. The Bureau of Transportation Statistics maintains data on all aspects of air travel, including flight delays at departure and arrival.<sup>1</sup>

LaGuardia Airport (LGA) is one of three major airports that serves the New York City metropolitan area. In 2008, over 23 million passengers and over 375 000 planes flew in or out of LGA. United Airlines and American Airlines are two major airlines that schedule services at LGA. The data set `FlightDelays` contains information on all 4029 departures of these two airlines from LGA during May and June 2009 (Tables 1.1 and 1.2).

Each row of the data set is an *observation*. Each column represents a *variable* – some characteristic that is obtained for each observation. For instance, on the first observation listed, the flight was a United Airlines plane, flight number 403, destined for Denver, and departing on Friday between 4 and 8 a.m. This data set consists of 4029 observations and 9 variables.

Questions we might ask include the following: Are flight delay times different between the two airlines? Are flight delay times different depending on the day

<sup>1</sup> <http://www.bts.gov/xml/ontimesummarystatistics/src/index.xml>.

**Table 1.1** Partial view of `FlightDelays` data.

Flight	Carrier	FlightNo	Destination	DepartTime	Day
1	UA	403	DEN	4–8 a.m.	Friday
2	UA	405	DEN	8–noon	Friday
3	UA	409	DEN	4–8 p.m.	Friday
4	UA	511	ORD	8–noon	Friday
		:			

**Table 1.2** Variables in data set `FlightDelays`.

Variable	Description
Carrier	UA = United Airlines, AA = American Airlines
FlightNo	Flight number
Destination	Airport code
DepartTime	Scheduled departure time in 4 h intervals
Day	Day of week
Month	May or June
Delay	Minutes flight delayed (negative indicates early departure)
Delayed30	Departure delayed more than 30 min?
FlightLength	Length of time of flight (minutes)

of the week? Are flights scheduled in the morning less likely to be delayed by more than 15 min?

## 1.2 Case Study: Birth Weights of Babies

The birth weight of a baby is of interest to health officials since many studies have shown possible links between this weight and conditions in later life, such as obesity or diabetes. Researchers look for possible relationships between the birth weight of a baby and the age of the mother or whether or not she smoked cigarettes or drank alcohol during her pregnancy. The Centers for Disease Control and Prevention (CDC) maintains a database on all babies born in a given year,<sup>2</sup> incorporating data provided by the US Department of Health and

---

<sup>2</sup> <http://wonder.cdc.gov/natality-current.html>.

**Table 1.3** Variables in data set NCBirths2004.

Variable	Description
MothersAge	Mother's age
Smoker	Mother smoker or non-smoker
Gender	Gender of baby
Weight	Weight at birth (grams)
Gestation	Gestation time (weeks)

Human Services, the National Center for Health Statistics, and the Division of Vital Statistics. We will investigate different samples taken from the CDC's database of births.

One data set that we will investigate consists of a random sample of 1009 babies born in North Carolina during 2004 (Table 1.3). The babies in the sample had a gestation period of at least 37 weeks and were single births (i.e. not a twin or triplet).

In addition, we will also investigate a data set, Girls2004, consisting of a random sample of 40 baby girls born in Alaska and 40 baby girls born in Wyoming. These babies also had a gestation period of at least 37 weeks and were single births.

The data set TXBirths2004 contains a random sample of 1587 babies born in Texas in 2004. In this case, the sample was not restricted to single births, nor to a gestation period of at least 37 weeks. The numeric variable Number indicates whether the baby was a single birth, or one of a twin, triplet, and so on. The variable Multiple is a factor variable indicating whether or not the baby was a multiple birth.

### 1.3 Case Study: Verizon Repair Times

Verizon is the primary local telephone company (incumbent local exchange carrier (ILEC)) for a large area of the Eastern United States. As such, it is responsible for providing repair service for the customers of other telephone companies known as competing local exchange carriers (CLECs) in this region. Verizon is subject to fines if the repair times (the time it takes to fix a problem) for CLEC customers are substantially worse than those for Verizon customers.

The data set Verizon contains a sample of repair times for 1664 ILEC and 23 CLEC customers (Table 1.4). The mean repair times are 8.4 h for ILEC customers and 16.5 h for CLEC customers. Could a difference this large be easily explained by chance?

**Table 1.4** Variables in data set **Verizon**.

Variable	Description
Time	Repair times (in hours)
Group	ILEC or CLEC

## 1.4 Case Study: Iowa Recidivism

When a person is released from prison, will he or she relapse into criminal behavior and be sent back? The state of Iowa tracks offenders over a 3-year period, and records the number of days until recidivism for those who are readmitted to prison. The Department of Corrections uses this recidivism data to determine whether or not their strategies for preventing offenders from relapsing into criminal behavior are effective.

The data set **Recidivism** contains all offenders convicted of either a misdemeanor or felony who were released from an Iowa prison during the 2010 fiscal year (ending in June) (Table 1.5). There were 17 022 people released in that period, of whom 5386 were sent back to prison in the following 3 years (through the end of the 2013 fiscal year).<sup>3</sup>

The recidivism rate for those under the age of 25 years was 36.5% compared to 30.6% for those 25 years or older. Does this indicate a real difference in the behavior of those in these age groups, or could this be explained by chance variability?

**Table 1.5** Variables in data set **Iowa Recidivism**.

Variable	Description
Gender	F, M
Age	Age at release: Under 25, 25–34, 35–44, 45–54, 55 and Older
Age25	Under 25, Over 25 (binary)
Offense	Original conviction: Felony or Misdemeanor
Recid	Recidivate? No, Yes
Type	New (crime), No Recidivism, Tech (technical violation, such as a parole violation)
Days	Number of days to recidivism; NA if no recidivism

<sup>3</sup> <https://data.iowa.gov/Public-Safety/3-Year-Recidivism-for-Offenders-Released-from-Pris/mw8r-vqy4>.

## 1.5 Sampling

In analyzing data, we need to determine whether the data represent a *population* or a *sample*. A *population* represents all the individual cases, whether they are babies, fish, cars, or coin flips. The data from the flight delays case study in Section 1.1 are *all* the flight departures of United Airlines and American Airlines out of LGA in May and June 2009; thus, this data set represents the population of all such flights. On the other hand, the North Carolina data set contains only a subset of 1009 births from over 100 000 births in North Carolina in 2004. In this case, we will want to know how representative statistics computed from this sample are for the entire population of North Carolina babies born in 2004.

Populations may be finite, such as births in 2004, or infinite, such as coin flips or births next year.

Throughout this book, we will talk about drawing random samples from a population. We will use capital letters (e.g.  $X$ ,  $Y$ ,  $Z$ , and so on) to denote random variables and lower-case letters (e.g.  $x_1$ ,  $x_2$ ,  $x_3$ , and so on) to denote actual values or data.

There are many kinds of random samples. Strictly speaking, a “random sample” is any sample obtained using a random procedure. However, we use *random sample* to mean a sample of independent and identically distributed (i.i.d.) observations from the population, if the population is infinite.

For instance, suppose you toss a fair coin 20 times and consider each head a “success.” Then your sample consists of the random variables  $X_1, X_2, \dots, X_{20}$ , each a Bernoulli random variable with success probability 1/2. We use the notation  $X_i \sim \text{Bern}(1/2)$ ,  $i = 1, 2, \dots, 20$ .

If the population of interest is finite  $\{x_1, x_2, \dots, x_N\}$ , we can choose a random sample as follows: Label  $N$  balls with the numbers 1, 2, …,  $N$  and place them in an urn. Draw a ball at random, record its value  $X_1 = x_{i_1}$ , and then replace the ball. Draw another ball at random, record its value,  $X_2 = x_{i_2}$ , and replace. Continue until you have a sample  $x_{i_1}, x_{i_2}, \dots, x_{i_n}$ . This is *sampling with replacement*. For instance, if  $N = 5$  and  $n = 2$ , then there are  $5 \times 5 = 25$  different samples of size 2 (where order matters). (*Note:* By “order matters” we do not imply that order matters in practice, rather we mean that we keep track of the order of the elements when enumerating samples. For instance, the set  $\{a, b\}$  is different from  $\{b, a\}$ .)

However, in most real situations, for example, in conducting surveys, we do not want to have the same person polled twice. So we would sample *without replacement*, in which case, we will not have independence. For instance, if you wish to draw a sample of size  $n = 2$  from a population of  $N = 10$  people, then the probability of any one person being selected is 1/10. However, after having chosen that first person, the probability of any one of the remaining people being chosen is now 1/9.

In cases where populations are very large compared to the sample size, calculations under sampling without replacement are reasonably approximated by calculations under sampling with replacement.

**Example 1.1** Consider a population of 1000 people, 350 of whom are smokers, and the rest are nonsmokers. If you select 10 people at random but with replacement, then the probability that 4 are smokers is  $\binom{10}{4} (350/1000)^4 (650/1000)^6 \approx 0.2377$ . If you select without replacement, then the probability is  $\binom{350}{4} \binom{650}{6} / \binom{1000}{10} \approx 0.2388$ .  $\square$

## 1.6 Parameters and Statistics

When discussing numeric information, we will want to distinguish between populations and samples.

**Definition 1.1** A *parameter* is a (numerical) characteristic of a population or of a probability distribution.

A *statistic* is a (numerical) characteristic of data.  $\parallel$

Any function of a parameter is also a parameter; any function of a statistic is also a statistic. When the statistic is computed from a random sample, it is itself random, and hence is a random variable.

**Example 1.2**  $\mu$  and  $\sigma$  are parameters of the normal distribution with pdf  $f(x) = (1/\sqrt{2\pi}\sigma)e^{-(x-\mu)^2/(2\sigma^2)}$ .

The variance  $\sigma^2$  and *signal-to-noise ratio*  $\mu/\sigma$  are also parameters.  $\square$

**Example 1.3** If  $X_1, X_2, \dots, X_n$  are a random sample, then the mean  $\bar{X} = 1/n \sum_{i=1}^n X_i$  is a statistic.  $\square$

**Example 1.4** Consider the population of all babies born in the United States in 2017. Let  $\mu$  denote the average weight of all these babies. Then  $\mu$  is a parameter. The average weight of a sample of 2500 babies born in that year is a statistic.  $\square$

**Example 1.5** If we consider the population of all adults in the United States today, the proportion  $p$  who approve of the president's job performance is a parameter. The fraction  $\hat{p}$  who approve in any given sample is a statistic.  $\square$

**Example 1.6** The average weight of 1009 babies in the North Carolina case study in Section 1.2 is 3448.26 g. This average is a statistic.  $\square$

**Example 1.7** If we survey 1000 adults and find that 60% intend to vote in the next presidential election, then  $\hat{p} = 0.60$  is a statistic – it estimates the parameter  $p$ , the proportion of all adults who intend to vote in the next election.  $\square$

## 1.7 Case Study: General Social Survey

The General Social Survey (GSS) is a major survey that has tracked American demographics, characteristics, and views on social and cultural issues since the 1970s. It is conducted by the National Opinion Research Center (NORC) at the University of Chicago. Trained interviewers meet face to face with the adults chosen for the survey and question them for about 90 min in their homes.

The GSS case study includes the responses of 2348 participants selected in 2018 to a subset of the questions asked, listed in Table 1.6 (Smith et al., 2014). For example, one of the questions (`Courts`) asked whether the respondent thinks that the courts in their area deal too harshly or not harshly enough with criminals.

One of the core variables that has been included in all GSS surveys since its inception is `Sex` with the values (Female/Male) coded by the interviewer.

**Table 1.6** Variables in data set `GSS2018`.

Variable	Description
Region	Interview location
GenderNow	Current gender
Age	Age of respondent ( <i>Note:</i> 89 = 89 or older)
Marital	Marital status
Degree	Highest level of education
Employed	Respondent employed? (Yes = full/part-time; No = unemployed, retired, in school, etc.)
Income	Respondents income
Polviews	Respondents political views
Pres16	Whom did you vote for in the 2016 presidential election?
DeathPenalty	Death penalty for murder?
Courts	Courts deal harshly with criminals?
Attend	Attendance at religious services
Postlife	Believe in life after death?
Happy	General happiness
Satfin	Satisfied with financial situation?
Energy	Amount of spending on alternative energy

Starting in 2018, GSS added a new question, “What is your current gender?”, to reflect recent research on the nature of gender identity which indicates that gender is not binary.<sup>4</sup> Definitions of terms and concepts are always changing over time so it is important to consider the context when analyzing data.

We will analyze the GSS data to investigate questions such as the following: Is there a relationship between a person’s marital status and whom they voted for in the 2016 presidential election? Are people who live in certain regions happier? Are there educational differences in support for the death penalty? These data can be obtained with the GSS Data Explorer.<sup>5</sup>

## 1.8 Sample Surveys

“Who do you plan to vote for in the next presidential election?” “Would you purchase our product again in the future?” “Do you smoke cigarettes? If yes, how old were you when you first started?” Questions such as these are typical of sample surveys. Researchers want to know something about a population of individuals, whether they are registered voters, online shoppers, or American teenagers, but to poll every individual in the population – that is, to take a *census* – is impractical and costly. Thus, researchers will settle for a sample from the target population. But if, say, 60% of those in your sample of 1000 adults intend to vote for candidate Wong in the next election, how close is this to the actual percentage who will vote for Wong? How can we be sure that this sample is truly representative of the population of all voters? We will learn techniques for *statistical inference*, drawing a conclusion about a population based on information about a sample.

When conducting a survey, researchers will start with a *sampling frame* – a list from which the researchers will choose their sample. For example, to survey all students at a college, the campus directory listing could be a sampling frame. For pre-election surveys, many polling organizations use a sampling frame of registered voters. Note that the choice of sampling frame could introduce the problem of *undercoverage*: omitting people from the target population in the survey. For instance, young people were missed in many pre-election surveys during the 2008 Obama–McCain presidential race because they had not yet registered to vote.

Once the researchers have a sampling frame, they will then draw a random sample from this frame. Researchers will use some type of *probability (scientific) sampling scheme*, that is, a scheme that gives everybody in the population a positive chance of being selected. For example, to obtain a sample of size 10 from a population of 100 individuals, write each person’s name on a

---

<sup>4</sup> This question was only asked of about two-thirds of the 2348 participants.

<sup>5</sup> <https://gssdataexplorer.norc.org>.

slip of paper, put the slips of paper into a basket, and then draw out 10 slips of paper. Nowadays, statistical software is used to draw random samples from a sampling frame.

Another basic survey design uses *stratified sampling*: The population is divided into nonoverlapping strata, and then random samples are drawn from each stratum. The idea is to group individuals who are similar in some characteristic into homogeneous groups, thus reducing variability. For instance, in a survey of university students, a researcher might divide the students by class: first year, sophomores, juniors, seniors, and graduate students. A market analyst for an electronics store might choose to stratify customers based on income levels.

In *cluster sampling*, the population is divided into nonoverlapping clusters, and then a random sample of clusters is drawn. Every person in a chosen cluster is then interviewed for the survey. An airport wanting to conduct a customer satisfaction survey might use a sampling frame of all flights scheduled to depart from the airport on a certain day. A random sample of flights (clusters) is chosen, and then all passengers on these flights are surveyed. A modification of this design might involve sampling in stages: For instance, the analysts might first choose a random sample of flights, and then from each flight choose a random sample of passengers.

The GSS uses a more complex sampling scheme in which the sampling frame is a list of counties and county equivalents (standard metropolitan statistical areas) in the United States. These counties are stratified by region, age, and race. Once a sample of counties is obtained, a sample of block groups and enumeration districts is selected, stratifying these by race and income. The next stage is to randomly select blocks and then interview a specific number of men and women who live within these blocks.

Indeed, all major polling organizations such as Gallup or Roper as well as the GSS use a *multistage* sampling design. In this book, we use the GSS data or polling results for examples as if the survey design used simple random sampling. Calculations for more complex sampling scheme are beyond the scope of this book, and we refer the interested reader to Lohr (1991) for details.

## 1.9 Case Study: Beer and Hot Wings

Carleton student Nicki Catchpole conducted a study of hot wings and beer consumption at the Williams Bar in the Uptown area of Minneapolis (N. Catchpole, private communication). She asked patrons at the bar to record their consumption of hot wings and beer over the course of several hours. She wanted to know if people who ate more hot wings would then drink more beer. In addition, she asked each person their gender to investigate whether or not gender had an impact on hot wings or beer consumption.

**Table 1.7** Variables in data set **Beerwings**.

Variable	Description
Gender	Male or female
Beer	Ounces of beer consumed
Hotwings	Number of hot wings eaten

The data for this study are in **Beerwings** (Table 1.7). There are 30 observations and 3 variables.

## 1.10 Case Study: Black Spruce Seedlings

Black spruce (*Picea mariana*) is a species of a slow-growing coniferous tree found across the northern part of North America. It is commonly found on wet organic soils. In a study conducted in the 1990s, a biologist interested in factors affecting the growth of the black spruce planted its seedlings on sites located in boreal peatlands in northern Manitoba, Canada (Camill et al., 2010).

The data set **Spruce** contains a part of the data from the study (Table 1.8). Seventy-two black spruce seedlings were planted in four plots under varying conditions (fertilizer—no fertilizer, competition—no competition), and their heights and diameters were measured over the course of 5 years.

The researcher wanted to see whether the addition of fertilizer or the removal of competition from other plants (by weeding) affected the growth of these seedlings.

**Table 1.8** Variables in data set **Spruce**.

Variable	Description
Tree	Tree number
Competition	C (competition), CR (competition removed)
Fertilizer	F (fertilized), NF (not fertilized)
Height0	Height (cm) of seedling at planting
Height5	Height (cm) of seedling at year 5
Diameter0	Diameter (cm) of seedling at planting
Diameter5	Diameter (cm) of seedling at year 5
Ht.change	Change (cm) in height
Di.change	Change (cm) in diameter

## 1.11 Studies

Researchers carry out studies to understand the conditions and causes of certain outcomes: Does smoking cause lung cancer? Do teenagers who smoke marijuana tend to move on to harder drugs? Do males eat more hot wings than females? Do black spruce seedlings grow taller in fertilized plots?

The beer and hot wings case study in Section 1.9 is an example of an *observational study*, a study in which researchers observe participants but do not influence the outcome. In this case, the student just recorded the gender and the number of hot wings eaten and beer consumed as reported by these patrons in the Williams bar.

**Example 1.8** The first Nurses' Health Study is a major observational study funded by the National Institutes of Health. Over 120 000 registered female nurses who were married, between the ages of 33 and 55 years, and lived in the 11 most populous states (all in 1976) have been responding every 2 years to written questions about their health and lifestyle, including smoking habits, hormone use, and menopause status. Many results on women's health have come out of this study, such as finding an association between taking estrogen after menopause and lowering the risk of heart disease, and determining that for nonsmokers there is no link between taking birth control pills and developing heart disease.

Because this is an observational study, no *cause and effect* conclusions can be drawn. For instance, we cannot state that taking estrogen after menopause will *cause* a lowering of the risk for heart disease. In an observational study, there may be many unrecorded or hidden factors that impact the outcomes. Also, because the participants in this study are registered nurses, we need to be careful about making inferences about the general female population. Nurses are more educated and more aware of health issues than the average person. □

On the other hand, the black spruce case study in Section 1.10 was an *experiment*. In an experiment, researchers will manipulate the environment in some way to observe the response of the objects of interest (people, mice, ball bearings, etc.). When the objects of interest in an experiment are people, we refer to them as *subjects*; otherwise, we call them *experimental units*. In this case, the biologist randomly assigned the experimental units – the seedlings – to plots subject to four *treatments*: fertilization with competition, fertilization without competition, no fertilization with competition, and no fertilization with no competition. He then recorded their height over a period of several years.

A key feature in this experiment was the *random assignment* of the seedlings to the treatments. The idea is to spread out the effects of unknown or uncontrollable factors that might introduce unwanted variability into the results. For instance, if the biologist had planted all the seedlings obtained from one particular nursery in the fertilized, no competition plot and subsequently recorded that these seedlings grew the least, then he would not be able to discern whether this was due to this particular treatment or due to some possible problem with seedlings from this nursery. With random assignment of treatments, the seedlings from this particular nursery would usually be spread out over the four treatments. Thus, the differences between the treatment groups should be due to the treatments (or chance).

**Example 1.9** Knee osteoarthritis (OA) that results in deterioration of cartilage in the joint is a common source of pain and disability for the elderly population. In a 2008 paper, “Tai Chi is effective in treating knee osteoarthritis: A randomized controlled trial,” Wang et al. (2009) at Tufts University Medical School describe an experiment they conducted to see whether practicing tai chi, a style of Chinese martial arts, could alleviate pain from OA. Forty patients over the age of 65 with confirmed knee OA but otherwise in good health were recruited from the Boston area. Twenty were randomly assigned to attend twice weekly 60 min sessions of tai chi for 12 weeks. The remaining 20 participants, the *control group*, attended twice weekly 60 min sessions of instructions on health and nutrition, as well as some stretching exercises.

At the end of the 12 weeks, those in the tai chi group reported a decrease in knee pain. Because the subjects were randomly assigned to the two treatments, the researchers can assert that the tai chi sessions lead to decrease in knee pain due to OA. Note that because the subjects were recruited, we need to be careful about making an inference about the general elderly population: People who voluntarily sign up to be in an experiment may be different from other people. □

Another important feature of a well-designed experiment is *blinding*: A *double-blind* experiment is one in which neither the researcher nor the subject knows who is receiving which treatment. An experiment is *single-blinded* if just the researcher or the subject (but not both) knows who is receiving which treatment. Blinding is important in reducing *bias*, the systematic favoring of one outcome over another.

For instance, suppose in a clinical trial to test the efficacy of a new drug for a disease, the subjects know whether they are receiving the drug or a placebo (a pill or drug with no therapeutic effect). Those on the placebo might feel that the trial is a waste of time and drop out, or perhaps seek additional treatment elsewhere. On the other hand, if the researcher knows that a subject received the drug, the researcher then might behave differently toward the subject, perhaps by asking leading questions that result in responses that appear to suggest relief from the disease.

## 1.12 Google Interview Question: Mobile Ads Optimization

The following question was posted on an internal Google statistics email list:

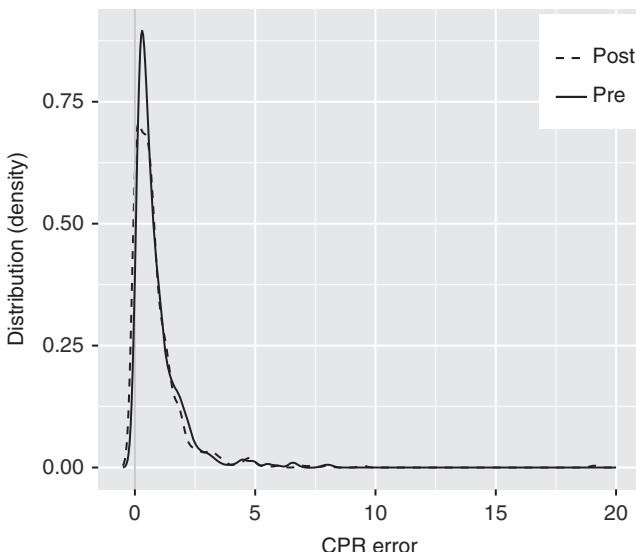
I have a pre v post comparison I'm trying to make where alternative hypothesis is  $\text{pre.mean.error} > \text{post.mean.error}$ . My distribution for these samples are both right-skewed as shown below. Anyone know what test method would be best suited for this type of situation?

When I (Tim) interview candidates for a quantitative analyst position at Google, I frequently ask applicants to “Imagine that you are consulting with this person. What would you ask or tell this person?”

Many applicants start off on the wrong track by proposing various statistical tests without first understanding the problem or data. Data analysis and statistics are ultimately about solving problems, not just applying techniques, so we need to begin by asking questions to gain insight about the data, the context, and the problem itself.

Important questions include: What is CPR error? What does pre versus post mean? How many observations are there? How were the data collected? How are the data related? Also, make sure to understand the distribution in Figure 1.1.

In this data set, the pre and post variables are paired, not independent – and an answer that does not take that into account is wrong.



**Figure 1.1** Density plot for CPR error.

It is the duty of the consultant to not only answer the questions that the client posed but also to think about whether those are the right questions to ask. In this case the client asked how to compare means (averages), but there are outliers, so comparing averages is inaccurate; we should consider other comparisons.<sup>6</sup> However, once we learn more about the data, it turns out that answering a slightly different question is probably better in both practical terms (does a better job of measuring what really matters) and statistical terms (more accurate, better signal-to-noise ratio). For this modified problem the outliers are much less of an issue; we discuss this in Section 8.7.

Once a consultant understands the data, she is in a position to suggest appropriate methods. In this book, we will provide guidelines for determining methods to handle different data scenarios.

Here is some more background.<sup>7</sup> *Google Adwords*<sup>TM</sup> lets advertisers bid for ads to appear when people search on Google. An auction determines which ads are shown, based on a combination of the bid and Google's estimate of how interested the searcher would be in an ad. If an ad places high in the auction, it is shown to the user; if the user clicks on the ad, then the advertiser pays Google.

Within Adwords, *Enhanced Campaigns* offered advertisers the ability to customize their bidding, e.g. to show one ad to someone searching for "pizza" at 1 p.m. on their PC at work (perhaps a link to an online order form or menu), and a different ad to someone searching for "pizza" at 8 p.m. on a smartphone a half-mile from the restaurant (perhaps a click-to-call phone number and restaurant locator). When this was launched, many advertisers did not understand how to bid appropriately on mobile phones.

Google crafted an experiment to help advertisers bid appropriately. The experiment was based on equalizing the return on the ad investment between desktop and mobile platforms because \$1 revenue on a mobile device is the same benefit to advertisers as \$1 revenue on a desktop computer. If the return on investment (the ratio of the amount of user purchases to the cost of advertising to the users) was higher on mobile devices than on desktops in the "pre period" (before the experiment), Google would recommend increasing the "mobile multiplier," the ratio between the mobile and desktop bids. This would result in more ads being shown to mobile customers with an increased average cost, and fewer ads shown to desktop customers with a decreased average cost, and consequently, tend to make the return on investment more equal. For example, if the mobile multiplier for a particular campaign was 1.2 before the recommendation, Google might recommend increasing it to 1.4.

<sup>6</sup> Do not worry if you do not understand the statistical terms in this section. We will discuss them in the chapters ahead!

<sup>7</sup> I give a simpler explanation during interviews – once people ask!

As a result of Google's recommendations, advertisers raised their bids in some cases and lowered them in others, resulting in both a lower cost for advertising and greater return for them.

Why would Google do this if it made less money? Two reasons – advertisers and users. In the long run, advertisers will use Google ads more if they get more bang for their buck. And it is better for Google's users if advertisers target their ads to people interested in their product and avoid showing their ads to people who are not.

However, this was not a pure randomized experiment where some advertisers were given recommendations and others not. Advertisers had to agree to participate and report the number of “conversions” (purchases) and the value of those conversions. The comparison is between before-and-after results, but some advertisers might have adjusted their bids even without the recommendations. They also were not obligated to follow the recommendations.

The experiment was designed as a pre versus post paired  $t$  test to compare results before and after the recommendations. However, the distributions of the data shown in the figure above are very long tailed making  $t$  tests questionable.

The data `MobileAds` are a subset of the experimental data for one advertiser. Each row corresponds to a single combination of campaign and ad group: These could be for different products, a different set of ads, target a different population, be shown for different searches, etc. Important variables are given in Table 1.9

Most variables have two versions,

- before the experiment (with a `_pre` suffix).
- during the experiment (with a `_post` suffix).

and two platforms:

- mobile (with an `m.` prefix)
- desktop+tablet (with a `d.` prefix).

For example, from the first row of the data set (Table 1.10), we see that on mobile devices, there are 155 impressions (ads shown) in the pre period and 255 in the post period, and on desktops or tablets, 1430 and 1466 impressions in the pre and post periods, respectively.

`error` is the difference in `cpr` (the reciprocal of return on investment) between mobile and desktop platforms; small values indicate parity and suggest efficient bidding. The analyst was interested in whether the experiment would result in reductions in `error`.

`mult_change` is the change in mobile multiplier; a negative number indicates a lower mobile multiplier in the post period. (This is what the advertiser actually did, not what Google recommended.)

**Table 1.9** Variables in data set **MobileAds**.

Variable	Description
impr	Number of ad impressions (ads shown)
click	Number of clicks
cost	What advertisers paid
conv	Number of conversions (purchases)
value	Value of conversions as reported by advertisers
cpm	Cost per impression (cost/impr)
cpc	Cost per click (cost/click)
cpa	Cost per conversion (cost/conv) (or 0, if conv is 0)
cpr	Cost per return (cost/value) (or 0 if value is 0)
Prefix indicates platform:	
m.*	Mobile, e.g. m.impr
d.*	Desktop/tablet, e.g. d.impr
Suffix indicates when:	
*_pre	Before experiment, e.g. m.impr_pre
*_post	In experiment, e.g. m.impr_post
error.cpr*	$m.cpr - d.cpr$ (pre, post)
mult.change	Change in mobile multiplier

**Table 1.10** Partial view of **MobileAds** data.

Campaign	m.impr_post	m.impr_pre	...	d.impr_post	d.impr_pre
1	255	155	...	1466	1430
2	18	4900		64 535	54 535
3	583	6857		119 831	86 900
		:			

## Exercises

- 1.1** For each of the following, describe the population and, if relevant, the sample. For each number presented, determine if it is a parameter or a statistic (or something else).

- (a) A survey of 1500 high school students finds that 47% watch the cable show “Game of Thrones.”
  - (b) The 2010 US Census reports that 9.6% of the US population was between the ages of 18 and 24 years.
  - (c) Based on the rosters of all National Basketball Association teams for the 2006–2007 season, the average height of the players was 78.93 in.
  - (d) A March 2016 Harris poll consisting of 2106 national adults, age 18 years or older, found that 19% strongly or somewhat disagree with the statement that the US has come a long way toward reaching gender equality.
- 1.2** Review the description of the Iowa recidivism case study in Section 1.4.
- (a) Does this data represent a population or a sample?
  - (b) In this data set, 19.4% of the offenders were originally convicted of a misdemeanor. Does this number represent a parameter or a statistic?
- 1.3** Researchers reported that strict rest after a concussion did not improve the outcome of patients between the age of 11 and 22 years (Thomas et al., 2015). Eighty-eight patients who went to a pediatric emergency department in Wisconsin within 24 h of a concussion were recruited for this study. They were randomly assigned to either strict rest for 5 days or the usual care of 1–2 days of rest followed by stepwise return to activity. Participants self-reported post-concussion symptoms in a diary. The patients who were assigned to strict rest for 5 days reported more daily symptoms and slower symptom resolution than those assigned to usual care.
- (a) In this experiment, what were the treatments?
  - (b) Was this a double-blind study?
  - (c) Can the researchers conclude that strict rest for 5 days causes more daily symptoms?
  - (d) Can we generalize these results to a population? Why or why not?
- 1.4** The journal *Molecular Psychiatry* reported on a study claiming that playing the video game Tetris reduces the formation of bad memories after a traumatic event (Iyadurai et al., 2017). Seventy-one patients who were involved in a motor vehicle accident and admitted to a British emergency room were recruited. After completing some baseline assessments, they were randomly assigned to either play Tetris for at least 10 uninterrupted minutes or fill out a simple log detailing their activities while waiting in the emergency room. The patients who played Tetris reported having fewer intrusive memories about their accident than the patients who had completed a log.

- (a) In this experiment, what were the treatments?
  - (b) Was this a double-blind study?
  - (c) Can the researchers conclude that playing Tetris causes the reduction of painful memories?
  - (d) Can we generalize these results to a population? Why or why not?
- 1.5** Researchers reported that moderate drinking of alcohol was associated with a lower risk of dementia (Mukamal et al., 2003). Their sample consisted of 373 people with dementia and 373 people without dementia. Participants were asked how much beer, wine, or shot of liquor they consumed. Those who consumed 1–6 drinks a week had a lower incidence of dementia than those who abstained from alcohol.
- (a) Was this study an observational study or an experiment?
  - (b) Can the researchers conclude that drinking alcohol causes a lower risk of dementia? Why or why not?
- 1.6** Researchers surveyed 959 ninth graders who attended 3 large US urban high schools and found that those who listened to music that had references to marijuana were almost twice as likely to have used marijuana as those who did not listen to music with references to marijuana (Primack et al., 2010).
- (a) Was this an observational study or an experiment?
  - (b) Can the researchers conclude that listening to music with references to marijuana causes students to use drugs?
  - (c) Can the researchers extend their results to all urban American adolescents? Why or why not?
- 1.7** Duke University researchers found that diets low in carbohydrates are effective in controlling blood sugar levels (Westman et al., 2008). Eighty-four volunteers with obesity and type 2 diabetes were randomly assigned to either a diet of less than 20 g of carbohydrates/day or a low-glycemic, reduced calorie diet (500 calories/day). 95% of those on the low-carbohydrate diet were able to reduce or eliminate their diabetes medications compared to 62% on the low-glycemic diet.
- (a) Was this study an observational study or an experiment?
  - (b) Can researchers conclude that a low-carbohydrate diet causes an improvement in type 2 diabetes?
  - (c) Can researchers extend their results to a more general population? Why or why not.
- 1.8** In the Google mobile ads case study (Section 1.12),
- (a) Why is this study described as an experiment and not an observational study?

- (b) Can Google claim that their recommendations “caused” the outcome of the study?
- (c) Can Google generalize their results to all advertisers who advertise on Google?
- 1.9** In a population of size  $N$ , the probability of any subset of size  $n$  being chosen is  $1/\binom{N}{n}$ . Show this implies that any one person in the population has a  $n/N$  probability of being chosen in a sample. Then, in particular, every person in the population has the same probability of being chosen.
- 1.10** A typical Gallup poll surveys about  $n = 1000$  adults. Suppose the sampling frame contains 100 million adults (including you). Now, select a random sample of 1000 adults.
- What is the probability that you will be in this sample?
  - Now suppose that 2000 such samples are selected, each independently of the others. What is the probability that you will *not* be in any of the samples?
  - How many samples must be selected for you to have a 0.5 probability of being in at least one sample?
- 1.11** In the mobile ads case study (Section 1.12), the variables `m.cpr` and `d.cpr`, which measure cost/value (how much it costs a company to advertise per how much they make), are recorded as 0 if value is 0. The error is defined by `error = m.cpr - d.cpr`. If `m.cpr=10` and `d.cpr=1`, then `error` is 9. However, if on the mobile, `m.value` is 0, that is, the company did not make any money, then `m.cpr` is defined to be 0. So the error is  $-1$  which is smaller in absolute value than the first case.
- Do you think that this accurately reflects the magnitude of the difference in these two scenarios?
  - If you were a consultant for Google, can you recommend other ways of defining `cpr` when the denominator `value` is 0?



## 2

# Exploratory Data Analysis

*Exploratory data analysis* (EDA) is an approach to examining and describing data to gain insight, discover structure, and detect anomalies and outliers. John Tukey (1915–2000), an American mathematician and statistician who pioneered many of the techniques now used in EDA, stated in his 1977 book *Exploratory Data Analysis* (Tukey, 1977) that “Exploratory data analysis is detective work – numerical detective work, counting detective work, or graphical detective work.” In this chapter, we will learn many of the basic techniques and tools for gaining insight into data.

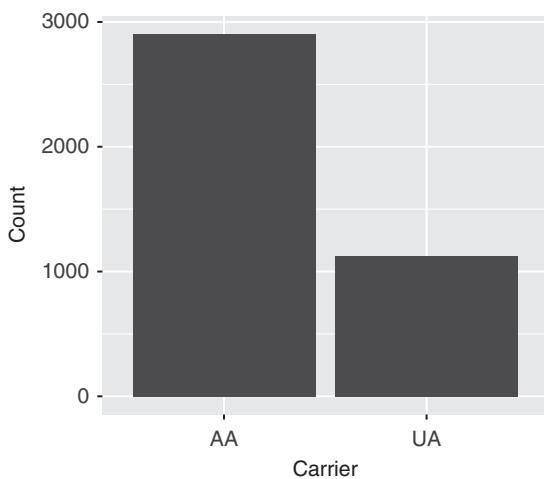
Statistical software packages can easily do the calculations needed for the basic plots and numeric summaries of data. We will use the software package R. We will assume that you have gone through the introduction to R available at the website <https://github.com/lchihara/MathStatsResamplingR>.

## 2.1 Basic Plots

In Chapter 1, we described data on the lengths of flight delays of airplanes of two major airlines flying from LaGuardia Airport in New York City in 2009. Some basic questions we might ask include: How many of these flights were flown by United Airlines (UA) and how many by American Airlines (AA)? How many flights flown by each of these airlines were delayed more than 30 min?

A *categorical* variable is one that places the observations into groups. For instance, in the `FlightDelays` data set, `Carrier` is a categorical variable (we will also call this a *factor* variable) with two *levels*, UA and AA. Other data sets might have categorical variables such as `gender` (with two or more levels) or `size` (with levels small, medium, and large).

A *bar chart* is used to describe the distribution of a categorical (factor) variable. Bars are drawn for each level of the factor variable, and the height of the bar is the number of observations in that level. For the `FlightDelays`



**Figure 2.1** Bar chart of Carrier variable.

data, there were 2906 AA and 1123 UA flights. The corresponding bar chart is shown in Figure 2.1.

We might also be interested in investigating the relationship between a carrier and whether or not a flight was delayed more than 30 min. A *contingency table* summarizes the counts in the different categories (Table 2.1).

From Table 2.2, we can compute the *conditional distribution* of delayed flights: 13.5% of AA flights were delayed more than 30 min, compared to 18.2% of UA flights. Could the difference in percentages be due to *natural variability*,

**Table 2.1** Counts for the Carrier variable.

	Carrier	
	American	United
Number of flights	2906	1123

**Table 2.2** Counts of Delayed Flights grouped by carrier.

Carrier	Delayed more than 30 min?		
	No	Yes	Total
American Airlines	2513	393	2906
(percent)	86.5%	13.5%	100%
United Airlines	919	204	1123
(percent)	81.8%	18.2%	100%

or is there a systematic difference between the two airlines? We will address this question in the following chapters.

With a numeric variable, we will be interested in its distribution: What is the range of values? What values are taken on most often? Where is the *center*? What is the *spread*?

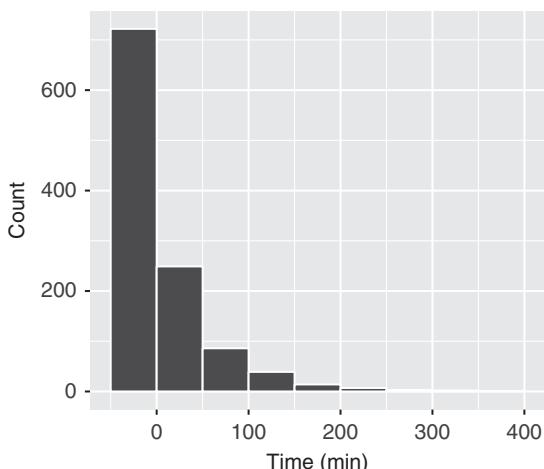
For the flight delay data, although we can inspect the distribution of the lengths of the delays with a table by partitioning the values into nonoverlapping intervals (Table 2.3), a visual representation is often more informative.

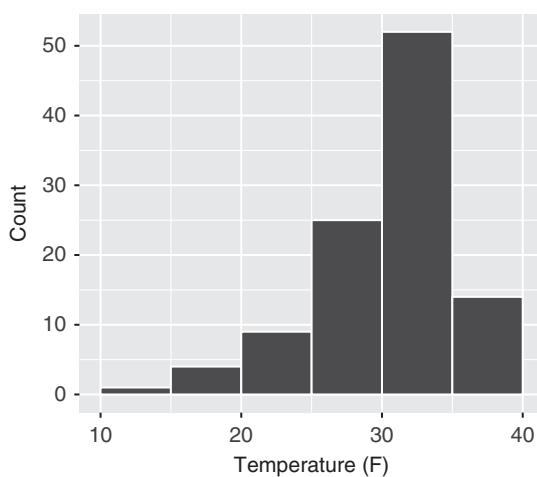
A *histogram* corresponding to Table 2.3 is shown in Figure 2.2. Note that the height of each bar reflects the frequency of flights whose delays fall in

**Table 2.3** Distribution of length of flight delays for United Airlines.

Time interval	Number of flights
(−50, 0]	722
(0, 50]	249
(50, 100]	86
(100, 150]	39
(150, 200]	14
(200, 250]	7
(250, 300]	3
(300, 350]	2
(350, 400]	2
(400, 450]	1

**Figure 2.2** Histogram of lengths of flight delays for United Airlines. The distribution is right-skewed.





**Figure 2.3** Histogram of average January temperatures in Washington state (1895–1999). The distribution is left-skewed.

the corresponding interval. For example, 722 flights departed on time or earlier than scheduled, while 249 flights were delayed by at most 50 min. Some software will give users the option to create bar heights equal to proportions or percentages.

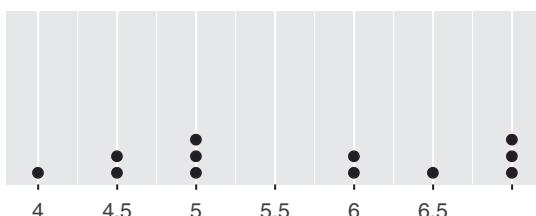
We describe this distribution as *right-skewed* (long right tail). Most of the flights departed on time (or were early), but there are some flights with long delays.

Average January temperatures in the state of Washington follow a *left-skewed distribution* (Figure 2.3): In most years the average temperature fell in the 30–35 °F interval, the warmer years aren't much larger, but the cold years are much lower.

**Remark** The exact choice of subintervals to use is discretionary, and different choices can give very different impressions. Different software packages utilize various algorithms for determining the length of the subintervals; also, some software packages may use subintervals of the form  $[a, b]$  instead of  $(a, b)$ . ||

For small data sets, a *dot plot* is an easy graph to create by hand. A dot represents one observation and is placed above the value it represents. The number of dots in a column represents the frequency of that value.

The dot plot for the data 4, 4.5, 4.5, 5, 5, 5, 6, 6, 6.5, 7, 7, 7 is shown in Figure 2.4.



**Figure 2.4** Example of a dot plot.

## 2.2 Numeric Summaries

It is often useful to have numerical descriptions of variables. Unfortunately, the old adage “a picture is worth a thousand words” cuts both ways – doing without a picture limits what we can say without thousands of words. So we focus on key characteristics – center, spread, and sometimes shape.

### 2.2.1 Center

First consider *center*. By eyeballing the histogram (Figure 2.2) of flight delay times, we might put the center at around 0. Two statistics commonly used to describe the *center* of a variable include the *mean* and *median*.

If  $x_1, x_2, \dots, x_n$  are  $n$  data values, then the mean is

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

The median is the middle value in a sorted arrangement of the values; that is, half the values are less than or equal to the median, and half are greater. If  $y_1 \leq y_2 \leq \dots \leq y_n$  denotes a sorted list of values and  $n$  is odd, the median is the middle value  $y_{(n+1)/2}$ . If  $n$  is even, then the median is the average of the two middle values,  $(1/2)(y_{n/2} + y_{(n/2)+1})$ .

A compromise between the mean and the median is a *trimmed mean*. The mean is the average of all observations, while the median is the average of the middle one or two observations. For a 25% trimmed mean, for example, you sort the data, omit 25% of the observations on each side, and take the mean of the remaining middle 50% of the observations. The 25% trimmed mean is also known as *midmean*.

**Example 2.1** The mean of the 12 values 1, 3, 3, 4, 4, 7, 8, 10, 14, 21, 24, 26 is 10.42, the median is the average of the sixth and seventh values,  $(7 + 8)/2 = 7.5$ , and the midmean is the average of fourth through ninth values, 7.83.

The mean of the 15 values 1, 3, 3, 4, 4, 7, 8, 10, 14, 21, 24, 28, 30, 30, 34 is 14.73, the median is the 8th value, 10, and the midmean is the average of the 4th through 12th values, 13.33.  $\square$

**Example 2.2** The mean departure delay for UA was 15.9831 min. The median length of a departure delay was  $-1.00$  min; that is, half of the flights left more than 1 min earlier than their scheduled departure time.  $\square$

**Remark** Software may differ in how it calculates trimmed means. In R, `mean(x, trim = 0.25)` rounds  $0.25n$  down; thus, for  $n = 15$ , three observations are omitted on each side.  $\parallel$

## 2.2.2 Spread

To describe spread, three common choices are the range, the interquartile range (IQR), and the standard deviation.

The *range* is the difference between the largest and smallest values.

The *IQR* is the difference between the third and the first quartiles. It gives a better measure of the spread of the middle of the data than the range does, and is not sensitive to outliers.

The *sample standard deviation*, or *standard deviation*, is

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}, \quad (2.1)$$

where  $x_1, x_2, \dots, x_n$  represent values from a sample. To motivate the standard deviation, we begin with a less common measure of spread, the *mean absolute deviation* (MAD),  $(1/n) \sum_{i=1}^n |x_i - \bar{x}|$ . This is the average distance from the mean and is a natural measure of spread. In contrast, the standard deviation is roughly the average squared distance from the mean, followed by a square root; the combination is roughly equal to the MAD, though usually a bit larger. The standard deviation tends to have better statistical properties.

There are a couple of versions of standard deviation. The *population standard deviation* is the square root of the *population variance*, which is the average of the squared distances from the population mean  $\mu$ ,  $(1/n) \sum_{i=1}^n (x_i - \mu)^2$ . For comparison, the *sample variance*  $s^2$  uses a divisor of  $n-1$ .

The population versions are appropriate when the data are the whole population. When the data are a sample from a larger population, we use the sample versions; in this case, the population versions tend to be too small – they are *biased*; we return to this point in Section 6.3.1. For  $n$  large, there is little practical difference between using  $n-1$  or  $n$ .

**Example 2.3** The standard deviation of the departure delay times for UA flights is 45.119 min. Since the observations represent a population (we compiled *all* UA flights for the months of May and June), we use the definition with the  $1/n$  term rather than the  $1/(n-1)$  term. Using Equation (2.1) gives 45.139.  $\square$

## 2.2.3 Shape

To describe the shape of a data set, we may use skewness and kurtosis (see Section 2.7). However, more common and intuitive is to use the *five-number summary*: the minimum, first quartile, median, third quartile, and maximum value.

**Example 2.4** Consider the 15 numbers 9, 10, 11, 11, 12, 14, 16, 17, 19, 21, 25, 31, 32, 41, 61.

The median is 17. Now, find the median of the numbers less than or equal to 17. This will be the first quartile  $Q_1 = 11.5$ . The median of the numbers greater than or equal to 17 is the third quartile  $Q_3 = 28$ . Thus, the five-number summary is 9, 11.5, 17, 28, 61.  $\square$

**Remark** Different software packages use different algorithms for computing quartiles, so do not be alarmed if your results do not match exactly.  $\parallel$

### R Note

Use the `summary` function on a data set to obtain numeric summaries, including counts for factor variables. For example, the `Recidivism` data set from the Iowa recidivism case study gives:

```
> summary(Recidivism)
   Gender          Age        Age25      ...
F : 2101  25-34     :6227  Over 25 :13942  ...
M :14918  35-44     :4035  Under 25: 3077  ...
NA's:    3  45-54     :2872  NA's       :     3
                  55 and Older: 808
                  Under 25   :3077
                  NA's       :     3
...
...
```

In particular, we see that several of the variables have missing values (NA), something we will need to remember for future analyses.

## 2.3 Boxplots

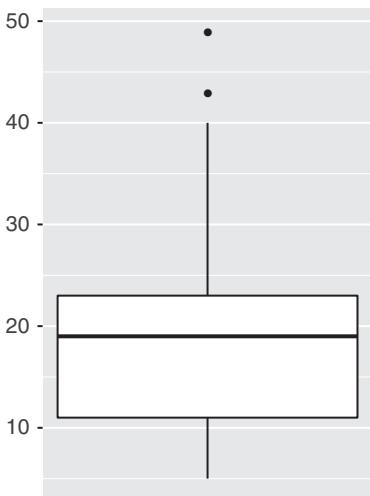
A boxplot is a type of graph that can be used to visualize the five-number summary of a set of numeric values.

**Example 2.5** Consider the following set of 21 values (Table 2.4).

The five-number summary for these data is 5, 11, 19, 23, 48, and the IQR is  $23 - 11 = 12$ . The corresponding boxplot is shown in Figure 2.5.  $\square$

**Table 2.4** A set of 21 data values.

5	6	6	8	9	11	11
14	17	17	19	20	21	21
22	23	24	32	40	43	49

**Figure 2.5** Boxplot for Table 2.4.

To create a boxplot:

- Draw a box with the bottom placed at the first quartile and the top placed at the third quartile. Draw a line through the box at the median.
- Compute the number  $Q_3 + 1.5 \times \text{IQR}$ , called the *upper fence*, and then place a cap at the largest observation that is less than or equal to this number.
- Similarly, compute the *lower fence*,  $Q_1 - 1.5 \times \text{IQR}$ , and place a cap at the smallest observation that is greater than or equal to this number.
- Extend *whiskers* from the edge of the box to the caps.
- The observations that fall outside the caps are considered *outliers*, and separate points are drawn to indicate these values.

In the above example, the upper fence is  $23 + 1.5 \times 12 = 41$ . The largest observation that falls below this fence is 40, so a cap is drawn at 40. The lower fence is  $11 - 1.5 \times 12 = -7$ . The smallest observation that falls above this fence is 5, so a cap is drawn at 5. The outliers are 43 and 49.

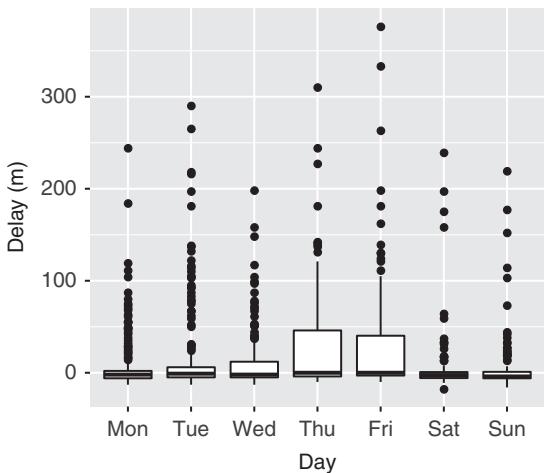
**Example 2.6** For the length of UA flight delays, the five-number summary is  $-17.00, -5.00, -1.00, 12.50, 377.00$ . Thus, the IQR is  $12.50 - (-5.00) = 17.50$ , and half of the 1123 values are contained in an interval of length 17.50.  $\square$

Boxplots are especially useful in comparing the distribution of a numeric variable across levels of a factor variable.

**Example 2.7** We can compare the lengths of the flight delays for UA across the days of the week for which the departure was scheduled.

For instance, we can see that the most variability in delays seems to occur on Thursdays and Fridays (Figure 2.6).

**Figure 2.6** Distribution of lengths of the flight delays for United Airlines across the days of the week.



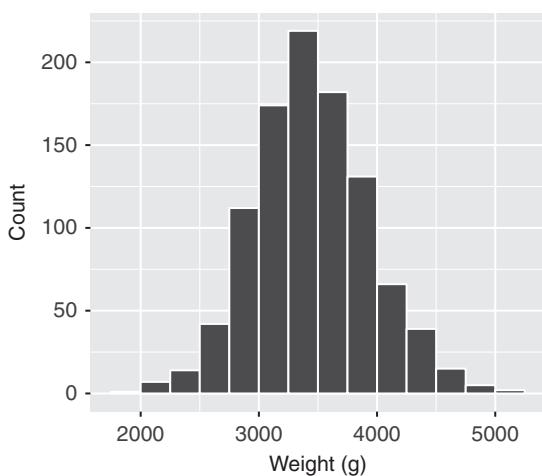
□

## 2.4 Quantiles and Normal Quantile Plots

In the UA flight data shown in Figure 2.6, it is clear that the distributions are not normal. In contrast, for the random sample of 1009 babies born in North Carolina (NC) in 2004, the distribution of their weights is unimodal and roughly symmetric (Figure 2.7), so could well be normal. But histograms are notorious for changing dramatically when bar widths are changed, and for hiding features of data. In this section, we introduce another type of graph that is more effective for comparing a data distribution with a normal distribution, a *normal quantile plot*. But first – what are *quantiles*?

**Definition 2.1** Let  $X$  denote a random variable. For  $0 \leq p \leq 1$ , the  $p$ th *quantile* of  $X$  is the value  $q_p$  such that  $P(X \leq q_p) = p$ . That is,  $q_p$  is the value at which the amount of area under the density curve (of  $X$ ) to the left of  $q_p$  is  $p$ , or  $p \times 100\%$  of the area under the curve is to the left of  $q_p$ .

A related term is *percentile*. For instance, the 0.3 quantile is the same as the 30th percentile. ||



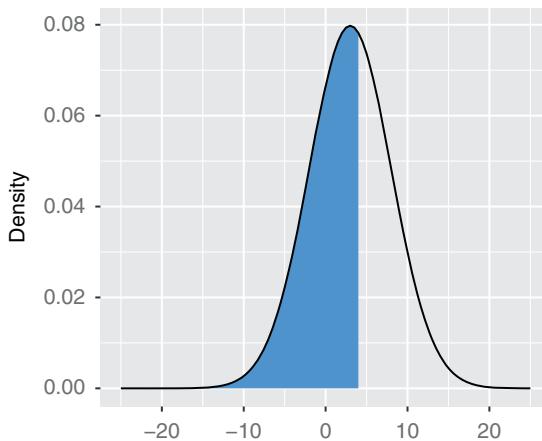
**Figure 2.7** Distribution of birth weights for North Carolina babies.

**Example 2.8** Let  $Z$  denote the standard normal distribution. Let  $p = 0.5$ . Then, the 0.5 quantile of  $Z$  is 0 since  $P(Z \leq 0) = 0.5$ . That is, 0 is the 50th percentile of the standard normal distribution.

Let  $p = 0.25$ . Then,  $q_{0.25} = -0.6744$  since  $P(Z \leq -0.6744) = 0.25$ . That is,  $-0.6744$  is the 25th percentile of the standard normal distribution.  $\square$

**Example 2.9** Let  $X$  be a normal random variable,  $N(3, 5^2)$ . Find the  $p = 0.6$  quantile.

We want  $q_p$  such that  $P(X \leq q_p) = 0.6$ . The desired value is  $q_p = 4.3$  (see Figure 2.8).



**Figure 2.8** Density for  $N(3, 5^2)$  with  $P(X \leq 4.27) = 0.6$ .

$\square$

**R Note**

Use the `qnorm` function to find normal quantiles:

```
> qnorm(.25)           # standard normal
[1] -0.6744898
> qnorm(.6, 3, 5)     # N(3, 5^2)
[1] 4.266736
```

We can also formulate quantiles in terms of the cumulative distribution function (cdf)  $F$  of a random variable  $X$  since

$$F(q_p) = P(X \leq q_p) = p \text{ implies } q_p = F^{-1}(p).$$

**Example 2.10** Let  $X$  be an exponential random variable with  $\lambda = 3$ . The cdf of  $X$  is given by  $F(x) = 1 - e^{-3x}$ . Since  $F^{-1}(y) = (-1/3) \ln(1 - y)$ , the  $p$ th quantile is given by  $q_p = (-1/3) \ln(1 - p)$ .

Alternatively, since we know the pdf of  $X$  is  $f(x) = 3e^{-3x}$ ,  $x \geq 0$ , we could also solve for  $q_p$  in

$$p = P(X \leq q_p) = \int_0^{q_p} 3e^{-3t} dt.$$

□

Now we're ready to use those quantiles to create the *normal quantile plot* – a scatter plot of the  $x$ 's against normal quantiles,  $(q_k, x_k)$  for  $k = 1, \dots, n$ , where  $q_k$  is the  $k/(n+1)$  quantile of the standard normal distribution and  $x_1 \leq x_2 \leq \dots \leq x_n$  are the sorted data. If these points fall (roughly) on a straight line, then we conclude that the data follow an approximate normal distribution.

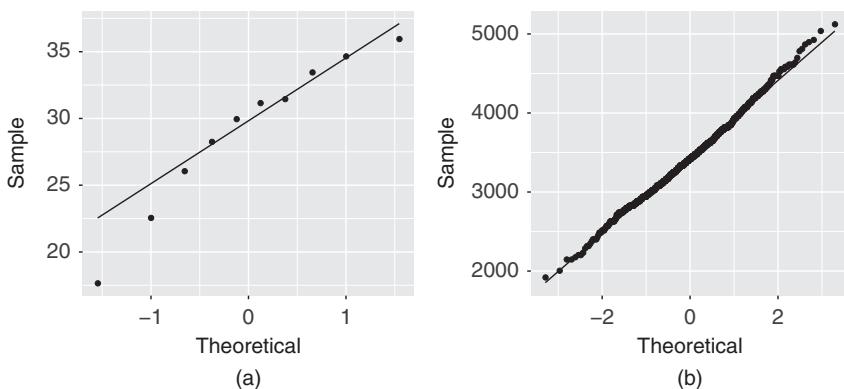
**Example 2.11** Here are  $n = 10$  points. We will look at the  $i/(n+1) = i/(11)$ th quantiles,  $i = 1, \dots, 10$ , of the standard normal.

$x$	17.7	22.6	26.1	28.3	30.0	31.2	31.5	33.5	34.7	36.0
$p_i$	1/11	2/11	3/11	4/11	5/11	6/11	7/11	8/11	9/11	10/11
$q_p$	-1.34	-0.91	-0.60	-0.35	-0.11	0.11	0.35	0.60	0.91	1.34

For instance, the  $q_p$  entry corresponding to  $p_5 = 5/11 = 0.455$  (the 45.5th percentile) is

$$q_{0.455} = -0.11 \text{ because } P(Z \leq -0.11) = 0.455.$$

To create a normal quantile plot, we graph the pairs  $(q_p, x)$ . A straight line is often drawn through the points corresponding to the first and third quartiles of each variable (see Figure 2.9).



**Figure 2.9** (a) Example of normal quantile plot for data in Example 2.11. (b) Normal quantile plot for weights of NC babies.

### R Note

Create normal quantile plots using `qqnorm` and `qqline` in base R, or `geom_qq` and `geom_qq_line` in `ggplot2`:

```
x <- c(17.7, 22.6, 26.1, 28.3, 30, 31.2, 31.5, 33.5, 34.7, 36)
df <- data.frame(x = x)
ggplot(df, aes(sample = x)) + geom_qq() + geom_qq_line() +
  labs(x = "Theoretical quantiles", y = "Sample quantiles")

ggplot(NCBirths2004, aes(sample = Weight)) + geom_qq() +
  geom_qq_line()
```

To obtain the normal quantile plots for the weights of babies born to mothers who smoke and mothers who do not smoke,

```
ggplot(NCBirths2004, aes(sample = Weight)) + geom_qq() +
  geom_qq_line() + facet_grid(~ Smoker)
```

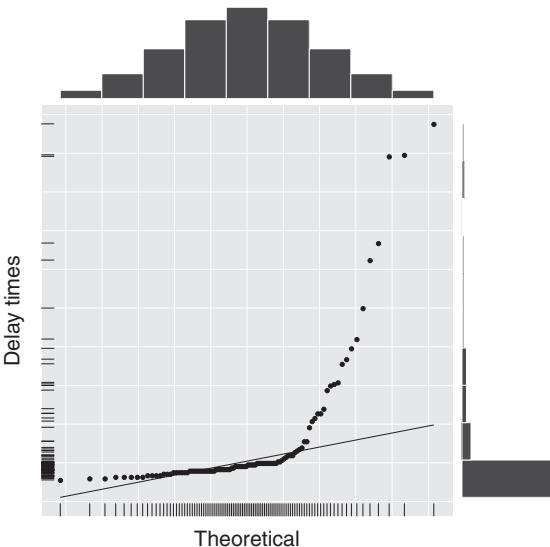
□

Recall that the distribution of the flight delay times for UA is strongly right-skewed (Figure 2.2). Normal quantile plots for these data and for the left-skewed distribution of average January temperatures in Washington state (Figure 2.3) are shown in Figures 2.10 and 2.11.

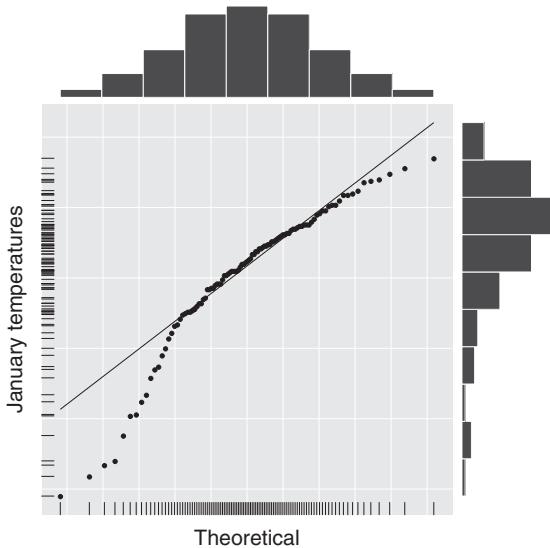
### Remark

- Even for samples drawn from a normal distribution, the points on a normal quantile plot do not lie *exactly* on a straight line. See Exercise 2.17.

**Figure 2.10** Normal quantile plot for a random sample of the flight delay times for United Airlines.



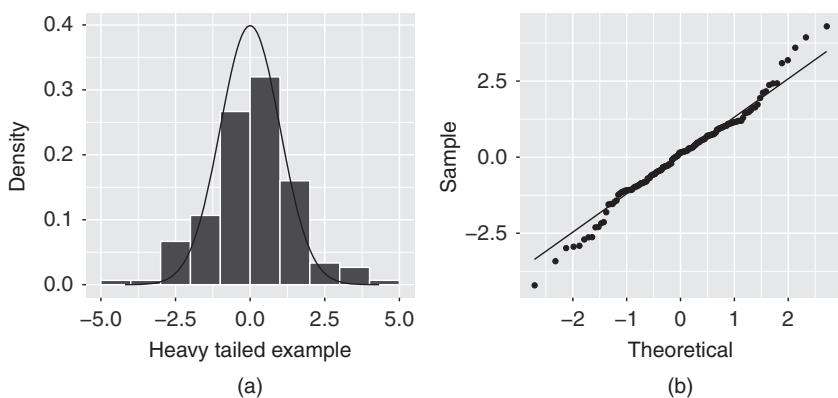
**Figure 2.11** Normal quantile plot for average January temperatures in Washington state.



- It is often difficult to tell from a histogram whether a sample has heavier (longer) or lighter (shorter) tails than a normal distribution. See Figure 2.12.

||

The normal quantile plot is one example of a *quantile–quantile plot*, or *qq plot* for short, in which quantiles of a data set are plotted against quantiles of a



**Figure 2.12** (a) Example of sample with symmetric, bell-shaped distribution with normal density superimposed. (b) Normal quantile plot indicating heavier (longer) tails than a normal distribution.

distribution or of another data set. Another example is a uniform quantile plot, plotting the data against uniform quantiles. That is basically our next plot, an empirical cumulative distribution function (ecdf) plot, albeit with axes switched.

## 2.5 Empirical Cumulative Distribution Functions

The *ecdf* is an estimate of the underlying cumulative distribution function (Appendix A) for a sample. The ecdf, denoted by  $\hat{F}$ , is a step function

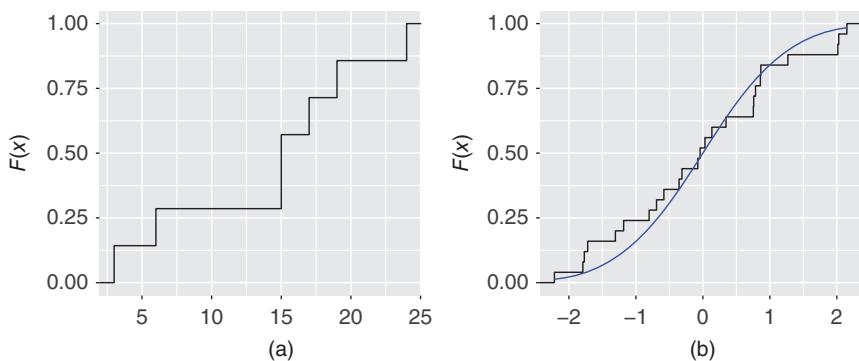
$$\hat{F}(x) = \frac{1}{n} (\text{number of values } \leq x),$$

where  $n$  is the sample size.

For instance, consider the set of values 3, 6, 15, 15, 17, 19, 24. Then,  $\hat{F}(18) = 5/7$  since there are five data values less than or equal to 18. More generally,

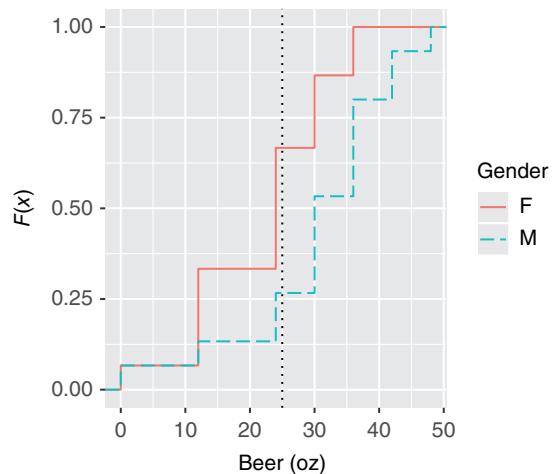
$$\hat{F}(x) = \begin{cases} 0, & x < 3, \\ 1/7, & 3 \leq x < 6, \\ 2/7, & 6 \leq x < 15, \\ 4/7, & 15 \leq x < 17, \\ 5/7, & 17 \leq x < 19, \\ 6/7, & 19 \leq x < 24, \\ 1, & x \geq 24. \end{cases}$$

Figure 2.13 displays the ecdf for this example as well as the ecdf for a random sample of size 25 from the standard normal distribution. The graph of the cdf for the standard normal  $\Phi(t)$  is added for comparison.



**Figure 2.13** (a) Empirical cumulative distribution function for the data 3, 6, 15, 15, 17, 19, 24. (b) Ecdf for a random sample from  $N(0, 1)$  with the cdf for the standard normal.

**Figure 2.14** Ecd's for male and female beer consumption. The vertical line is at 25 oz.



The ecdf is useful for comparing two distributions. Figure 2.14 shows the ecdfs of beer consumption for males and females from the beer and hot wings case study in Section 1.9. With the vertical line at 25 oz, we can see that about 30% of the males and nearly 70% of the females have consumed 25 or fewer ounces of beer.

#### R Note

```
x <- c(3, 6, 15, 15, 17, 19, 24)
df <- data.frame(x = x)
ggplot(df, aes(x = x)) + stat_ecdf()
```

```
df <- data.frame(x = rnorm(25))      # random sample from N(0,1)
ggplot(df, aes(x = x)) + stat_ecdf() +
  stat_function(fun = pnorm, color = "blue") +
  labs(x = "", y = "F(x)")
```

For the beer and hot wings case study, we first create vectors that hold the data for the men and women separately.

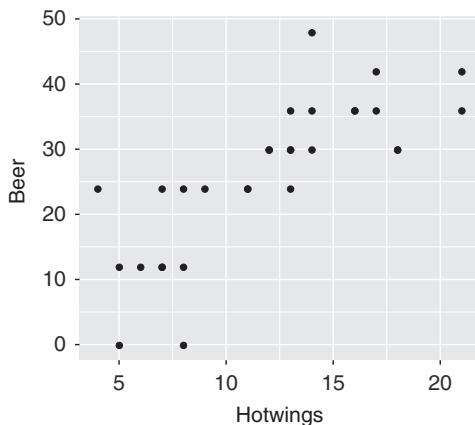
```
ggplot(Beerwings, aes(x = Beer, linetype = Gender, color = Gender)) +
  stat_ecdf() + labs(x = "Beer (oz)", y = "F(x)") +
  geom_vline(xintercept = 25, lty = 3) +
  scale_linetype_manual(values = c(1, 5))
```

## 2.6 Scatter Plots

In the beer and hot wings case study in Section 1.9, one question that the student asked was whether there was a relationship between the number of hot wings eaten and the amount of beer consumed. A way to visualize the relationship between two numeric variables is with a *scatter plot*; see Figure 2.15.

Each point in the scatter plot represents a single observation, that is, a single person who took part in the study. From the graph, we note that there is a positive, roughly linear, association between hot wings and beer: As the number of hot wings eaten increases, the amount of beer consumed also increases.

**Remark** In statistics, the convention is to put the variable of primary interest on the  $y$ -axis, and the variable that may help predict or explain that variable as  $x$ , and to “plot  $y$  against  $x$ .”



**Figure 2.15** A scatter plot of Beer against Hotwings.

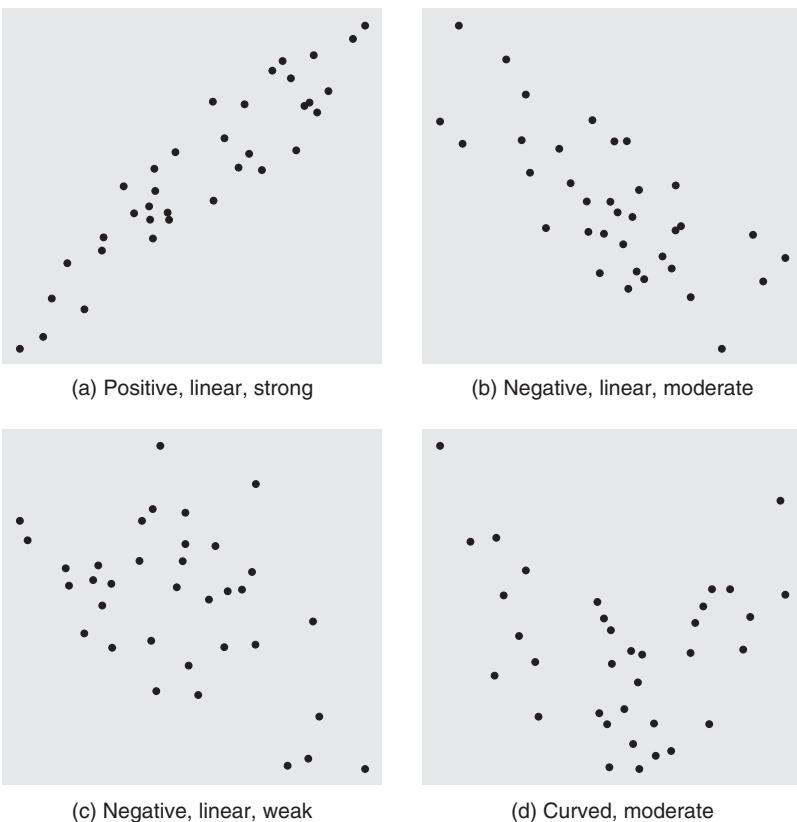


Figure 2.16 Examples of scatter plots.

Further examples are shown in Figure 2.16. In general, when describing the relationship between two numeric variables, we will look for *direction*, *form*, and *strength*. In Chapter 9, we will investigate the relationship between two numeric variables in more detail.

#### R Note

```
ggplot(Beerwings, aes(x = Hotwings, y = Beer)) + geom_point()
```

We can also distinguish the two genders by adding the `color` aesthetic:

```
ggplot(Beerwings, aes(x = Hotwings, y = Beer, color = Gender)) +  
  geom_point()
```

## 2.7 Skewness and Kurtosis

Asymmetry and peakedness are often measured using *skewness* and *kurtosis*, which are defined using third and fourth central moments (Section A.8).

**Definition 2.2** Let  $X$  be a random variable with mean  $\mu$  and standard deviation  $\sigma$ . The *skewness* of  $X$  is

$$\gamma_1 = E \left[ \left( \frac{X - \mu}{\sigma} \right)^3 \right] = \frac{\mu_3}{\sigma^3} \quad (2.2)$$

and the *kurtosis* of  $X$  is

$$\gamma_2 = E \left[ \left( \frac{X - \mu}{\sigma} \right)^4 \right] - 3 = \frac{\mu_4}{\sigma^4} - 3. \quad (2.3)$$

||

A variable with positive skewness typically has a longer or heavier tail on the right than on the left; the opposite holds for negative skewness. A variable with positive kurtosis typically has a higher central peak and a longer or heavier tail on at least one side than a normal distribution, while a variable with negative kurtosis is flatter in the middle and has shorter tails. Figure 2.17 shows some examples.

**Example 2.12** Let  $Z$  be the standard normal variable with  $\mu = 0$  and  $\sigma = 1$ . Then the skewness of  $Z$  is

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^3 e^{-z^2/2} dz = 0$$

and the kurtosis is

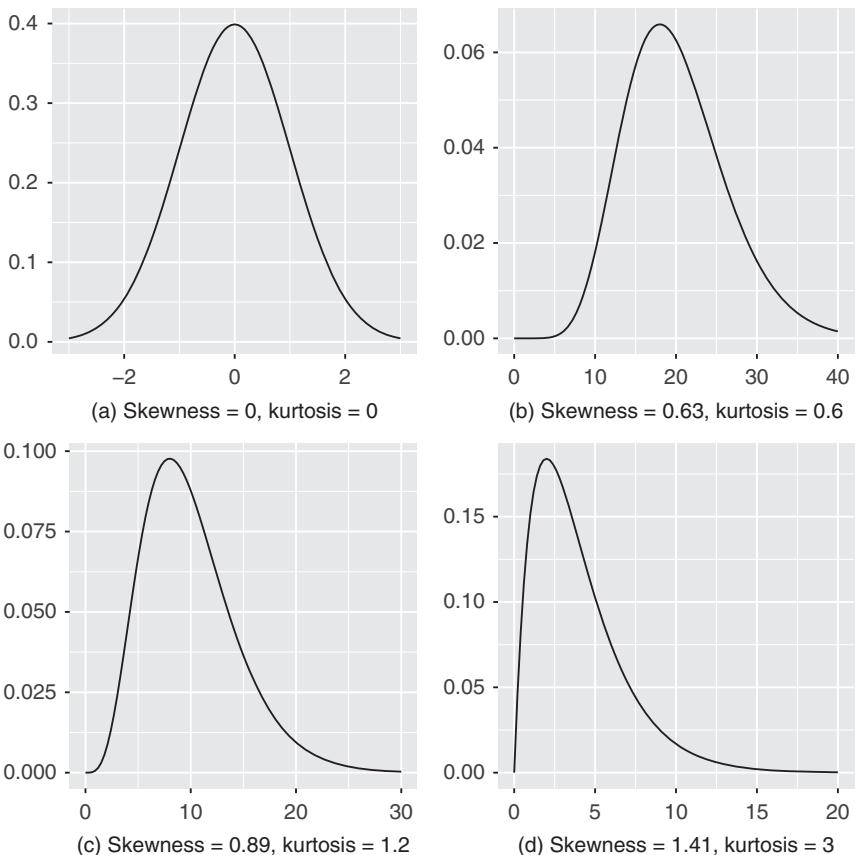
$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^4 e^{-z^2/2} dz - 3 = 0. \quad \square$$

**Example 2.13** Let  $X$  be an exponential random variable with parameter  $\lambda = 1$ . Then  $\mu = 1 = \sigma$ , the skewness of  $X$  is

$$\int_0^{\infty} (x - 1)^3 e^{-x} dx = 2$$

and the kurtosis is

$$\int_0^{\infty} (x - 1)^4 e^{-x} dx - 3 = 6. \quad \square$$



**Figure 2.17** Examples of skewness and kurtosis for four distributions, including the standard normal (a).

**Example 2.14** Let  $X$  be the standard uniform random variable,  $f(x) = 1$  for  $0 < x < 1$ . Then  $\mu = 0.5$ ,  $\sigma^2 = 1/12$ , the skewness is zero, and the kurtosis is

$$\frac{\int_0^1 (x - 0.5)^4 dx}{\left(\int_0^1 (x - 0.5)^2 dx\right)^2} - 3 = -1.2.$$

□

## Exercises

- 2.1** Compute the mean  $\bar{x}$  and median  $m$  of the seven numbers 3, 5, 8, 15, 20, 21, 24. Apply the logarithm to the data, and then compute the mean  $\bar{x}'$  and median  $m'$  of the transformed data. Is  $\ln(\bar{x}) = \bar{x}'$ ? Is  $\ln(m) = m'$ ?

- 2.2** Compute the mean  $\bar{x}$  and median  $m$  of the eight numbers 1, 2, 4, 5, 6, 8, 11, 15. Let  $f(x) = \sqrt{x}$ . Apply this transformation to the data, and then compute the mean  $\tilde{x}$  and the median  $\tilde{m}$  of the transformed data. Is  $f(\bar{x}) = \tilde{x}$ ? Is  $f(m) = \tilde{m}$ ?
- 2.3** Let  $\bar{x}$  and  $m$  denote the mean and median, respectively, of  $x_1 < x_2 < \dots < x_n$ . Let  $f$  be a real-valued function.
- Is  $f(\bar{x})$  the mean of  $f(x_1), f(x_2), \dots, f(x_n)$ ?
  - Is  $f(m)$  the median of  $f(x_1), f(x_2), \dots, f(x_n)$ ?
  - Are there any conditions that would ensure that  $f(\bar{x})$  is the mean of the transformed data?
  - Are there any conditions that would ensure that  $f(m)$  is the median of the transformed data?
- 2.4** Import the flight delays case study data in Section 1.1 into R.
- Create a table and a bar chart of the departure times (`DepartTime`).
  - Create a contingency table of the variables `Day` and `Delayed30`. For each day, what is the proportion of flights delayed at least 30 min?
  - Create side-by-side boxplots of the lengths of the flights, grouped by whether or not the flight was delayed at least 30 min.
  - Do you think that there is a relationship between the length of a flight and whether or not the departure is delayed by at least 30 min?
- 2.5** Import the data from the General Social Survey case study (Section 1.7) into R.
- Create a table and a bar chart of the responses to the question about the death penalty.
  - Use the `table` function and add the argument `exclude = NULL` in R on the `Courts` variable (“Do you think the courts deal too harshly with criminals?”) What additional information does this provide?
  - Create a contingency table displaying the relationship between opinions about the courts to that about the death penalty.
  - What proportion of those who think the courts are not harsh enough with criminals favor the death penalty? Does it appear to be different from the proportion among those who think the courts are too harsh?
- 2.6** Import the data from the recidivism case study in Section 1.4 into R.
- Create a table and bar chart of the `Recid` variable.
  - Create a contingency table summarizing the relationship between recidivism (`Recid`) by age (`Age25`). Of those under 25 years of age,

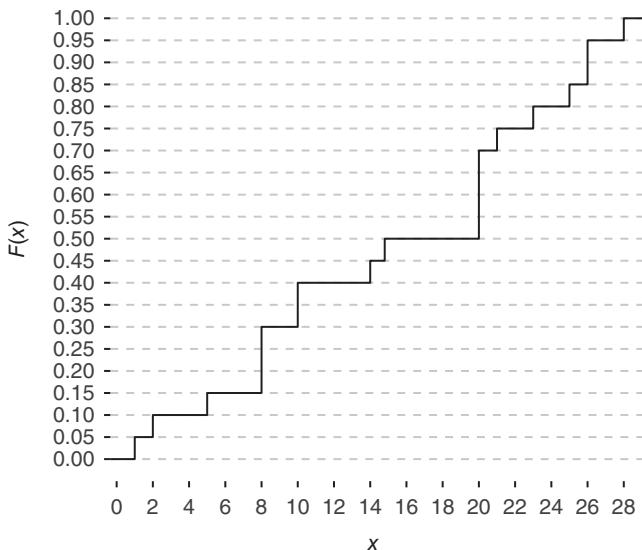
- what proportion were sent back to prison? Of those over 25 years of age, what was this proportion?
- (c) Create side-by-side boxplots of the number of days to recidivism grouped by type of violation (Type), and give three comparative statements about the distributions.
  - (d) Use the `quantile` function to obtain the quartiles of the number of days to recidivism. Since there are missing values (NA) for those released offenders who had not recidivated, you will need to add the argument `na.rm = TRUE` to the `quantile` command to exclude those observations.
  - (e) Create ecdfs of days to recidivism for those under 25 years of age and those 25 years of age or older. Approximately what proportion in each age group were sent back to prison 400 days after release?  
*Note:* `stat_ecdf` automatically removes the missing values (that is, the NA's, so your sentence should reflect that: "of those who relapsed...").
- 2.7** Import data from the black spruce case study in Section 1.10 into R.
- (a) Compute the numeric summaries for the height changes (`Ht.change`) of the seedlings.
  - (b) Create a histogram and normal quantile plot for the height changes of the seedlings. Is the distribution approximately normal?
  - (c) Create a boxplot to compare the distribution of the change in diameters of the seedlings (`Di.change`) grouped by whether or not they were in fertilized plots.
  - (d) Use the `tapply` function to find the numeric summaries of the diameter changes for the two levels of fertilization.
  - (e) Create a scatter plot of the height changes against the diameter changes, and describe the relationship.
- 2.8** Import the data `MobileAds` for the mobile ads case study, Section 1.12. We will investigate the variable `m.cpc`, the cost per click for the mobile platform.
- (a) Create histograms of the variables `m.cpc_pre` and `m.cpc_post`, and describe their distributions.
  - (b) Compute the difference between these two variables, create a histogram, and describe this distribution.
  - (c) Create a normal quantile plot of the difference. Does it appear to be normally distributed?
- 2.9** Let  $x_1 < x_2 < \dots < x_n$  and  $y_1 < y_2 < \dots < y_n$  be two sets of data with means  $\bar{x}, \bar{y}$  and medians  $m_x, m_y$ , respectively. Let  $w_i = x_i + y_i$  for  $i = 1, 2, \dots, n$ .

- (a) Prove or give a counterexample:  $\bar{x} + \bar{y}$  is the mean of  $w_1, w_2, \dots, w_n$ .  
 (b) Prove or give a counterexample:  $m_x + m_y$  is the median of  $w_1, w_2, \dots, w_n$ .
- 2.10** Find the median  $m$  and first and third quartiles for the random variable  $X$  having  
 (a) the exponential distribution with pdf  $f(x) = \lambda e^{-\lambda x}$ .  
 (b) the Pareto distribution with parameter  $\alpha > 0$  with pdf  $f(x) = \alpha/x^{\alpha+1}$  for  $x \geq 1$ .
- 2.11** Let the random variable  $X$  have a Cauchy distribution with pdf  $f(x) = 1/(\pi(1 + (x - \theta)^2))$  for  $-\infty < x < \infty$ .  
 (a) Show that the mean of  $X$  does not exist.  
 (b) More generally, will  $E[X^k]$  exist? ( $k = 1, 2, 3, \dots$ )  
 (c) Show that  $\theta$  is the median of the distribution.
- 2.12** Find  
 (a) the 30th and 60th percentiles for  $N(10, 17^2)$ .  
 (b) the 0.10 and 0.90 quantile for  $N(25, 32^2)$ .  
 (c) the point that marks off the upper 25% in  $N(25, 32^2)$ .
- 2.13** The cdf of the exponential distribution is  $F(t) = 1 - e^{-\lambda t}$ .  
 (a) Find an expression for the 0.05 quantile  $q_{0.05}$ .  
 (b) Let  $\lambda = 4$ , and use your answer from (a) to find  $q_{0.05}$ , then check your answer in R using the `qexp` function.
- 2.14** Let  $X$  be a random variable with cdf  $F(x) = x^2/a^2$  for  $0 \leq x \leq a$ . Find an expression for the  $\alpha/2$  and  $(1 - \alpha/2)$  quantiles, where  $0 < \alpha < 1$ .
- 2.15** Let  $X$  be a random variable with cdf  $F(x) = 1 - 9/x^2$  for  $x \geq 3$ . Find an expression for the  $p$ th quantile of  $X$ .
- 2.16** Let  $X \sim \text{Binom}(20, 0.3)$  and let  $F$  denote its cdf. Does there exist a  $q$  such that  $F(q) = 0.05$ ?
- 2.17** In this exercise, we investigate normal quantile plots using R.  
 (a) Draw a random sample of size  $n = 15$  from  $N(0, 1)$ , and plot both a normal quantile plot and a histogram. Do the points on the quantile plot appear to fall on a straight line? Is the histogram symmetric, unimodal, and bell shaped? Do this several times.  
 (b) Repeat part (a) for samples of size  $n = 30, n = 60$ , and  $n = 100$ .  
 (c) What lesson do you draw about using graphs to assess whether or not a data set follows a normal distribution?

**2.18** Plot by hand the empirical cumulative distribution function for the set of values 4, 7, 8, 9, 9, 13, 18, 18, 18, 21.

**2.19** The ecdf for a data set with  $n = 20$  values is given in Figure 2.18.

- How many values are less than or equal to 7?
- How many times does the value 8 occur?
- In a histogram of these values, how many values fall in the bin  $(20, 25]$ ?



**Figure 2.18** Empirical cdf for a data set,  $n = 20$ .

**2.20** The data set ChiMarathonMen has a sample of times for men between 20 and 39 years of age who completed the Chicago Marathon in 2015. Graph the ecdf's of the times for men in the 25–29-age division and men in the 35–39-age division. Approximately what proportion of men in these two divisions finished in 160 min or less?



# 3

## Introduction to Hypothesis Testing: Permutation Tests

### 3.1 Introduction to Hypothesis Testing

Suppose scientists invent a new drug that supposedly will inhibit a mouse's ability to run through a maze. The scientists design an experiment in which three mice are randomly chosen to receive the drug and another three mice serve as controls by ingesting a placebo. The time each mouse takes to go through a maze is measured in seconds. Suppose the results of the experiment are as follows:

Drug			Control		
30	25	20	18	21	22

The average time for the drug group is 25 s and the average time for the control group is 20.33 s. The mean difference in times is  $25 - 20.33 = 4.67$  s.

The average time for the mice, given the drug is greater than the average time for the control group, but this could be due to random variability rather than a real drug effect. We cannot tell for sure whether there is a real effect. What we do instead is to estimate how easily pure random chance would produce a difference this large. If that probability is small, then we conclude there is something other than pure random chance at work and conclude that there is a real effect.

If the drug really does not influence times, then the split of the six observations into two groups was essentially random. The outcomes could just as easily be distributed

Drug			Control		
30	25	18	20	21	22

In this case, the mean difference is  $((30 + 25 + 18)/3) - ((20 + 21 + 22)/3) = 3.33$ .

There are  $\binom{6}{3} = 20$  ways to distribute 6 numbers into two sets of size 3, ignoring any ordering with each set. Of the 20 possible differences in means, 3 are as large or larger than the observed 4.67, so the probability that pure chance would give a difference this large is  $3/20 = 0.15$ .

Fifteen percent is small, but not small enough to be remarkable. It is plausible that chance alone is the reason the mice in the drug group ran slower (had larger times) through the maze.

For comparison, suppose a friend claims that he can control the flip of a coin, producing a head at will. You are skeptical; you give him a coin, and he indeed flips a head, three times. Are you convinced? I hope not; that could easily occur by chance, with a 12.5% probability.

This is the core idea of classical *hypothesis testing* – to calculate how often pure random chance would give an effect as large as that observed in the data, in the absence of any real effect. If that probability is small enough, we conclude that the data provide convincing evidence of a real effect.

If the probability is not small, we do not make that conclusion. This is not the same as concluding that there is no effect; it is only that the data available do not provide convincing evidence that there is an effect. In practice, there may be just too little data to provide convincing evidence. If the drug effect is small, it may be possible to distinguish the effect from random noise with 60 mice, but not 6. More flips might make your friend's claim convincing, though it would be prudent to check for a two-headed coin. (One of us had one, and had a former magician professor who could flip whichever side he wanted.<sup>1</sup>)

## 3.2 Hypotheses

We formalize the core idea using the language of statistical *hypothesis testing*, also known as *significance testing*.

**Definition 3.1** The *null hypothesis*, denoted  $H_0$ , is a statement that corresponds to no real effect. This is the status quo, in the absence of the data providing convincing evidence to the contrary.

The *alternative hypothesis*, denoted by  $H_A$ , is a statement that there is a real effect. The data may provide convincing evidence that this hypothesis is true.

A hypothesis normally involves a statement about a population parameter or parameters, commonly referred to as  $\theta$ ; the null hypothesis is  $H_0: \theta = \theta_0$  for some  $\theta_0$ . A *one-sided alternative hypothesis* is of the form  $H_A: \theta > \theta_0$  or  $H_A: \theta < \theta_0$ ; a *two-sided alternative hypothesis* is  $H_A: \theta \neq \theta_0$ . ||

---

<sup>1</sup> <http://news-service.stanford.edu/news/2004/june9/diaconis-69.html>.

There are some cases where hypothesis tests are not statements about larger populations. For example, if we randomly assign a treatment or a control to some people with a disease, we can test the hypothesis that the treatment has no effect, even if these subjects are the only people in the world with a disease.

**Example 3.1** Consider the mice example in Section 3.1. Let  $\mu_d$  denote the true mean time that a randomly selected mouse that received the drug takes to run through the maze; let  $\mu_c$  denote the true mean time for a control mouse. Then  $H_0: \mu_d = \mu_c$ . That is, on average, there is no difference in the mean times between mice who receive the drug and mice in the control group.

The alternative hypothesis is  $H_A: \mu_d > \mu_c$ . That is, on average, mice who receive the drug have slower times (larger values) than the mice in the control group.

The hypotheses may be rewritten as follows:  $H_0: \mu_d - \mu_c = 0$  and  $H_A: \mu_d - \mu_c > 0$ ; thus  $\theta = \mu_d - \mu_c$  (any function of parameters is itself a parameter).  $\square$

The next two ingredients in hypothesis testing are a numerical measure of the effect and the probability that chance alone could produce that measured effect.

**Definition 3.2** A *test statistic* is a numerical function of the data and parameter whose value determines the result of the test. The function itself is generally denoted  $T = T(X, \theta)$  where  $X$  represents the data, e.g.  $T = T(X_1, X_2, \dots, X_n, \theta)$  in a one-sample problem, or  $T = T(X_1, X_2, \dots, X_m, Y_1, \dots, Y_n, \theta_1, \theta_2)$  in a two-sample problem. After being evaluated for the sample data  $x$ , the result is called an *observed test statistic* and is written in lower-case,  $t = T(x, \theta)$ .  $\parallel$

**Remark** In this chapter, the test statistics can be written without  $\theta$ .  $\parallel$

**Definition 3.3** The *P-value* is the probability that chance alone would produce a test statistic as extreme as the observed test statistic if the null hypothesis were true. For example, if large values of the test statistic support the alternative hypothesis, the *P-value* is the probability  $P(T \geq t)$ .  $\parallel$

**Example 3.2** In the mice example (Section 3.1), we let the test statistic be the difference in means,  $T = T(X_1, X_2, X_3, Y_1, Y_2, Y_3) = \bar{X} - \bar{Y}$  with observed value  $t = \bar{x} - \bar{y} = 4.67$ . Large values of the test statistic support the alternative hypothesis, so the *P-value* is  $P(T \geq 4.67) = 3/20$ .  $\square$

**Definition 3.4** A result is *statistically discernible* if it would rarely occur by chance. How rarely? It depends on context, but for example a *P-value* of 0.0002 would indicate that if the null hypothesis is true, then the observed outcome would occur just 2 out of 10 000 times by chance alone, which in

most circumstances seems pretty rare; we would conclude that the evidence supports the alternative hypothesis. ||

**Example 3.3** Suppose public health officials are concerned about lead levels in drinking water due to old pipes throughout a city. The officials measure lead levels in a sample of households and test the hypotheses that lead levels are at a safe level versus the alternative that the lead levels are at an unsafe level. They find that the mean value of lead found in these households is at an unsafe level, with a  $P$ -value of 0.06. If lead levels in the city are truly safe, should we consider an outcome that occurs 6 out of 100 by chance a rare event? Considering the consequences of being wrong, officials might conclude that this result is statistically discernible and something other than chance variability accounts for the mean lead level they obtained; they would conclude that lead levels in the city are indeed unsafe.

On the other hand, suppose you want to prepare for the College Board SAT Math exam. An online company provides intense tutoring at a cost of \$1000. You find the results of an experiment conducted by an independent researcher that tested the hypotheses that with this tutoring, the mean SAT math score will stay the same versus the mean SAT math score will increase. From their data, they estimate that the mean score increases by 10 points, with a  $P$ -value of 0.06. So if the tutoring is not effective (mean score stays the same), then 6 out of 100 times, we'd obtain the observed result by chance. Is that enough evidence to convince you that the mean increase is unlikely to be due to chance, and it is the intense tutoring that explains the increase? At a cost of \$1000, would you sign up for the tutoring? What if the cost of the tutoring was \$5? □

The smaller you require the  $P$ -value to be to declare the outcome statistically discernible, the more conservative you are being – you are requiring stronger evidence to reject the status quo (the null hypothesis). We will discuss  $P$ -values in more detail in Chapter 8.

Rather than just calculating the probability, we often begin by answering a larger question: What is the distribution of the test statistic when there is no real effect? For example, Table 3.1 gives all values of the test statistic in the mice example; each value has the same probability if there is no drug effect.

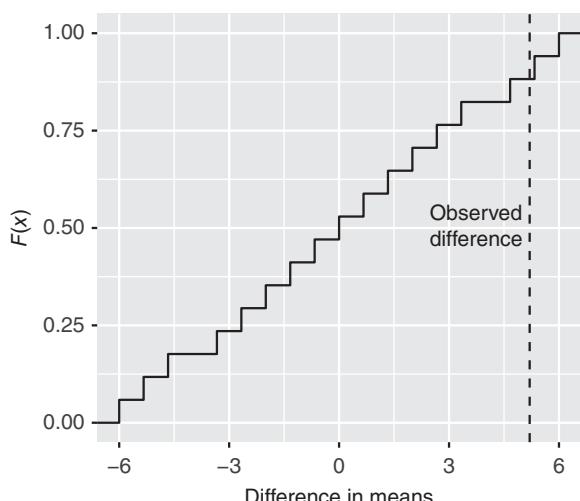
**Definition 3.5** The *null distribution* is the distribution of the test statistic if the null hypothesis is true. ||

You can think of the null distribution as a reference distribution; we compare the observed test statistic to this reference to determine how unusual the observed test statistic is. Figure 3.1 shows the cumulative distribution function of the null distribution in the mice example.

**Table 3.1** All possible partitions of {30, 25, 20, 18, 21, 22} into two sets.

Drug				Control		$\bar{X}_D$	$\bar{X}_C$	Difference in means
18	20	21	22	25	30	19.67	25.67	-6.00
18	20	22	21	25	30	20	25.33	-5.33
18	20	25	21	22	30	21	24.33	-3.33
18	20	30	21	22	25	22.67	22.67	0.00
18	21	22	20	25	30	20.33	25	-4.67
18	21	25	20	22	30	21.33	24	-2.67
18	21	30	20	22	25	23	22.33	0.67
18	22	25	20	21	30	21.67	23.67	-2.00
18	22	30	20	21	25	23.33	22	1.33
18	25	30	20	21	22	24.33	21	3.33
20	21	22	18	25	30	21	24.33	-3.33
20	21	25	18	22	30	22	23.33	-1.33
20	21	30	18	22	25	23.67	21.67	2.00
20	22	25	18	21	30	22.33	23	-0.67
20	22	30	18	21	25	24	21.33	2.67
20	25	30	18	21	22	25	20.33	4.67 *
21	22	25	18	20	30	22.67	22.67	0.00
21	22	30	18	20	25	24.33	21	3.33
21	25	30	18	20	22	25.33	20	5.33 *
22	25	30	18	20	21	25.67	19.67	6.00 *

Rows where the difference in means exceeds the original value are highlighted.

**Figure 3.1** Empirical cumulative distribution function of the null distribution for difference in means for mice.

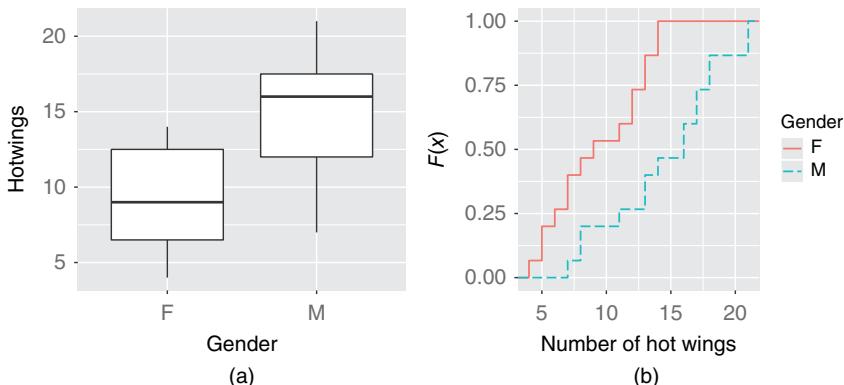
There are different ways to calculate exact or approximate null distributions and  $P$ -values. For now we focus on one method – permutation tests.

### 3.3 Permutation Tests

In the mice example in Section 3.1, we compared the test statistic to a reference distribution using permutations of the observed data. We investigate this approach in more detail.

Recall the beer and hot wings case study in Section 1.9. The mean number of wings consumed by females and males were 9.33 and 14.53, respectively, while the standard deviations were 3.56 and 4.50, respectively. See Figure 3.2 and Table 3.2.

The sample means for the males and females are clearly different, but the difference ( $14.53 - 9.33 = 5.2$ ) could have arisen by chance. Can the difference *easily* be explained by chance alone? If not, we will conclude that there are genuine gender differences in hot wings consumption.



**Figure 3.2** Number of hot wings consumed, by gender.

**Table 3.2** Hot wings consumption.

Females					Males				
4	5	5	6	7	7	8	8	11	13
7	8	9	11	12	13	14	16	16	17
12	13	13	14	14	17	18	18	21	21

For a hypothesis test, let  $\mu_M$  denote the mean number of hot wings consumed by males and  $\mu_F$  denote the mean number of hot wings consumed by females. We test

$$H_0: \mu_M = \mu_F \quad \text{versus} \quad H_A: \mu_M > \mu_F$$

or equivalently

$$H_0: \mu_M - \mu_F = 0 \quad \text{versus} \quad H_A: \mu_M - \mu_F > 0.$$

We use  $T = \bar{X}_M - \bar{X}_F$  as a test statistic, with observed value  $t = 5.2$ .

Suppose there really is no gender influence in the number of hot wings consumed by bar patrons. Then the 30 numbers come from a single population, the way they were divided into two groups (by labeling some as male and others as female) is essentially random, and any other division is equally likely. For instance, the distribution of hot wings consumed might have been as below:

Females					Males				
5	6	7	7	8	4	5	7	8	9
8	11	12	13	14	11	12	13	13	13
14	14	16	16	21	17	17	18	18	21

In this case, the difference in means is  $12.4 - 11.47 = 0.93$ .

We could proceed, as in the mice example, calculating the difference in means for *every* possible way to split the data into two samples of size 15 each. This would result in  $\binom{30}{15} = 155\,117\,520$  differences! In practice, such exhaustive calculations are impractical unless the sample sizes are small, so we resort to sampling instead.

We create a *permutation resample*, or *resample* for short, by drawing  $n_1 = 15$  observations *without* replacement from the pooled data to be one sample (the males), leaving the remaining  $n_2 = 15$  observations to be the second sample (the females). We calculate the statistic of interest, for example difference in means of the two samples. We repeat this many times (1000 or more). The *P*-value is then the fraction of times the random statistic exceeds<sup>2</sup> the original statistic.

---

<sup>2</sup> In hypothesis testing, “exceeds” means  $\geq$  rather than  $>$ .

We follow this algorithm:

### Two-Sample Permutation Test

Pool the  $m + n$  values.

**repeat**

    Draw a resample of size  $n_1$  without replacement.

    Use the remaining  $n_2$  observations for the other sample.

    Calculate the difference in means or another statistic that compares samples.

**until** we have enough samples.

Calculate the  $P$ -value as the fraction of times the random statistics exceed the original statistic. Multiply by 2 for a two-sided test.

Optionally, plot a histogram of the random statistic values.

The distribution of this difference across all permutation resamples is the *permutation distribution*. This may be exact (calculated exhaustively) or approximate (implemented by sampling). In either case, we usually use statistical software for the computations. Here is code that performs the test in R.

### R Note

We first compute the observed mean difference in the number of hot wings consumed by males and females.

```
> Beerwings %>% group_by(Gender) %>% summarize(mean(Hotwings))
# A tibble: 2 x 2
  Gender `mean(Hotwings)`
  <fct>     <dbl>
1 F          9.33
2 M         14.5
> observed <- 14.5333 - 9.3333 # store observed mean difference
> observed
[1] 5.2
```

Since we will use the hot wings variable repeatedly, we create a vector holding these values. Then we draw a random sample of size 15 from the numbers 1 through 30 (there are 30 observations total). The hot wing values corresponding to these positions correspond to males and the others to females. We store the difference in means in `result`. We repeat those steps many times.

```
hotwings <- Beerwings$Hotwings
# Alternative syntax using the dplyr package:
# hotwings <- Beerwings %>% pull(Hotwings)
```

```
N <- 10^5 - 1          # number of times to repeat this process
result <- numeric(N) # space to save the random differences
for (i in 1:N)
{
  # sample of size 15, from 1 to 30, without replacement
  index <- sample(30, size = 15, replace = FALSE)
  result[i] <- mean(hotwings[index]) - mean(hotwings[-index])
}
```

We create a histogram of the permutation distribution and add a vertical line at the observed mean difference (Figure 3.3).

```
ggplot() + geom_histogram(aes(result), bins = 8) +
  geom_vline(xintercept = observed, linetype="dashed")
```

We determine how likely it is to obtain an outcome as larger or larger than the observed value.

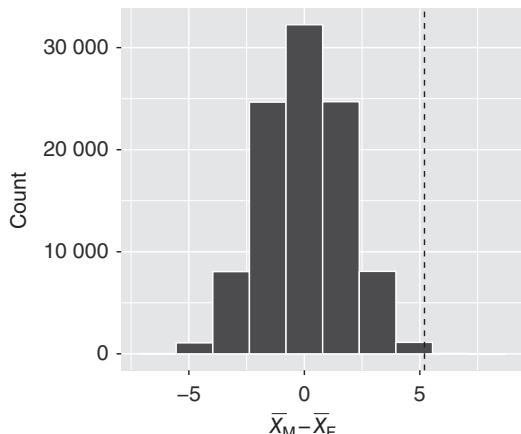
```
> (sum(result >= observed) + 1)/(N + 1) # P-value
[1] 0.00081                         # results will vary
```

The code snippet `result >=observed` gives a vector of TRUE's and FALSE's, and converting those to numerical values gives 1s and 0s. Hence, `sum(result >= observed)` counts the number of TRUE's.

Thus, the computed *P*-value is the proportion of statistics (including the original) that are as large or larger than the original mean difference.

From the output, we see that the observed difference in means is 5.2. The *P*-value is 0.00081. Of the  $10^5 - 1$  resamples computed by R, less than 0.1% of the resampled difference in means were as large or larger than 5.2. There are two possibilities – either there is a real difference, or there is no real effect, but

**Figure 3.3** Permutation distribution of the difference in means, male – female, in the beer and hot wings example.



a miracle occurred giving a difference well beyond the range of normal chance variation. We cannot rule out the miracle, but the evidence does support the hypothesis that females in this study consume fewer hot wings than males.

The participants in this study were a convenience sample – they were chosen because they happened to be at the bar when the study was conducted. Thus, we cannot make any inference about a population.

### 3.3.1 Implementation Issues

We note here some implementation issues for permutation tests. The first (choice of test statistic) applies to both the exhaustive and sampling implementations, while the final three (add one to both numerator and denominator, sample with replacement from null distribution, and more samples for better accuracy) are specific to sampling.

#### *Choice of Test Statistic*

In the examples above, we used the difference in means. We could have equally well used  $\bar{X}_1$  (the mean of the first sample),  $n_1\bar{X}_1$  (the sum of the observations in the first sample), or a variety of other test statistics. For example, in Table 3.1, the same three rows have test statistics that exceed the observed test statistic, whether the test statistic is difference in means or  $\bar{X}_D$  (the mean of the sample in the drug group).

Here is the result that states this more formally:

**Theorem 3.1** In permutation testing, if two test statistics  $T_1$  and  $T_2$  are related by a strictly increasing function,  $T_1(X^*) = f(T_2(X^*))$ , where  $X^*$  is any permutation resample of the original data  $x$ , then they yield exactly the same  $P$ -values, for both the exhaustive and resampling versions of permutation testing.

*Proof.* For simplicity, we consider only a one-sided (greater) test. Let  $X^*$  be any permutation resample. Then

$$\begin{aligned} p_1 &= P(T_2(X^*) \geq T_2(x)) \\ &= P(f(T_2(X^*)) \geq f(T_2(x))) \quad \text{since } f \text{ is strictly increasing} \\ &= P(T_1(X^*) \geq T_1(x)) \quad \text{by hypothesis.} \end{aligned}$$

Furthermore, in the sample implementation, exactly the same permutation resamples have  $T_2(X) \geq T_2(x)$  as have  $T_1(X) \geq T_1(x)$ , so counting the number or fraction of samples that exceed the observed statistic yields the same results.  $\square$

**Remark** One subtle point is that the transformation must be strictly monotone *for the observed data*, not for all possible sets of data. For example, in the mice example, both samples are size 3 and together total 136, so  $3\bar{x}_1 + 3\bar{x}_2 = 136$  for every resample, and  $\bar{x}_2 = 136/3 - \bar{x}_1$ . Let  $T_1 = \bar{X}_1^* - \bar{X}_2^*$  and  $T_2 = \bar{X}_1^*$ , then for these data  $T_1 = \bar{X}_1^* - \bar{X}_2^* = 2\bar{X}_1^* - 136/3 = 2T_2 - 136/3$  is a monotone function of  $T_2$ , so qualifies under the theorem. ||

#### Add One to Both Numerator and Denominator

When computing the  $P$ -value in the sampling implementation, we add one to both numerator and denominator. This corresponds to including the original data as an extra resample. This is a bit conservative and avoids reporting an impossible  $P$ -value of 0.0 – since there is always at least one resample that is as extreme as the original data, namely, the original data itself.

#### Sample with Replacement from the Null Distribution

In the sampling implementation, we do not attempt to ensure that the resamples are unique. In effect, we draw resamples *with replacement* from the population of  $\binom{m+n}{m}$  possible resamples, and hence obtain a sample with replacement from the  $\binom{m+n}{m}$  test statistics that make up the exhaustive null distribution. Sampling without replacement would be more accurate, but it is not feasible, requiring too much time and memory to check that a new sample does not match any previous sample.

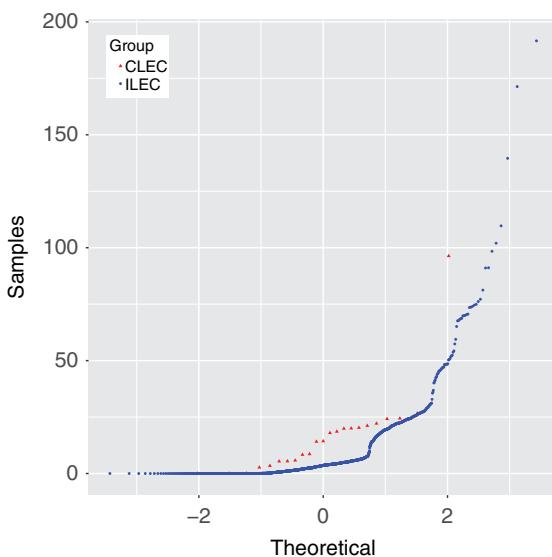
#### More Samples for Better Accuracy

In the hot wings example, we resampled 99 999 times. In general, the more resamples, the better. If the true  $P$ -value is  $p$ , the estimated  $P$ -value has variance approximately equal to  $p(1-p)/N$ , where  $N$  is the number of resamples.

**Remark** Just as the original  $n$  data values are a sample from the population, so too the  $N$  resampled statistics are a sample from a population (in this case, the null distribution). ||

The next example features highly skewed distributions and unbalanced sample sizes, as well as the need for high accuracy.

**Example 3.4** Recall the Verizon case study in Section 1.3. Whether Verizon is judged to be making repairs slower for competitors' customers is determined using hypothesis tests, as mandated by the New York Public Utilities



**Figure 3.4** Normal quantile plots of the ILEC and CLEC data.

Commission (PUC). Thousands of tests are performed to compare the speed of different types of repairs, over different time periods, relative to different competitors. If substantially more than 1% of the tests give  $P$ -values below 1%, then Verizon is deemed to be discriminating.

The mean of 1664 repairs for incumbent local exchange carrier (ILEC) customers is 8.4 h, whereas the mean for 23 repairs for competing local exchange carrier (CLEC) customers is 16.5 h. Could a difference that large easily be explained by chance? Figure 3.4 shows the normal quantile plots for the raw data for one of these tests. The distributions of both samples are skewed and there appears to be one outlier in the CLEC data set; perhaps that explains the difference in means? However, it would not be reasonable to throw out that observation as faulty – it is clear from the larger data set that large repair times do occur fairly frequently. Furthermore, even in the middle of both distributions, the CLEC times do appear to be longer. There are curious bends in the normal quantile plot, due to 24-h cycles.

Let  $\mu_1$  denote the mean repair time for the ILEC customers and  $\mu_2$  the mean repair time for the CLEC customers. We test

$$H_0: \mu_1 = \mu_2 \quad \text{versus} \quad H_A: \mu_1 < \mu_2.$$

We use a one-sided test because the alternative of interest to the PUC is that the CLEC customers are receiving worse service (longer repair times) than the ILEC customers.

## R Note

```
> Verizon %>% group_by(Group) %>% summarize(mean(Time))
# A tibble: 2 x 2
  Group `mean(Time)`
  <fct>     <dbl>
1 CLEC      16.5
2 ILEC      8.41
```

We create three vectors, one containing the times for all the customers, one for the ILEC customers, and one for the CLEC customers.

```
Time <- Verizon$Time
TimeILEC <- Verizon %>% filter(Group == "ILEC") %>% pull(Time)
TimeCLEC <- Verizon %>% filter(Group == "CLEC") %>% pull(Time)
```

Now, we compute the mean difference in repair times and store in the vector observed.

```
> observed <- mean(TimeILEC) - mean(TimeCLEC)
> observed
[1] -8.09752
```

We draw a random sample of size 1664 (size of ILEC group) from 1, 2, ..., 1687. The times that correspond to these observations go into the ILEC group; the remaining times go into the CLEC group.

```
N <- 10^4-1
result <- numeric(N)
for (i in 1:N)
{
  index <- sample(1687, size = 1664, replace = FALSE)
  result[i] <- mean(Time[index]) - mean(Time[-index])
}
```

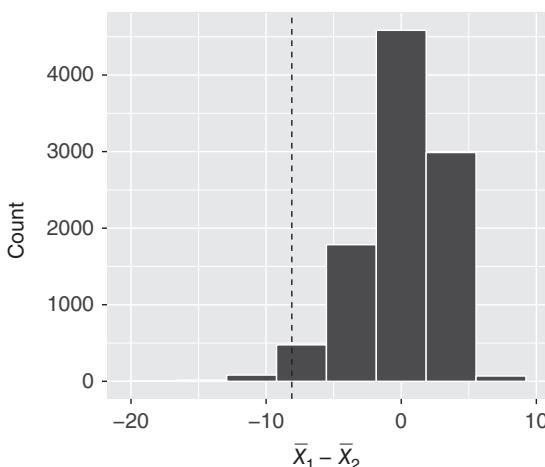
First, plot the histogram

```
ggplot() + geom_histogram(aes(result), bins = 8) +
  geom_vline(xintercept = observed, linetype = "dashed")
```

Note that we want to find the proportion of times the resampled mean difference is *less than or equal* to the observed mean difference.

```
(sum(result <= observed) + 1) / (N + 1)
```

One run of the simulation results in a  $P$ -value of 0.0178 indicating that a difference in means as small or smaller than the observed difference of  $-8.097$  would occur less than 2% of the time if the mean times were truly equal.



**Figure 3.5** Permutation distribution of difference of means (ILEC – CLEC) for the Verizon repair time data.

In the above simulation, we used  $10^4 - 1$  resamples to speed up the calculations. For higher accuracy, we should use a half-million resamples; this was negotiated between Verizon and the PUC. The goal is to have only a small chance of a test wrongly being declared statistically discernible or not, due to random sampling.

The permutation distribution is shown in Figure 3.5. The  $P$ -value is the fraction of the distribution that falls to the left of the observed value.

This test works fine even with unbalanced sample sizes of 1664 and 23 and even for very skewed data. The permutation distribution is left-skewed, but that doesn't matter; both the observed statistic and the permutation resamples are affected by the size imbalance and skewness in the same way.  $\square$

### 3.3.2 One-Sided and Two-Sided Tests

For the hypothesis test with alternative  $H_A: \mu_1 - \mu_2 < 0$ , we compute a  $P$ -value by finding the fraction of resample statistics that are less than or equal to the observed test statistic (or greater than or equal to for the alternative  $\mu_1 - \mu_2 > 0$ ).

For a two-sided test, we calculate both one-sided  $P$ -values, multiply the smaller by 2, and finally (if necessary) round down to 1.0 (because probabilities can never be larger than 1.0).

In the mice example with observed test statistic  $t = 4.67$ , the one-sided  $P$ -values are 3/20 for  $H_A: \mu_d - \mu_c > 0$  and 18/20 for  $H_A: \mu_d - \mu_c < 0$ . Hence the two-sided  $P$ -value is 6/20 = 0.30 (recall Table 3.1).

Two-sided  $P$ -values are the default in statistical practice – you should perform a two-sided test unless there is a clear reason to pick a one-sided

alternative hypothesis. It is not fair to look at the data before deciding to use a one-sided hypothesis.

**Example 3.5** We return to the Beerwings data set, and the comparison of the mean number of hot wings consumed by males and females. Suppose prior to this study, we had no preconceived idea of which gender would consume more hot wings. Then our hypotheses would be

$$H_0: \mu_M = \mu_F \quad \text{versus} \quad H_A: \mu_M \neq \mu_F.$$

We find the one-sided  $P$ -value (for alternative “greater”) to be 0.000831, so for a two-sided test, we double 0.000831 to obtain the  $P$ -value 0.00166.

If gender does not influence average hot wings consumption, then a difference as or more extreme than the observed difference would occur only about 0.2% of the time. We conclude that males and females do not consume, on average, the same number of hot wings.  $\square$

#### To Obtain $P$ -Values in the Two-Sided Case We Multiply by 2

We multiply the smaller of the one-sided  $P$ -values by 2, using the observed test statistic. Multiplying by 2 has a deeper meaning. Because we are open to more than one alternative to the null hypothesis, it takes stronger evidence for any one of these particular alternatives to provide convincing evidence that the null hypothesis is incorrect. With two possibilities, the evidence must be stronger by a factor of 2, measured on the probability scale.

### 3.3.3 Other Statistics

We noted in Section 3.3.1 the possibility of using a variety of statistics and getting equivalent results, provided the statistics are related by a monotone transformation.

Permutation testing actually offers considerably more freedom than that; the basic procedure works with any test statistic. We compute the observed test statistic, resample, compute the test statistics for each resample, and compute the  $P$ -value (see the algorithm in Section 3.3). Nothing in the process requires that the statistic be a mean or equivalent to a mean.

This provides the flexibility to choose a test statistic that is more suitable to the problem at hand. Rather than using means, for example we might base the test statistic on *robust statistics*, that is, statistics that are not sensitive to outliers. Two examples of robust statistics are the median and the trimmed mean. We have already encountered the median. The trimmed mean is just a variant of the mean: We sort the data, omit a certain fraction of the low and high values, and calculate the mean of the remaining values. In addition, permutation tests could also compare proportions or variances. We illustrate each of these cases next.

**Example 3.6** In the Verizon example, we observed that the data have a long tail – there are some very large repair times (Figure 3.4). We may wish to use a test statistic that is less sensitive to these observations. There are a number of reasons we might do this. One is to get a better measure of what is important in practice – how inconvenienced customers are by the repairs. After a while, each additional hour probably does not matter as much, yet a sample mean treats an extra 1 h on a repair time of 100 h the same as an extra 1 h on a repair time of 1 h. Second, a large recorded repair time might just be a blunder; for example a repair time of  $10^6$  h must be a mistake. Third, a more robust statistic could be more sensitive at detecting real differences in the distributions – the mean is so sensitive to large observations that it pays less attention to moderate observations, whereas a statistic more sensitive to moderate observations could detect differences between populations that show up in the moderate observations.

Here is the R code for permutation tests using medians and trimmed means.

#### R Note for Verizon, Cont

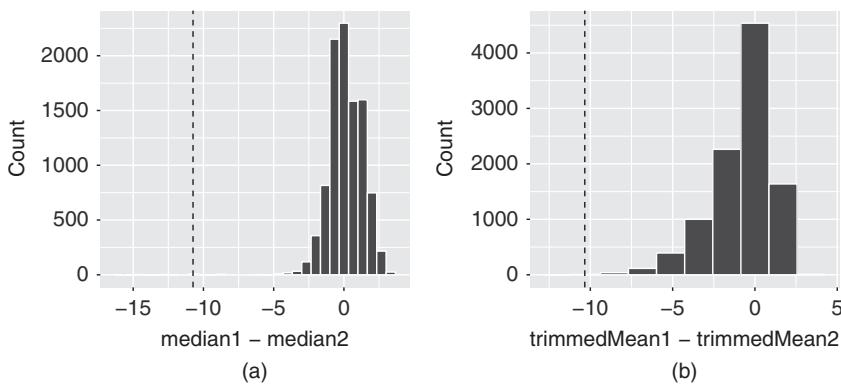
```
observed <- median(TimeILEC) - median(TimeCLEC)
N <- 10^4-1
result <- numeric(N)
for (i in 1:N)
{
  index <- sample(1687, size = 1664, replace = FALSE)
  result[i] <- median(Time[index]) - median(Time[-index])
}
(sum(result <= observed) + 1)/(N + 1) # P-value
```

To obtain the results for trimmed means, we add the option `trim=.25` to the `mean` command. Substitute the following in the above:

```
observed <- (mean(TimeILEC, trim = .25) -
              mean(TimeCLEC, trim = .25))
result[i] <- (mean(Time[index], trim = .25) -
               mean(Time[-index], trim = .25))
```

It seems apparent that these more robust statistics are more sensitive to a possible difference between the two populations; we would conclude the outcomes are statistically discernible with estimated *P*-values of 0.0015 and 0.0004, respectively. Figure 3.6 also suggests that the observed statistics are well outside the range of normal chance variation.

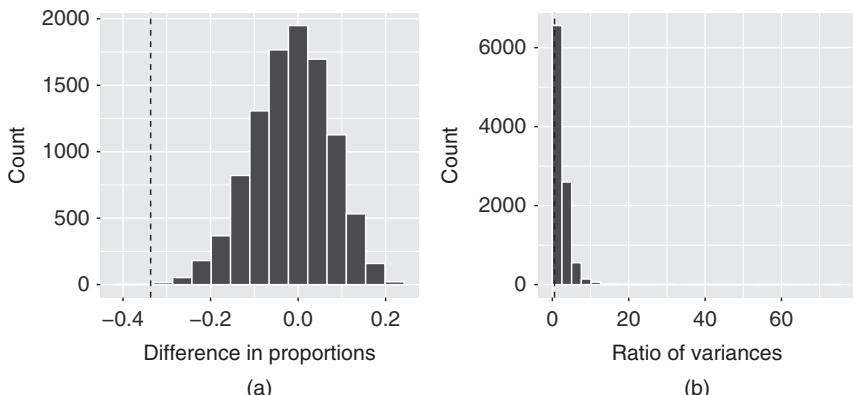
One caveat is in order – it is wrong to try many different tests, possibly with minor variations, until you obtain a statistically discernible outcome. If you try enough different things, eventually one result will be statistically discernible, whether or not there is a real difference.



**Figure 3.6** Repair times for Verizon data. (a) Permutation distribution for difference in medians. (b) Permutation distribution for difference in 25% trimmed means.

There are ways to guard against this. In Section 8.5.5.2, we will learn about different corrections to avoid these excessive positives when doing multiple tests.

We can also apply permutation tests to questions other than comparing the centers of two populations, for example the difference in the proportion of repair times that exceed 10 h or the ratio of variances of the two populations. Using the R code below, it appears that the proportions do differ ( $P\text{-value} = 0.0006$ , one-sided), while the variances do not ( $P\text{-value} = 0.267$ , two-sided). The permutation distributions are very different (see Figure 3.7), but this does not affect the validity of the method.



**Figure 3.7** Repair times for Verizon data. (a) Difference in proportion of repairs exceeding 10 h. (b) Ratio of variances (ILEC/CLEC).

### R Note for Verizon, Cont

The command `mean(TimeILEC > 10)` computes the proportion of times the ILEC times are greater than 10.

```
> observed <- mean(TimeILEC > 10) - mean(TimeCLEC > 10)
> observed
[1] -0.336852
```

Thus, about 33.7% fewer ILEC customers had repair times exceeding 10 h.

We reuse the previous code for trimmed means but with the following modification to compute the difference in proportions:

```
result[i] <- mean(Time[index] > 10) - mean(Time[-index] > 10)
```

To perform the test for the ratio of variances, substitute:

```
observed <- var(TimeILEC) / var(TimeCLEC)
result[i] <- var(Time[index]) / var(Time[-index])
```

□

### 3.3.4 Conditions

Under what conditions can we use the permutation test? First, the permutation test does not require particular distributions, such as normal populations.

In fact, permutation testing does not even require that the data be drawn by random sampling from two populations. A study for the treatment of a rare disease could include all patients with the disease in the world. In this case, it does require that subjects be assigned to the two groups randomly.

In the usual case that the two groups are samples from two populations, pooling the data does require that the two *populations* have the same distribution when the null hypothesis is true. They must have the same mean, spread, and shape. This does not mean that the two *samples* must have the same mean, spread, and shape – there will always be some chance variation in the data.

In practice, the permutation test is usually robust when the two populations have different distributions. The major exception is when the two populations have different spreads and the sample sizes are dissimilar. This exception is rarely a concern in practice, unless you have other information (besides the data) that the spreads are different. For example, one of us consulted for a large pharmaceutical company testing a new procedure for measuring a certain quantity; the new procedure was substantially cheaper, but not as accurate. The higher variability was acceptable, provided that the means matched. This is a case where permutation testing would be doubtful because it would pool data from different distributions. Even then, it would usually work fine if the sample sizes were equal.

**Example 3.7** Let us look at an extreme case of different variances and sample sizes. Suppose population A is normal with mean 0 and variance  $\sigma_A^2 = 10^6$ , and population B is normal with mean 0 and variance  $\sigma_B^2 = 1$ . Both populations have mean 0, and the null hypothesis that the difference in means is zero is true. Draw a sample of size  $n_A = 10^2$  from population A and a sample of size  $n_B = 10^6$  from population B. Let the test statistic be  $T = \bar{X}_A$ . When drawing the original sample,  $T$  has variance  $\sigma_A^2/n_A = 10^4$  (by Theorem A.7). What is the probability that this statistic  $T$  is greater than, say, 5? By standardizing, we find

$$P(T \geq 5) = P\left(\frac{T}{100} \geq \frac{5}{100}\right) = P(Z \geq 0.05) = 0.48.$$

Thus, with its huge variance of  $10^4$ , there is nearly a 50% chance of  $T$  being greater than 5.

When we pool the two samples, the variance of the pooled data is approximately  $(n_A\sigma_A^2 + n_B\sigma_B^2)/(n_A + n_B) \approx 101$  (plus some random variation), and the resampled  $T$ 's have variance around  $101/n_A \approx 1.01$ , or equivalently, a standard deviation about 1.005 (again, by Theorem A.7). So almost none of the permutation  $T$ 's will be larger than 5:

$$P(T \geq 5) = P\left(\frac{T}{1.005} \geq \frac{5}{1.005}\right) = P(Z \geq 4.975) \approx 0.$$

Thus, there is nearly a 50% chance of reporting a  $P$  value near 0 and erroneously concluding that the means are not the same.  $\square$

**Example 3.8** In the Iowa recidivism case study in Section 1.4, we have the population of offenders, convicted in Iowa of either a felony or a misdemeanor, who were released from prison in 2010. Of these, 36.5% of those under 25 years of age were sent back to prison compared to 30.6% of those 25 years of age or older, so the observed difference in proportions is 0.059. Is this difference statistically discernible? We can perform a permutation test to check.

### R Note

The `Age25` variable has some missing values so we first remove those observations, using the `drop_na` function from the `tidyverse` package.

```
library(tidyverse)
data <- Recidivism %>% drop_na(Age25) %>%
  select(Age25, Recid)
table(data$Age25)
proportions(table(data$Age25, data$Recid), 1)
```

Note that there were 3077 offenders under the age of 25 and 13 942 offenders who were 25 years of age or older.

```

Recid <- data$Recid    # create vector
observed <- .365 - .306
N <- 10^4 - 1
result <- numeric(N)
for (i in 1:N)
{
  index <- sample(17019, size = 3077, replace = FALSE)
  result[i] <- mean(Recid[index]=="Yes") -
    mean(Recid[-index]=="Yes")
}
2*(sum(result >= observed)+1) / (N+1)

```

For a two-sided test, the  $P$ -value is  $2 \times 10^{-4}$ , so we conclude that there is a statistically discernible difference in recidivism between those under 25 years of age and those 25 years of age or older.

This test tells us that if recidivism was a random occurrence, unrelated to age group, then the chance of observing an outcome as or more extreme than the observed difference in proportions of 0.059 is  $2 \times 10^{-4}$ .  $\square$

**Example 3.9** The Pew Research Center is a nonpartisan organization that conducts studies on “the issues, attitudes and trends shaping the world.” In February 2021, they published a report on “Faith Among Black Americans.”<sup>3</sup> Part of the report looked at differences between Blacks from different generations: Generation Z born between 1997 and 2002, Millennials born between 1981 and 1996, and Generation X born between 1965 and 1980. Of the 257 Black participants from Generation Z 41% responded that religion was very important, while of the 2094 Black Millennials 46% responded that religion was very important. Could this 5% point difference be due to chance variability?

In this example, we can think of the responses as being a collection of 1s and 0s – a 1 corresponding to the response that religion was important, 0 otherwise. Thus, we have a collection of 105 1s and  $257 - 105 = 152$  0s for Generation Z, and 963 1s and  $2094 - 963 = 1131$  0s for the Millennials. To conduct the permutation test, we pool the data which gives us 1068 1s and 1283 0s.

Since we are not provided with a data set, we use the `rep` function to create a vector of 1s and 0s.

### R Note

```

pooled.data <- rep(c(1,0), c(1068, 1283))  # create vector
observed <- (963/2094) - (105/257)  # observed difference
                                         # (Mill-Gen Z)
N <- 10^4-1
result <- numeric(N)

```

---

3 [https://www.pewforum.org/2021/02/16/faith-among-black-americans.](https://www.pewforum.org/2021/02/16/faith-among-black-americans/)

```

for (i in 1:N)
{
  index <- sample(2351, 2094, replace = FALSE)
  result[i] <- mean(pooled.data[index]) -
    mean(pooled.data[-index])
}
2 * (sum(result >= observed)+1) / (N+1)

```

One run of the algorithm gives a  $P$ -value of 0.1284. Thus, if the proportions truly are the same, then 12.8% of resamples would result in an outcome as or more extreme than the observed difference of 5%. Hence, chance variability easily accounts for the observed difference.  $\square$

### 3.3.5 Remark on Terminology

Why is the two-sample permutation test above called *permutation* testing? It seems like all we are doing is splitting the data into two samples, with no hint of a permutation. Well, imagine storing the data in a table with two columns and  $m + n$  rows; the first column contains labels, for example,  $m$  copies of “M” and  $n$  copies of “F,” whereas the second contains the numerical data. We may permute the rows of either column, randomly; this is equivalent to splitting the data into two groups randomly.

Table 3.3 illustrates one such permutation of one of the columns in the beerwings data.

**Table 3.3** Partial view of Beerwings data set.

Gender		Hot wings	Gender		Hot wings
1	F	4	11	F	9
2	F	5	26	F	17
3	F	5	25	F	17
4	F	6	2	F	5
5	F	7	4	F	6
6	F	7	8	F	8
7	M	7	3	M	5
8	F	8	20	F	14
9	M	8	10	M	8
10	M	8	18	M	13
:			:		

The Gender column is held fixed and the rows of the Hotwings variable are permuted. The first column indicates which rows of the hot wing values were permuted.

This idea of permuting the rows of one column generalizes to other situations, including the analysis of contingency tables, which we will encounter in Chapter 10.

### 3.4 Matched Pairs

Divers competing in the FINA 2017 World Championships perform five dives in each of several rounds.<sup>4</sup> The sum of the five scores determines who moves on to the next round. Do divers tend to get the same score, on average, in the semifinal and final rounds of a competition? Or might the scores in the final round be different, due to fatigue, or heightened effort, or a strategy to perform more difficult dives in the final round? We have the scores from the semifinal and final round of the 10 m platform for the top 12 female divers (Table 3.4). The average score in the semifinal is 338.50 and the final is 350.475. Is this a real difference or could this be attributed to chance variability?

Now, it may be tempting to proceed as we did in investigating the mean number of hot wings consumed by men and women, by comparing the mean scores in the semifinal and final rounds. But note that the data here are *not independent!* The scores that any particular diver receives in the semifinal and final rounds are related, in the sense that how well she dives depends on her training and genetics. Thus, the data are called *matched pairs* or *paired data*.

So, for instance if there is no true difference in how Qian Ren of China performs in the last two rounds, then the fact that she received a score of 367.5 in the semifinal and a 391.95 in the final is due to chance. In another circumstance, she might have received the 391.95 in the semifinal and the 367.5 in the final. For a permutation test, we randomly select some of the divers and transpose their two scores, leaving the other divers scores the same.

**Table 3.4** Partial view of diving scores in file Diving2017.

Name	Country	Semifinal	Final
Cheog Jun Hoong	Malaysia	325.5	397.5
Si Yajie	China	382.8	396.00
Ren Qian	China	367.5	391.95
Kim Mi Rae	North Korea	346.00	385.55
:			

<sup>4</sup> Fédération Internationale de Natation.

### R Note

Since the effect of transposing the semifinal and final score for a diver results in a sign change in the difference, we will draw 12 random values from  $\{-1, 1\}$ . A draw of  $-1$  indicates to transpose and multiply the difference by  $-1$ , while a  $1$  keeps the original order and value.

```
Diff <- Diving 2017 Final Diving 2017$Semifinal #difference in two scores
observed <- mean(Diff) #mean of difference

N <- 10^5-1
result <- numeric(N)

for (i in 1:N)
{
  Sign <- sample(c(-1,1), 12, replace=TRUE) #random vector of 1's or -1's
  Diff2 <- Sign*Diff #random pairs (a-b) -> (b-a)
  result[i] <- mean(Diff2) #mean of difference
}

ggplot() + geom_histogram(aes(result), bins = 8) +
  geom_vline(xintercept = mean(observed), linetype="dashed")

2 * (sum(result >= observed)+1) / (N+1) #P-value
```

We obtain a  $P$ -value of 0.26, which suggests that chance alone easily accounts for the observed difference in the mean diving scores between the semifinal and final rounds.

If we had performed a permutation test assuming that the final scores were independent of the semifinal scores, we would have obtained (in one simulation) a  $P$ -value of 0.37, a substantially larger probability. Although in this example, we would have reached the same conclusion, that is not always true. Thus, when you have two variables, it is important to think carefully about whether or not they represent data from two independent populations.

## 3.5 Cause and Effect

So far we have learned techniques for checking whether the difference between two groups can easily be explained by chance. But suppose it cannot – does that mean that the difference was *caused* by the issue we were considering? Or could there be some other cause? For example if repair times were longer

for CLEC customers than for Verizon customers, is that because Verizon is discriminating or could there be another explanation? Perhaps more CLEC customers ask for repairs on weekends.

The question of causality is a fundamental issue in statistics. Suppose that bicycle riders have more muscle mass – is that because of biking? If students in classrooms with higher teacher salaries learn more – is that because of the salaries?

What exactly do we mean by cause and effect, or causation?

Judea Pearl, who has made important and influential contributions to the theory of causality, gives an informal definition (Pearl et al., 2016):

A variable  $X$  is a *cause* of a variable  $Y$  if  $Y$  in any way relies on  $X$  for its value...think of causation as a form of listening;  $X$  is a cause of  $Y$  if  $Y$  listens to  $X$  and decides its value in response to what it hears.

Determining causation in observational studies or experiments that do not randomize treatments is hampered by the problem of unrecorded or hidden factors that may impact the outcome. In particular, there may be a *confounding* variable which is related to both the treatment and the outcome. We use a simple diagram to visualize this relationship. Suppose the variable (treatment)  $X$  is a cause of variable (outcome)  $Y$ : we draw an arrow from  $X$  to  $Y$ . See Figure 3.8.

For the mice experiment mentioned at the start of this chapter,  $X$  would be the drug treatment and  $Y$  would be the outcome of time through maze.

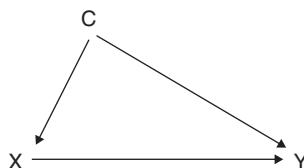
If there is a confounding variable ( $C$ ) affecting both the treatment ( $X$ ) and the outcome ( $Y$ ), we can depict this in a diagram by extending a (directed) edge from  $C$  to each of  $X$  and  $Y$  (Figure 3.9).

When confounding is present, we cannot conclude that (changes in)  $X$  is causing (changes in)  $Y$  – it might be that a third ( $C$ ) or other (possibly unobserved) variables are causing  $Y$ .

For instance, does spending time on Instagram ( $X$ ) cause mental anxiety ( $Y$ ) in teenagers? A confounding variable here might be family dynamics ( $C$ ):



**Figure 3.8** A diagram to indicate that  $X$  causes  $Y$ .



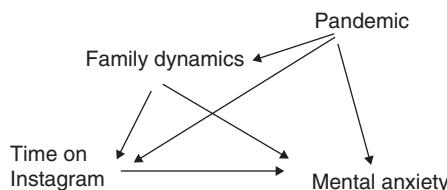
**Figure 3.9** Confounding variable  $C$  affecting both treatment  $X$  and outcome  $Y$ .

a teenager with a stressful family dynamic might be more likely to spend time on Instagram, and a teenager with a stressful family dynamic might have mental anxiety. So it might be that time on Instagram is causing the family dynamics that lead to mental anxiety.

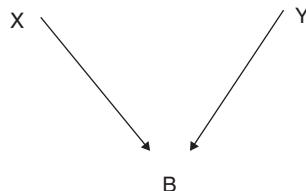
Confounding variables are a problem in observational studies or nonrandomized experiments, and researchers need to adjust for them during analysis. For example we would want to adjust for family dynamics in any analysis of the time on Instagram and mental anxiety study. Can you think of other possible confounding variables in this study (Figure 3.10)? The difficulty is that it may be impossible to account for all confounding variables. In particular, there might be factors that the researchers are unaware of.

Another possible variable impacting a study is a *collider*: both the treatment and the outcome cause the collider. In the case of a collider, it is possible that  $X$  has no relationship with  $Y$ , but the two variables can be associated by conditioning on the variable collider ( $B$ ) (Figure 3.11). To see how that can happen, consider the acts of tossing a red die ( $X$ ) and a green die ( $Y$ ). Let  $B$  denote the sum of the two dice. The results of tossing the two dice are independent; that is, if the outcome of tossing the red die is a 3, that tells us nothing about the value on the green die. But now condition on the sum: suppose we know that the sum of the dice is 5 (Figure 3.12). Then once we know the value on the red die, this will tell us the value on the green die, so  $X$  and  $Y$  are no longer independent.

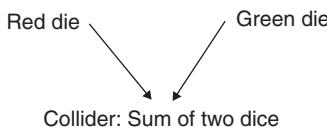
Consider a study to determine if there is an association between working the night shift at a job and restless leg syndrome. A researcher goes through hospital records to find patients who have been treated for sleep disorders.



**Figure 3.10** Possible confounders in a study of time on Instagram and mental anxiety in teenagers. Can you think of others?



**Figure 3.11** Both treatment  $X$  and outcome  $Y$  cause the collider  $B$ .



**Figure 3.12** Outcome on green die is not independent of red die if conditioning on the sum.

Since working night shifts and restless leg syndrome both cause problems with sleeping, sleep disorders is a collider. Thus, in a sample of hospital patients, by conditioning on only those patients with sleep disorders, the researcher may find an association between working the night shift and restless leg syndrome, even if these two variables are independent.

In some cases, the variable X may have a relationship with Y, but conditioning on the collider results in no relationship or a reverse relationship (e.g. a positive relationship changes to a negative relationship).

For example, consider a 1-year study to determine if a certain exercise regime will help patients lower their blood pressure. In longitudinal studies, retention of participants is a major problem. Perhaps the most obese patients are prone to dropping out of this study. If the patients with the highest blood pressures also find the exercise regime difficult and drop out, then the sample – only those who finish the study – would be a collider. It might be that this exercise regime is effective, but in the sample of (possibly healthier) patients who completed the entire study, it could turn out that there is not much of an effect.

More recently, researchers have been searching for the risk factors associated with infection from the SARS-CoV-2 virus. At least during the early days of the pandemic, because of the scarcity of diagnostic tests, much of the data on patients came from those with a severe case of COVID-19, not necessarily a representative sample of the general population. Again, the collider is the sample. By conditioning only on the part of the population that had a positive COVID-19 test, researchers may have observed a spurious association between certain risk factors and infection. The wealth of data from the COVID-19 pandemic has led to many observational studies which may be flawed due to collider bias (Griffith et al., 2020).

Currently, there are several frameworks for studying causal inference. One approach is the structural causal model (SCM) championed by Judea Pearl et al. (Pearl et al., 2016, Pearl 2000). Associated with each SCM is a causal graph of which Figures 3.8 and 3.9 are examples. These graphs with vertices that include the treatments, response, confounders, and colliders as well as other (possibly unobserved) factors that guide the researcher in identifying variables that must be controlled for in the analysis.

Another approach is the Rubin causal model (RCM) which is based on potential outcomes (Rubin, 1974). For instance, in a randomized experiment, suppose every participant gets either a drug or a placebo. Before the experiment, each

person has a potential outcome depending on whether or not this person gets the drug or the placebo. For each participant, the causal effect is the comparison between these two outcomes. Of course, the fundamental problem is that one of the outcomes is not known: that is, if Claire had received the drug, then how she would have reacted under the placebo is not unobserved. The RCM approach considers ways to model these unobserved potential outcomes. We will apply this framework in Section 13.4.2.

## Exercises

- 3.1** Suppose you conduct an experiment and inject a drug into three mice. Their times for running a maze are 8, 10, and 15 s; the times for two control mice are 5 and 9 s.
- Compute the difference in mean times between the treatment group and the control group.
  - Write out all possible ways to split these times into the two groups and calculate the corresponding differences in means.
  - What proportion of the differences are as large or larger than the observed difference in mean times?
  - For each permutation, calculate the mean of the treatment group only. What proportion of these means are as large or larger than the observed mean of the treatment group?
- 3.2** Your statistics professor comes to class with a big urn that she claims contains 9999 blue marbles and 1 red marble. You draw out one marble at random and find that it is red. Would you be willing to tell your professor that you think she is wrong about the distribution of colors? Why or why not? What are you assuming in making your decision? What if instead, she had claimed there are 9 blue marbles and 1 red one?
- 3.3** In a hypothesis test comparing two population means,  $H_0: \mu_1 = \mu_2$  versus  $H_A: \mu_1 > \mu_2$ ,
- Which  $P$ -value, 0.03 or 0.006 provides stronger evidence for the alternative hypothesis?
  - Which  $P$ -value, 0.095 or 0.04 provides stronger evidence that chance alone might account for the observed result?
- 3.4** In the algorithms for conducting a permutation test, why do we add 1 to the number of replications  $N$  when calculating the  $P$ -value?
- 3.5** In the flight delays case study in Section 1.1, the data contain flight delays for two airlines, American Airlines (AA) and United Airlines (UA).

- (a) Conduct a two-sided permutation test to see if the difference in mean delay times between the two carriers is statistically discernible.
  - (b) The flights took place in May and June of 2009. Conduct a two-sided permutation test to see if the difference in mean delay times between the 2 months is statistically discernible.
- 3.6** In the flight delays case study in Section 1.1, the data contains flight delays for two airlines, AA and UA.
- (a) Compute the proportion of times that each carrier's flights was delayed more than 20 min. Conduct a two-sided test to see if the difference in these proportions is statistically discernible (see the R Note in Example 3.6).
  - (b) Compute the variance in the flight delay lengths for each carrier. Conduct a test to see if the variances for UA and AA differ.
- 3.7** In the flight delays case study in Section 1.1, repeat Exercise 3.5 part (a) using three test statistics, (i) the mean of the UA delay times, (ii) the sum of the UA delay times, and (iii) the difference in means, and compare the *P*-values. Make sure all three test statistics are computed within the same `for` loop. What do you observe?
- 3.8** In the flight delays case study in Section 1.1,
- (a) Find the trimmed mean of the delay times for UA and AA by trimming 10% on either side of the distribution.
  - (b) Conduct a two-sided test to see if the difference in trimmed means is statistically discernible.
- 3.9** In the flight delays case study in Section 1.1,
- (a) Compute the proportion of times the flights in May and in June were delayed more than 20 min, and conduct a two-sided test to see if the difference between months is statistically discernible.
  - (b) Compute the ratio of the variances in the flight delay times in May and in June. Is this evidence that the true ratio is not equal to 1, or could this be due to chance variability? Conduct a two-sided test to check.
- 3.10** In the black spruce case study in Section 1.10, seedlings were planted in plots that were either subject to competition (from other plants) or not. Use the data set `Spruce` to conduct a test to see if the mean difference in how much the seedlings grew (in height) over the course of the study under these two treatments is statistically discernible.
- 3.11** In the Iowa recidivism case study in Section 1.4, for those offenders who recidivated, we have data on the number of days until they reoffended.

For those offenders who did recidivate, determine if the difference in the mean number of days (Days) until recidivism between those under 25 years of age and those 25 years of age and older is statistically discernible.

*Note:* Data on recidivism were collected for only 3 years from time of release from prison since studies suggest that most relapses occur within that time period. Thus, it is possible that some of the offenders who had not relapsed in that time period, might be convicted of another crime at a later point in time. The variable Days is *right censored*.

- 3.12** The file `Phillies2009` contains data from the 2009 season for the baseball team the Philadelphia Phillies.
- Compare the empirical distribution functions of the number of strikeouts per game (`StrikeOuts`) for games played at home and games played away (`Location`).
  - Find the mean number of strikeouts per game for the home and the away games.
  - Perform a permutation test to see if the difference in means is statistically discernible.
- 3.13** The data set `Oscars` contains the names of the Academy Awards best actor and best actress winners from 1928 through 2021.
- Compute summary statistics and create a graph of the ages grouped by gender and comment on what you observe.
  - Is it appropriate to perform a permutation test to determine if the difference in mean ages between the actors and actresses is statistically discernible? If yes, perform the test and interpret the result. If not, why not?
- 3.14** The file `Cafeteria` contains measurements on ingredients in a sample of dishes served in a college cafeteria (Stephenson, private communication); the dishes are also classified by type: meat or vegetarian.
- Create a plot to compare the distribution of fiber (in grams) between the meat and vegetarian dishes. Also, compute the mean and standard deviation.
  - Perform a permutation test to see if the difference in means is statistically discernible.
  - Note an outlier in the meat sample. If this observation is removed, do you think this would increase or decrease the difference in means? Try to answer this without doing a calculation, then check your answer.
  - Repeat the permutation test without the outlier and note how it changes the  $P$ -value.

- 3.15** Knee pain is a common problem for which there is no agreement on the best treatment. Researchers in the Netherlands conducted a study (van Linschoten et al., 2009) in which participants between 14 and 40 years of age with patellofemoral pain syndrome were randomized to either exercise therapy (supervised by a physical therapist) or “usual care” (typically, wait-and-see). Patients were excluded if they had knee osteoarthritis, previous knee injuries or surgery or other defined pathological conditions of the knee. After 12 months, 36 out of the 58 patients who followed the exercise program reported they had recovered compared to 30 out of the 59 patients in the control group. Perform a permutation test to see if the difference in proportions is statistically discernible.
- 3.16** Referring to the study in Example 3.9, another survey question asked participants how often they attended religious services. Of the 2475 Blacks from Generation X, 31% responded weekly or more compared to 23% of the 2094 Black Millennials. Is this difference of 8% statistically discernible?
- 3.17** Patients with the bacteria *Staphylococcus aureus* in the nose are at an increased risk for infection. Researchers conducted a study to determine if by rapidly identifying nasal carriers of this bacteria and treating these patients with a mupirocin nasal ointment and chlorhexidine soap is effective in the risk of hospital-associated *S. aureus* infection. In a double-blind trial (Bode et al., 2010), 917 patients were identified with carrying this bacteria in their nose. Of these 504 were randomly assigned to receive the mupirocin/chlorhexidine treatment, while the remaining 413 received a placebo ointment and soap. At the end of the study, 17 (3.4%) in the treatment group and 32 (7.7%) in the placebo group had hospital-acquired *S. aureus* infections. Perform a permutation test to see if the difference in proportions could be due to chance variability.
- 3.18** In the Iowa recidivism case study in Section 1.4, offenders had originally been convicted of either a felony or a misdemeanor.
- Use R to create a table displaying the proportion of felons who recidivated and the proportion of those convicted of a misdemeanor who recidivated.
  - Determine whether or not the difference in recidivism proportions computed in (a) is statistically discernible.
- 3.19** According to the Centers for Disease Control and Prevention (CDC), the 10th percentile of birth weights for baby girls is 2747 g. Referring to the data set Girls2004 (see the case study in Section 1.2),

what proportion of the babies born in Alaska are under this weight? In Wyoming? Conduct a permutation test to see if this difference in proportions could be explained by chance variability.

- 3.20** Does chocolate ice cream have more calories than vanilla ice cream? The data set `IceCream` contains calorie information for a sample of brands of chocolate and vanilla ice cream.
- Inspect the data set, then explain why this is an example of matched pairs data.
  - Compute summary statistics of the number of calories for the two flavors.
  - Conduct a permutation test to determine whether or not chocolate ice cream has more calories on average than vanilla ice cream.
- 3.21** Is there a difference in the mean price of groceries sold by Target and Walmart? The data set `Groceries` contains a sample of grocery items and their prices advertised on their respective websites on one specific day.
- Inspect the data set, then explain why this is an example of matched pairs data.
  - Compute summary statistics of the prices for each store.
  - Conduct a permutation test to determine whether or not there is a statistically discernible difference in the mean prices.
  - Create a histogram of the difference in prices. What is unusual about Quaker Oats Life cereal?
  - Redo the hypothesis test without this observation. Do you reach the same conclusion?
- 3.22** When you get fitted for glasses, the optometrist will measure the pupillary distance (PD) of each eye, the distance of the pupil to the middle of the bridge of your nose. Are people symmetric with respect to their eyes? The data set `Eyes` contains PD measurements (in mm) for a sample of volunteers. The variables `hand` and `eye` indicate which hand or eye is the dominant one.
- Inspect the data set. Are the PD measurements an example of matched pairs or independent data? Explain.
  - Compute summary statistics of the PD measures for the left eye and the right eye.
  - Conduct a permutation test to determine whether or not the mean PD measurements are the same for each eye.
- 3.23** Suppose you want to conduct an observational study to see if living in a nursing home causes high blood pressure. Draw a causal diagram and indicate a possible confounding variable.

- 3.24** Suppose researchers want to use the baby's birth weight data (case study in Section 1.2) to see if smoking during pregnancy causes a mother's baby to have a low birth weight. Draw a causal diagram and indicate at least two confounding variables.
- 3.25** A student wants to explore the link between height and athletic ability. Explain how a sample of collegiate basketball players is a collider.
- 3.26** Suppose researchers think there is an association between having a broken leg and kidney failure. Is hospitalization a confounding variable or a collider? Explain.
- 3.27** In the sampling version of permutation testing, the one-sided  $P$ -value is  $\hat{P} = (X + 1)/(N + 1)$ , where  $X$  is the number of permutation test statistics that are as large or larger than the observed test statistic. Suppose the true  $P$ -value (for the exhaustive test, conditional on the observed data) is  $p$ .
- What is the variance of  $\hat{P}$ ?
  - What is the variance of  $\hat{P}_2$  for the two-sided test (assuming that  $p$  is not close to 0.5, where  $p$  is the smaller of the two true one-sided  $P$ -values?)

**4**

## Sampling Distributions

During an election year, pollsters may want to gauge how the voters feel about a particular issue. For instance, what proportion  $p$  of registered voters in a state intend to vote “Yes” on a referendum? From one random sample of size  $n = 1000$ , the pollsters might calculate a sample proportion of  $\hat{p} = 0.47$ . However, a different sample of the same size might have yielded  $\hat{p} = 0.49$ , or maybe even 0.37. To gauge the accuracy of the original estimate, we need to understand how these proportions  $\hat{p}$  vary from sample to sample.

In Chapter 3, we saw examples of a permutation distribution: all possible values of a test statistic obtained by permuting the data among two groups. By comparing the observed test statistic to the null distribution, we could quantify how unusual the observed test statistic was. There are many other situations where we want to know something about how a statistic varies due to random sampling.

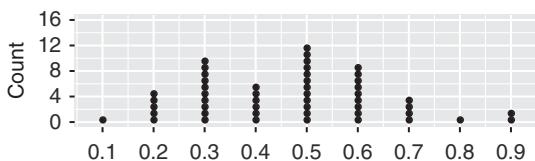
### 4.1 Sampling Distributions

Toss a fair coin  $n = 10$  times and note the proportion of heads  $\hat{p}$ . If you repeat the experiment, you probably would not get exactly the same proportion of heads. If you toss 50 sets of 10 coin flips, you might see outcomes (i.e. proportion of heads,  $\hat{p}$ ) such as those shown in Figure 4.1.

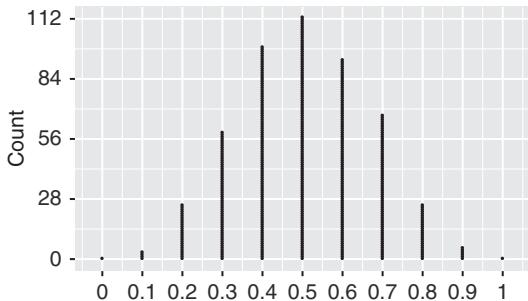
Although proportions between 0.2 and 0.7 occur most often, we see there is a proportion as low as 0.1 heads or as high as 0.9. If you do this yourself, you might be “lucky” and obtain a sample proportion as low as 0, or as high as 1!

Figure 4.2 shows the result from 500 sets of 10 tosses, done using a computer.

This distribution is (an approximation to) the *sampling distribution* of  $\hat{p}$ . The most likely outcome is 0.5, followed by 0.4 and 0.6, and so on – values farther from 0.5 are less likely. The distribution is bell-shaped and centered approximately at 0.50; the sample mean is 0.504. The standard deviation is 0.167; we call the standard deviation of a statistic a *standard error*.



**Figure 4.1** Distribution of  $\hat{p}$  after 50 sets of 10 tosses.



**Figure 4.2** Distribution of  $\hat{p}$  after 500 sets of 10 tosses.

**Definition 4.1** Let  $X$  be a random sample and let  $T = h(X)$  denote some statistic. The *sampling distribution* of  $T$  is its probability distribution. ||

(Here  $X$  may be a vector  $(X_1, \dots, X_n)$ ; it may also represent samples from multiple groups or populations.)

The permutation distributions in Chapter 3 are sampling distributions, as is the above example. The key point is that a sampling distribution is the distribution of a *statistic* that summarizes a data set and represents how the statistic varies across many random data sets. A histogram of one set of observations drawn from a population does not represent a sampling distribution. A histogram of permutation means, each from one sample, does represent a sampling distribution.

**Definition 4.2** A statistic which estimates a quantity of interest is an *estimator*. If  $X_1, X_2, \dots, X_n$  are random variables from a distribution with parameter  $\theta$  and  $g(X_1, X_2, \dots, X_n)$  an expression used to estimate  $\theta$ , then we call this function an *estimator*. ||

For example, if  $X$  is the number of heads in 50 throws of a coin, then  $\hat{p} = X/50$  is an estimator for the true proportion of head  $p$ , whereas  $X$  is a statistic that is not an estimator.

**Definition 4.3** A *standard error* can be either the standard deviation of a sampling distribution of an estimator, or (more commonly) an estimate of that standard deviation. We use the notation  $SE_T$  to denote the standard error for the sampling distribution of an estimator  $T$ , or  $SE$  for short. ||

Let us consider another example.

**Example 4.1** Suppose a population consists of four numbers,  $w_1 = 3$ ,  $w_2 = 4$ ,  $w_3 = 6$ , and  $w_4 = 6$ ; the population mean and standard deviation are  $\mu = 4.75$  and  $\sigma = 1.299$ , respectively. If we draw samples of size  $n = 2$  (with replacement), there are 16 unique samples, with the following sample means:

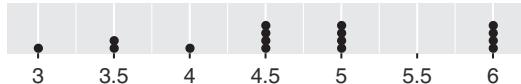
Sample	$w_1, w_1$	$w_1, w_2$	$w_1, w_3$	$w_1, w_4$	$w_2, w_1$	$w_2, w_2$	$w_2, w_3$	$w_2, w_4$
Mean	3	3.5	4.5	4.5	3.5	4	5	5
Sample	$w_3, w_1$	$w_3, w_2$	$w_3, w_3$	$w_3, w_4$	$w_4, w_1$	$w_4, w_2$	$w_4, w_3$	$w_4, w_4$
Mean	4.5	5	6	6	4.5	5	6	6

The sampling distribution for the mean of samples of size 2 (with replacement) from the given population is shown in Figure 4.3. The range is from 3 to 6, with mean 4.75 and standard deviation (standard error)  $1.299/\sqrt{2} = 0.9186$  (both agreeing with Theorem A.7).  $\square$

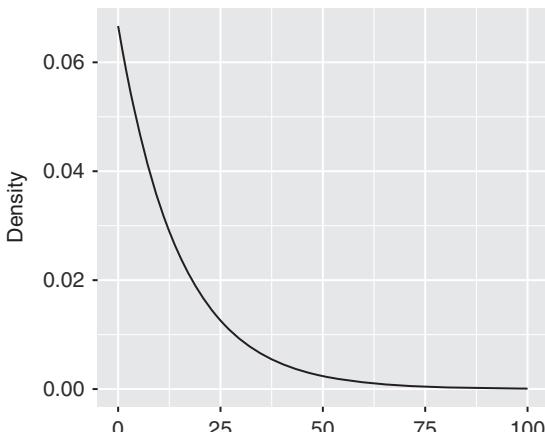
**Example 4.2** Let us simulate the sampling distribution for the mean of an exponential distribution with pdf  $f(x) = (1/15)e^{-x/15}$  ( $\lambda = 1/15$ ) (see Figure 4.4). Recall that the mean and standard deviation are both  $1/\lambda = 15$ .

We draw samples of size  $n = 100$  from this distribution. From Theorem A.7, the mean of the sampling distribution of  $\bar{x}$  is 15 and the standard error is  $15/\sqrt{100} = 1.5$ .

**Figure 4.3** Dot plot for sampling distribution of sample means.



**Figure 4.4** Density for the exponential distribution with  $\lambda = 1/15$ .



### R Note

The following commands draw 1000 random samples of size 100 from the exponential distribution with  $\lambda = 1/15$ , compute the mean of each sample, and store this mean in the vector `Xbar`.

```
Xbar <- numeric(1000)      # space for results (vector of 0's)
for (i in 1:1000)
{
  x <- rexp(100, rate = 1/15) # draw random sample of size 100
  Xbar[i] <- mean(x)         # compute mean, save in position i
}

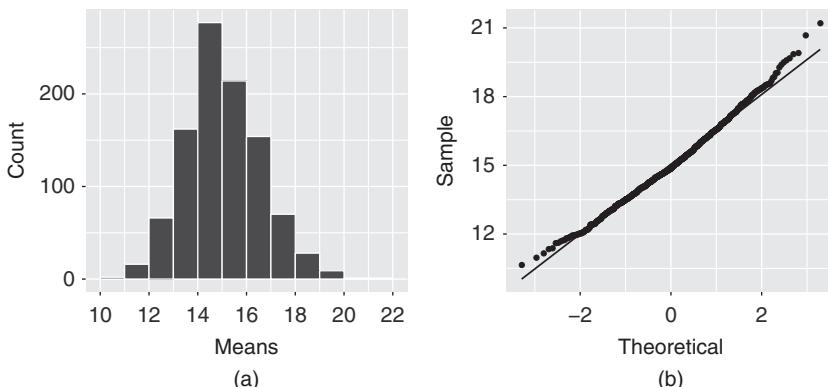
df <- data.frame(Xbar)
ggplot(df, aes(Xbar)) + geom_histogram(bins = 10)
ggplot(df, aes(sample = Xbar)) + geom_qq() + geom_qq_line()
```

We can see how close the simulation-based mean and standard deviation are to Theorem A.7:

```
> mean(Xbar)
[1] 15.0489
> sd(Xbar)
[1] 1.567628
```

In contrast to the original distribution, the sampling distribution of  $\bar{X}$  seen in Figure 4.5 is nearly bell-shaped, with the normal quantile plot indicating a hint of skewness. From Theorem A.7, the mean of the sampling distribution is 15. Does the mean obtained by our simulation approximate this reasonably well?

Also, compare the estimated standard error (the standard deviation of the sampling distribution) to the theoretical standard error,  $\sigma/\sqrt{n} = 15/10 = 1.5$ ,



**Figure 4.5** (a) Histogram and (b) quantile-normal plots for simulated sampling distribution of  $\bar{x}$  for  $n = 100$  from  $\text{Exp}(1/15)$ .

and the standard deviation of the population, 15. The standard error measures how much the sample means deviate from the population mean, when drawing samples of size 100 from the exponential distribution with  $\lambda = 1/15$ .  $\square$

Let us look at the sampling distribution of a different statistic.

**Example 4.3** We draw random samples of size 12 from the uniform distribution on the interval [0, 1] and take the maximum value of each sample. We simulate the sampling distribution of the maximum by taking 1000 samples of size 12 from the uniform distribution Unif[0, 1].

### R Note

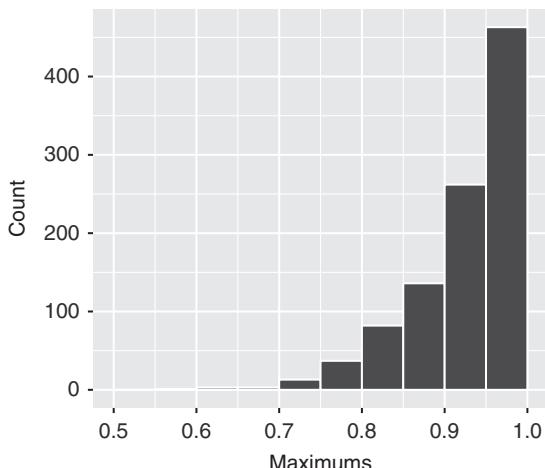
In R, the function `runif` gives random samples from Unif[0, 1].

```
maxY <- numeric(1000)
for (i in 1:1000) {
  y <- runif(12)      # draw random sample of size 12
  maxY[i] <- max(y)  # find max, save in position i
}

df <- data.frame(maxY)
ggplot(df, aes(maxY)) + geom_histogram(bins = 10)
```

The sampling distribution is shown in Figure 4.6. For samples of size 12 from the Unif[0, 1] distribution, the maximum is usually larger than 0.8. Rarely is the maximum less than 0.6.  $\square$

**Figure 4.6** Simulated sampling distribution of the maximum of a sample.



## 4.2 Calculating Sampling Distributions

There are three basic approaches for calculating sampling distributions and standard errors – exact calculations (by exhaustive calculation or formulas), simulation, and formula approximations. The mice example in Section 3.1 was small enough to calculate the exact permutation distribution by exhaustive calculation, but we approximated the Verizon (Example 3.4) and hot wings (Section 3.3) permutation distributions using simulation. Earlier in this chapter, we used simulation and exact calculation for coin tosses and sampling from a small population, respectively. In some cases, we can obtain exact answers by formulas rather than exhaustive calculation.

**Example 4.4** In Example 4.3, we simulated the sampling distribution of the maximum for a sample of size 12 from a uniform distribution. We can also obtain the distribution as follows:

First note that for  $X_i \sim \text{Unif}[0, 1]$ , if  $0 \leq a \leq 1$ , then  $P(X_i \leq a) = a$ . Therefore, for  $X_1, X_2, \dots, X_{12} \stackrel{\text{i.i.d.}}{\sim} \text{Unif}[0, 1]$ , the cdf of the maximum of  $X_1, X_2, \dots, X_{12}$  is

$$\begin{aligned} F(a) &= P(\max\{X_1, X_2, \dots, X_{12}\} \leq a) \\ &= P(X_1 \leq a, X_2 \leq a, \dots, X_{12} \leq a) \\ &= P(X_1 \leq a)P(X_2 \leq a) \dots P(X_{12} \leq a) \quad \text{by independence} \\ &= a^{12} \quad \text{for } 0 \leq a \leq 1 \end{aligned}$$

with  $F(a) = 0$  for  $a \leq 0$  and  $F(a) = 1$  for  $1 \leq a$ . Thus, the pdf of the sampling distribution of the maximum of  $X_1, X_2, \dots, X_{12}$  is  $f(a) = F'(a) = 12a^{11}$  for  $0 \leq a < 1$ , undefined at  $a = 1$ ; otherwise, 0.  $\square$

The maximum or minimum of a set of random variables comes up frequently, so we state this result more formally.

**Theorem 4.1** Suppose that continuous random variables  $X_1, X_2, \dots, X_n$  are i.i.d. with pdf  $f$  and cdf  $F$ . Define their minimum and maximum to be the random variables

$$X_{\min} = \min\{X_1, X_2, \dots, X_n\},$$

$$X_{\max} = \max\{X_1, X_2, \dots, X_n\}.$$

Then, the pdfs for  $X_{\min}$  and  $X_{\max}$  are

$$f_{\min}(x) = n(1 - F(x))^{n-1}f(x), \tag{4.1}$$

$$f_{\max}(x) = nF^{n-1}(x)f(x). \tag{4.2}$$

*Proof.* Exercise.  $\square$

In particular,

**Corollary 4.1** Let  $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Unif}[0, \beta]$ . The pdfs for the minimum and maximum of  $X_1, X_2, \dots, X_n$  are

$$f_{\min}(x) = n \left(1 - \frac{x}{\beta}\right)^{n-1} \frac{1}{\beta}, \quad (4.3)$$

$$f_{\max}(x) = \frac{n}{\beta^n} x^{n-1}. \quad (4.4)$$

*Proof.* The cdf for random variables from  $\text{Unif}[0, \beta]$  is  $F(x) = x/\beta$ .  $\square$

**Remark** Write the random variables  $X_1, X_2, \dots, X_n$  in increasing order, say,  $X_{(1)} < X_{(2)} < \dots < X_{(i)} < \dots < X_{(n-1)} < X_{(n)}$ .  $X_{(i)}$  is called the *i*th *order statistic*. The “extremes”  $X_{(1)} = X_{\min}$  and  $X_{(n)} = X_{\max}$  are just special cases. See Exercise 32.  $\parallel$

### Example 4.5

Let  $X_1, X_2, \dots, X_{10}$  be a random sample from a distribution with  $\text{pdf } f(x) = 2/x^3$ ,  $x \geq 1$ . Let  $X_{\min}$  denote the minimum of this sample. Find the probability that  $X_{\min}$  is less than or equal to 1.2.

### Solution

First, we compute the cdf  $F$  corresponding to  $f$ :

$$F(x) = \int_1^x \frac{2}{t^3} dt = 1 - \frac{1}{x^2}.$$

By Theorem 4.1, we have

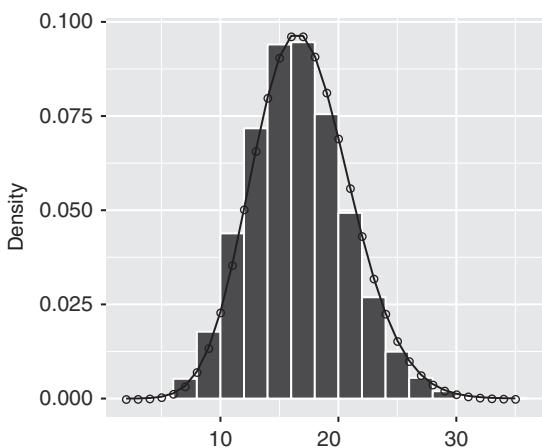
$$f_{\min}(x) = 10 \left(1 - \left(1 - \frac{1}{x^2}\right)\right)^9 \frac{2}{x^3} = \frac{20}{x^{21}},$$

for  $x \geq 1$ . Thus,

$$P(X_{\min} \leq 1.2) = \int_1^{1.2} \frac{20}{x^{21}} dx = 1 - \frac{1}{1.2^{20}} = 0.974. \quad \square$$

**Example 4.6** Let  $X_1, X_2, \dots, X_n$  be independent Poisson random variables with parameters  $\lambda_1, \lambda_2, \dots, \lambda_n$ , respectively. Then by Theorem B.6, the sampling distribution of  $X = X_1 + X_2 + \dots + X_n$  is  $\text{Pois}(\sum_{i=1}^n \lambda_i)$ .

For example, suppose  $X \sim \text{Pois}(5)$ ,  $Y \sim \text{Pois}(12)$ , with  $X$  and  $Y$  drawn independently. Then  $X + Y \sim \text{Pois}(17)$ . A simulation for this is given below (Figure 4.7).



**Figure 4.7** Simulated sampling distribution of the sum of two Poissons.

### R Note

```
X <- rpois(10^4, 5) # Draw 10^4 values from Pois(5)
Y <- rpois(10^4, 12) # Draw 10^4 values from Pois(12)
W <- X + Y

df1 <- data.frame(W)
df2 <- data.frame(x = 2:35, y = dpois(2:35,17))
ggplot(df1, aes(W)) +
  geom_histogram(aes(y=stat(density)), color = "white",
                 breaks=seq(2, 36, by = 2)) +
  geom_line(data = df2, aes(x = x, y = y)) +
  geom_point(data = df2, aes(x = x, y = y), pch = 1) + xlab("")

mean(W)      #compare to theoretical, lambda = 17
var(W)
```

□

For another example of a sampling distribution, let  $X_1, X_2, \dots, X_n$  be a random sample from a normal distribution  $N(\mu, \sigma^2)$ , and let  $\bar{X}$  denote the sample mean. The sampling distribution of  $\bar{X}$  is  $N(\mu, \sigma^2/n)$  by Corollary A.2.

In other cases, we estimate sampling distributions by approximations; we will see examples of this with the chi-square approximation for chi-square statistics for contingency tables in Section 10.1 and goodness-of-fit in Sections 10.5.1 and 10.5.2. But the most common approximations are based on normal distributions, for which the central limit theorem (CLT) plays a central role.

## 4.3 The Central Limit Theorem

The sampling distributions were approximately normally distributed in a number of previous examples. That is not a fluke. A wide variety of statistics have approximately normal sampling distributions, if sample sizes are large enough and some other conditions are met. Here, we look at the most common statistic, the mean.

We already know that if populations are normal, then the sampling distribution of  $\bar{X}$  is normal, by Corollary A.2—if  $X_1, X_2, \dots, X_n$  are a sample of independent observations from a normal distribution with mean  $\mu$  and standard deviation  $\sigma$ , the sampling distribution of  $\bar{X}$  is also normal with mean  $\mu$  and standard deviation  $\sigma/\sqrt{n}$ . We also observe that for exponential populations, the sampling distribution is approximately normal in Example 4.2. This is true for many other distributions, by the CLT:

**Theorem 4.2 The Central Limit Theorem** Let  $X_1, X_2, \dots, X_n$  be independent, identically distributed random variables with mean  $\mu$  and variance  $\sigma^2$ , both finite. Then for any constant  $z$ ,

$$\lim_{n \rightarrow \infty} P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z\right) = \Phi(z),$$

where  $\Phi$  is the cdf of the standard normal distribution (Equation (A.6)).

See Casella and Berger (2001), Ghahramani (2004), or Ross (2009) for a proof.

The CLT means that for  $n$  “sufficiently large,” the sampling distribution of  $\bar{X}$  is approximately normal with mean  $\mu$  and standard error  $\sigma/\sqrt{n}$ , regardless of the distribution from which the sample was drawn. Thus, the standardized random variable

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} - E[\bar{X}]}{SE_{\bar{X}}}$$

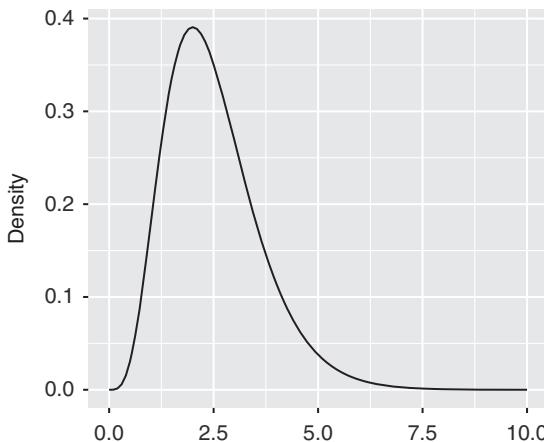
is approximately standard normal. How large is enough? It depends primarily on how skewed the population is, whether the population is continuous or discrete, and how accurate the answers should be. We will come back to this in Sections 4.3.2 and 4.3.3.

### Standard Error for a Sample Mean

The (true) standard error for a mean of  $n$  i.i.d. random variables is  $\sigma/\sqrt{n}$ . In practice, we usually estimate this using  $SE_{\bar{X}} = s/\sqrt{n}$ .

**Example 4.7**

Suppose  $X_1, X_2, \dots, X_{30}$  are a random sample from the gamma distribution with parameters  $r = 5$  and  $\lambda = 2$  (Figure 4.8). Use the CLT to estimate the probability  $P(\bar{X} > 3)$ .



**Figure 4.8** Density for the gamma distribution  $r = 5$  and  $\lambda = 2$ .

**Solution**

From Theorem B.10, we have  $E[X_i] = 5/2$  and  $SD[X_i] = \sqrt{5/2^2}$ ,  $i = 1, 2, \dots, 30$ . The sampling distribution of  $\bar{X}$  is approximately normal with mean  $E[\bar{X}] = 5/2$  and standard error  $SE_{\bar{X}} = \sqrt{5/2^2}/\sqrt{30} = 0.204$ . Hence,

$$\begin{aligned} P(\bar{X} > 3) &= P\left(\frac{\bar{X} - 5/2}{\sqrt{(5/2^2)/\sqrt{30}}} > \frac{3 - 5/2}{\sqrt{(5/2^2)/\sqrt{30}}}\right) \\ &\approx P(Z > 2.4495) \\ &= 0.0072. \end{aligned}$$

We can also simulate the sampling distribution in R.

**R Note**

```
Xbar <- numeric(1000)
for (i in 1:1000)
{
  x <- rgamma(30, shape = 5, rate = 2)
  Xbar[i] <- mean(x)
}

df <- data.frame(Xbar)
ggplot(df, aes(x = Xbar)) +
```

```
geom_histogram(aes(y = stat(density)), color = "white", bins = 10) +
  stat_function(fun = dnorm, args = list(mean = 5/2, s = 0.204)) +
  labs(x = "Means", y = "Density")
ggplot(df, aes(sample = Xbar)) + geom_qq() + geom_qq_line()
```

The approximate mean, standard error, and an empirical check of the probability that the sample mean is greater than 3 are given by:

```
> mean(Xbar)      # output will vary
[1] 2.484278
> sd(Xbar)
[1] 0.2030441
> mean(Xbar > 3) # empirical check of P(mean > 3)
[1] 0.007
```

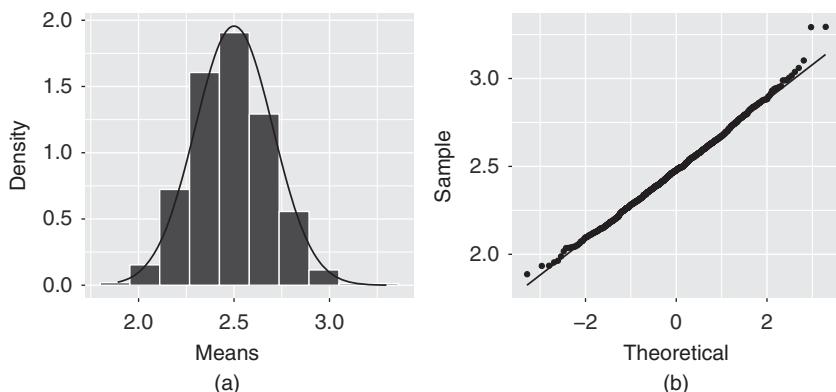
The sampling distribution is shown in Figure 4.9. □

### Example 4.8

A friend has a set of 17 numbers that he claims were randomly drawn from a gamma distribution with parameters  $r = 100$  and  $\lambda = 5$ . You compute the sample mean to be 21.9. Should you be suspicious of his claim?

### Solution

For  $X \sim \text{Gamma}(100, 5)$  the mean and variance are  $\mu = 100/5 = 20$  and  $\sigma^2 = 100/5^2 = 4$ . Using the CLT, we compare  $z = (21.9 - 20)/(2/\sqrt{17}) = 3.917$  to a standard normal distribution. Since  $P(Z \geq 3.917) = 0.00004$ , a random sample of size 17 with mean 21.9 or larger from  $\text{Gamma}(100, 5)$  is relatively rare. Yes, you are justified in being suspicious of your friend's claim. □



**Figure 4.9** (a) Histogram and (b) quantile-normal plots for the sampling distribution of  $\bar{X}$  for  $\text{Gamma}(5, 2)$  with normal density.

### 4.3.1 CLT for Binomial Data

An important special case is the use of the CLT for binomial data.

#### Example 4.9

Toss a fair coin 300 times. Find the approximate probability of getting at most 160 heads.

#### Solution

Let  $X_1, X_2, \dots, X_{300}$  denote Bernoulli random variables (1 for heads, 0 for tails). Then  $E[X_i] = 1/2$  and  $\text{Var}[X_i] = 1/4$ ,  $i = 1, 2, \dots, 300$ . Now  $\bar{X}$ , which gives the proportion of heads, is approximately normal with mean 1/2 and standard error  $(1/2)/\sqrt{300} = 0.029$ .

Thus,

$$\begin{aligned} P(\bar{X} \leq 160/300) &= P\left(\frac{\bar{X} - 0.5}{0.0289} \leq \frac{0.5333 - 0.5}{0.0289}\right) \\ &= P(Z \leq 1.1522) = 0.875. \end{aligned}$$

□

We generally take a shortcut in problems like these. Note that  $X = \sum_{i=1}^n X_i$  is a binomial random variable, and  $\hat{p} = \bar{X} = X/n$ . So rather than expressing the problem as a sum of Bernoulli random variable, we work directly with the binomial variable. Furthermore, we use the following corollary to the CLT:

**Corollary 4.2** Let  $X \sim \text{Binom}(n, p)$  be a binomial random variable and  $\hat{p} = X/n$ , the proportion of successes. Then if both  $np$  and  $n(1-p)$  are large, the sampling distribution of  $\hat{p}$  is approximately normal with mean  $p$  and standard deviation  $\sqrt{p(1-p)/n}$ .

Similarly, the sampling distribution of  $X$  is approximately normal with mean  $np$  and standard deviation  $\sqrt{np(1-p)}$ .

*Proof.* A binomial variable  $X$  can be written as  $X = \sum_{i=1}^n X_i$ , where  $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Bern}(p)$  with  $E[X_i] = p$  and  $\text{Var}[X_i] = p(1-p)$ . Then  $\bar{X} = (X_1 + X_2 + \dots + X_n)/n$  is the proportion of 1's, more commonly denoted by  $\hat{p}$ . By the CLT, for  $n$  sufficiently large, the sampling distribution of  $\bar{X} = \hat{p}$  is approximately normal with mean  $p$  and standard error  $\sqrt{p(1-p)/n}$ . In practice, “sufficiently large” requires that both  $np$  and  $n(1-p)$  be large.

Furthermore, if  $\hat{p}$  is approximately normal, then  $X = n\hat{p}$  is also approximately normal, since any linear transformation of a normal variable is also normal (Theorem A.9). □

#### Standard Error for a Proportion

The (true) standard error for a proportion estimated from a sample of  $n$  i.i.d. draws is  $\sqrt{p(1-p)/n}$ , which we estimate using  $\sqrt{\hat{p}(1-\hat{p})/n}$ .

**Example 4.10**

According to the 2004 American Community Survey, 28% of adults over 25 years old in Utah have completed a bachelor's degree. In a random sample of 64, adults over 25 years old from Utah, what is the probability that at least 30% have a bachelor's degree?

**Solution**

The sampling distribution of  $\hat{p}$  is approximately normal with mean 0.28 and standard error  $\sqrt{0.28(1 - 0.28)/64} = 0.056$ . Standardizing, we find

$$P(\hat{p} \geq 0.30) = P\left(Z \geq \frac{0.30 - 0.28}{0.056}\right) = P(Z \geq 0.356) = 0.361.$$

That is, the probability that at least 30% of those in the sample have a bachelor's degree is 0.361.  $\square$

**Example 4.11**

Let  $X$  be a binomial random variable,  $X \sim \text{Binom}(120, 0.3)$ . Compute  $P(X \leq 25)$ .

**Solution**

An exact solution would require calculating  $P(X \leq 25) = \sum_{k=0}^{25} \binom{120}{k} 0.3^k 0.7^{120-k}$ . We use Corollary 4.2 instead.

$$\begin{aligned} P(X \leq 25) &= P\left(\frac{X}{120} \leq \frac{25}{120}\right) \\ &= P\left(\frac{\hat{p} - 0.3}{\sqrt{0.3(1 - 0.3)/120}} \leq \frac{0.2083 - 0.3}{\sqrt{0.3(1 - 0.3)/120}}\right) \\ &\approx P(Z \leq -2.1913) = 0.014. \end{aligned}$$

The exact probability is 0.0159, so the CLT approximation underestimates the exact probability by about 0.0017. This is a large error, over 10% of the exact answer. We need to do better, so we turn to a continuity correction.

**R Note**

To compute  $\binom{n}{k} p^k (1-p)^{n-k}$ , use the `dbinom` function.

For instance,  $\binom{120}{25} 0.3^{25} 0.7^{95}$ ,

```
> dbinom(25, 120, .3)
```

```
[1] 0.006807598
```

The `pbinom` function computes cumulative probabilities for the binomial distribution known as the cumulative mass function. For instance, to calculate  $\sum_{k=0}^{25} \binom{120}{k} 0.3^k 0.7^{120-k}$ ,

```
> pbinom(25, 120, .3)
```

```
[1] 0.01593170
```

$\square$

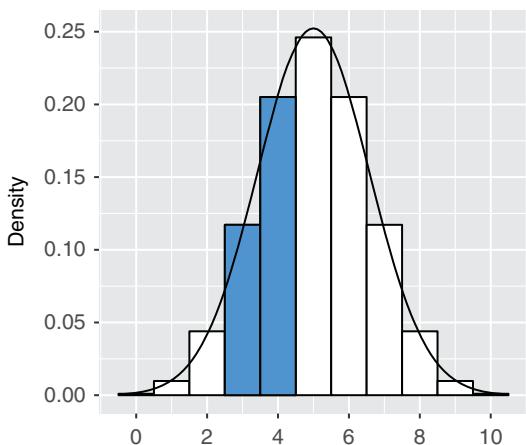
### 4.3.2 Continuity Correction for Discrete Random Variables

A binomial variable  $X$  is a discrete random variable, but the CLT approximation uses a continuous density. We can improve the approximation of the CLT for binomial and other discrete data using a *continuity correction*. This is best explained visually.

Consider the case of tossing 10 coins and estimating the probability that either 3 or 4 coins come up.  $X \sim \text{Binom}(10, 0.5)$ , and our goal is to estimate  $P(3 \leq X \leq 4)$ . Figure 4.10 shows the exact probabilities for the binomial distribution, overlaid with the CLT normal approximation density. The desired probability corresponds to the area of the two shaded bars, each with width 1.0 and heights  $f(3)$  and  $f(4)$ , respectively. To approximate the shaded region using an area under the normal curve, it would be better to use the integral  $\int_{2.5}^{4.5} g(x)dx$  rather than  $\int_3^4 g(x)dx$ , where  $g$  is the corresponding normal density.

To do the corresponding calculations, we express each inequality in two different ways and split the difference, ultimately, adding or subtracting 0.5 to each end. In this case, there are two inequalities inside the probability,  $3 \leq X$  and  $X \leq 4$ ; we handle these separately. Remember that  $X$  is discrete, so  $P(3 \leq X) = P(2 < X) = P(2.5 < X)$ ; similarly,  $P(X \leq 4) = P(X < 5) = P(X < 4.5)$ . In each case, we split the difference between the two ways to express the boundaries using whole numbers. This yields

$$\begin{aligned} P(3 \leq X \leq 4) &= P(2.5 < X < 4.5) \\ &\approx P\left(\frac{2.5 - 5}{1.58} \leq Z \leq \frac{4.5 - 5}{1.58}\right) = 0.319. \end{aligned}$$



**Figure 4.10** Binomial distribution with  $n = 10$  and  $p = 0.5$ , with CLT approximation, and  $P(X = 3 \text{ or } 4)$  highlighted.

This is not far from the exact answer of 0.322. In contrast, without the continuity correction, the estimate would be 0.161, only about half as large as desired.

In the American Community Survey in Example 4.10, 30% of a sample of size 64 is 19.2 people, so the example is really asking for the probability of at least 20 people having a bachelor's degree.

$$\begin{aligned} P(X \geq 19.2) &= P(X \geq 20) \\ &= P(X > 19.5) \\ &\approx P\left(Z > \frac{19.5 - (64)(0.28)}{\sqrt{64(0.28)(0.72)}}\right) \\ &= P(Z > 0.440) = 0.330. \end{aligned}$$

The probability that at least 20 people have a bachelor's degree is about 33%; in contrast, the estimate without the correction is 36%, and the exact answer is 32.4%.

In Example 4.11,

$$\begin{aligned} P(X \leq 25) &= P(X < 25.5) \\ &\approx P\left(Z < \frac{25.5 - 36}{\sqrt{120(0.3)(0.7)}}\right) \\ &= P(Z \leq -2.092) = 0.018. \end{aligned}$$

This is even further off than without the continuity correction! It turns out that the CLT approximation is not very accurate for skewed distributions unless sample sizes are much larger, especially in the tails of distributions. In this example with a  $z$ -score of  $-2.092$ , we are fairly far in the tail. We will discuss the accuracy of the CLT in more detail later. In this case, we fixed a small error in a direction that exacerbated a larger error. Applying the same method elsewhere in the distribution would generally not be so unlucky.

In short, use the CLT approximation of the binomial distribution for a quick and dirty estimate, but in situations where accuracy is important, apply the continuity correction, or use exact binomial calculations.

### Example 4.12

According to the Centers for Disease Control and Prevention, in 2019, about 23.6% of adults in Kentucky are everyday smokers. In a random sample of 700 adults in Kentucky, what is the probability that between 150 and 170 of them are everyday smokers?

**Solution**

Let  $X$  denote the number of everyday smokers in the sample. Then  $X \sim \text{Binom}(700, 0.236)$ , with expected value  $np = 700 \times 0.236 = 165.2$  and standard deviation  $\sqrt{700(0.236)(1 - 0.236)} = 11.234$ . We find

$$\begin{aligned} P(150 \leq X \leq 170) &= P(149.5 < X < 170.5) \\ &= P\left(\frac{149.5 - 165.2}{11.234} \leq Z \leq \frac{170.5 - 165.2}{11.234}\right) \\ &\approx \Phi(0.4273) - \Phi(-1.3975) = 0.584. \quad \square \end{aligned}$$

**4.3.3 Accuracy of the Central Limit Theorem\***

The usual rule of thumb, found in most textbooks, is that the CLT is reasonably accurate if  $n \geq 30$ , unless the data are quite skewed. For binomial data, the common rule of thumb is to use the CLT, with continuity correction, if both  $np \geq 10$  and  $n(1 - p) \geq 10$ .

These rules are wishful thinking, dating to a precomputer age when one had few realistic alternatives to using the CLT because most other methods were computationally infeasible. We can obtain better approximations using simulation-based methods, including permutation tests (Chapter 3) and bootstrapping (Chapter 5); in addition, these methods provide a way to check the accuracy of the CLT, based on a set of data. And for binomial distributions, we can do exact calculations.

The CLT is exact if the population is normal (Corollary A.2). For nonnormal populations, the biggest problem is skewness, followed by discreteness. If the population is symmetric, the sampling distribution of the mean may be very close to normal for quite small  $n$ . If the population is not symmetric, then the sampling distribution may be nonnormal even for large  $n$ .

There is an expanded version of the CLT approximation for continuous data (obtained using a particular Taylor series, known as an Edgeworth approximation):

$$P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z\right) \approx \Phi(z) - \frac{\kappa_3}{6\sqrt{n}}(z^2 - 1)\Phi'(z), \quad (4.5)$$

where  $\Phi(z)$  is the standard normal cdf,  $\Phi'(z)$  is the standard normal density, and  $\kappa_3 = E[(X - \mu)^3]/\sigma^3$  is the population skewness. This can be used in its own right, or the term

$$-\frac{\kappa_3}{6\sqrt{n}}(z^2 - 1)\Phi'(z) \quad (4.6)$$

can be used to estimate the error of the CLT approximation. We can use this approximation to estimate necessary sample sizes for prescribed accuracy for two examples considered above, the exponential and binomial.

We will focus on the error of the CLT approximation at  $z = 2.33$  since the probability  $P(X > x) \approx 0.01$  is important in statistical practice

$(x = \mu + 2.33\sigma/\sqrt{n})$ . What size  $n$  ensures an error of at most 10%, that is, that the true probability be between 0.009 and 0.011? For exponential distributions,  $\kappa_3 = 2$ . Setting  $0.001 = \frac{2}{6\sqrt{n}}(1 - z^2)\Phi'(z)$  and solving for  $n$  in Equation (4.6) results in  $n \geq 1536$ , a far cry from the rule  $n \geq 30$ ! The approximation Equation (4.5) is not far off; exact calculations indicate that  $n \geq 1541$  is required.

In the binomial case,  $\kappa_3 = (1 - 2p)/\sqrt{p(1 - p)}$ , the skewness of a Bernoulli variable. The distribution is symmetric when  $p = 0.5$ , but is left-skewed if  $p > 0.5$  and right-skewed if  $p < 0.5$ . For comparison, the skewness of a Bernoulli when  $p = 0.146$  is the same as the skewness of an exponential variable; we saw above that we need  $n \geq 1536$  for CLT accuracy there, which implies  $np \geq 224$ , a far cry from the rule  $np \geq 10$ !

Even that is not enough when  $p$  is closer to zero. For  $p$  near zero,  $\kappa_3 \approx 1/\sqrt{p}$ . Let  $\alpha$  be the normal tail probability for  $z$ , e.g.  $\alpha = 0.01$  for  $z = 2.33$ . We set  $\alpha/10 = |1/(6\sqrt{np})(z^2 - 1)\Phi'(z)|$  to obtain the rule  $np \geq ((z^2 - 1)\Phi'(z)/(6\alpha/10))^2$  which requires  $np \geq 385$  when  $\alpha = 0.01$ . Similarly, when  $1 - p$  is small, we need  $n(1 - p) \geq 385$ . Hence, in order for a rule to hold across all values of  $p$ , 10 expected successes and failures are not nearly enough, we need  $np \geq 385$  and  $n(1 - p) \geq 385$ , when  $\alpha = 0.01$ . For  $\alpha = 0.25$ ,  $np \geq 123$  and  $n(1 - p) \geq 123$  would suffice. We get the same answers for the Poisson distribution, for which the skewness is  $1/\sqrt{\lambda}$ ; we need  $\lambda > 385$  for the Edgeworth correction to be smaller than  $\alpha/10$  when  $\alpha = 0.01$ .

#### 4.3.4 CLT for Sampling Without Replacement

When working with finite populations, one typically samples without replacement. Thus, a random sample  $X_1, X_2, \dots, X_n$  is not independent, and we cannot invoke the usual CLT to say that  $\bar{X}$  is asymptotically normally distributed. In fact,  $n \rightarrow \infty$  is impossible, because the sample size  $n$  is bounded above by the size  $N$  of the population. However, in 1960, Jaroslav Hajek proved a version of the CLT for sampling without replacement. The accuracy depends on both  $n$  and the number of nonsampled observations  $(N - n)$ . The SE for the mean of  $n$  i.i.d. observations chosen with replacement from  $N$  is  $(s/\sqrt{n})\sqrt{(N - n)/(N - 1)}$ ; the *finite population correction factor*  $\sqrt{(N - n)/(N - 1)}$  is small when the sample size  $n$  is large relative to the population size  $N$ . We may skip this if it is close to 1 (see Exercise 35). Courses on survey sampling discuss issues related to finite populations in more detail.

## Exercises

- 4.1 Consider the population  $\{1, 2, 5, 6, 10, 12\}$ . Find (and plot) the sampling distribution of medians for samples of size 3 without replacement. Compare the median of the population to the mean of the medians.

- 4.2** Consider the population  $\{3, 6, 7, 9, 11, 14\}$ . For samples of size 3 without replacement, find (and plot) the sampling distribution of the minimum. What is the mean of the sampling distribution? The statistic is an estimate of some parameter – what is the value of that parameter?
- 4.3** Let  $A$  denote the population  $\{1, 3, 4, 5\}$  and  $B$  the population  $\{5, 7, 9\}$ . Let  $X$  be a random value from  $A$ , and  $Y$  a random value from  $B$ .
- Find the sampling distribution of  $X + Y$ .
  - In this example, does the sampling distribution depend on whether you sample with or without replacement? Why or why not?
  - Compute the mean of the values for each of  $A$  and  $B$ . Compute the mean of the values in the sampling distribution of  $X + Y$ . How are the means related?
  - Suppose you draw a random value from  $A$  and a random value from  $B$ . What is the probability that the sum is 13 or larger?
- 4.4** Consider the population  $\{3, 5, 6, 6, 8, 11, 13, 15, 19, 20\}$ .
- Compute the mean and standard deviation and create a dot plot of its distribution.
  - Simulate the sampling distribution of  $\bar{X}$  by taking random samples of size 4 and plot your results. Compute the mean and standard error and compare to the population mean and standard deviation.
  - Use the simulation to find  $P(\bar{X} < 11)$ .

```

pop <- c(3, 5, 6, 6, 8, 11, 13, 15, 19, 20)
N <- 10^4
Xbar <- numeric(N)
for (i in 1:N)
{
  samp <- sample(pop, 4, replace = TRUE)
  Xbar[i] <- mean(samp)
}

df <- data.frame(Xbar)
ggplot(df, aes(Xbar)) + geom_histogram(bins = 10)
mean(Xbar < 11)

```

- 4.5** Consider two populations  $A = \{3, 5, 7, 9, 10, 16\}$  and  $B = \{8, 10, 11, 15, 18, 25, 28\}$ .
- Using R, draw random samples (without replacement) of size three from each population and simulate the sampling distribution of the sum of their maximum. Describe the distribution.
  - Use your simulation to estimate the probability that the sum of the maximums is less than 20.

- (c) Draw random samples of size 3 from each population and find the maximum of the union of these two sets. Simulate the sampling distribution of the maximum of this union. Compare the distribution to part (a). In R , `max(union(a, b))` returns the maximum of the union of sets a and b.
- (d) Use simulation to find the probability that the maximum of the union is less than 20.
- 4.6** The data set `Recidivism` contains the population of all Iowa offenders convicted of either a felony or a misdemeanor who were released in 2010 (case study in Section 1.4.) Of these, 31.6% recidivated and were sent back to prison. Simulate the sampling distribution of  $\hat{p}$ , the sample proportion of offenders who recidivated, for random samples of size 25.

```
N <- 10^4
phat <- numeric(N)
for (i in 1:N)
{
  samp <- sample(Recidivism$Recid, 25)
  phat[i] <- mean(samp == "Yes") # proportion yes
}
```

- (a) Create a histogram and describe the simulated sampling distribution of  $\hat{p}$ . Estimate the mean and standard error.
- (b) Compare your estimate of the standard error with the theoretical standard error (Corollary 4.2).
- (c) Repeat the above using samples of size 250 and compare to the  $n = 25$  case.
- 4.7** The data set `FlightDelays` contains the population of all flights departures by United Airlines and American Airlines out of LaGuardia Airport (LGA) during May and June 2009 (case study in Section 1.1.)
- (a) Create a histogram of `Delay` and describe the distribution. Compute the mean and standard deviation.
- (b) Simulate the sampling distribution of  $\bar{x}$  and the sample mean of the length of the flight delays (`Delay`) for samples of size 25. Create a histogram and describe the simulated sampling distribution of  $\bar{x}$ . Estimate the mean and standard error.
- (c) Compare your estimate of the standard error with the theoretical standard error (Corollary A.1).
- (d) Repeat the above using samples of size 250 and compare to the  $n = 25$  scenario.

- 4.8** Let  $X_1, X_2, \dots, X_{25}$  be a random sample from some distribution and  $W = T(X_1, X_2, \dots, X_{25})$  be a statistic. Suppose the *sampling distribution* of  $W$  has a pdf given by  $f(x) = 2/x^2$ ,  $1 < x < 2$ . Find the probability that  $W < 1.5$ .
- 4.9** Let  $X_1, X_2, \dots, X_n$  be a random sample from some distribution and suppose  $Y = T(X_1, X_2, \dots, X_n)$  is a statistic. Suppose the sampling distribution of  $Y$  has pdf  $f(y) = (3/8)y^2$  for  $0 \leq y \leq 2$ . Find  $P(0 \leq Y \leq 1/5)$ .
- 4.10** Suppose the heights of boys in a certain large city follows a distribution with mean 48 in and variance  $9^2$ . Use the CLT approximation to estimate the probability that in a random sample of 30 boys, the mean height is more than 51 in.
- 4.11** Let  $X_1, X_2, \dots, X_{36} \sim \text{Bern}(0.55)$  be independent and let  $\hat{p}$  denote the sample proportion. Use the CLT approximation with continuity correction to find the probability that  $\hat{p} \leq 0.50$ .
- 4.12** A random sample of size  $n = 20$  is drawn from a distribution with mean 6 and variance 10. Use the CLT approximation to estimate  $P(\bar{X} \leq 4.6)$ .
- 4.13** A random sample of size  $n = 244$  is drawn from a distribution with pdf  $f(x) = (3/16)(x - 4)^2$ ,  $2 \leq x \leq 6$ . Use the CLT approximation to estimate  $P(\bar{X} \geq 4.2)$ .
- 4.14** According to the 2000 census, 28.6% of the US adult population received a high school diploma. In a random sample of 800 US adults, what is the probability that between 220 and 230 (inclusive) people have a high school diploma? Use the CLT approximation with continuity correction and compare to the exact probability (use `pbinom` in R).
- 4.15** If  $X_1, \dots, X_n$  are i.i.d. from  $\text{Unif}[0, 1]$ , how large should  $n$  be so that  $P(|\bar{X} - 1/2| < 0.05) \geq 0.90$ , that is, there is at least a 90% chance that the sample mean is within 0.05 of 1/2? Use the CLT approximation.
- 4.16** Maria claims that she has drawn a random sample of size 30 from the exponential distribution with  $\lambda = 1/10$ . The mean of her sample is 12.
- What is the expected value of a sample mean?
  - Run a simulation by drawing 1000 random samples, each of size 30, from  $\text{Exp}(1/10)$ , and compute the mean for each sample. What proportion of the sample means are as large or larger than 12?
  - Is a mean of 12 unusual for a sample of size 30 from  $\text{Exp}(1/10)$ ?

- 4.17** Let  $X \sim N(15, 3^2)$  and  $Y \sim N(4, 2^2)$  be independent random variables.
- What is the exact sampling distribution of  $W = X - 2Y$ ?
  - Use R to simulate the sampling distribution of  $W$  and plot your results. Check that the simulated mean and standard error are close to the theoretical mean and standard error.
  - Use the simulated sampling to estimate  $P(W \leq 10)$  and then check your estimate with an exact calculation.
- 4.18** Let  $X \sim \text{Pois}(4)$ ,  $Y \sim \text{Pois}(12)$ ,  $U \sim \text{Pois}(3)$  be independent random variables.
- What is the exact sampling distribution of  $W = X + Y + U$ ?
  - Use R to simulate the sampling distribution of  $W$  and plot your results. Check that the simulated mean and standard error are close to the theoretical mean and standard error.
  - Use the simulated sampling distribution to estimate  $P(W \leq 14)$  and then check your estimate with an exact calculation.

- 4.19** Let  $\underline{X}_1, \underline{X}_2, \dots, \underline{X}_{10} \stackrel{\text{i.i.d.}}{\sim} N(20, 8^2)$  and  $\underline{Y}_1, \underline{Y}_2, \dots, \underline{Y}_{15} \stackrel{\text{i.i.d.}}{\sim} N(16, 7^2)$ . Let  $W = \bar{X} + \bar{Y}$ .
- Give the exact sampling distribution of  $W$ .
  - Simulate the sampling distribution in R and plot your results. Check that the simulated mean and standard error are close to the exact mean and standard error.
  - Use your simulation to find  $P(W < 40)$ . Calculate an exact answer and compare.

```

W <- numeric(1000)
for (i in 1:1000)
{
  x <- rnorm(10, 20, 8)      # draw 10 from N(20, 8^2)
  y <- rnorm(15, 16, 7)      # draw 15 from N(16, 7^2)
  W[i] <- mean(x) + mean(y) # save sum of means
}
df <- data.frame(W)
ggplot(df, aes(W)) + geom_histogram(bins = 10)
mean(W < 40)

```

- 4.20** Let  $\underline{X}_1, \underline{X}_2, \dots, \underline{X}_9 \stackrel{\text{i.i.d.}}{\sim} N(7, 3^2)$  and  $\underline{Y}_1, \underline{Y}_2, \dots, \underline{Y}_{12} \stackrel{\text{i.i.d.}}{\sim} N(10, 5^2)$ . Let  $W = \bar{X} - \bar{Y}$ .
- Give the exact sampling distribution of  $W$ .
  - Simulate the sampling distribution of  $W$  in R and plot your results (adapt code from the previous exercise). Check that the simulated

mean and standard error are close to the theoretical mean and standard error.

- (c) Use your simulation to find  $P(W < -1.5)$ . Calculate an exact answer and compare.

**4.21** Let  $X_1, X_2, \dots, X_n$  be a random sample from  $N(0, 1)$ . Let  $W = X_1^2 + X_2^2 + \dots + X_n^2$ . Describe the sampling distribution of  $W$  by running a simulation, using  $n = 2$ . What is the mean and variance of the sampling distribution of  $W$ ? Repeat using  $n = 4, n = 5$ . What observations or conjectures do you have for general  $n$ ?

**4.22** Let  $X$  be a uniform random variable on the interval  $[40, 60]$  and  $Y$  a uniform random variable on  $[45, 80]$ . Assume that  $X$  and  $Y$  are independent.

- (a) Compute the expected value and variance of  $X + Y$  (see Theorem B.7).  
 (b) The following code simulates the sampling distribution of  $X + Y$ :

```
X <- runif(1000, 40, 60) # 1000 values from Unif[40,60]
Y <- runif(1000, 45, 80) # 1000 values from Unif[45,80]
total <- X + Y           # Add them coordinate-wise
df <- data.frame(total)
ggplot(df, aes(total)) +
  geom_histogram(bins = 10) # Distribution of the sums
```

Describe the distribution of  $X + Y$ . Compute the mean and variance of the sampling distribution (i.e. of `total`) and compare this to the theoretical mean and variance.

- (c) Suppose the time (in minutes) Andy takes to complete his statistics homework is  $\text{Unif}[40, 60]$  and the time Adam takes is  $\text{Unif}[45, 80]$ . Assume their times are independent. One day they announce that their total time to finish an assignment was less than 90 min. How likely is this? (Use the simulated sampling distribution in part (b)).

**4.23** Let  $X_1, X_2, \dots, X_{20} \stackrel{\text{i.i.d.}}{\sim} \text{Exp}(2)$ . Let  $X = \sum_{i=1}^{20} X_i$ .

- (a) Simulate the sampling distribution of  $X$  in R.  
 (b) From your simulation, find  $E[X]$  and  $\text{Var}[X]$ .  
 (c) From your simulation, find  $P(X \leq 10)$ .

**4.24** Let  $X_1, X_2, \dots, X_{30} \stackrel{\text{i.i.d.}}{\sim} \text{Exp}(1/3)$  and let  $\bar{X}$  denote the sample mean.

- (a) Simulate the sampling distribution of  $\bar{X}$  in R.  
 (b) Find the mean and standard error of the sampling distribution and compare to the theoretical results.

- (c) From your simulation, find  $P(\bar{X} \leq 3.5)$ .  
 (d) Estimate  $P(\bar{X} \leq 3.5)$  by assuming the CLT approximation holds. Compare this result with the one in part (c).
- 4.25** Consider the exponential distribution with density  $f(x) = (1/20)e^{-x/20}$ , with mean and standard deviation 20.
- Calculate the median of this distribution.
  - Using R, draw a random sample of size 50 and graph the histogram. Describe the distribution of your *sample*. What are the mean and the standard deviation of your sample?
  - Run a simulation to find the (approximate) sampling distribution for the median of samples of size 50 from the exponential distribution and describe it. What is the mean and the standard error of this sampling distribution?
  - Repeat the above but use sample sizes  $n = 100, 500$ , and 1000. How does sample size affect the sampling distribution?
- 4.26** Prove Theorem 4.1.
- 4.27** Let  $X_1, X_2 \stackrel{\text{i.i.d.}}{\sim} F$  with corresponding pdf  $f(x) = 2/x^2, 1 \leq x \leq 2$ .
- Find the pdf of  $X_{\max}$ .
  - Find the expected value of  $X_{\max}$ .
- 4.28** Let  $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} F$  with corresponding pdf  $f(x) = 3x^2, 0 \leq x \leq 1$ .
- Find the pdf for  $X_{\min}$ .
  - Find the pdf for  $X_{\max}$ .
  - If  $n = 10$ , find the probability that the largest value,  $X_{\max}$ , is greater than 0.92.
- 4.29** Compute the pdf of the sampling distribution of the maximum of samples of size 10 from a population with an exponential distribution with  $\lambda = 12$ .
- 4.30** Let  $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Exp}(\lambda)$  with pdf  $f(x) = \lambda e^{-\lambda x}, \lambda > 0, x > 0$ .
- Find the pdf  $f_{\min}(x)$  for the sample minimum  $X_{\min}$ . Recognize this as the pdf of a known distribution.
  - Simulate in R the sampling distribution of  $X_{\min}$  of samples of size  $n = 25$  from the exponential distribution with  $\lambda = 7$ . Compare the theoretical expected value of  $X_{\min}$  to the simulated expected value.
- 4.31** Let  $X_1, X_2, \dots, X_{12}$  be a random sample from a distribution with pdf  $f(x)$ . Let  $m$  denote the median of this distribution.

- (a) Find the probability that  $X_{\max}$  is less than  $m$ .  
 (b) Suppose  $f(x) = (1/6)e^{-1/6x}$ , the exponential distribution. Compute the median and then run a simulation to check your answer in (a).

- 4.32** Let  $X_1, X_2, \dots, X_n$  be a random sample from a distribution with pdf  $f(x)$  and cdf  $F(x)$ .

- (a) Show that the number of observations less than or equal to  $t$  is binomial,  $\text{Binom}(n, F(t))$ .  
 (b) Let  $E_i(t)$  denote the event that exactly  $i$  of the  $X_1, X_2, \dots, X_n$  are less than or equal to  $t$ . Show that

$$P(E_i(t)) = \binom{n}{i} (F(t))^i (1 - F(t))^{n-i}.$$

- (c) Let  $X_{(i)}$  denote the  $i$ th order statistic. Show that the cdf of  $X_{(i)}$  is

$$F_{(i)}(t) = \sum_{k=i}^n \binom{n}{k} F(t)^k (1 - F(t))^{n-k}.$$

- (d) \*Show that the pdf of  $X_{(i)}$  is given by

$$f_{(i)}(t) = \frac{n!}{(i-1)!(n-i)!} F(t)^{i-1} (1 - F(t))^{n-i} f(t).$$

*Hint:* One way to view the problem is as follows: For the  $i$ th largest observation to equal  $t$  splitting the data into three groups:  $(i-1)$  observations that are  $< t$ , one right at  $t$ , and  $n-i$  that are  $> t$ . The corresponding probabilities are  $F(t)^{i-1}, f(t)^1$ , and  $(1 - F(t))^{n-i}$  (well, the middle one is a likelihood, not a probability). The number of ways to split  $n$  observations into sets of size  $(i-1, 1, n-i)$  is  $\frac{n!}{(i-1)!1!(n-i)!}$ .

- 4.33** Let  $X_1, X_2, \dots, X_{10} \stackrel{\text{i.i.d.}}{\sim} \text{Pois}(3)$ . Let  $X = \sum_{i=1}^{10} X_i$ . Find the pdf for the sampling distribution of  $X$ .

- 4.34** Let  $X_1$  and  $X_2$  be independent exponential random variables, both with parameter  $\lambda > 0$ . Find the cumulative distribution function for the sampling distribution of  $X = X_1 + X_2$ .

- 4.35** This simulation illustrates the CLT for a finite population.

```
N <- 400 # population size
n <- 5 # sample size
finpop <- rexp(N, 1/10) # Create a finite pop. of size N=400
# from Exp(1/10)
df <- data.frame(finpop)
ggplot(df, aes(finpop)) +
  geom_histogram(bins = 10) # distribution of your finite pop.
```

```

mean(finpop)           # mean (mu) of your pop.
sd(finpop)             # stdev (sigma) of your pop.

sd(finpop)/sqrt(n)   # theoretical standard error of sampling
                      # dist. of mean(x), with replacement
sd(finpop)/sqrt(n) * sqrt((N-n)/(N-1)) # without replacement

Xbar <- numeric(1000)
for (i in 1:1000)
{
  x <- sample(finpop, n) # Random sample of size n w/o replacement
  Xbar[i] <- mean(x)      # Save mean of sample
}

df <- data.frame(Xbar)
ggplot(df, aes(Xbar)) + geom_histogram(bins = 10)
ggplot(df, aes(sample = Xbar)) + geom_qq() + geom_qq_line()

mean(Xbar)
sd(Xbar)   # estimated standard error of sampling distribution

```

- (a) Does the sampling distribution of sample means appear approximately normal?
- (b) Compare the mean and standard error of your simulated sampling distribution to the theoretical ones.
- (c) Calculate  $(\sigma/\sqrt{n})\sqrt{(N-n)/(N-1)}$ , where  $\sigma$  is the standard deviation of the finite population and compare with the (estimated) standard error of the sampling distribution.
- (d) Repeat for larger  $n$ , say  $n = 20$ , and  $n = 100$ , and discuss.
- 4.36** Let  $X_1, X_2, \dots, X_n$  be independent random variables from  $N(\mu, \sigma^2)$ . We are interested in the sampling distribution of the variance. Run the following script in R, which draws random samples of size 20 from  $N(25, 7^2)$  and calculates the variance for each sample.

```

W <- numeric(1000)
for (i in 1:1000)
{
  x <- rnorm(20, 25, 7)
  W[i] <- var(x)
}

mean(W)
var(W)

df <- data.frame(W)
ggplot(df, aes(W)) + geom_histogram(bins = 10)
ggplot(df, aes(sample = W)) + geom_qq() + geom_qq_line()

```

Does the sampling distribution appear to be normally distributed?  
Repeat with  $n = 50$  and  $n = 200$ .

- 4.37** A random sample of size  $n = 100$  is drawn from a distribution with pdf  $f(x) = 3(1 - x)^2$ ,  $0 \leq x \leq 1$ .
- Use the CLT approximation to estimate  $P(\bar{X} \leq 0.27)$ .
  - Use the expanded CLT to estimate the same probability. The function `dnorm` computes the normal density in R. *Hint:*  $E[(X - \mu)^3] = 1/160$ .
  - If  $X_1, X_2, X_3 \stackrel{\text{i.i.d.}}{\sim} \text{Unif}[0, 1]$ , then the minimum has density  $f$  given above. Use simulation to estimate the probability. *Hint:* `pmin(runif(100), runif(100), runif(100))` gives 100 values from the density.

# 5

## Introduction to Confidence Intervals: The Bootstrap

In Chapter 4, we learned about sampling distributions and some ways to compute or estimate them. A common feature of previous examples is that the relevant populations were *known* – for example a binomial distribution with specified  $p$  or exponential distribution with specified  $\lambda$ . In the case of permutation testing, we estimated the sampling distribution of the test statistic conditioned on the pooled data – in other words, we treated the pooled data as a known population.

We now move from the realm of probability to statistics and from situations where the population is known to where it is unknown. If all we have are data and a statistic estimated from the data, we need to estimate the sampling distribution of the statistic. In this chapter, we introduce one way to do so, the bootstrap.

### 5.1 Introduction to the Bootstrap

For the North Carolina (NC) data (Case Study in Section 1.2), the mean weight of the 1009 babies in the sample is 3448.26 g. We don't know  $\mu$ , the true mean birth weight for all North Carolina babies born in 2004 and have no way to determine it. What we can do is estimate how far from 3448.26 it is likely to be, given that our estimate was based on 1009 observations.

If we knew the sampling distribution of sample means for samples of size 1009 from the population of all 2004 North Carolina births, then we could calculate the standard deviation of that distribution to measure how far sample means typically deviate from the population mean  $\mu$ . But, of course, since we do not have *all* the birth weights, we cannot generate that sampling distribution (and if we did have all the weights, we would know the true  $\mu$ !).

We would like to draw samples from the population, but we cannot. Instead, we draw samples from the data. In the bootstrap, we draw samples of size  $n$  with replacement from the data of size  $n$ . For example if a sample consisted of the numbers  $\{1, 2, 3, 4, 5\}$ , then two *bootstrap samples* (or *resamples*) are  $\{1, 3, 3, 4, 2\}$  and  $\{2, 3, 3, 2, 5\}$ . We compute the statistic of interest, say the mean, for each resample. We repeat that many times, to get many *bootstrap statistics*. The collection of these bootstrap statistics is the bootstrap distribution.

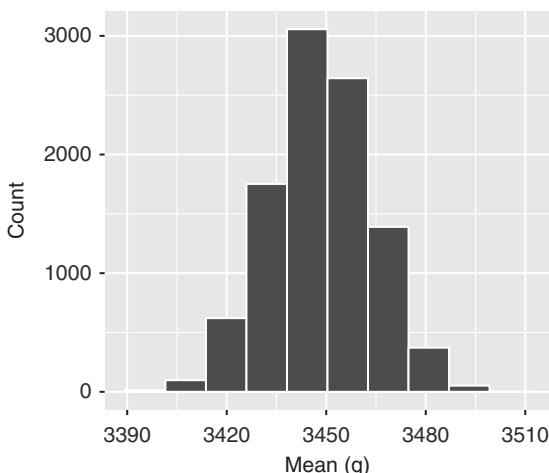
### Bootstrap for a Single Population

Given a sample of size  $n$  from a population,

1. Draw a resample of size  $n$  with replacement from the sample. Compute a statistic that describes the sample, such as the sample mean.
2. Repeat this resampling process many times, say 10 000.
3. Construct the bootstrap distribution of the statistic. Inspect its center, spread, and shape.

For the NC data, we draw say 10 000 samples of size  $n = 1009$  from the data and compute the mean for each bootstrap sample to give the distribution shown in Figure 5.1.

In Figure 5.1, we see that the bootstrap distribution is approximately normal. This suggests that the sampling distribution is approximately normal (the CLT!) Second, with a mean of 3448.409, the bootstrap distribution is centered at approximately the same location as the original sample mean, 3448.26.



**Figure 5.1** Bootstrap distribution of means for the North Carolina birth weights.

Third, we get a rough idea of the variability. We can quantify the variability by computing the standard deviation of the bootstrap distribution, in this case 15.324. This is the *bootstrap standard error*.

For comparison, the standard deviation of the data is 487.736. The bootstrap standard error (SE) is smaller – this reflects the fact that an average of 1009 observations is more accurate (less variable) than a single observation is.

### Bootstrap Standard Error

The *bootstrap standard error* of an estimator is the standard deviation of the bootstrap distribution of that estimator.

In general, a *standard error* is an estimate of the standard deviation of the sampling distribution of an estimator.

The idea behind the bootstrap is that if the original sample is like the population, then the bootstrap distribution of the mean will look approximately like the sampling distribution of the mean, that is, the bootstrap distribution will have roughly the same spread and shape. However, the mean of the bootstrap distribution will match the mean of the *original sample* (Theorem A.7), rather than the mean of the original population.

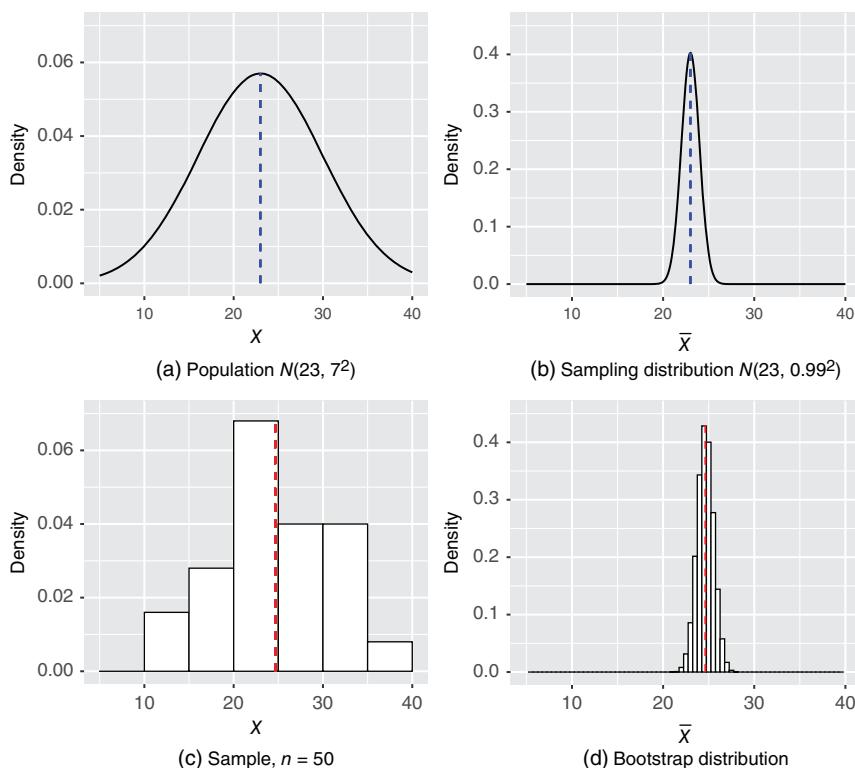
### The Bootstrap Idea

The original sample approximates the population from which it was drawn. Thus, resamples from this sample approximate what we would get if we took samples from the population. The spread and shape of a bootstrap distribution of a statistic approximate the spread and shape of the sampling distribution of the statistic.

To highlight some key features of the bootstrap distribution, we begin with two examples where the sampling distribution of the mean is known.

**Example 5.1** Consider a random sample of size 50 drawn from  $N(23, 7^2)$ . From Corollary A.2, we know the sampling distribution of the sample means is normal with mean 23 and standard error  $\sigma/\sqrt{n} = 7/\sqrt{50} = 0.99$ . Figure 5.2 shows the distribution of one such random sample with sample mean and standard deviation  $\bar{x} = 24.13$ ,  $s = 6.69$ , respectively.

For comparison, to bootstrap, we draw say 1000 resamples of size 50 from the original sample and compute the mean of each resample. The 1000 sample means comprise the bootstrap distribution. In Figure 5.2, we can see that the



**Figure 5.2** (a) The population distribution,  $N(23, 7^2)$ . (b) The theoretical sampling distribution of  $\bar{X}$ ,  $N(23, 7^2/50)$ . (c) The distribution of one sample of size 50 from  $N(23, 7^2)$ . (d) The bootstrap distribution based on that sample. Vertical lines mark the means.

**Table 5.1** Summary of center and spread for the normal distribution example.

	Mean	Standard deviation
Population	23	7
Sampling distribution of $\bar{X}$	23	0.99
One random sample	24.64	6.43
Bootstrap distribution	24.64	0.91

bootstrap distribution has roughly the same spread and shape as the theoretical sampling distribution, but the centers are different; the mean of the bootstrap distribution matches  $\bar{x}$  rather than  $\mu$ . In Table 5.1, we quantify that the standard deviation of the bootstrap distribution is close to the standard deviation of  $\bar{X}$ .  $\square$

This example illustrates some important features of the bootstrap that hold for other statistics besides the mean: The bootstrap distribution of an estimator  $\hat{\theta}$  typically has approximately the same spread and shape as the sampling distribution of  $\hat{\theta}$ , but the center of the bootstrap distribution is at  $\hat{\theta}$  of the original sample. Hence, we do not use the center of the bootstrap distribution as a new estimator, but we do compare the center of the bootstrap distribution with the observed statistic; if they differ (more than random variation), it indicates *bias* (Section 5.6).

### Bootstrap Distributions and Sampling Distributions

For most statistics, bootstrap distributions approximate the spread, bias, and shape of the actual sampling distribution.

**Example 5.2** We now consider an example where neither the population nor the sampling distribution is normal (Table 5.2).

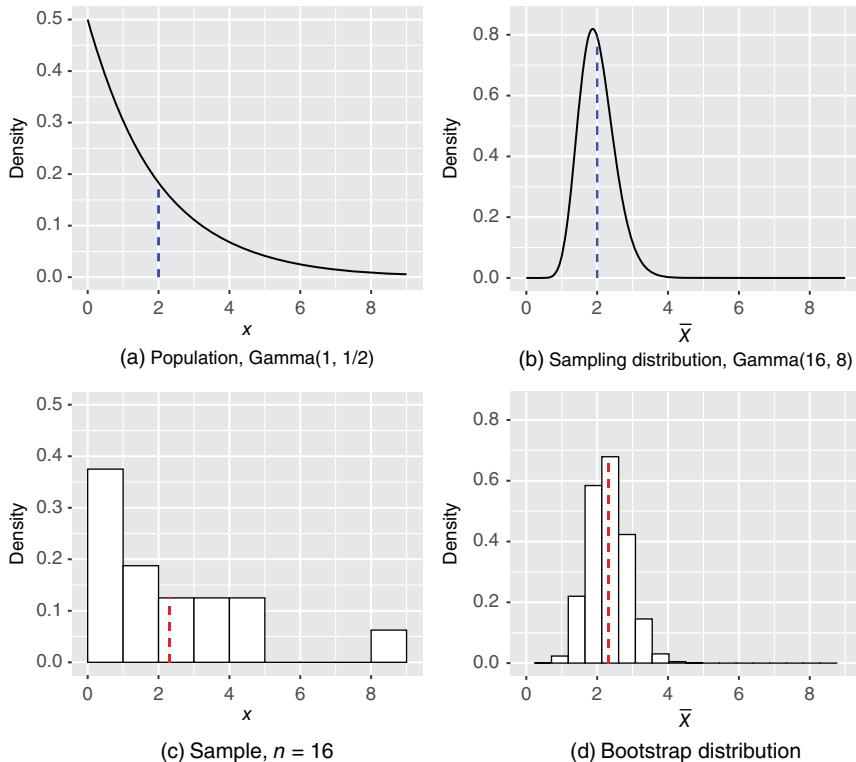
Recall that if  $X$  is a gamma random variable,  $\text{Gamma}(r, \lambda)$ , then  $E[X] = r/\lambda$  and  $\text{Var}[X] = r/\lambda^2$  (Theorem B.10). Let  $X_1, \dots, X_n \sim \text{Gamma}(r, \lambda)$ . It is a fact that the sampling distribution of the mean  $\bar{X}$  is  $\text{Gamma}(nr, n\lambda)$  (a consequence of Theorem B.11 and Proposition B.3).

We draw a random sample of size  $n = 16$  from the gamma distribution  $\text{Gamma}(1, 1/2)$ . The left side of Figure 5.3 shows a graph of the population (mean 2, standard deviation 2) and the distribution of one random sample ( $\bar{x} = 2.73$  and  $s = 2.61$ ). The right side shows the theoretical sampling distribution of the means,  $\text{Gamma}(16, 8)$ , and the bootstrap distribution, based on 10 000 resamples from that sample.

Even though the distribution of the sample does not exactly match the population distribution, the bootstrap distribution is similar to the sampling distribution: it has roughly the same shape, a slightly larger spread (because the data have a slightly larger standard deviation than does the population), and the mean of the bootstrap distribution matches the empirical distribution rather than the population.

**Table 5.2** Summary of center and spread for the gamma distribution example.

	Mean	Standard deviation
Population	2	2
Sampling distribution of $\bar{X}$	2	0.5
Sample	2.304	2.292
Bootstrap distribution	2.308	0.549



**Figure 5.3** Sampling and bootstrap distribution from a gamma distribution. (a) The population distribution  $\text{Gamma}(1, 1/2)$ . (b) The theoretical sampling distribution of  $\bar{X}$ . (c) A single sample of size 16 from  $\text{Gamma}(1, 1/2)$ . (d) The bootstrap distribution for means of size 16, drawn from the sample.

### R Note

Draw a random sample of size 16 from  $\text{Gamma}(1, 1/2)$ :

```
gamSample <- rgamma(16, 1, 1/2)
```

The following simulates a bootstrap distribution based on  $10^5$  resamples.

```
N <- 10^5
mean.boot <- numeric(N)
for (i in 1:N)
{
  x <- sample(gamSample, 16, replace = TRUE) # draw resample
  mean.boot[i] <- mean(x) # compute mean, store in mean.boot
}
```

```
mean(mean.boot)
sd(mean.boot)

df <- data.frame(mean.boot)
ggplot(df, aes(mean.boot)) +
  geom_histogram(bins = 20, color = "white")
```

□

## 5.2 The Plug-in Principle

We hinted that we use the bootstrap to estimate the sampling distribution or at least some things about the sampling distribution. Let us talk about what the bootstrap does, why it works, and what we can and cannot do with it.

The idea behind the bootstrap is the *plug-in principle* – that if something is unknown, we plug in an estimate for it.

### The Plug-in Principle

To estimate a parameter, a quantity that describes the population, use the statistic that is the corresponding quantity for the sample.

This principle is often used in statistics. For example, the standard error for  $\bar{X}$  calculated from i.i.d. observations from a population with standard deviation  $\sigma$  is  $\sigma/\sqrt{n}$ ; when  $\sigma$  is unknown, we plug in an estimate  $s$  to obtain the usual standard error  $s/\sqrt{n}$ .

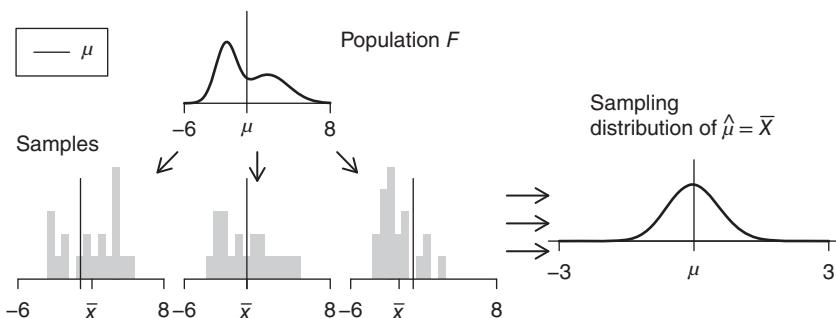
What is different in the bootstrap is that we plug in an estimate for the whole population, not just for a numerical summary of the population. We use the observed data as an estimate of the whole population; we will come to this in Section 5.2.1, and alternatives in Chapter 13, but for now, we will continue with the main idea, that we plug in an estimate, and what follows from that.

Our goal is to estimate a sampling distribution of some statistic. The sampling distribution depends on

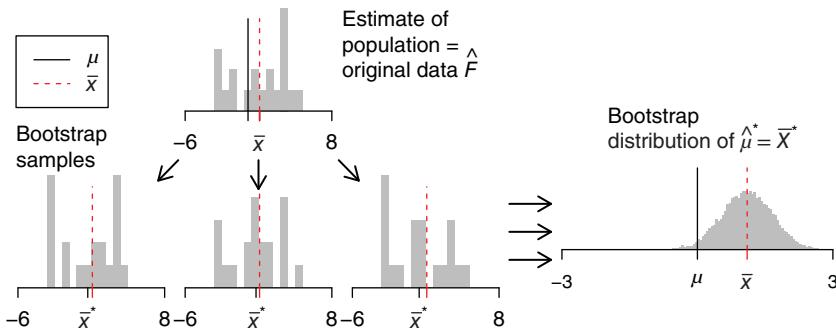
1. the underlying population(s),
2. the sample size,
3. the sampling procedure (e.g. sampling with or without replacement), and
4. the statistic, such as  $\bar{X}$ .

Figure 5.4 contains a diagram of this process.

The sampling distribution of a statistic is the result of drawing many samples from the population and calculating the statistic for each. The problem in most statistical applications is that the population is unknown.



**Figure 5.4** Diagram of the process of creating a sampling distribution. Many (infinitely many) samples are drawn from the population, a statistic like  $\bar{X}$  is calculated for each. The distribution of the statistics is the sampling distribution.



**Figure 5.5** Diagram of the process of creating a bootstrap distribution. This is like Figure 5.4, except that the original data take the place of the population. We draw many samples from the original data, calculate  $\bar{X}^*$  or another statistic for each, and collect the statistics to form the bootstrap distribution.

The bootstrap principle is to plug in an estimate for the population and then mimic the real-life sampling procedure and statistic calculation (Figure 5.5). The bootstrap distribution depends on

1. an estimate for the population(s),
2. the sample size,
3. the sampling procedure, and
4. the statistic, such as  $\bar{X}$ .

### 5.2.1 Estimating the Population Distribution

In this chapter, we use the empirical distribution as an estimate for the population. Let us look at this more closely.

Let  $F$  and  $f$  denote the cdf and pdf for some unknown distribution, and let  $x_1, x_2, \dots, x_n$  denote a random sample.

If we were willing to make assumptions about the population, say that it followed an exponential distribution, we could estimate the parameter  $\lambda$  from the data and then draw bootstrap samples from an exponential distribution with the estimated  $\lambda$ . This would be a parametric bootstrap, discussed in Section 13.2.

But most often when bootstrapping, we want to make as few assumptions as possible about the population. We want the data to tell us what it can, not introduce bias by making assumptions that may be wrong. So we resort to the empirical distribution, introduced in Section 2.5:

$$\hat{F}(s) = \frac{1}{n} \{\text{number of points } \leq s\}.$$

This is a discrete distribution, with probability  $1/n$  at each observed data point, and empirical mass function

$$\hat{f}(s) = \frac{1}{n} \{\text{number of points } = s\}.$$

For instance, if the sample is 5, 5, 6, 10, 11, 11, 11, 12, then  $\hat{f}(5) = 2/8$ ,  $\hat{f}(6) = 1/8$ ,  $\hat{f}(10) = 1/8$ ,  $\hat{f}(11) = 3/8$ ,  $\hat{f}(12) = 1/8$ , and  $\hat{f}(s) = 0$  for all other values  $s$ .

When bootstrapping we rarely bother to define or write down  $\hat{F}$  and  $\hat{f}$  – instead, we just need to know how to draw samples, which we do by sampling from the original observations, with equal probabilities  $1/n$  for each.

In some cases, we need the mean and variance of  $\hat{F}$ . Recall that the mean for  $F$  is

$$E_F[X] = \mu_F = \int_{-\infty}^{\infty} x f(x) dx$$

or

$$E_F[X] = \mu_F = \sum_x x f(x),$$

depending whether the population is continuous or discrete, where the subscript  $F$  indicates expected value based on  $F$ . Since  $\hat{F}$  is discrete, we calculate the expected value using summation,

$$\begin{aligned} E_{\hat{F}}[X] &= \mu_{\hat{F}} \\ &= \sum_x x \hat{f}(x) \\ &= \sum_{i=1}^n x_i \frac{1}{n} = \bar{x}. \end{aligned}$$

Similarly, the population variance under  $\hat{F}$  is

$$\begin{aligned} \text{Var}_{\hat{F}}[X] &= \sigma_{\hat{F}}^2 \\ &= E_{\hat{F}}[(X - \mu_{\hat{F}})^2] \\ &= \sum_{i=1}^n (x_i - \bar{x})^2 \frac{1}{n}. \end{aligned}$$

This is like the sample variance  $s^2$  (Equation (2.1)), but with a denominator of  $n$  instead of  $n - 1$ .

### 5.2.2 How Useful Is the Bootstrap Distribution?

A fundamental question is how well the bootstrap distribution approximates the sampling distribution. We discuss this question in greater detail in Section 5.8, but note a few key points here.

First, the statistics that we bootstrap are usually *estimators*, statistics that estimate a parameter. For example the sample mean  $\bar{X}$  is an estimator for the population mean  $\mu$ . We will also bootstrap  $t$ -statistics in Sections 5.8.4 and 7.5.2;  $t = (\bar{X} - \mu)/(s/\sqrt{n})$  is not an estimator.

For most common estimators and under fairly general distribution conditions, the following hold:

*Center:* The center of the bootstrap distribution is *not* an accurate approximation for the center of the sampling distribution. For example the center of the bootstrap distribution for  $\bar{X}$  is centered at approximately  $\bar{x} = \mu_F$ , the mean of the sample, whereas the sampling distribution is centered at  $\mu$ .

*Spread:* The spread of the bootstrap distribution does reflect the spread of the sampling distribution.

*Bias:* The bootstrap bias estimate (see Section 5.6) does reflect the bias of the sampling distribution. Bias occurs if a sampling distribution is not centered at the parameter.

*Skewness:* The skewness of the bootstrap distribution does reflect the skewness of the sampling distribution.

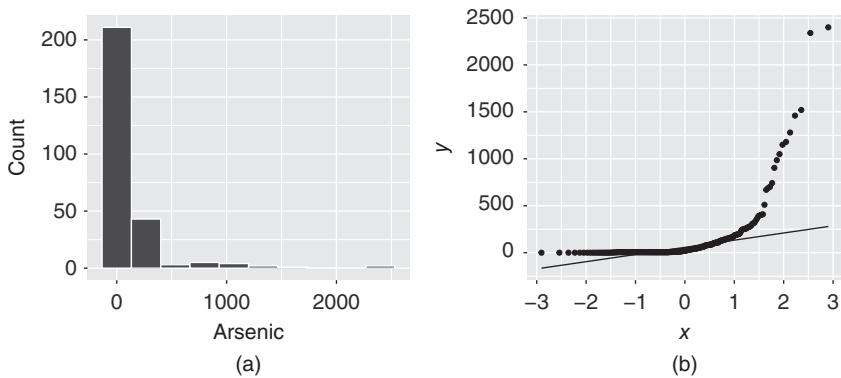
The first point bears emphasis. It means that *the bootstrap is not used to get better parameter estimates* because the bootstrap distributions are centered around statistics  $\hat{\theta}$  calculated from the data (e.g.  $\bar{x}$ ) rather than the unknown population values (e.g.  $\mu$ ). Drawing bootstrap observations from the original data is not like drawing observations from the underlying population, it does not create new data.

Instead, the bootstrap sampling is useful for *quantifying the behavior of a parameter estimate*, such as its standard error, skewness, bias, or for calculating confidence intervals.

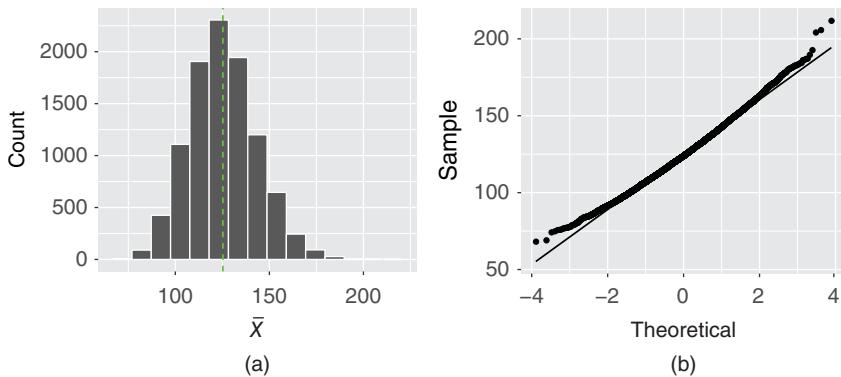
**Example 5.3** Arsenic is a naturally occurring element in the groundwater of Bangladesh. However, much of this groundwater is used for drinking water by rural populations, so arsenic poisoning is a serious health issue. Figure 5.6 displays the distribution of arsenic concentrations from 271 wells in Bangladesh<sup>1</sup>.

---

<sup>1</sup> Data provided solely for illustrative purposes and to enable statistical analysis. Reproduced with the permission of the British Geological Survey © UKRI. All Rights Reserved. <https://www2.bgs.ac.uk/groundwater/health/arsenic/Bangladesh/data.html>



**Figure 5.6** (a) Histogram and (b) normal quantile plot of arsenic levels in 271 wells in Bangladesh.



**Figure 5.7** (a) Histogram and (b) normal quantile plot of the bootstrap distribution for mean arsenic concentration.

The sample mean and standard deviation are  $\bar{x} = 125.31$  and  $s = 297.98$ , respectively (measured in micrograms per liter, or in parts per billion (ppb)). The US Environmental Protection Agency (EPA) set an arsenic maximum contaminant level (MCL) for public water supplies at 10 ppb. A total of about 57% of the samples exceed that level.

We draw resamples of size 271 with replacement from the data and compute the mean for each resample. Figure 5.7 shows a histogram and normal quantile plot of the bootstrap distribution. The bootstrap distribution looks quite normal, with some skewness. This is the central limit theorem (CLT) at work – when the sample size is large enough, the sampling distribution for the mean is approximately normal, even if the population is not normal.

## R Note

Import the data set Bangladesh into R, then:

```
ggplot(Bangladesh, aes(Arsenic)) +
  geom_histogram(bins = 10, color = "white")
ggplot(Bangladesh, aes(sample = Arsenic)) +
  geom_qq() + geom_qq_line()

Arsenic <- Bangladesh$Arsenic

n <- length(Arsenic)
N <- 10^4
mean.boot <- numeric(N)
for (i in 1:N)
{
  x <- sample(Arsenic, n, replace = TRUE)
  mean.boot[i] <- mean(x)
}

df <- data.frame(mean.boot)
ggplot(df, aes(mean.boot)) +
  geom_histogram(bins = 15, color = "white") +
  geom_vline(xintercept = mean(mean.boot), color = "red", lty = 2)
ggplot(df, aes(sample = mean.boot)) + geom_qq() + geom_qq_line()
```

The mean of the bootstrap means is 125.54, quite close to the sample mean  $\bar{x}$  (the difference is 0.22). The *bootstrap standard error* is the standard deviation of the bootstrap distribution; in this case, the bootstrap standard error is 18.26.

For the normal distribution, we know that the 2.5 and 97.5 percentiles are at the mean  $\pm 1.96$  standard deviations. But for this particular bootstrap distribution, we find that 1.5% of the resample means are below the bootstrap mean  $-1.96$  SE, and 3.4% of the resample means are above the bootstrap mean  $+1.96$  SE (see below). In this case, relying on the CLT would be inaccurate.

## R Note (Arsenic, Continued)

We find the numeric summaries:

```
> mean(mean.boot)                      # bootstrap mean
[1] 125.5375
```

```
> mean(mean.boot)-mean(Arsenic) # bias
[1] 0.2175773
> sd(mean.boot) # bootstrap SE
[1] 18.25759
```

Compute the points that are 1.96 SE from the mean of the bootstrap distribution:

```
> 125.5375-1.96*18.25759 # mark of 1.96SE from mean
[1] 89.75262
> 125.5375+1.96*18.25759
[1] 161.3224
> sum(arsenic.mean > 161.3224)/N
[1] 0.0337
> sum(arsenic.mean < 89.75262)/N
[1] 0.0153
```

□

## 5.3 Bootstrap Percentile Intervals

The sample mean  $\bar{x}$  gives an estimate of the true mean  $\mu$ , but it probably does not hit it exactly. It would be nice to have a *range* of values for the true  $\mu$  that we are 95% sure includes the true  $\mu$ .

In the NC birth weights case study, the bootstrap distribution (Figure 5.1) shows roughly how sample means vary for samples of size 1009. If most of the sample means are concentrated within a certain interval of the bootstrap distribution, it seems reasonable to assume that the true mean is most likely somewhere in that same interval. Thus, we can construct what is called a 95% confidence interval by using the 2.5 and 97.5 percentiles of the bootstrap distribution as endpoints. We would then say that we are 95% confident that the true mean lies within this interval. These are *bootstrap percentile confidence intervals*.<sup>2</sup>

### Bootstrap Percentile Confidence Intervals

The interval between the 2.5 and 97.5 percentiles of the bootstrap distribution of a statistic is a 95% *bootstrap percentile confidence interval* for the corresponding parameter.

---

<sup>2</sup> We will discuss the logic of confidence intervals more formally in Chapter 7.

For the NC birth weights, the interval marked by the 2.5 and 97.5 percentiles is (3419, 3478). Thus, we would state that we are 95% confident that the true mean weight of NC babies born in 2004 is between 3419 and 3478 g.

In the arsenic example, the 2.5% and 97.5% points of the bootstrap distribution give the interval (92.95, 164.44), so we are 95% confident that the true mean arsenic level is in this interval. This can be written as  $(\bar{x} - 32.37, \bar{x} + 39.12)$ ; in particular, and this interval is *not* symmetric about the mean, reflecting the asymmetry of the bootstrap distribution.

#### R Note

```
> quantile(mean.boot, c(0.025, 0.975)) # conf int
  2.5%    97.5%
92.9515 164.4418
```

The arsenic data illustrate an interesting point. A good confidence interval for the mean need not necessarily be symmetric: an endpoint will be farther from the sample mean in the direction of any outliers. A confidence interval is an insurance policy: rather than relying on a single statistic, the sample mean, as an estimate of  $\mu$ , we give a range of plausible values for  $\mu$ . We can see that there are some extremely large arsenic measurements: of the 271 observations, 8 are above 1000  $\mu\text{g/l}$  and 2 are above 2200  $\mu\text{g/l}$  (remember, the sample mean is only 125.31!). What we do not know is just how huge arsenic levels in the population can be, or how many huge ones there are. It could be that huge observations are *underrepresented* in our data set. In order to protect against this – that is, to have only a 2.5% chance of missing a true big mean, the interval of plausible values for  $\mu$  must stretch far to the right. Conversely, there is less risk of missing the true mean on the low side, so the left endpoint need not be as far away from the sample mean.

## 5.4 Two Sample Bootstrap

We now turn to the problem of comparing two samples. In general, bootstrapping should mimic how the data were obtained. So if the data correspond to independent samples from two populations, we should draw two samples that way. Then we proceed to compute the same statistic comparing the samples as for the original data, for example difference in means or ratio of proportions.

### Bootstrap for Comparing Two Populations

Given independent samples of sizes  $n_1$  and  $n_2$  from two populations,

1. Draw a resample of size  $n_1$  with replacement from the first sample and a separate resample of size  $n_2$  from the second sample. Compute a statistic that compares the two groups, such as the difference between the two sample means.
2. Repeat this resampling process many times, say 10 000.
3. Construct the bootstrap distribution of the statistic. Inspect its spread, bias, and shape.

**Example 5.4** Do men take more physical risks in the presence of an attractive woman? Two psychologists in Australia conducted an experiment to explore this question (Ronay and von Hippel, 2010). Male skateboarders between the ages of 18 and 35 years were randomly assigned to perform tricks in the presence of an attractive 18-year female experimenter or a male experimenter.<sup>3</sup> The two experimenters, both of whom were blind to the hypotheses, videotaped these sessions. At the end of the experiment, the researchers collected saliva samples from the participants and measured testosterone levels.

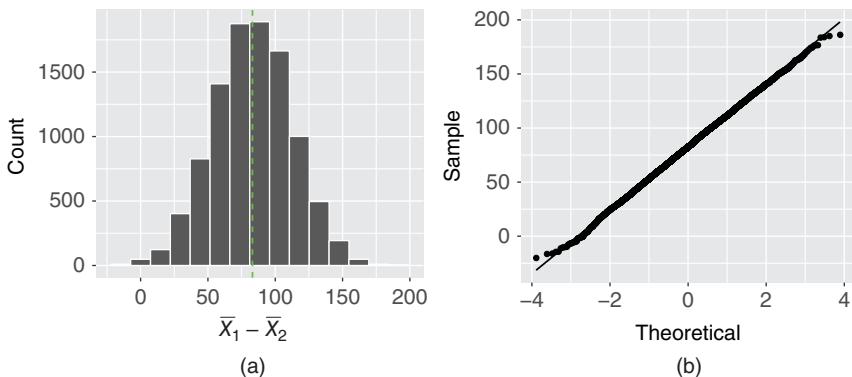
In normal adult males, testosterone levels range from 270 to 1070 nanograms per deciliter (ng/dl). From this study, the mean (sd) of testosterone levels of the 49 men who skateboarded in front of the female experimenter was 295.95 (143.69) ng/dl, compared to 212.88 (101.62) ng/dl for the 22 men who skateboarded in front of the male experimenter. Is the difference of 83.07 ng/dl statistically discernible?

We draw a bootstrap sample from the testosterone data of the 49 skateboarders who performed in front of the female experimenter and independently draw a bootstrap sample from the testosterone data of the 22 skateboarders who performed in front of the male experimenter.

Figure 5.8 shows the bootstrap distribution of the difference of sample means. As in the single-sample case, we see that the bootstrap distribution is approximately normal and centered at the original statistic (the difference in sample means). We also get a quick idea of how much the difference in sample means varies due to random sampling. We may quantify this variation by computing the bootstrap standard error, which is 29.38. Again, the bootstrap standard error is the standard error of the sampling distribution.

---

<sup>3</sup> The authors of the study claim that attractiveness was determined independently by 20 male raters who viewed this experimenter's photograph.



**Figure 5.8** (a) Histogram and (b) normal quantile plot of the bootstrap distribution for the difference in mean testosterone levels between skateboarders who performed in front of a female or male experimenter. The vertical line in the histogram marks the observed mean difference.

### R Note

Import the Skateboard data into R.

```
testF <- Skateboard %>% filter(Experimenter == "Female") %>%
  pull(Testosterone)
testM <- Skateboard %>% filter(Experimenter == "Male") %>%
  pull(Testosterone)

observed <- mean(testF) - mean(testM)           #observed difference
observed

nf <- length(testF)   #sample size
nm <- length(testM)   #sample size

N <- 10^4
mean.boot <- numeric(N)

for (i in 1:N)
{
  resampleF <- sample(testF, nf, replace = TRUE)
  resampleM <- sample(testM, nm, replace = TRUE)
  mean.boot[i] <- mean(resampleF)-mean(resampleM)
}

df <- data.frame(mean.boot)
ggplot(df, aes(mean.boot)) +
  geom_histogram(bins = 15, color = "white") +
```

```
geom_vline(xintercept = observed, color = "green", lty = 2)
ggplot(df, aes(sample = mean.boot)) + geom_qq() + geom_qq_line()
```

We find the numeric summaries.

```
> mean(testF) - mean(testM)
[1] 83.0692

> mean(mean.boot)
[1] 83.27132

> sd(mean.boot)
[1] 29.32129

> quantile(mean.boot, c(.025,.975))
[1] 26.36154 140.16294

> mean(mean.boot) - (mean(testF) - mean(testM)) #bias
[1] 0.2021206
```

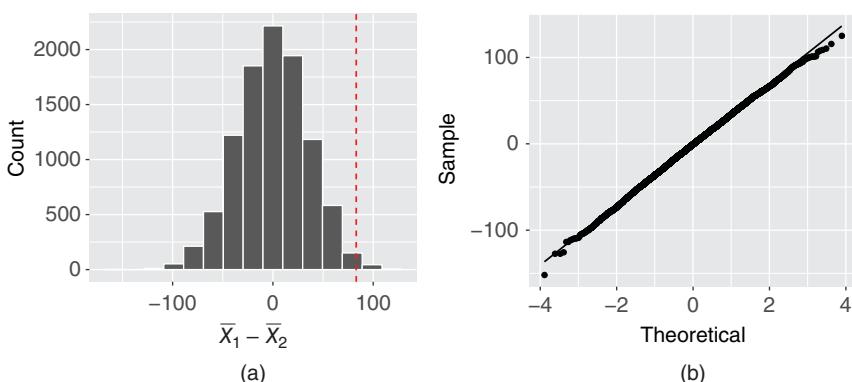
We will discuss bias in Section 5.6.

The 95% bootstrap percentile confidence interval for the difference in means (female–male) is (26.36, 140.16). Thus, we are 95% confident that testosterone levels of men who skateboard in front of a female experimenter are, on average, between 26.36 and 140.16 ng/dl higher than that of the men who skateboard in front of a male experimenter.

We can also conduct a permutation test of the hypothesis that the mean testosterone levels for the two experimenter options are the same versus the hypothesis that mean levels are not. Figure 5.9 shows the permutation distribution for the difference in mean testosterone levels between the two experimenters.

Recall that in permutation testing, we sample *without* replacement from the pooled data. The permutation distribution corresponds to sampling in a way that is consistent with the null hypothesis that the population means are the same. Thus, the permutation distribution is centered at 0. But in bootstrapping, we sample *with* replacement from the individual sample. However, the bootstrap has no restriction in regard to any null hypothesis, so its distribution is centered at the original difference in means.

The permutation distribution is used for a single purpose to calculate a  $P$ -value to see how extreme an observed statistic is if the null hypothesis is true. The bootstrap is used for estimating standard errors and for answering some other questions we will raise below.



**Figure 5.9** (a) Histogram and (b) normal quantile plot of the permutation distribution for the difference in mean testosterone levels of male skateboarders in the presence of a female versus male experimenter. The vertical line in the histogram marks the observed mean difference.

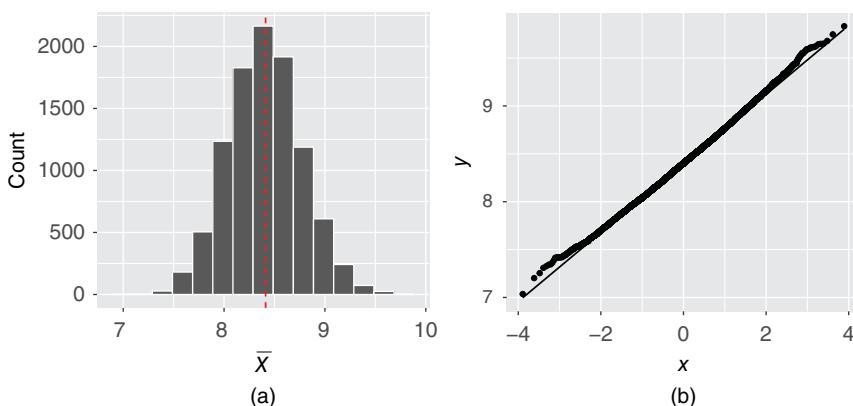
The permutation test for this example results in a  $P$ -value of 0.008; thus, we conclude that the mean testosterone levels are not the same for skateboarders who perform in front of a female experimenter and those who perform in front of a male experimenter.

**Remark** In the study, 53 participants were assigned to the female experimenter and 43 were assigned to the male experimenter. However, some of the participants declined to have their saliva sampled, and some of the saliva samples were contaminated. Thus, the researchers had testosterone measurements for 49 skateboarders in the female experimenter group and only 22 skateboarders from the male experimenter group. Participants were not asked if they were gay or straight. See Exercise 5.5.  $\square$

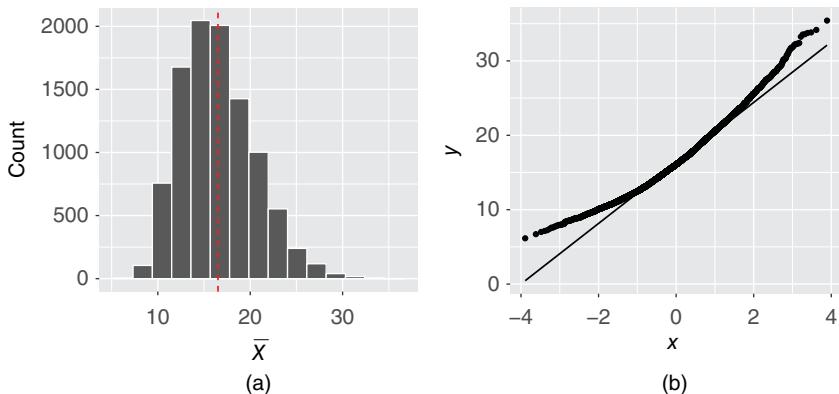
**Example 5.5** We return again to the Verizon example in Example 3.4. The distribution of the original data is shown in Figure 3.4, the permutation distribution for the difference in means is shown in Figure 3.5, and a permutation test of the difference in medians and trimmed means shown in Figure 3.6.

The bootstrap distribution for the larger ILEC data set ( $n = 1664$ ) is shown in Figure 5.10. The distribution is centered around the sample mean of 8.4, has a relatively narrow spread primarily due to the large sample size, with a bootstrap SE of 0.36 and a 95% bootstrap percentile interval of (7.7, 9.1). The distribution is roughly symmetric, with little skewness.

The bootstrap distribution for the smaller CLEC data set ( $n = 23$ ) is shown in Figure 5.11. The distribution is centered around the sample mean of 16.5, has a much larger spread due to the small sample size, with a bootstrap SE of 3.99



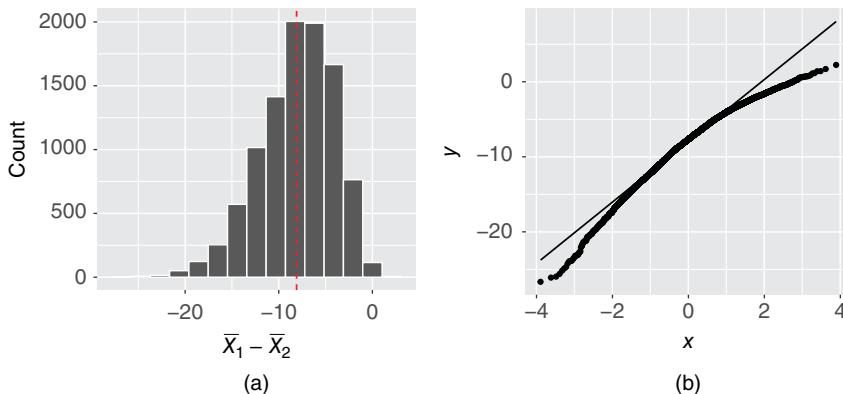
**Figure 5.10** (a) Histogram and (b) quantile normal plots of the bootstrap distribution for the sample mean of the Verizon ILEC data set,  $n = 1664$ . The vertical line in the histogram is at the observed mean.



**Figure 5.11** Bootstrap distribution for the sample mean of the Verizon CLEC data set,  $n = 23$ . The vertical line is at the observed mean.

and a 95% bootstrap percentile interval of  $(10.1, 25.4)$ . The distribution is very skewed.

The bootstrap distribution for the difference in means is shown in Figure 5.12. Note the strong skewness in the distribution. The mean of the bootstrap distribution is  $-8.096$  with a standard error of  $4.006$ . A 95% bootstrap percentile confidence interval for the difference in means (ILEC–CLEC) is given by  $(-16.97, -1.69)$  and so we would say, with 95% confidence, that mean repair times for ILEC customers are from 1.69 to 16.97 h shorter than mean repair times for CLEC customers.  $\square$



**Figure 5.12** Bootstrap distribution for the difference in means, ILEC–CLEC. The vertical line in the histogram is at the observed mean difference.

#### 5.4.1 Matched Pairs

In Section 3.4, we compared the mean semifinal and final scores of 12 female divers competing in the FINA World Championships in 2017. We performed a permutation test to determine that the observed difference could be explained by chance alone. Here, let us suppose that these 12 divers are a representative sample of all elite divers, and we are interested in a confidence interval for the true mean difference in scores.

In this case, for each diver, we compute the difference in scores between the two rounds (final round score – semifinal round score). We then have one variable – the score differences – and we are back to the one-sample setting described in Section 5.1.

Performing a one sample bootstrap with  $10^5$  resamples, we find a 95% bootstrap percentile interval for the mean score difference to be  $(-6.65, 31.06)$ . Since 0 is contained in the interval, we cannot conclude that the mean scores for divers differ between the final and semifinal rounds.

**Remark** Suppose we want to compare the median scores, rather than means. The difference of medians is not the same as the median of differences. For the latter, we could take differences of the two variables and then do a one-sample bootstrap on that difference. But we want the former, so we resample the pairs – rows of Table 3.4 – to get a bootstrap sample with two columns, compute the median for each column, then take the difference of medians. See Exercise 5.24. ||

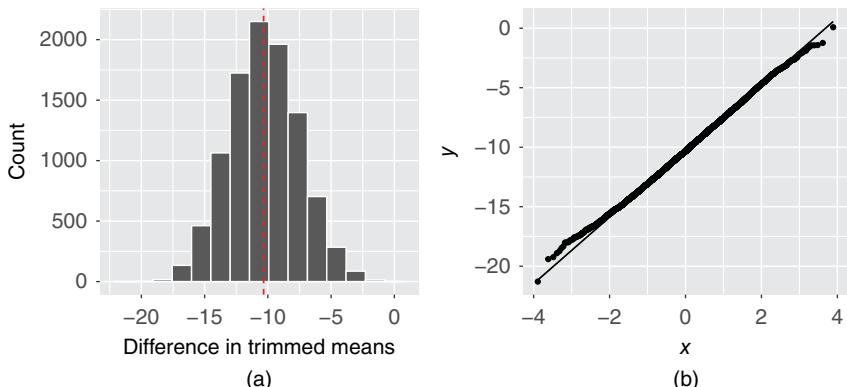
## 5.5 Other Statistics

When bootstrapping, we are not limited to simple statistics like the simple mean. Once we have drawn a bootstrap sample, we can calculate any statistic for that sample.

For example, instead of a sample mean, we can use more-robust statistics that are less sensitive to extreme observations. Figure 5.13 shows the bootstrap distribution for the difference in trimmed means, in this case 25% trimmed means, also known as the *midmean*, the mean of the middle 50% of observations. (The vertical line in the histogram is at the observed difference of trimmed means,  $-10.34$ .) This distribution has a much smaller spread than the bootstrap difference in ordinary means (see Figure 5.12).

The bootstrap procedure may be used with a wide variety of statistics – means, medians, trimmed means, proportions, correlation coefficients, and so on – using the same procedure. This is a major advantage of the bootstrap. It allows statistical inferences such as confidence intervals to be calculated even for statistics for which there are no easy formulas. It offers hope of reforming statistical practice – away from simple but nonrobust estimators like a sample mean or least-squares regression (Chapter 9), in favor of robust alternatives.

**Example 5.6** In the Verizon data, rather than looking at the difference in means, suppose we look at the ratio of means. The sample ratio is 0.51, so for ILEC customers, repair times are about half that of CLEC customers.



**Figure 5.13** Bootstrap distribution for the difference in 25% trimmed means for the Verizon data.

### R Note

```

TimeILEC <- Verizon %>% filter(Group=="ILEC") %>% pull(Time)
TimeCLEC <- Verizon %>% filter(Group=="CLEC") %>% pull(Time)

observed <- mean(TimeILEC)/mean(TimeCLEC)
observed

nILEC <- length(TimeILEC)
nCLEC <- length(TimeCLEC)

N <- 10^4
ratio.boot <- numeric(N)

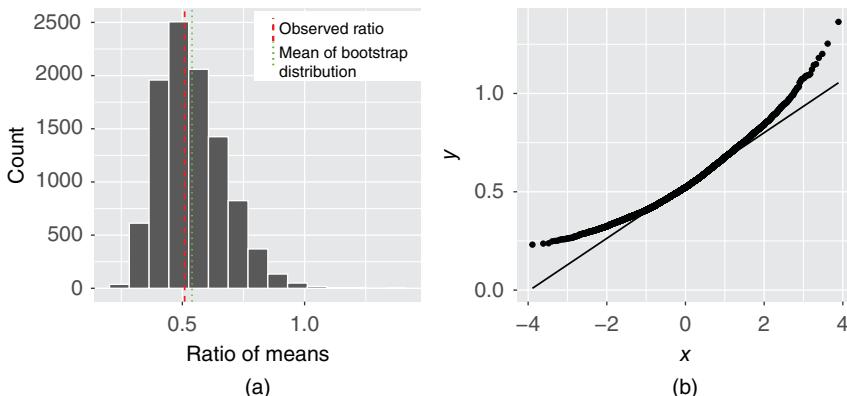
for (i in 1:N)
{
  resampleILEC <- sample(TimeILEC, nILEC, replace = TRUE)
  resampleCLEC <- sample(TimeCLEC, nCLEC, replace = TRUE)
  ratio.boot[i] <- mean(resampleILEC)/mean(resampleCLEC)
}

df <- data.frame(ratio.boot)
ggplot(df, aes(ratio.boot)) +
  geom_histogram(bins = 15, color="white") +
  xlab("Ratio of means") +
  geom_vline(xintercept = observed, lty = 2, color = "red") +
  geom_vline(xintercept = mean(ratio.boot), lty = 3, color = "blue")

ggplot(df, aes(sample = ratio.boot)) +
  geom_qq() + geom_qq_line()

```

As in the difference of means example, the bootstrap distribution of the ratio of means exhibits skewness (Figure 5.14).



**Figure 5.14** Bootstrap distribution for the ratio of means.

The 95% bootstrap percentile confidence interval for the ratio of means (ILEC/CLEC) is (0.328, 0.834), so with 95% confidence, the true mean repair time for ILEC customers is between 0.33 and 0.83 times the true mean for CLEC customers.

### R Note (Verizon, Continued)

For the numeric summaries:

```
> mean(ratio.boot)
[1] 0.5395905
> sd(ratio.boot)
[1] 0.1334427
> quantile(ratio.boot, c(0.025, 0.975))
 2.5%    97.5%
0.3276632 0.8344401
> mean(ratio.boot) - mean(TimeILEC)/mean(TimeCLEC)
[1] 0.0300779
```

The last calculation above estimates bias, which we will discuss in Section 5.6.

□

**Example 5.7** We can also bootstrap binary data to obtain confidence intervals for proportions. Continuing with the Verizon data set, we find that 148 out of 1664 repair times for ILEC customers took longer than 24 h. We will compute a bootstrap confidence interval for the proportion:

### R Note (Verizon, Continued)

Recall that `TimeILEC` is the vector of repair times for the ILEC customers.

```
N <- 10^4

prop.boot <- numeric(N)
for (i in 1:N)
{
  resampleILEC <- sample(TimeILEC, nILEC, replace = TRUE)
  prop.boot[i] <- mean(resampleILEC > 24)
}

quantile(prop.boot, c(0.025, 0.975))
```

With 95% confidence, the percentage of ILEC customers whose repairs take more than 24 h is between 7.6% and 10.3%. □

**Example 5.8** We refer again to the survey in Example 3.9. Of the 257 Blacks who were from Generation Z, 118 responded that they seldom or never attend religious services. Of the 2475 participants who were from Generation X, 965 responded that they seldom or never attend religious services.

The observed difference in proportions is  $0.46 - 0.39 = 0.07$ . Compute a 95% bootstrap confidence interval for the true difference in proportions.

### R Note

Since we do not have the data set, we use the `rep()` function to create the two samples. For generation Z, we create a vector with 118 1's and  $257 - 118 = 139$  0's; similarly, the vector for generation X will consist of 965 1's and  $2475 - 965 = 1510$  0's. (Recall that in a permutation test, we pool all of the data!)

```
genZ <- rep(c(1, 0), c(118, 139))
genX <- rep(c(1, 0), c(965, 1510))
```

The rest of the R code is similar to the previous examples.

```
observed <- mean(genZ) - mean(genX) # observed diff.
observed

N <- 10^4
prop.boot <- numeric(N)
for (i in 1:N)
{
  resampleZ <- sample(genZ, 257, replace = TRUE)
  resampleX <- sample(genX, 2475, replace = TRUE)
  prop.boot[i] <- mean(resampleZ) - mean(resampleX)
}
quantile(prop.boot, c(0.025, 0.975))
```

We are 95% confident that the percentage of generation Z Blacks who seldom or never attend religious services is from 0.45% to 13.4% higher than the percentage of generation X Blacks who seldom or never attend religious services.  $\square$

## 5.6 Bias

An estimator  $\hat{\theta}$  is biased if, on average, it tends to be too high or too low, relative to the true value of  $\theta$ . Formally, this is defined using expected values:

**Definition 5.1** The *bias* of an estimator  $\hat{\theta}$  is

$$\text{Bias}[\hat{\theta}] = E[\hat{\theta}] - \theta.$$

The bootstrap estimate of bias is

$$\text{Bias}_{\text{boot}}[\hat{\theta}^*] = E[\hat{\theta}^*] - \hat{\theta},$$

the mean of the bootstrap distribution, minus the estimate from the original data. ||

### Bias

A statistic used to estimate a parameter is *biased* when the mean of its sampling distribution is not equal to the true value of the parameter. The bias of a statistic  $\hat{\theta}$  is  $\text{Bias}[\hat{\theta}] = E[\hat{\theta}] - \theta$ . A statistic is **unbiased** if its bias is zero.

The bootstrap method allows us to check for bias by seeing whether the bootstrap distribution of a statistic is centered at the statistic of the original random sample. The bootstrap estimate of bias is the mean of the bootstrap distribution minus the statistic for the original data,  $\text{Bias}_{\text{boot}}[\hat{\theta}] = \hat{E}[\hat{\theta}^*] - \hat{\theta}$ .

You probably learned in a probability course that the sample mean is an unbiased estimator of the population mean  $\mu$  (Theorem A.7). In addition, the difference of sample means is also an unbiased estimator of the difference of population means. However, the ratio of sample means is not generally an unbiased estimator of the ratio of population means. The bootstrap distribution for the ratio of means has a long right tail, large observations that occur when the denominator is small. Consequently, the mean of the resample ratio of means is large, causing positive mean bias.

Let us compare the bias in different examples; we will standardize by SE, i.e. compare bias/SE, so we are measuring bias relative to uncertainty. For the arsenic example (Section 5.3), the ratio is only  $0.218/18.258 = 0.0119$  and for the skateboarders example (Section 5.4), the ratio is only  $0.202/29.321 = 0.0069$ ; in both cases, the bias is less than 2% of the standard error. On the other hand, for the Verizon ratio of means (Section 5.6), the ratio is  $0.030/0.133 = 0.2255$ , so the bias is about 22.6% of the standard error.

If the ratio of bias/SE exceeds  $\pm 0.02$ , then it is large enough to potentially have a substantial effect on the accuracy of confidence intervals. In applications where accuracy matters, there are other bootstrap confidence intervals that are more accurate than the relatively quick-and-dirty bootstrap percentile intervals. (As it turns out, bootstrap percentile intervals are actually reasonably accurate for the ratio of means.)

**Remark** Here is how we obtain that  $\pm 0.02$  value mentioned above. The bootstrap percentile interval is affected by bias twice –  $\hat{\theta}$  is biased for  $\theta$ , and

$\hat{\theta}^*$  is biased for  $\hat{\theta}$ . Suppose the bootstrap distribution is approximately normal, and that sample sizes are large enough that the SE is accurate. Let  $b = \text{bias}/\text{SE}$ . Then the actual noncoverage probabilities on each side are approximately  $\Phi(2b \pm z_{\alpha/2})$  for a two-sided  $(1 - \alpha)$  interval and  $z_{\alpha/2}$  denotes the  $\alpha/2$ -quantile of the standard normal distribution. For a 95% interval, if  $b = 0.0205$  then one of those probabilities is 0.0225. That is 10% less than the desired 0.025 probability, which is our rule of thumb in this book for reasonable accuracy. We round 0.0205 to 0.02.

||

The next example also shows noticeable bias.

**Example 5.9** A major study of the association between blood pressure and cardiovascular disease found that 55 out of 3338 ( $\hat{p}_1 = 0.0165$ ) men with high blood pressure died of cardiovascular disease during the study period, compared to 21 out of 2676 ( $\hat{p}_2 = 0.0078$ ) with low blood pressure. The estimated *relative risk* is  $\hat{\theta} = \hat{p}_1/\hat{p}_2 = 0.0165/0.0078 = 2.12$ . Thus, we would say that the risk of cardiovascular disease for men with high blood pressure is 2.12 times greater than the risk for men with low blood pressure.

To bootstrap the relative risk, we draw samples of size  $n_1 = 3338$  with replacement from the first group, independently draw samples of size  $n_2 = 2676$  from the second group, and calculate the relative risk  $\hat{\theta}^*$ . In addition, we record the individual proportions  $\hat{p}_1^*$  and  $\hat{p}_2^*$ . The bootstrap distribution for relative risk is shown in Figure 5.15. It is highly skewed, with a long right tail caused by denominator values relatively close to zero. The standard error is 0.6188, based on  $10^4$  pairs of resamples.

The average of the resample relative risks is larger than the sample relative risk, indicating bias. The estimated bias is  $2.199 - 2.12 = 0.079$ , so the ratio of

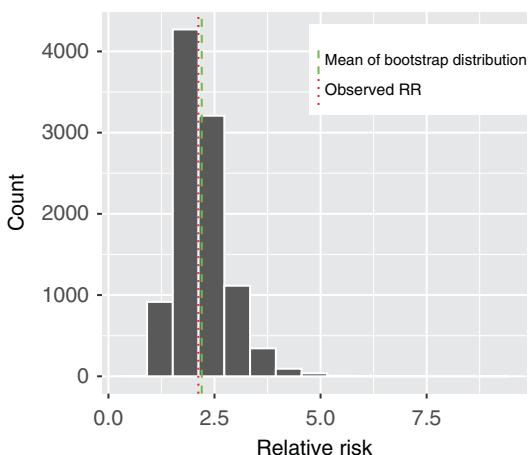
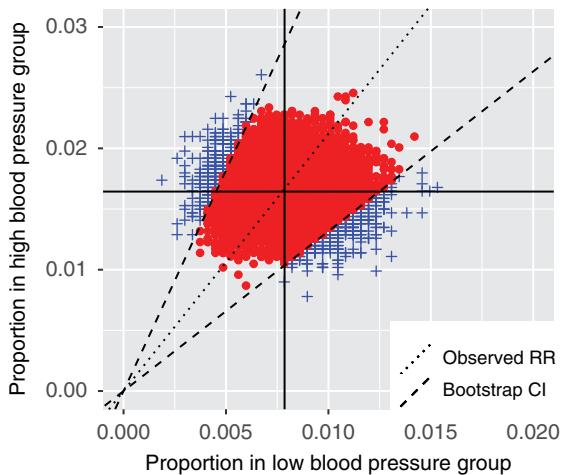


Figure 5.15 Bootstrap distribution of relative risk.

**Figure 5.16** Bootstrapped proportions of the high blood pressure group against the proportions of the low blood pressure group.



bias to the standard error is 0.129. While the bias does not appear large in the figure, this amount of bias can have a huge impact on formula-based confidence intervals. While the bootstrap percentile interval is fine, some common symmetric confidence intervals would miss by falling under the true value about twice as often as they should.

Figure 5.16 shows the joint bootstrap distribution of  $\hat{p}_1^*$  and  $\hat{p}_2^*$ . Each point corresponds to one bootstrap resample, and the relative risk is the slope of the line between the origin and the point. The original data are at the intersection of horizontal and vertical lines. The 95% bootstrap confidence interval for the true relative risk is (1.317, 3.638). Thus, for one bootstrap sample, if the relative risk satisfies  $\hat{p}_1^*/\hat{p}_2^* < 1.317$ , then  $\hat{p}_1^* < 1.317\hat{p}_2^*$ . Similarly, if  $\hat{p}_1^*/\hat{p}_2^* > 3.638$ , then  $\hat{p}_1^* > 3.638\hat{p}_2^*$ . This is shown by the points outside of the region bounded by the dashed lines of slopes 1.317 and 3.638. □

### R Note

Here is the code to bootstrap the relative risk.

```
highbp <- rep(c(1,0), c(55,3283)) #high bp sample
lowbp <- rep(c(1,0), c(21,2655)) #low bp sample

N <- 10^4
rr.boot <- numeric(N)

for (i in 1:N)
{
  resampleHigh <- sample(highbp, 3338, replace = TRUE)
```

```

resampleLow <- sample(lowbp, 2676, replace = TRUE)

rr.boot[i] <- mean(resampleHigh)/mean(resampleLow) #rel. risk
}

quantile(rr.boot, c(0.025, 0.975))

```

## 5.7 Monte Carlo Sampling

The key bootstrap idea is the plug-in principle – to replace the unknown population by the data, and draw samples from that.

A key detail is that we (usually) implement this by random sampling, known as “Monte Carlo sampling.”

The name *Monte Carlo* dates from the 1940s, when physicists and applied mathematicians working on the Manhattan Project at Los Alamos Laboratory in New Mexico encountered difficult integrals with no closed form solutions. Stanislaw Ulam and John von Neumann proposed using computer simulations to estimate these integrals. Their conceptual leap was in using a random method to solve a deterministic problem. Because of the use of randomness, they named the method after the casino in Monaco.

Given that we are drawing i.i.d. samples of size  $n$  from the observed data, there are at most  $n^n$  possible samples ( $\binom{2n-1}{n}$ , if we disregard the order of observations), and ties in the data can further reduce the number of unique samples. In small samples, we could create all possible bootstrap samples, deterministically. In practice,  $n$  is usually too large for that to be feasible, so we use random sampling.

Let  $N$  be the number of bootstrap samples used, for example,  $N = 10^4$ . The resulting  $N$  resample statistic values represent a random sample of size  $N$  with replacement from the *theoretical bootstrap distribution* (or *exhaustive bootstrap distribution*) consisting of  $n^n$  values.

In some cases, we can calculate some aspects of the sampling distribution without simulation. For example, for the ordinary one-sample bootstrap of a sample mean, the mean and standard deviation of the theoretical bootstrap distribution are  $\bar{x}$  and  $\hat{\sigma}/\sqrt{n}$ , respectively, where  $\hat{\sigma}^2 = (1/n) \sum (x_i - \bar{x})^2$  is the variance of the data distribution.

Monte Carlo sampling causes unwanted variability, that may be reduced by increasing the value of number of resamples  $N$ . We discuss how large  $N$  should be in Section 5.9.

## 5.8 Accuracy of Bootstrap Distributions

How accurate is the bootstrap? This entails two questions:

- How accurate is the theoretical bootstrap?
- How accurately does the Monte Carlo implementation approximate the theoretical bootstrap?

### Sources of Variation in a Bootstrap Distribution

Bootstrap distributions and conclusions based on them include two sources of random variation:

1. The original sample is chosen at random from the population.
2. Bootstrap resamples are chosen at random from the original sample.

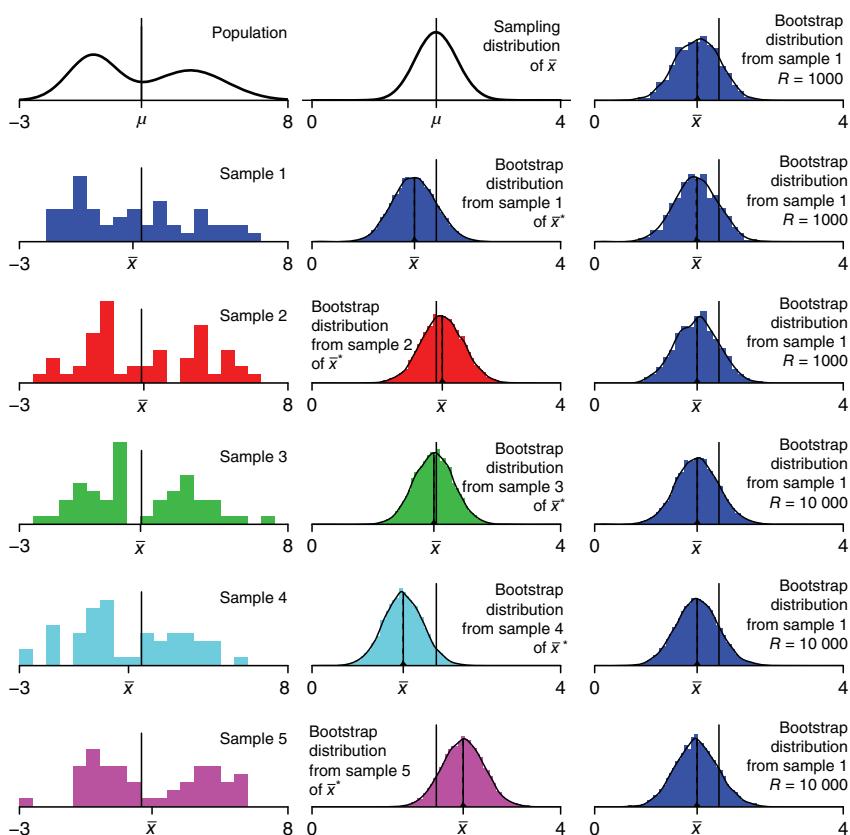
We begin this section with a series of pictures intended to illustrate both questions. We conclude this section with a discussion of cases where the theoretical bootstrap is not accurate and remedies. In Section 5.9, we return to the question of Monte Carlo accuracy.

### 5.8.1 Sample Mean, Large Sample Size

Figure 5.17 shows a population and five samples of size 50 from the population in the left column. The middle column shows the sampling distribution for the mean and bootstrap distributions from each sample, based on  $N = 10^4$  bootstrap samples. Each bootstrap distribution is centered at the statistic ( $\bar{x}$ ) from the corresponding sample rather than being centered at the population mean  $\mu$ . The spreads and shapes of the bootstrap distributions vary a bit but not a lot.

This informs what the bootstrap distributions may be used for. The bootstrap does not provide a better estimate of the population parameter  $\mu$ , because no matter how many bootstrap samples are used, they are centered at  $\bar{x}$  (plus random variation), not  $\mu$ . On the other hand, the bootstrap distributions are useful for estimating the spread and shape of the sampling distribution.

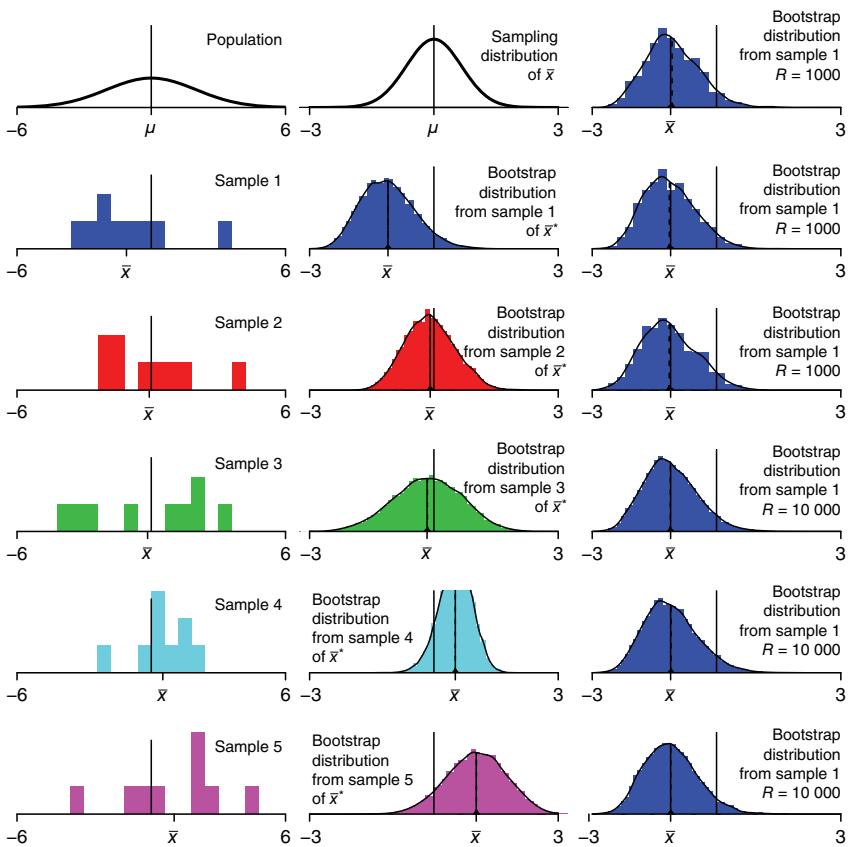
The right column shows more bootstrap distributions from the first sample, three each using 1000 and  $10^4$  resamples. These illustrate the Monte Carlo variation in the bootstrap. This variation is much smaller than the variation due to different original samples. For many uses, such as quick-and-dirty estimation of standard errors or approximate confidence intervals, 1000 resamples is adequate. However, there is noticeable variability, and the distributions are less accurate, especially in the tails; so when accuracy matters,  $10^4$  or more samples should be used.



**Figure 5.17** Bootstrap distribution for the mean,  $n = 50$ . The left column shows the population and five samples. The middle column shows the sampling distribution, and bootstrap distributions from each sample. The right column shows five more bootstrap distributions from the first sample, with  $N = 1000$  or  $N = 10^4$ .

### 5.8.2 Sample Mean: Small Sample Size

Figure 5.18 is similar to Figure 5.17, but for a smaller sample size,  $n = 9$  (and a different population). As before, the Monte Carlo variation is small and can be reduced with more samples. As before, the bootstrap distributions are centered at the corresponding sample means, but now the spreads and shapes of the bootstrap distributions vary substantially, because the spreads and shapes of the samples vary substantially. As a result, bootstrap confidence interval widths vary substantially (this is also true of nonbootstrap confidence intervals).



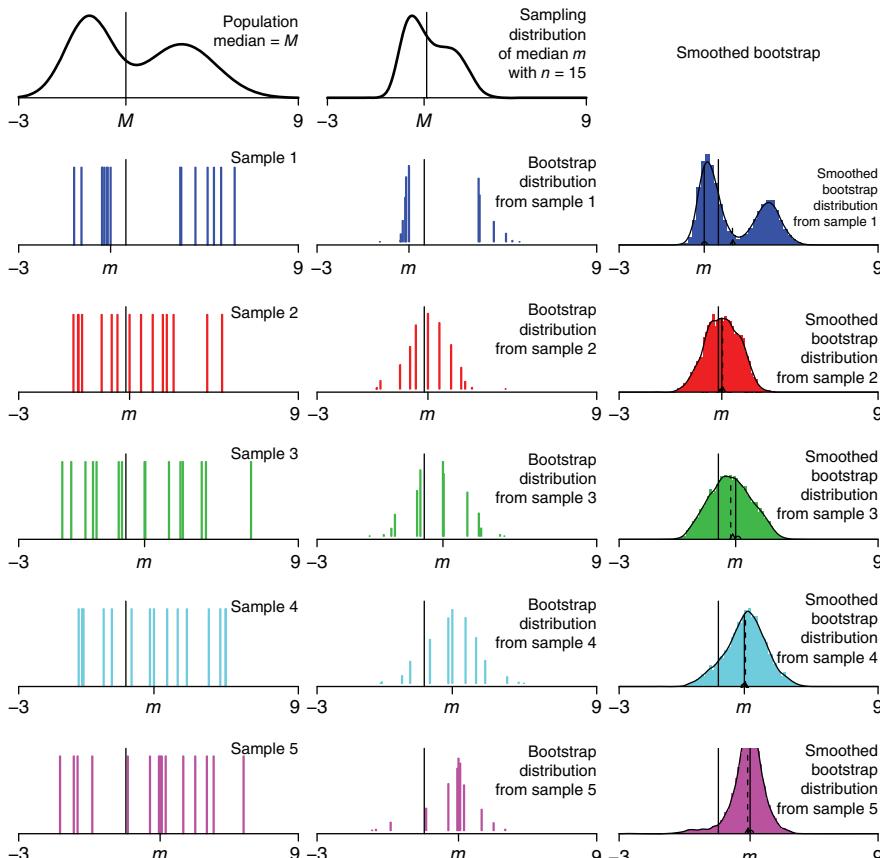
**Figure 5.18** Bootstrap distributions for the mean,  $n = 9$ . The left column shows the population and five samples. The middle column shows the sampling distribution and bootstrap distributions from each sample. The right column shows five more bootstrap distributions from the first sample, with  $N = 1000$  or  $N = 10^4$ .

**Remark** While not apparent in the pictures, bootstrap distributions tend to be too narrow on average, by a factor of  $\sqrt{(n-1)/n}$  for the sample mean, and approximately that for many other statistics. This goes back to the plug-in principle; the variance of the empirical distribution is the population variance (Section 2.2.2)  $\hat{\sigma}^2 = (1/n) \sum (x_i - \bar{x})^2$ , not the sample variance  $s^2$  with divisor of  $n-1$ . The theoretical bootstrap SE =  $\hat{\sigma}/\sqrt{n}$  is smaller than the usual formula standard error  $s/\sqrt{n}$  by a factor of  $\sqrt{(n-1)/n}$ . For instance, for the CLEC mean, the bootstrap SE is 3.96 which is smaller than  $s/\sqrt{n} = 4.07$ . ||

### 5.8.3 Sample Median

Now turn to Figure 5.19 where the statistic is the sample median. Here, the bootstrap distributions are poor approximations of the sampling distribution. The sampling distribution is continuous, but the bootstrap distributions are discrete, with the only possible values being values in the original sample (here  $n$  is odd). The bootstrap distributions are very sensitive to the sizes of gaps among the observations near the center of the sample (see Exercise 5.8).

The ordinary bootstrap tends not to work well for statistics such as the median or other quantiles that depend heavily on a small number of observations out of a larger sample.

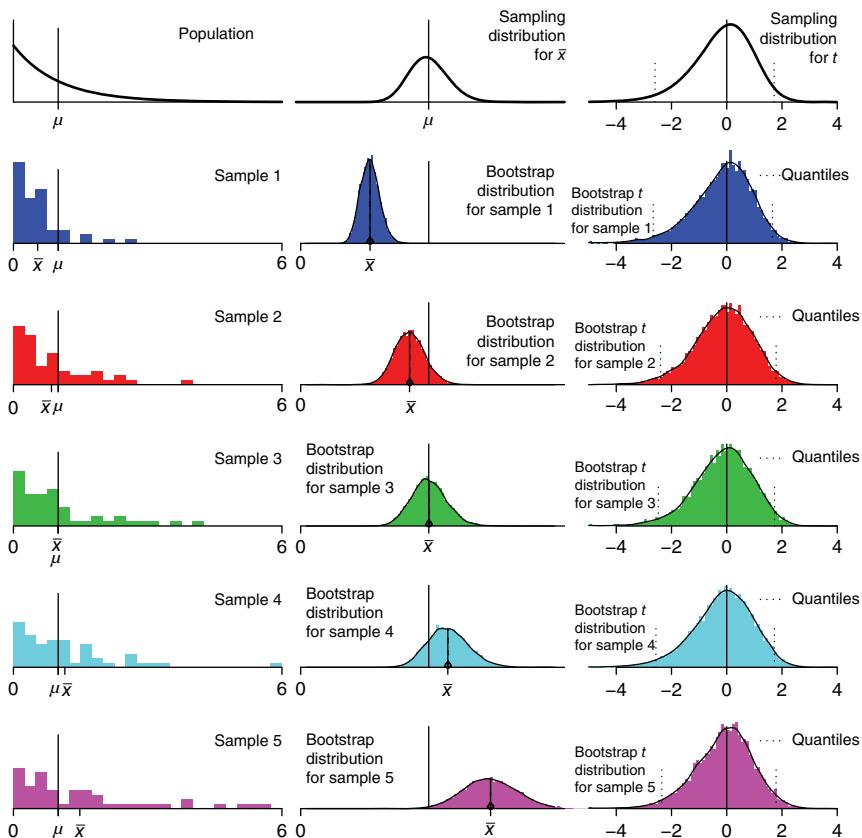


**Figure 5.19** Bootstrap distributions for the median,  $n = 15$ . The left column shows the population and five samples. The middle column shows the sampling distribution and bootstrap distributions from each sample. The right column shows smoothed bootstrap distributions, with kernel sd  $s/\sqrt{n}$ , see Sections 13.1.

### 5.8.4 Mean–Variance Relationship

In many applications, the spread or shape of the sampling distribution depends on the parameter of interest. For example, the binomial distribution spread and shape depend on  $p$ . Similarly, for an exponential distribution, the standard deviation of the sampling distribution of  $\bar{x}$  is proportional to  $\mu$ .

This mean–variance relationship is reflected in bootstrap distributions. Figure 5.20 shows samples and bootstrap distributions for an exponential population. There is a strong dependence between  $\bar{x}$  and the corresponding bootstrap SE. This relationship has important implications for confidence



**Figure 5.20** Bootstrap distributions for the mean,  $n = 50$ , exponential population. The left column shows the population and five samples. (These samples are selected from a larger set of random samples, to have means spread across the range of sample means, and average standard deviations conditional on the means.) The middle column shows the sampling distribution and bootstrap distributions for each sample. The right column shows bootstrap  $t$  distributions, see Section 7.5.2.

intervals; intervals (bootstrap or otherwise) that ignores the relationship are inaccurate, they miss too often on one side. We discuss this more in Section 7.6.

Here there is a bright spot. The right column of Figure 5.20 shows the sampling distribution and bootstrap distributions of the  $t$  statistic,  $t = (\bar{X} - \mu) / (s/\sqrt{n})$ . These distributions are much less sensitive to the original sample. We use these bootstrap  $t$  distributions in Section 7.5.2 to construct accurate confidence intervals.

### Variation in Bootstrap Distributions

For most statistics, almost all the variation in bootstrap distributions comes from randomly selecting the original sample from the population. Reducing this variation requires collecting a larger original sample.

Bootstrapping does not overcome the weakness of small samples as a basis for inference. Some bootstrap procedures are more accurate than others (we will discuss this later) and more accurate than common nonbootstrap procedures, but still they may not be accurate for very small samples. Use caution in any inference – including bootstrap inference – from a small sample.

About a total of 1000 resamples is enough for approximate answers, but for good accuracy use 10 000 or more.

## 5.9 How Many Bootstrap Samples Are Needed?

We see in Figures 5.17–5.19 that using 1000 or 10 000 bootstrap samples give similar bootstrap distributions, but that 10 000 samples are more accurate. We elaborate on this here. The focus here is on Monte Carlo accuracy – how well the usual random sampling implementation of the bootstrap approximates the theoretical bootstrap distribution.

A bootstrap distribution based on  $N$  random samples corresponds to drawing  $N$  observations with replacement from the theoretical bootstrap distribution.

Brad Efron, inventor of the bootstrap, suggested in 1993 that  $N = 200$ , or even as few as  $N = 25$ , suffices for estimating standard errors and that  $N = 1000$  is enough for confidence intervals (Efron and Tibshirani, 1993).

We argue that more resamples are appropriate, on two grounds. First, those criteria were developed when computers were much slower; with faster computers it is much easier to take more resamples.

Second, those criteria were developed using arguments that combine the random variation due to the original sample with the random variation due to bootstrap sampling. We prefer to treat the data as given and look just at

the variability due to bootstrap sampling. Two people analyzing the same data set should not get substantially different results due to random bootstrap sampling. For typical 95% bootstrap percentile confidence intervals, to reduce Monte Carlo variability to the point that a supposed 95% confidence interval has a high probability of missing between 2.25% and 2.75% on each side requires about 15 000 bootstrap samples. (See Exercise 5.29.)

Thus, for routine practice, we recommend at least 10 000 bootstrap resamples, and more when accuracy matters.

## Exercises

For all exercises that ask you to perform exploratory data analysis (EDA), you should plot the data (histogram, normal quantile plots), describe the shape of the distribution (bell-shaped, symmetric, skewed, etc.), and provide summary statistics (mean, standard deviation). For bootstrapping questions, always provide plots and describe the shape, spread, and bias of the distribution.

- 5.1** Consider the sample 2, 4, 5, 9 from some distribution. Which of the following are bootstrap samples from this sample?
  - (a) 2, 2, 2, 2
  - (b) 2, 5, 6, 4
  - (c) 4, 9, 4, 9
  - (d) 4, 5, 5, 9, 9
- 5.2** Consider the sample 1, 3, 4, 6 from some distribution. Give three different bootstrap samples and their corresponding means.
- 5.3** Consider the sample 1, 2, 3. How many different bootstrap samples are there? (Calculate this as if order matters.)
- 5.4** Consider the sample 1, 3, 8, 5 from some distribution. How many different bootstrap samples are there? (Calculate this as if order matters.)
- 5.5** Refer to Example 5.4 and the remark at the end of the example.
  - (a) What might account for the fact that there were more missing values for the men who skateboarded in front of the male experimenter? How might this bias the outcome?
  - (b) Why do you suppose it was important that the two experimenters were blinded to the purpose of the study?
  - (c) Participants were not asked if they were heterosexual or homosexual. Discuss how this might bias the results.

- 5.6** Consider a population that has a normal distribution with mean  $\mu = 36$ , standard deviation  $\sigma = 8$ .
- The sampling distribution of  $\bar{X}$  for samples of size 200 will have what mean, standard error, and shape?
  - Use R to draw a random sample of size 200 from this population. Conduct EDA on your sample.
  - Compute the bootstrap distribution for your sample and note the bootstrap mean and standard error.
  - Compare the bootstrap distribution to the theoretical sampling distribution by creating a table like Table 5.1.
  - Repeat for sample sizes of  $n = 50$  and  $n = 10$ . Carefully describe your observations about the effects of sample size on the bootstrap distribution.
- 5.7** Consider a population that has a gamma distribution with parameters  $r = 5, \lambda = 1/4$ .
- Use simulation (with  $n = 200$ ) to generate an approximate sampling distribution of the mean; plot and describe the distribution.
  - Now, draw one random sample of size 200 from this population. Create a histogram of your sample and find the mean and standard deviation.
  - Compute the bootstrap distribution of the mean for your sample, plot it, and note the bootstrap mean and standard error.
  - Compare the bootstrap distribution to the approximate theoretical sampling distribution by creating a table like Table 5.1.
  - Repeat (a)–(e) for sample sizes of  $n = 50$  and  $n = 10$ . Describe carefully your observations about the effects of sample size on the bootstrap distribution.
- 5.8** We investigate the bootstrap distribution of the median. Create random samples of size  $n$  for various  $n$  and bootstrap the median. Describe the bootstrap distribution.

```

ne <- 14 # n even
no <- 15 # n odd

wwe <- rnorm(ne) # draw random sample of size ne
wwo <- rnorm(no) # draw random sample of size no

N <- 10^4
even.boot <- numeric(N) # save space
odd.boot <- numeric(N)
for (i in 1:N)

```

```

{
  x.even <- sample(wwe, ne, replace = TRUE)
  x.odd <- sample(wwo, no, replace = TRUE)
  even.boot[i] <- median(x.even)
  odd.boot[i]  <- median(x.odd)
}

df <- data.frame(even.boot, odd.boot)
p1 <- ggplot(df, aes(even.boot)) +
  geom_histogram(bins = 10, color = "white") +
  xlim(c(-2,2))
p2 <- ggplot(df, aes(odd.boot)) +
  geom_histogram(bins = 10, color = "white") +
  xlim(c(-2,2))

library(gridExtra)
grid.arrange(p1, p2)

```

Change the sample sizes to 36 and 37; 200 and 201; 10 000 and 10 001. Note the similarities/dissimilarities, trends, and so on. Why does the parity of the sample size matter? (Note: Adjust the  $x$  limits in the plots as needed.)

- 5.9** Import the data from data set Bangladesh. In addition to arsenic concentrations for 271 wells, the data set contains cobalt and chlorine concentrations.

- (a) Conduct EDA on the chlorine concentrations and describe the salient features.
- (b) Bootstrap the mean.
- (c) Find and interpret the 95% bootstrap percentile confidence interval.
- (d) What is the bootstrap estimate of the bias? What fraction of the bootstrap standard error does it represent?

The Chlorine variable has some missing values. The following code will remove these entries, resulting in a vector:

```

library(tidyr)
chlorine <- drop_na(Bangladesh, Chlorine) %>% pull(Chlorine)

```

- 5.10** Consider Bangladesh chlorine (concentration). Bootstrap the trimmed mean (say, trim the upper and lower 25%) and compare your results with the usual mean (previous exercise).

- 5.11** The data set MnGroundwater contains measurements of various chemicals found in 895 randomly selected wells in Minnesota.

- (a) Perform some EDA of arsenic levels (ppb) and describe the distribution.

- (b) Compute 95% bootstrap percentile intervals of the ordinary mean, the 10% trimmed mean, and the midmean (25% trimmed mean). Which of these intervals is widest? Narrowest? Why do you think this is?
- (c) Compute bootstrap distributions for two measures of scale:  $s$  and IQR (inter-quartile range). The IQR may be preferred here because it is less sensitive to outliers. Assess this by comparing the *coefficient of variation* for each – the standard error divided by the mean of the bootstrap distribution. *Note:* We do not just compare the standard errors for these statistics because they do not measure exactly the same thing, e.g. for a normal distribution  $\text{IQR} = 1.35\sigma$ .
- 5.12** The data set `FishMercury` contains mercury levels (parts per million) for 30 fish caught in lakes in Minnesota.
- Create a histogram or boxplot of the data. What do you observe?
  - Bootstrap the mean and record the bootstrap standard error and the 95% bootstrap percentile interval.
  - Remove the outlier and bootstrap the mean of the remaining data. Record the bootstrap standard error and the 95% bootstrap percentile interval.
  - What effect did removing the outlier have on the bootstrap distribution, in particular, on the standard error?
- 5.13** In Section 3.3, we performed a permutation test to determine if men and women consumed, on average, different amounts of hot wings.
- Bootstrap the difference in means and describe the bootstrap distribution.
  - Find a 95% bootstrap percentile confidence interval for the difference of means and give a sentence interpreting this interval.
  - How do the bootstrap and permutation distributions differ?
- 5.14** A high school student was curious about the total number of minutes devoted to commercials during any given half-hour time period on basic and extended cable TV channels (B. Rodgers and T. Robinson, private communication). Import the data `TV`.
- Perform some exploratory data analysis and obtain summary statistics on the commercial times on basic and extended cable TV channels (do separate analyses for each type of channel).
  - Bootstrap the difference in mean times, plot the distribution and give summary statistics of the bootstrap distribution. Obtain a 95% bootstrap percentile confidence interval and interpret this interval.
  - What is the bootstrap estimate of the bias? What fraction of the bootstrap standard error does this represent?

- (d) Conduct a permutation test to see if the difference in mean commercial times is statistically discernible and state your conclusion.
- 5.15** Researchers conducted a study of primary and early secondary school children in Italy to examine gender differences in math anxiety (Hill et al., 2016). One of the measures used to understand math anxiety is the *abbreviated math anxiety scale (AMAS)*, a self-reported math anxiety questionnaire. A higher score indicates more math anxiety. The data set `MathAnxiety` contains the results for a subset of the children in the original study.<sup>4</sup>
- (a) Perform some exploratory analysis and obtain summary statistics of the AMAS scores for the boys and girls (do separate analyses for each gender).
  - (b) Bootstrap the difference in mean scores, plot the Distribution, and give summary statistics of the bootstrap distribution. Obtain a 95% bootstrap percentile confidence interval and interpret this interval.
  - (c) What is the bootstrap estimate of the bias? What fraction of the bootstrap standard error does this represent?
  - (d) Conduct a permutation test to see if the difference in mean AMAS scores is statistically discernible, and state your conclusion.
- 5.16** Import the data from `Girls2004` (see Section 1.2).
- (a) Perform some exploratory data analysis and obtain summary statistics on the weights of baby girls born in Wyoming and Alaska (do separate analyses for each state).
  - (b) Bootstrap the difference in means, plot the distribution, and give the summary statistics. Obtain a 95% bootstrap percentile confidence interval and interpret this interval.
  - (c) What is the bootstrap estimate of the bias? What fraction of the bootstrap standard error does it represent?
  - (d) Conduct a permutation test to see if the difference in mean weights is statistically discernible and state your conclusion.
  - (e) For what population(s), if any, does this conclusion hold? Explain.
- 5.17** The data set `MnGroundwater` contains measurements of various chemicals for 895 randomly chosen wells in Minnesota. The variable `Arsenic` measures the amount of arsenic in parts per billion (ppb).
- (a) Obtain summary statistics for the arsenic levels.
  - (b) The US Environmental Protection Agency sets the maximum arsenic level in public water supplies to be 10 ppb. What proportion of wells exceed this threshold?

---

<sup>4</sup> We have the data for five of the seven school types originally studied.

- (c) Find a 95% bootstrap percentile confidence interval of the true proportion of wells with arsenic levels that exceed 10 ppb.
- 5.18** In 2021, the Harris Poll conducted a survey of 1600 teenagers (ages 13–19 years) and found that 74% of respondents believe that “the government should provide high-speed internet access to Everyone.”<sup>5</sup> Find a 95% bootstrap percentile interval for the true proportion of teenagers who believe that high-speed Internet should be provided by the government.
- 5.19** Refer to Exercise 3.19 where we investigated the proportion of newborn babies in Alaska and in Wyoming weighing less than 2747 g (the 10 percentile mark for baby girls). Compute a 95% bootstrap percentile interval for the difference in proportions and give a sentence interpreting this interval.
- 5.20** The *Toxoplasma gondii* parasite, a common parasite often found in chickens, can result in toxoplasmosis, a disease that may cause flu-like symptoms in humans. Humans typically contract the disease by eating undercooked meat. Researchers in China tested for *T. gondii* antibodies in 160 free-range chickens and 450 caged chickens (Xu et al., 2012). They found that 18.8% of the free-range chickens and 5.6% of the caged chickens tested positive. Compute a 95% bootstrap percentile interval for the difference in proportions and give a sentence interpreting this interval.
- 5.21** Refer to Exercise 3.17 where we describe a randomized, double-blind study in which patients identified with a nasal *S. aureus* bacteria were assigned to a treatment of mupirocin/chlorhexidine or a placebo. Find a 95% bootstrap percentile confidence interval, for the difference in the proportions of those who had hospital-acquired *S. aureus* infections between the two groups. Give a sentence interpreting this interval.
- 5.22** Is there a difference in the price of groceries sold by Target and Walmart? The data set *Groceries* contain a sample of grocery items and their prices advertised on their respective websites on one specific day.
- Compute summary statistics of the prices for each store.
  - Use the bootstrap to determine whether or not there is a difference in the mean prices.
  - Create a histogram of the difference in prices. What is unusual about Quaker Oats Life cereal?

---

<sup>5</sup> <https://4-h.org/wp-content/uploads/2021/08/4-H-Digital-Divide-Survey-Report-2021.pdf>.

- (d) Recompute the bootstrap percentile interval without this observation. What do you conclude?
- 5.23** Do chocolate and vanilla ice creams have the same number of calories? The data set `IceCream` contains calorie information for a sample of brands of chocolate and vanilla ice cream.
- Compute summary statistics of the calories for the two flavors.
  - Use the bootstrap to determine whether or not there is a difference in the mean number of calories.
- 5.24** In a remark at the end of Section 5.4.1, we mentioned that the procedure for bootstrapping the difference of medians is different than for the mean. Import the data set `Diving2017`.
- Compute the difference between the median scores in the final and semifinal rounds.
  - Run the code below to obtain a 95% bootstrap percentile interval of the median of the difference in scores.

```
N <- 10^5
result <- numeric(N)
for (i in 1:N)
{
  index <- sample(12, replace = TRUE)      #resample row numbers
  Dive.boot <- Diving2017[index, ]          #pairs in resample
  result[i] <- median(Dive.boot$Final) - median(Dive.boot$Semifinal)
}
df <- data.frame(result)
ggplot(df, aes(result)) + geom_histogram(bins = 10, color = "white")
quantile(result, c(0.025, 0.975))
```

- 5.25** Two college students collected data on the price of hardcover textbooks from two disciplinary areas: Mathematics and the Natural Sciences, and the Social Sciences (R. Hien and S. Backer, private communication). The data are in the file `BookPrices`.
- Perform some exploratory data analysis on book prices for each of the two disciplinary areas.
  - Bootstrap the mean of book price for each area separately and describe the distributions.
  - Bootstrap the ratio of means. Provide a plot of the bootstrap distribution and comment.
  - Find the 95% bootstrap percentile interval for the ratio of means. Interpret this interval.

- (e) What is the bootstrap estimate of the bias? What fraction of the bootstrap standard error does it represent?
- 5.26** Import the data from flight delays case study in Section 1.1 data into R. Although the data represent all UA and AA flights in May and June 2009, we will assume they represent a sample from a larger population of UA and AA flights flown under similar circumstances. We will consider the ratio of means of the flight delay lengths,  $\mu_{UA}/\mu_{AA}$ .
- Perform some exploratory data analysis on flight delay lengths for each of UA and AA flights.
  - Bootstrap the mean of flight delay lengths for each airline separately and describe the distribution.
  - Bootstrap the ratio of means. Provide plots of the bootstrap distribution and describe the distribution.
  - Find the 95% bootstrap percentile interval for the ratio of means. Interpret this interval.
  - What is the bootstrap estimate of the bias? What fraction of the bootstrap standard error does it represent?
  - For inference in this text, we assume that the observations are independent. Is that condition met here? Explain.
- 5.27** The file Cafeteria contains measurements on ingredients in a sample of dishes served in a college cafeteria (R. Stephenson, private communication); the dishes are also classified by type: meat or vegetarian.
- Perform some exploratory data analysis and obtain summary statistics on the amount of protein in the meat and vegetarian dishes.
  - Bootstrap the ratio of the mean protein amounts, plot the distribution, and give summary statistics. Obtain a 95% bootstrap percentile confidence interval and interpret this interval.
  - What is the bootstrap estimate of the bias? What fraction of the bootstrap standard error does this represent?
- 5.28** During the COVID-19 pandemic, the Johnson & Johnson vaccine was one of three vaccines approved for emergency use in the United States. Janssen Biotech, Inc., the company that developed the vaccine, submitted the results of their clinical trial to the Food and Drug Administration (FDA) in 2021. In the US part of the trial, after at least 28 days from the shot, 32 of the 8959 participants who received the vaccine became infected with a moderate or severe case of the virus compared to 112 of the 8835 who received the placebo. Thus,  $(32/8959)/(112/8835) = 0.282$  so that a vaccinated person is 0.282 times less likely to become infected than a vaccinated person. The *efficacy* of

the vaccine is defined to be 1 – this relative risk, or  $1 - 0.282 = 0.718$ , so we state that the vaccine is 71.8% effective against infection.

- (a) Compute a 95% bootstrap percentile interval of the efficacy of the Johnson & Johnson vaccine.

Refer to Example 5.9 for the R code and check you answer on page 37 of the FDA Briefing Document <https://www.fda.gov/media/146217/download>.

- (b) What is the bootstrap estimate of the bias? What fraction of the bootstrap standard error does it represent?

- 5.29** This exercise provides an explanation of the fact stated in Section 5.9 that for a 95% bootstrap percentile interval, we should use at least 15 000 resamples if we want a high probability of missing between  $2.5\% \pm 0.25\%$  on each side of the interval.

Let  $\alpha$  satisfy  $P(\hat{\theta}^* > \alpha) = 0.025$ ; that is,  $\alpha$  is the upper endpoint of the theoretical bootstrap percentile confidence interval. We want the simulation to be accurate enough that the estimated bootstrap probability of exceeding  $\alpha$  is close to 0.025, where “close” means within 10% of the true value, i.e. between  $0.025 - 0.0025$  and  $0.025 + 0.0025$ .

Let  $r$  be the number of resamples, and  $X$  the number of resamples satisfying  $(\hat{\theta}_i^* > \alpha)$ .

- (a) Explain why  $X$  is binomial with parameters  $(r, 0.025)$ .  
 (b) Explain why the sample proportion  $X/r$  has mean 0.025 and variance  $\sigma^2 = (0.025 \times 0.975)/r$ .  
 (c) Verify that

$$\begin{aligned} P(0.025 - 0.0025 < X/r < 0.025 + 0.0025) \\ \approx P(-0.0025/\sigma < Z < 0.0025/\sigma) \\ = 1 - 2 \cdot P(Z > .0025/\sigma). \end{aligned}$$

- (d) If “high probability” is 95%, then verify that  $r \geq 1.96^2 \times 0.025 \times 0.975 / 0.0025^2 = 14\,982$ . The latter number is about 15 000.



# 6

## Estimation

In earlier chapters, we used the sample mean to estimate the true population mean and the sample proportion to estimate the true population proportion. These are very reasonable choices. They are examples of “plug-in” estimators; to estimate something about the population, we may use the corresponding statistic from the sample. But it is more complicated than it appears at first glance. For a symmetric distribution, the sample mean and sample median are both plug-in estimators for the middle of the distribution, but they give different answers. Which should we use? We also usually use the sample standard deviation (Equation (2.1)) defined using a divisor of  $n - 1$  instead of the plug-in estimator with a divisor of  $n$ . Why?

In this chapter, we will examine these choices, discuss what makes good estimators, introduce general procedures for estimating parameters, and discuss some practical issues.

### 6.1 Maximum Likelihood Estimation

We begin with a very general procedure, maximum likelihood estimation. This has a number of nice properties — among others, the answers it gives are reasonable, it never gives impossible answers, and it typically gives the answers that just make sense, and it is optimal in some ways.

Suppose a friend tells you she has a total of 25 chocolate chip or oatmeal raisin cookies in a bag. She also tells you that the number of chocolate chip cookies is either 2 or 20. If you draw out a cookie at random from the bag and see it is chocolate chip, what do you think is more likely the truth, that there are 2 or 20 chocolate cookies? Based solely on your one data point (the chocolate chip cookie), it seems that 20 chocolate cookies are more likely than 2 chocolate cookies. This is the idea behind maximum likelihood estimation — to choose the value of a parameter that is most consistent with the data.

### 6.1.1 Maximum Likelihood for Discrete Distributions

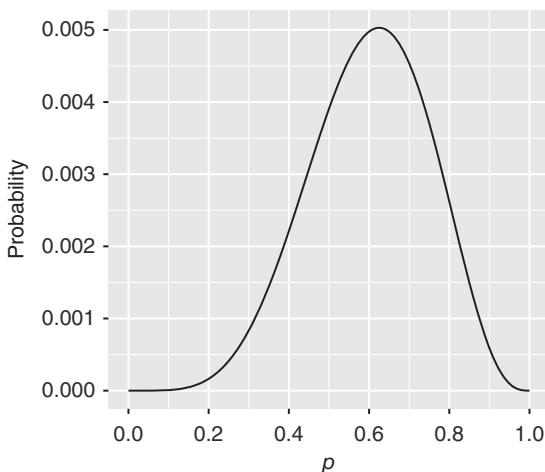
Suppose you buy a weighted coin in a magic shop. It looks like a typical coin with heads and tails, but the coin is not necessarily fair. You flip the coin eight times and observe the sequence HHHTHTHT. What would be a good estimate for  $p$ , the probability of heads, based on this sequence? Let  $X_i \sim \text{Bern}(p)$ , where  $X_i = 1$ ,  $i = 1, 2, \dots, 8$ , indicates heads. We will assume that the flips are independent. Then,

$$\begin{aligned} P(X_1 = 1, X_2 = 1, X_3 = 1, X_4 = 0, X_5 = 1, X_6 = 0, X_7 = 1, X_8 = 0) \\ = P(X_1 = 1)P(X_2 = 1)P(X_3 = 1)P(X_4 = 0) \dots P(X_8 = 0) \\ = p^5(1 - p)^3. \end{aligned}$$

We want to find the value of  $p$  that is most consistent with the observed data. One way to do that is to find the  $p$  that has the highest probability of giving the observed data—that is, the  $p$  that maximizes the likelihood  $L(p) = p^5(1 - p)^3$ . Using calculus, we set the derivative equal to zero,  $L'(p) = 5p^4(1 - p)^3 + p^53(1 - p)^2(-1) = 0$  and solve for  $p$  to obtain  $p = 5/8$  as the most likely candidate for  $p$ .

The function  $L(p) = p^5(1 - p)^3$  is called the *likelihood function* for the parameter  $p$  (Figure 6.1) and the estimate  $\hat{p} = 5/8$  is the *maximum likelihood estimate* for  $p$ .

**Definition 6.1** Let  $f(x; \theta)$  denote the probability mass function for a discrete distribution with associated parameter  $\theta$ . Suppose  $X_1, X_2, \dots, X_n$  are a random sample from this distribution and  $x_1, x_2, \dots, x_n$  are the actual observed values.



**Figure 6.1** Likelihood for  $p$ , after five heads and three tails.

Then, the *likelihood function* of  $\theta$  is

$$\begin{aligned} L(\theta \mid x_1, x_2, \dots, x_n) &= P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) \\ &= P(X_1 = x_1)P(X_2 = x_2) \dots P(X_n = x_n) \\ &= f(x_1; \theta)f(x_2; \theta) \dots f(x_n; \theta) \\ &= \prod_{i=1}^n f(x_i; \theta). \end{aligned}$$

The likelihood function is a function of  $\theta$  and is sometimes written as  $L(\theta) = L(\theta \mid x_1, x_2, \dots, x_n)$ .

A maximum likelihood estimate is a value  $\hat{\theta}$  such that  $L(\hat{\theta}) \geq L(\theta)$  for all  $\theta$ . ||

In practice, we usually maximize the log-likelihood

$$\ln(L) = \sum_{i=1}^n \ln(f(x_i; \theta)).$$

### Likelihood Function, Maximum Likelihood Estimate

The likelihood function  $L(\theta) = L(\theta \mid x_1, x_2, \dots, x_n)$  gives the likelihood of  $\theta$ , given the data. A maximum likelihood estimate (MLE)  $\hat{\theta}_{\text{MLE}}$  is a value of  $\theta$  that maximizes the likelihood, or equivalently that maximizes the log-likelihood  $\ln(L)$ .

Even though  $L(\theta \mid x_1, x_2, \dots, x_n)$  is equal to an expression involving  $f(x_i; \theta)$ , we think of the two functions differently. When we consider the density  $f(x; \theta)$ , we think of  $x$  as variable and  $\theta$  as fixed, whereas when we consider the likelihood  $L(\theta \mid x_1, x_2, \dots, x_n)$ , we think of  $\theta$  as variable and the  $x_i$ 's as fixed.

**Proposition 6.1** Let  $X_1, X_2, \dots, X_n$  denote  $n$  independent Bernoulli random variables,  $\text{Bern}(p)$ ,  $0 < p < 1$ . Let  $X = \sum_{i=1}^n X_i$ , the number of 1's. The maximum likelihood estimator of  $p$  is  $\hat{p} = X/n$ .

*Proof.* Assume  $X_i = 1$  with probability  $p$ ,  $i = 1, 2, \dots, n$  and let  $x = \sum_{i=1}^n x_i$ .

$$\begin{aligned} L(p) &= P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) \\ &= \prod_{i=1}^n P(X_i = x_i) \\ &= \prod_{x_i=1} p^{x_i} \prod_{x_i=0} (1-p)^{x_i} \\ &= p^{\sum_{i=1}^n x_i} (1-p)^{n - \sum_{i=1}^n x_i} \\ &= p^x (1-p)^{n-x}. \end{aligned}$$

Solving  $L'(p) = 0$  for  $p$  yields  $\hat{p} = x/n$  as the only critical value of  $L$ . A check of the graph of  $L(p)$  or the first derivative test confirms that  $\hat{p}$  is the value that maximizes  $L(p)$ . Thus,  $\hat{p} = X/n$  is the maximum likelihood estimator of  $p$ .  $\square$

**Remark** We use the term *estimator* for a function of random variables (see Definition 4.2), but we call the result for a set of observations,  $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$ , an *estimate*.

An estimator like  $\bar{X}$  is a rule that can be applied to new random data; an estimate like  $\bar{x}$  is a specific value for a given set of data.  $\parallel$

**Example 6.1** Suppose  $x_1 = 3, x_2 = 4, x_3 = 3, x_4 = 7$  come from a Poisson distribution with unknown  $\lambda$ . The probability mass function is  $f(x; \lambda) = \lambda^x e^{-\lambda} / x!$ ,  $x = 0, 1, 2, \dots$ , so the likelihood function is

$$\begin{aligned} L(\lambda) &= f(x_1; \lambda) f(x_2; \lambda) f(x_3; \lambda) f(x_4; \lambda) \\ &= \frac{\lambda^3 e^{-\lambda}}{3!} \frac{\lambda^4 e^{-\lambda}}{4!} \frac{\lambda^3 e^{-\lambda}}{3!} \frac{\lambda^7 e^{-\lambda}}{7!} \\ &= \frac{\lambda^{17} e^{-4\lambda}}{3! 4! 3! 7!}. \end{aligned}$$

Take the logarithm of each side, then differentiate with respect to  $\lambda$ :

$$\ln(L(\lambda)) = 17 \ln(\lambda) - 4\lambda - \ln(3! 4! 3! 7!)$$

$$(\ln(L(\lambda)))' = 17 \frac{1}{\lambda} - 4.$$

Setting the derivative equal to 0, we find the maximum occurs at  $\hat{\lambda} = 17/4$ . The first derivative test confirms that this is a global maximum.

Notice that  $17/4 = \bar{x}$ ; this is not a coincidence!  $\square$

More generally, the maximum likelihood estimator for the Poisson is always the mean.

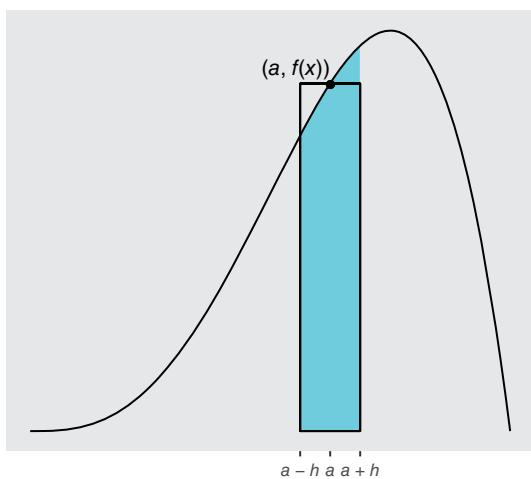
**Proposition 6.2** Let  $x_1, x_2, \dots, x_n$  be a random sample from a Poisson distribution with unknown parameter  $\lambda$ . Then the maximum likelihood estimate of  $\lambda$  is  $\hat{\lambda} = \bar{x}$ , the sample mean.

*Proof.* Exercise.  $\square$

### 6.1.2 Maximum Likelihood for Continuous Distributions

Now, consider the case of data  $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$  from a continuous distribution. If we were to mimic the discrete case, we would compute

**Figure 6.2** Estimate of area under curve by rectangle.



$P(X_1 = x_1)P(X_2 = x_2) \dots P(X_n = x_n)$ . But  $P(X = a) = 0$  for continuous random variables.

Let  $f(x)$  denote the pdf of the continuous random variable  $X$  and recall the interpretation of the integral as area under the graph of  $y = f(x)$ . Then, for small  $h > 0$ ,

$$P(a - h < X < a + h) = \int_{a-h}^{a+h} f(x)dx \approx 2hf(a).$$

Thus,  $P(a - h < X < a + h)$  is approximately proportional to  $f(a)$  (Figure 6.2).

**Definition 6.2** Let  $f(x; \theta)$  denote the pdf of a continuous random variable with associated parameter  $\theta$ . Suppose  $X_1, X_2, \dots, X_n$  are a random sample from this distribution and  $x_1, x_2, \dots, x_n$  are the corresponding observed values. Then, the *likelihood function* of  $\theta$  is

$$L(\theta | x_1, x_2, \dots, x_n) = f(x_1; \theta)f(x_2; \theta) \dots f(x_n; \theta). \quad (6.1)$$

||

A maximum likelihood estimate is a statistic  $\hat{\theta}$  that maximizes  $L$ :  $L(\hat{\theta}) \geq L(\theta)$  for all  $\theta$ . Equivalently, it maximizes the log-likelihood  $\ln(L)$ .

**Example 6.2** Let  $X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4$  be independent from an exponential distribution with pdf  $f(x; \lambda) = \lambda e^{-\lambda x}$ ,  $x \geq 0$ . Find a maximum likelihood estimate for  $\lambda$ .

**Solution**

$$\begin{aligned} L(\lambda \mid x_1, x_2, x_3, x_4) &= f(x_1; \lambda) f(x_2; \lambda) f(x_3; \lambda) f(x_4; \lambda) \\ &= \lambda e^{-\lambda x_1} \lambda e^{-\lambda x_2} \lambda e^{-\lambda x_3} \lambda e^{-\lambda x_4} \\ &= \lambda^4 e^{-\lambda(x_1+x_2+x_3+x_4)}. \end{aligned}$$

Take the log and differentiate with respect to  $\lambda$ :

$$\begin{aligned} \ln(L) &= 4 \ln(\lambda) - \lambda(x_1 + x_2 + x_3 + x_4) \\ (\ln(L))' &= \frac{4}{\lambda} - (x_1 + x_2 + x_3 + x_4). \end{aligned}$$

Solving  $(\ln(L))' = 0$  for  $\lambda$  yields  $\hat{\lambda} = 1/((1/4) \sum_{i=1}^4 x_i) = 1/\bar{x}$ . □

**Example 6.3** Let  $X_1, X_2, \dots, X_n$  denote a random sample from a distribution with pdf  $f(x; \theta) = \theta x^{\theta-1}$  for  $0 < x < 1$ , where  $\theta > 0$ . Find the MLE of  $\theta$ .

**Solution**

$$\begin{aligned} L(\theta) &= f(X_1; \theta) f(X_2; \theta) \dots f(X_n; \theta) \\ &= \prod_{i=1}^n \theta X_i^{\theta-1} \\ &= \theta^n \prod_{i=1}^n X_i^{\theta-1}. \end{aligned}$$

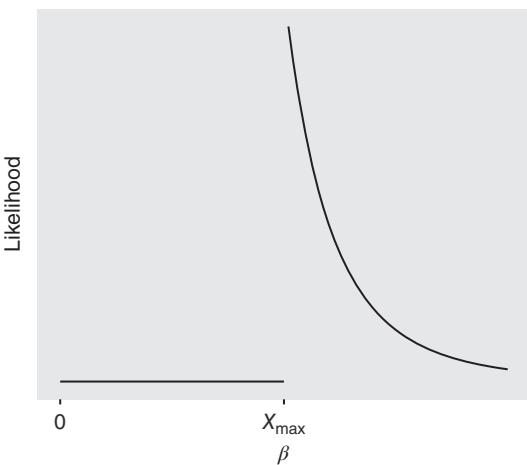
Take the log and differentiate with respect to  $\theta$ :

$$\begin{aligned} \ln(L(\theta)) &= n \ln(\theta) + (\theta - 1) \sum_{i=1}^n \ln(X_i) \\ (\ln(L(\theta)))' &= \frac{n}{\theta} + \sum_{i=1}^n \ln(X_i). \end{aligned}$$

Setting this equal to 0 and solving for  $\theta$  yields the estimator  $\hat{\theta} = -n / \sum_{i=1}^n \ln(X_i)$ .

Thus, for example if  $X_1 = 0.35, X_2 = 0.28, X_3 = 0.41$ , then the maximum likelihood estimate of  $\theta$  is  $\hat{\theta} = -3 / (\ln(0.35) + \ln(0.28) + \ln(0.41)) = 0.9333$ . □

**Example 6.4** Let  $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Unif}[0, \beta]$  with pdf  $f(x; \beta) = 1/\beta$ , where  $0 \leq x \leq \beta$ . Find the MLE of  $\beta$ .

**Figure 6.3** Likelihood for  $\beta$ .**Solution**

For  $0 \leq X_i \leq \beta$ , we have

$$\begin{aligned} L(\beta) &= f(X_1; \beta)f(X_2; \beta) \dots f(X_n; \beta) \\ &= \frac{1}{\beta} \frac{1}{\beta} \dots \frac{1}{\beta} \\ &= \left(\frac{1}{\beta}\right)^n \end{aligned} \tag{6.2}$$

(and  $L(\beta) = 0$  if any  $X_i > \beta$ ).

This function is positive as long as  $X_1, X_2, \dots, X_n \in [0, \beta]$  – in other words, as long as  $\beta \geq \max\{X_1, X_2, \dots, X_n\}$ . The function is zero if  $\beta < X_{\max}$ , jumps at  $\beta = X_{\max}$ , and then is positive. Where the function is positive, the derivative is  $L'(\beta) = -n(1/\beta^{n+1})$ . This negative derivative means the likelihood function is decreasing after the jump, so the maximum of the function occurs right at the jump,  $\hat{\beta} = \max\{X_1, X_2, \dots, X_n\}$  (see Figure 6.3).  $\square$

This next example shows that a maximum likelihood estimate may not always exist.

**Example 6.5** Suppose instead, we consider  $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Unif}(0, \beta)$  – that is, uniform on the open interval  $(0, \beta)$ . The likelihood function is nearly the same as Equation (6.2), except we need to assume strict inequalities,  $0 < X_i < \beta$  for  $i = 1, 2, \dots, n$ .

Again,  $L$  is as large as possible when  $\beta$  is as small as possible, so  $\beta$  must be bigger than *but not equal to* the maximum of  $X_1, X_2, \dots, X_n$ . Since  $\beta$  can be made arbitrarily close to this maximum, an MLE solution does not exist.  $\square$

**Example 6.6** The lifetime of an automobile tire is measured in miles rather than time. Suppose a tire company produces three versions of a tire: a standard tire whose lifetime  $X_s$  has an exponential distribution with mean  $1/\lambda > 0$ , an economy version whose lifetime  $X_e$  has an exponential distribution with mean  $0.77/\lambda$ , and a premium tire with a lifetime  $X_p$  whose distribution is exponential with mean  $1.25/\lambda$ . Suppose one tire of each type is chosen randomly and independently and tested to find its lifetime and the lifetime of each is  $x_s = 28$ ,  $x_e = 25$ , and  $x_p = 31$  (in thousands of miles). Find the MLE of  $\lambda$ .

### Solution

We have  $X_s \sim \text{Exp}(\lambda)$ ,  $X_e \sim \text{Exp}(1.3\lambda)$ , and  $X_p \sim \text{Exp}(0.8\lambda)$ , so the likelihood is

$$\begin{aligned} L(\lambda) &= f_s(X_s; \lambda) f_e(X_e; \lambda) f_p(X_p; \lambda) \\ &= \lambda e^{-\lambda X_s} (1.3\lambda e^{-1.3\lambda X_e}) (0.8\lambda e^{-0.8\lambda X_p}) \\ &= 1.04\lambda^3 e^{-(X_s + 1.3X_e + 0.8X_p)\lambda}. \end{aligned}$$

The log-likelihood and its derivative with respect to  $\lambda$  are

$$\begin{aligned} \ln(L(\lambda)) &= \ln(1.04) + 3\ln(\lambda) - (X_s + 1.3X_e + 0.8X_p)\lambda, \\ (\ln(L(\lambda)))' &= \frac{3}{\lambda} - (X_s + 1.3X_e + 0.8X_p). \end{aligned}$$

Setting the derivative equal to 0 yields the estimator  $\hat{\lambda} = 3/(X_s + 1.3X_e + 0.8X_p)$ ; for the given data  $\hat{\lambda} = 0.0352$ .

For standard tires, the estimated average lifetime is  $\hat{E}[X_s] = 1/\hat{\lambda} = 28.4$  thousand miles; for economy tires,  $\hat{E}[X_e] = 1/(1.3\hat{\lambda}) = 21.9$  thousand miles; for premium tires,  $\hat{E}[X_p] = 1/(0.8\hat{\lambda}) = 35.5$  thousand miles.  $\square$

In all of the examples so far where a MLE existed, we were able to find a closed-form expression for the MLE. In many situations, this is not possible.

**Example 6.7** Suppose  $X_1, X_2, \dots, X_n$  are a random sample from the Cauchy distribution with pdf  $f(x; \theta) = 1/(\pi(1 + (x - \theta)^2))$  for  $-\infty < x < \infty$ ,  $-\infty < \theta < \infty$ . The likelihood function for  $\theta$  is

$$L(\theta) = \frac{1}{\pi^n \prod_{i=1}^n [1 + (X_i - \theta)^2]}.$$

Thus,  $L(\theta)$  will be a maximum when  $\prod_{i=1}^n [1 + (X_i - \theta)^2]$  is a minimum, or equivalently when  $\sum_{i=1}^n \ln(1 + (X_i - \theta)^2)$  is a minimum. The value of  $\theta$  that minimizes this expression must usually be determined by numerical methods.

For instance, suppose  $X_1 = 1, X_2 = X_3 = 2, X_4 = 3$ . Then, to maximize  $L(\theta)$ , we minimize

$$\ln(1 + (1 - \theta)^2) + \ln(1 + (2 - \theta)^2) + \ln(1 + (2 - \theta)^2) + \ln(1 + (3 - \theta)^2).$$

### R Note

The `optimize` function minimizes functions of a single variable. We can use it either for the expression above, or for log likelihood:

```
> x <- c(1, 2, 2, 3)
> g <- function(theta) sum(log(1 + (x-theta)^2))
> optimize(g, interval = c(0, 4))
$minimum
[1] 2

$objective
[1] 1.386294
> logL <- function(theta) sum(log(dcauchy(x, theta)))
> optimize(logL, interval = c(0, 4), maximum = TRUE)
$maximum
[1] 2

$objective
[1] -5.965214
```

The solution of 2 is the desired estimate of  $\hat{\theta}$ .

□

**Remark** The mathematical approach for maximizing a function is to set the derivative equal to 0 and solve. When we cannot solve that analytically, we may use software, using either of two approaches: (i) use software to optimize the log-likelihood, e.g. using `optimize` in R, or (ii) take the derivative of the log-likelihood, and use software to solve that, e.g. using `uniroot` in R.

See Section 13.8, for another approach, the EM algorithm, for estimating an MLE. ||

### 6.1.3 Maximum Likelihood for Multiple Parameters

We can find MLEs for distributions with more than one parameter using multivariable calculus, or software for numerical optimization.

**Theorem 6.1** Let  $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$  be a random sample from a normal distribution with mean  $\mu$  and standard deviation  $\sigma$ . The maximum likelihood estimates of  $\mu$  and  $\sigma$  are

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}, \quad (6.3)$$

$$\hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}. \quad (6.4)$$

*Proof.* We form the likelihood and log-likelihood

$$\begin{aligned} L(\mu, \sigma) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x_1-\mu}{\sigma}\right)^2} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x_2-\mu}{\sigma}\right)^2} \dots \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x_n-\mu}{\sigma}\right)^2} \\ &= \left( \frac{1}{\sqrt{2\pi}\sigma} \right)^n e^{-\frac{1}{2} \sum_{i=1}^n \left( \frac{x_i-\mu}{\sigma} \right)^2}. \end{aligned} \quad (6.5)$$

$$\ln(L(\mu, \sigma)) = -\frac{n}{2} \ln(2\pi) - n \ln(\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2.$$

Setting the partial derivatives of the log-likelihood with respect to  $\mu$  and  $\sigma$  equal to 0 gives a system of equations:

$$\frac{\partial(\ln(L(\mu, \sigma)))}{\partial\mu} = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0, \quad (6.6)$$

$$\frac{\partial(\ln(L(\mu, \sigma)))}{\partial\sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^n (x_i - \mu)^2 = 0. \quad (6.7)$$

Solving for  $\mu$  in Equation (6.6), we find the maximum likelihood estimate of  $\mu$  is  $(1/n) \sum_{i=1}^n x_i = \bar{x}$ , the sample mean. We substitute this into Equation (6.7) to obtain the estimate of  $\sigma$  to be  $\sqrt{(1/n) \sum_{i=1}^n (x_i - \bar{x})^2}$ .  $\square$

### Case Study: Wind Energy

Concerns about climate change and rising costs of fossil fuels and interest in sustainability have made renewable energy from sources such as tides, wind, and the sun more attractive. Wind turbines harness the kinetic energy from the wind to produce electricity.<sup>1</sup> For instance, Carleton College in Northfield Minnesota owns a 1.65 MW wind turbine that has been operational since 2004. In 2008, the turbine produced 3965 MW h of electricity that was then sold to Xcel Energy, a utility company.

Wind speeds are highly variable, affected by the time of day and time of year. Since the amount of energy output from a turbine depends on wind speed, understanding the characteristics of wind speed is important. Engineers use wind speed information to determine suitable locations to build a wind turbine or to optimize the design of a turbine. Utility companies use this information to make predictions on energy availability during peak demand periods (say, during a heat wave) or to estimate yearly revenue.

The Weibull distribution is the most commonly used probability distribution used to model wind speed (Justus et al., 1978; Seguro and Lambert, 2000; Weisser, 2003; Zhou et al., 2010). The Weibull distribution has a density

---

<sup>1</sup> <https://www.awea.org/wind-power-101>.

**Table 6.1** Sample of wind speeds (m/s) from Carleton College turbine.

Feb 14	Feb 15	Feb 16	Feb 17	Feb 18	Feb 19
7.8	8.9	9.7	7.7	6.4	3.1

function with two parameters, a shape parameter  $k > 0$ , and a scale parameter  $\lambda > 0$ ,

$$f(x; k, \lambda) = \frac{k}{\lambda^k} x^{k-1} e^{-(x/\lambda)^k}, \quad x \geq 0. \quad (6.8)$$

We will use this distribution to model the average wind speeds (m/s) at the site of Carleton's wind turbine for 168 days from February 14 to August 1, 2010 (there were no data for July 2) (Table 6.1).

Given the data  $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$ , the likelihood function is

$$\begin{aligned} L(k, \lambda; x_1, x_2, \dots, x_n) &= L(k, \lambda) = \prod_{i=1}^n \frac{k}{\lambda^k} x_i^{k-1} e^{-(x_i/\lambda)^k} \\ &= \frac{k^n}{\lambda^{kn}} \prod_{i=1}^n x_i^{k-1} e^{-\sum_{i=1}^n (x_i/\lambda)^k}. \end{aligned}$$

Thus, the log-likelihood is

$$\ln(L(k, \lambda)) = n \ln(k) - kn \ln(\lambda) + (k-1) \sum_{i=1}^n \ln(x_i) - \sum_{i=1}^n \left(\frac{x_i}{\lambda}\right)^k.$$

We next compute the partial derivatives of  $\ln(L(k, \lambda))$  with respect to  $k$  and  $\lambda$  and set them equal to 0:

$$\frac{\partial(\ln(L(k, \lambda)))}{\partial k} = \frac{n}{k} - n \ln(\lambda) + \sum_{i=1}^n \ln(x_i) - \sum_{i=1}^n \left(\frac{x_i}{\lambda}\right)^k \ln\left(\frac{x_i}{\lambda}\right) = 0, \quad (6.9)$$

$$\frac{\partial(\ln(L(k, \lambda)))}{\partial \lambda} = \frac{-kn}{\lambda} + \frac{k}{\lambda^{k+1}} \sum_{i=1}^n x_i^k = 0. \quad (6.10)$$

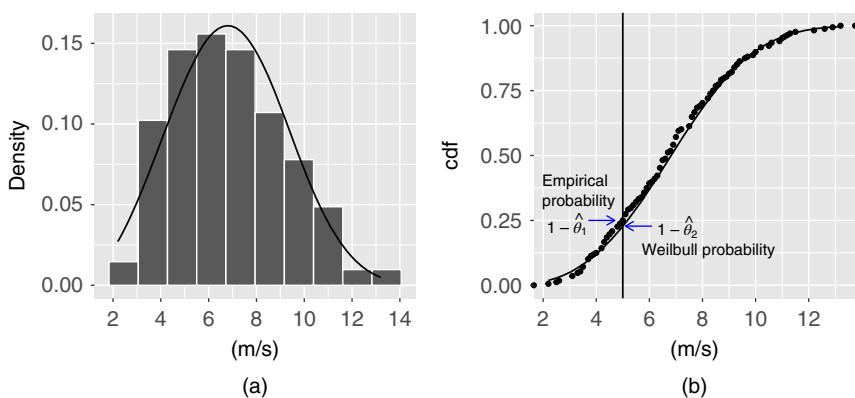
From Equation (6.10), we find

$$\lambda^k = \frac{1}{n} \sum_{i=1}^n x_i^k \quad (6.11)$$

and substituting this into Equation (6.9), we obtain

$$\frac{1}{k} + \frac{1}{n} \sum_{i=1}^n \ln(x_i) - \frac{1}{\alpha} \sum_{i=1}^n x_i^k \ln(x_i) = 0, \quad (6.12)$$

where  $\alpha = \sum_{i=1}^n x_i^k$ .



**Figure 6.4** (a) Histogram of wind speeds (m/s) with the pdf for the Weibull distribution superimposed,  $\hat{k} = 3.169$ ,  $\hat{\lambda} = 7.661$ . (b) Empirical cumulative distribution function, with cdf for Weibull superimposed. Also shown are the cdf and Weibull cdf values at wind=5;  $\hat{\theta}_1$  and  $\hat{\theta}_2$  are the corresponding estimated probabilities of wind exceeding 5.

Numerical methods must be used to find an approximate value for  $k$  in Equation (6.12). Using the wind data from the Carleton turbine, we obtain an estimate of  $\hat{k} = 3.169$  and, thus, from Equation (6.11), we get  $\lambda^{3.169} = (1/168) \sum_{i=1}^{168} x_i^{3.169}$ , which yields  $\hat{\lambda} = 7.661$ .

From Figure 6.4, it appears that the Weibull distribution models the data quite well. In Section 10.5.2, we will learn how to check this more formally using a goodness of fit test.

The R code for this analysis is provided on github: <https://github.com/Ichihara/MathStatsResamplingR>.

See Exercise 6.19 for a problem that uses the Weibull distribution to model time between earthquakes.

## 6.2 Method of Moments

Another approach for finding estimates for parameters is the *method of moments*. Let  $f(x)$  denote a probability density function (either continuous or discrete) for a random variable  $X$ . For a positive integer  $k$ , the  $k$ th (theoretical) moment of  $X$  is

$$\mu_k = E[X^k] = \int_{-\infty}^{\infty} x^k f(x) dx \quad (6.13)$$

and the  $k$ th sample moment is

$$M_k = \frac{1}{n} \sum_{i=1}^n X_i^k. \quad (6.14)$$

Suppose that there are  $K$  unknown parameters  $\theta_1, \dots, \theta_K$ , and the  $\mu_k$  are functions of these parameters. The method of moments estimate is obtained by solving the series of  $K$  equations in  $K$  unknowns:

$$M_k = \mu_k, k = 1, \dots, K. \quad (6.15)$$

In other words, we solve the system of  $K$  equations in  $K$  unknowns:

$$\begin{aligned} \int_{-\infty}^{\infty} xf(x; \theta_1, \theta_2, \dots, \theta_K) &= \frac{1}{n} \sum_{i=1}^n X_i, \\ \int_{-\infty}^{\infty} x^2 f(x; \theta_1, \theta_2, \dots, \theta_K) &= \frac{1}{n} \sum_{i=1}^n X_i^2, \\ &\vdots \\ \int_{-\infty}^{\infty} x^K f(x; \theta_1, \theta_2, \dots, \theta_K) &= \frac{1}{n} \sum_{i=1}^n X_i^K. \end{aligned}$$

**Remark** For discrete random variables, we replace the integrals with summations. ||

**Example 6.8** Let  $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Unif}[0, \beta]$ . The first theoretical moment is  $E[X_i] = \beta/2$ , while the sample mean  $\bar{X}$  is the first sample moment. Thus, setting  $\beta/2 = \bar{X}$  yields the method of moments estimator  $\hat{\beta} = 2\bar{X}$ .

Note that this estimator can give impossible results – for example if  $n = 4$ ,  $x_1 = x_2 = x_3 = 1$ ,  $x_4 = 9$ , then  $\bar{x} = 3$  and  $\hat{\beta} = 6$ , which makes  $x_4 = 9$  impossible. □

**Example 6.9** Suppose  $X_1, X_2, \dots, X_n$  are a random sample from the exponential distribution,  $f(x; \lambda) = \lambda e^{-\lambda x}$ ,  $x > 0$ ,  $\lambda > 0$ . Use the method of moments to find the estimator for  $\lambda$ .

### Solution

The first theoretical moment is  $E[X] = \int_0^\infty xe^{-\lambda x} dx = 1/\lambda$ . The sample first moment is the sample mean  $\bar{x} = (1/n) \sum_{i=1}^n X_i$ . Equating the two yields  $1/\lambda = \bar{X}$ ; thus, the estimator is  $\hat{\lambda} = 1/\bar{X}$ . □

**Example 6.10** Suppose  $x_1 = 1.3, x_2 = 1.8, x_3 = 2.1, x_4 = 2.25$  are a random sample from a distribution with pdf  $f(x; \theta) = 2(\theta - x)/\theta$  for  $0 < x < \theta$ . Find the method of moments estimate of  $\theta$ .

### Solution

The first theoretical moment is  $E[X] = \int_0^\theta x \cdot 2(\theta - x)/\theta dx = \theta^2/3$ . The first sample moment is the sample mean  $\bar{x} = 1.8625$ . Thus,  $\theta^2/3 = 1.8625$  yields an estimate of  $\hat{\theta} = 2.3638$ . □

**Example 6.11** Suppose  $X_1, X_2, \dots, X_n$  are a random sample from a distribution with pdf  $f(x; \lambda, \delta) = \lambda e^{-\lambda(x-\delta)}$ ,  $x > \delta$ , where  $\lambda, \delta > 0$ . Find estimators for  $\lambda$  and  $\delta$ .

### Solution

Use integration by parts to find

$$E[X] = \int_{\delta}^{\infty} x \lambda e^{-\lambda(x-\delta)} dx = \delta + \frac{1}{\lambda}.$$

Then equating this to the first sample moment, we have  $\delta + (1/\lambda) = \bar{X}$ .

Again, using integration by parts,

$$E[X^2] = \int_{\delta}^{\infty} x^2 \lambda e^{-\lambda(x-\delta)} dx = \delta^2 + \frac{2\delta}{\lambda} + \frac{2}{\lambda^2}.$$

Let  $m_2 = (1/n) \sum_{i=1}^n X_i^2$ , the second sample moment. Equating the second sample and theoretical moments and completing the square gives

$$\begin{aligned} m_2 &= \delta^2 + \frac{2\delta}{\lambda} + \frac{2}{\lambda^2} \\ &= \left( \delta + \frac{1}{\lambda} \right)^2 + \frac{1}{\lambda^2} \\ &= \bar{X}^2 + \frac{1}{\lambda^2}. \end{aligned}$$

Thus,  $1/\lambda = \sqrt{m_2 - \bar{X}^2}$ , and hence,  $\delta = \bar{X} - \sqrt{m_2 - \bar{X}^2}$ .

For instance, suppose  $X_1 = 3.5, X_2 = 3.9, X_3 = 4, X_4 = 4.7$ . Then the first and second sample moments are  $\bar{x} = 4.025$  and  $m_2 = 16.386$ , which results in the estimates  $\hat{\lambda} = 2.3133$  and  $\hat{\delta} = 3.5927$ .  $\square$

## 6.3 Properties of Estimators

We now have two very general methods of estimating parameters – maximum likelihood and method of moments – and in any given problem, we may be able to think of other more *ad hoc* methods. This raises the question of which method is best. Here, we discuss several criteria for comparing methods and properties that we think good methods should satisfy. The first three of these criteria – unbiasedness, efficiency, and mean square error – are fairly natural. The last two – consistency and transformation invariance – are more mathematical in how we state them, but more visceral in their application; each is a “sniff test,” and a procedure that fails these just doesn’t smell right.

### 6.3.1 Unbiasedness

In Section 5.6, we first introduced the idea of bias: We like an estimator to be, on average, equal to the parameter it is estimating. That is, we want the estimator to be unbiased, or equivalently,  $\text{Bias}[\hat{\theta}] = \text{E}[\hat{\theta}] - \theta = 0$ .

We have already seen that the sample mean is an unbiased estimator of the population mean  $\mu$  (Theorem A.7).

The sample proportion is also an unbiased estimator of the population proportion.

**Proposition 6.3** If  $X_1, X_2, \dots, X_n$  are Bernoulli random variables with parameter  $p$ , then  $\text{E}[\hat{p}] = p$ .

*Proof.* Let  $X = \sum_{i=1}^n X_i$ . Then

$$\text{E}[\hat{p}] = \text{E}\left[\frac{X}{n}\right] = \frac{np}{n} = p. \quad \square$$

The case of variance is less straightforward. The maximum likelihood estimate for  $\sigma^2$  in a normal setting with unknown mean and variance, given by Equation (6.4), is biased. But a minor variation is unbiased.

**Theorem 6.2** Let  $X_1, X_2, \dots, X_n$  be independent random variables from a distribution with unknown  $\text{Var}[X_i] = \sigma^2 < \infty$ . Then an unbiased estimator of  $\sigma^2$  is  $S^2 = (1/(n-1)) \sum_{i=1}^n (X_i - \bar{X})^2$ .

*Proof.* We utilize Proposition A.2, Theorem A.7, and the following algebraic fact:

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \left( \sum_{i=1}^n X_i^2 \right) - n\bar{X}^2. \quad (6.16)$$

Let  $\hat{\sigma}^2 = (1/n) \sum_{i=1}^n (X_i - \bar{X})^2$  (note that this is the MLE of  $\sigma^2$  in the normal case, Theorem 6.1). Then

$$\begin{aligned} \text{E}[\hat{\sigma}^2] &= \text{E}\left[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\right] \\ &= \frac{1}{n} \text{E}\left[\sum_{i=1}^n X_i^2 - n\bar{X}^2\right] \\ &= \frac{1}{n} \left[ \sum_{i=1}^n \text{E}[X_i^2] - n\text{E}[\bar{X}^2] \right] \\ &= \frac{1}{n} \left[ \sum_{i=1}^n (\sigma^2 + \text{E}[X_i]^2) - n \left( \frac{\sigma^2}{n} + \text{E}[\bar{X}]^2 \right) \right] \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{n} \left[ \sum_{i=1}^n (\sigma^2 + \mu^2) - \sigma^2 - n\mu^2 \right] \\
&= \frac{1}{n} [n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2] \\
&= \frac{n-1}{n} \sigma^2.
\end{aligned}$$

Thus,  $\hat{\sigma}^2$  is a biased estimator of  $\sigma^2$ . However, we can “unbias” it:

$$\frac{n}{n-1} E[\hat{\sigma}^2] = \sigma^2 \quad \text{or} \quad E\left[\frac{n}{n-1} \hat{\sigma}^2\right] = \sigma^2.$$

Hence, an unbiased estimator of  $\sigma^2$  is

$$\frac{n}{n-1} \hat{\sigma}^2 = \frac{n}{n-1} \cdot \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2. \quad \square$$

**Definition 6.3** Let  $X_1, X_2, \dots, X_n$  be independent random variables from a distribution with unknown variance  $\sigma^2 < \infty$ . The (*sample*) *variance* of  $X_1, X_2, \dots, X_n$  is

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2,$$

which by Theorem 6.2, is an unbiased estimator of  $\sigma^2$ . ||

In general, if a biased estimator is off by a multiplicative constant, we can “unbias” the estimator by dividing it by that constant. That is, if  $E[\hat{\theta}] = C \times \theta$  where  $C$  does not depend on  $\theta$ , then  $\hat{\theta}/C$  is an unbiased estimator of  $\theta$ .

**Example 6.12** Let  $X_1, X_2, \dots, X_n$  be i.i.d. from  $\text{Unif}[0, \beta]$ . We have already seen that the MLE of  $\beta$  is  $\hat{\beta}_{\text{mle}} = X_{\max}$ .

With  $f_{\max}(x) = (n/\beta^n)x^{n-1}$  (see Corollary 4.1), we have

$$\begin{aligned}
E[X_{\max}] &= \int_0^\beta x \frac{n}{\beta^n} x^{n-1} dx \\
&= \frac{n}{n+1} \beta,
\end{aligned}$$

so  $\hat{\beta}_{\text{mle}} = X_{\max}$  is a biased estimator of  $\beta$ .

However,  $((n+1)/n) \times X_{\max}$  is unbiased.

We also have the method of moments estimator  $\hat{\beta}_{\text{mom}} = 2\bar{X}$ . Computing the expectation, we find  $E[\hat{\beta}_{\text{mom}}] = E[2\bar{X}] = 2(\beta/2) = \beta$ . Thus,  $2\bar{X}$  is an unbiased estimator of  $\beta$ .

In this case, we have two unbiased estimators of  $\beta$ . To choose between them, we turn to other criteria below. But first we note some limitations of unbiasedness. □

### Limitations of Unbiasedness

In practice, we are generally satisfied with estimates that are approximately unbiased; estimates that are exactly unbiased may be impossible to obtain, or in some cases, are unreasonable.

For example, suppose the  $X$  is an observation from a geometric distribution with  $f(x) = p(1-p)^{(x-1)}$  for  $x = 1, 2, \dots$ , the time until the first success when the probability of a success is  $p$ . An unbiased estimator  $\hat{p} = g(X)$  satisfies  $\sum_{x=1}^{\infty} g(x)f(x) = p$ , which has the unique solution  $g(x) = 1$  if  $x = 1$ ; otherwise, 0, but an estimate of 0 is unreasonable, once we observe a success. A reasonable estimate is  $\hat{p} = 1/x$ , the fraction of successes.

Unbiased estimates may disagree with common sense. In Theorem 6.2, we saw that the sample variance  $S^2 = (1/(n-1)) \sum (X_i - \bar{X})^2$  is an unbiased estimator of the variance  $\sigma^2$ . However, the sample standard deviation  $S$  is not an unbiased estimator of the standard deviation  $\sigma$ . For normal distributions, for  $n = 10$ ,  $E[S] \approx 0.973\sigma$ . It just seems wrong to use  $S^2$  when estimating  $\sigma^2$ , but to use  $S/0.973$  when estimating  $\sigma$ . Furthermore, the correction factor needs to be recomputed for every  $n$ . In addition,  $S^2$  is unbiased for the variance for any distribution, not just normal distributions, but the correction factors for  $S$  depend on the distribution, which in practice is unknown. So in practice, we use  $S$  as an estimate of  $\sigma$ . It is approximately unbiased, which is good enough.

If  $\hat{\theta}$  is an unbiased estimator of  $\theta$ , and  $h$  is a function, then  $h(\hat{\theta})$  is not in general an unbiased estimator of  $h(\theta)$ . A notable exception is when  $h$  is a linear transformation,  $h(\theta) = a + b\theta$ . (Exercise 6.30).

For example, consider again the unbiased estimator  $\hat{\beta} = 2\bar{X}$  for  $\beta$  in  $\text{Unif}[0, \beta]$ .  $(\hat{\beta})^2 = 4(\bar{X})^2$  is not unbiased for  $\beta^2$ :

$$\begin{aligned} E[4\bar{X}^2] &= 4E[\bar{X}^2] = 4\left(\text{Var}[\bar{X}] + E[\bar{X}]^2\right) \\ &= 4\left(\frac{\beta^2}{12n} + \left(\frac{\beta}{2}\right)^2\right) = \beta^2 + \frac{\beta^2}{3n}. \end{aligned}$$

### Asymptotic Bias

In lieu of expecting the bias of an estimator to be zero, we may be satisfied with an estimator whose bias disappears as the sample size increases.

In Theorem 6.2, we saw that the MLE of  $\sigma^2$  for a sample of size  $n$  drawn from  $N(\mu, \sigma^2)$  satisfies  $E[\hat{\sigma}^2] = ((n-1)/n)\sigma^2$ , so that  $\hat{\sigma}^2$  is a biased estimator of  $\sigma^2$ . But note that

$$\lim_{n \rightarrow \infty} E[\hat{\sigma}^2] = \lim_{n \rightarrow \infty} \frac{n-1}{n} \sigma^2 = \sigma^2,$$

so we say that  $\hat{\sigma}^2$  is *asymptotically unbiased*.

Similarly, the MLE,  $X_{\max}$  of  $\beta$  for a random sample from the uniform distribution  $\text{Unif}[0, \beta]$  is also asymptotically unbiased:

$$\lim_{n \rightarrow \infty} E[X_{\max}] = \lim_{n \rightarrow \infty} \frac{n}{n+1} \beta = \beta.$$

### 6.3.2 Efficiency

Earlier we found two unbiased estimators for  $\beta$  for  $\text{Unif}[0, \beta]$ . Here we look at one criterion for comparing them – efficiency – that depends on their variance.

We begin with an example comparing two unbiased estimators for a mean. Let  $X_1, X_2, X_3$  be independent random variables from a distribution with mean  $\mu$  and variance  $\sigma^2$ . Then  $\bar{X}$  is an estimator of  $\mu$ . Now, consider  $Y = (1/6)X_1 + (1/3)X_2 + (1/2)X_3$ , a weighted average of  $X_1, X_2, X_3$ . Then

$$\begin{aligned} E[Y] &= \frac{1}{6}E[X_1] + \frac{1}{3}E[X_2] + \frac{1}{2}E[X_3] \\ &= \frac{1}{6}\mu + \frac{1}{3}\mu + \frac{1}{2}\mu = \mu. \end{aligned}$$

Thus,  $Y$  is also an unbiased estimator of  $\mu$ .

We now have two unbiased estimators of  $\mu$ , the sample mean  $\bar{X}$  and  $Y$ . We will compare their variances. Recall that  $\text{Var}[\bar{X}] = \sigma^2/3$  (Theorem A.7). On the other hand, since the  $X_i$ 's are independent,

$$\begin{aligned} \text{Var}[Y] &= \text{Var}[(1/6)X_1] + \text{Var}[(1/3)X_2] + \text{Var}[(1/2)X_3] \\ &= \frac{1}{36}\sigma^2 + \frac{1}{9}\sigma^2 + \frac{1}{4}\sigma^2 = \frac{7}{18}\sigma^2. \end{aligned}$$

Since  $\text{Var}[\bar{X}] < \text{Var}[Y]$ , we see that  $\bar{X}$  is less variable than  $Y$ .

**Definition 6.4** If  $\hat{\theta}_1$  and  $\hat{\theta}_2$  are both unbiased estimators of  $\theta$  and  $\text{Var}[\hat{\theta}_1] < \text{Var}[\hat{\theta}_2]$ , then  $\hat{\theta}_1$  is said to be more *efficient* than  $\hat{\theta}_2$ . ||

**Example 6.13** We saw earlier that if  $X_1, X_2, \dots, X_n$  are a random sample from the uniform distribution  $\text{Unif}[0, \beta]$ , then  $((n+1)/n)X_{\max}$  and  $2\bar{X}$  are two unbiased estimators of  $\beta$ .

We will run a simulation to see how these two estimators perform, in particular which has the smallest variance. We draw random samples of size 25 from  $\text{Unif}[0, 12]$ . For each sample, we compute  $2\bar{x}$  and  $26/25 \times \max\{x_1, x_2, \dots, x_{25}\}$  and record these values. We repeat this 1000 times.

#### R Note

```
N <- 10^4
Xbar2 <- numeric(N)
Max <- numeric(N)
for (i in 1:N)
{
  x <- runif(25, 0, 12)      # draw 25 from Unif[0, 12]
  Xbar2[i] <- 2 * mean(x)
  Max[i] <- 26/25 * max(x)
}
```

```
mean(Xbar2)
sd(Xbar2)
mean(Max)
sd(Max)
```

Here we scale the axes to be the same. You may need to adjust this setting for your simulation.

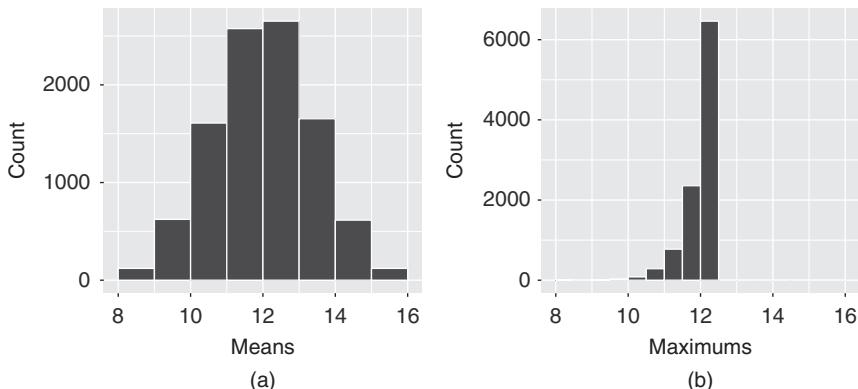
```
df <- data.frame(Xbar2, Max)
library(gridExtra)
p1 <- ggplot(df, aes(Xbar2)) +
  geom_histogram(binwidth = .5, center = .25, color = "white") +
  labs(x = "Means") + xlim(c(8,16))

p2 <- ggplot(df, aes(Max)) +
  geom_histogram(binwidth = .5, center = .25, color = "white") +
  labs(x = "Maximums") + xlim(c(8,16))
grid.arrange(p1, p2, nrow = 2)
```

In one run of this simulation, we obtained a mean of 12.007 and standard deviation of 1.383 for the estimates of  $\beta$  based on the method of moments  $2\bar{X}$  compared to a mean of 12.027 and standard deviation of 0.456 for the estimates based on  $((n+1)/n)X_{\max}$ . Histograms are shown in Figure 6.5.

The simulation shows that the estimates based on  $((n+1)/n)X_{\max}$  have less variability. We can also compute the variances exactly:

$$\text{Var}[2\bar{X}] = 2^2 \text{Var}[\bar{X}] = 2^2 \frac{\beta^2}{12n} = \frac{\beta^2}{3n}.$$



**Figure 6.5** Sampling distribution of estimators for  $\beta$ . (a) Distribution of  $2\bar{X}$ . (b) Distribution of  $(26/25)X_{\max}$ .

On the other hand,

$$\begin{aligned}\text{Var}\left[\frac{n+1}{n}X_{\max}\right] &= \frac{(n+1)^2}{n^2}\text{Var}[X_{\max}] \\ &= \frac{(n+1)^2}{n^2}\left\{\text{E}[X_{\max}^2] - \text{E}[X_{\max}]^2\right\} \\ &= \frac{(n+1)^2}{n^2}\left\{\int_0^\beta x^2 \frac{n}{\beta^n}x^{n-1}dx - \left(\int_0^\beta x \frac{n}{\beta^n}x^{n-1}dx\right)^2\right\} \\ &= \frac{\beta^2}{n(n+2)}.\end{aligned}$$

Since  $\frac{\beta^2}{n(n+2)} < \frac{\beta^2}{3n}$  for  $n > 1$ ,  $((n+1)/n)X_{\max}$  is more efficient than  $2X$ .  $\square$

**Remark** Suppose we have an unbiased estimators  $\hat{\theta}$  of a parameter  $\theta$ . Are there other unbiased estimators of  $\theta$  that are more efficient than  $\hat{\theta}$ ? There is a theorem that provides a lower bound for the variance of any unbiased estimator of  $\theta$ .

**Cramer–Rao Inequality** If  $X_1, X_2, \dots, X_n$  are a random sample from a distribution with continuous pdf  $f(x; \theta)$  and  $f$  satisfies certain smoothness criteria, then any unbiased estimator  $\hat{\theta}$  of  $\theta$  satisfies

$$\text{Var}[\hat{\theta}] \geq \frac{1}{n E[(\partial/\partial\theta(\ln(f(X; \theta))))^2]}.$$

Thus, if the variance of an unbiased estimator  $\theta$  achieves this lower bound, then in some sense, it is a “best” estimator. However, there are cases where *no* unbiased estimator of a parameter  $\theta$  achieves the lower bound in the Cramer–Rao inequality.

The expression

$$I(\theta) = E[(\partial/\partial\theta(\ln(f(X; \theta))))^2] \tag{6.17}$$

in the denominator is called the *Fisher information*.  $\parallel$

In practice, we may use efficiency to compare estimators, even if we do not know if the estimators are unbiased.

**Example 6.14** We continue the wind energy case study from Section 6.1.3. We are interested in the fraction of time that the wind speed exceeds 5 m/s.

Two ways to estimate this are the empirical fraction of the 168 measurements that exceed 5, and the probability based on the Weibull distribution with parameters estimated using maximum likelihood:

$$\hat{\theta}_1 = 126/168 = 0.75,$$

$$\hat{\theta}_2 = 1 - F(5; k = 3.169, \lambda = 7.661) = 0.772,$$

where  $F(\cdot; k, \lambda)$  is the cdf of a Weibull distribution with the specified parameters. Bootstrapping these two estimates gives standard errors:

$$s_{\hat{\theta}_1} = 0.033,$$

$$s_{\hat{\theta}_2} = 0.023.$$

The estimated *relative efficiency*, the ratio of squared standard errors, is 2.06.

The Weibull procedure appears to be much more accurate. However, the empirical fraction  $\theta_1$  is an unbiased procedure, while the Weibull procedure is biased because the true population distribution may not be Weibull, and even if it is, the  $\theta_2$  procedure is biased. Our earlier diagnostics suggested that the Weibull distribution fits the data well and the second effect is usually small, so we believe the bias is small and hence, willing to use the Weibull-based estimate.

The relative efficiency is amazingly large. In practice, statisticians may put great effort into developing estimation procedures with a relative efficiency gain of 1%, because data may be very expensive to collect, and this would cut data requirements by 1%.

The code for this example is provided on github: <https://github.com/lchihara/MathStatsResamplingR>.  $\square$

### 6.3.3 Mean Square Error

We now turn to a criterion that combines bias and variance. This is useful for comparing estimators that are not both unbiased. We may prefer an estimator with small bias and small variance over one that is unbiased but with larger variance. This criterion provides a way to quantify the preference.

**Definition 6.5** The *Mean Square Error* of an estimator is  $MSE[\hat{\theta}] = E[(\hat{\theta} - \theta)^2]$ .  $\parallel$

MSE measures the average squared distance between the estimator and the parameter. It combines bias and variance.

**Proposition 6.4**  $\text{MSE}[\hat{\theta}] = \text{Var}[\hat{\theta}] + \text{Bias}[\hat{\theta}]^2$ .

*Proof.*

$$\begin{aligned}
 \text{MSE}[\hat{\theta}] &= \text{E}[(\hat{\theta} - \theta)^2] \\
 &= \text{E}[(\hat{\theta} - \text{E}[\hat{\theta}] + \text{E}[\hat{\theta}] - \theta)^2] \\
 &= \text{E}[((\hat{\theta} - \text{E}[\hat{\theta}]) + (\text{E}[\hat{\theta}] - \theta))^2] \\
 &= \text{E}[(\hat{\theta} - \text{E}[\hat{\theta}])^2] + 2\text{E}[\hat{\theta} - \text{E}[\hat{\theta}]](\text{E}[\hat{\theta}] - \theta) + (\text{E}[\hat{\theta}] - \theta)^2 \\
 &= \text{Var}[\hat{\theta}] + (\text{Bias}[\hat{\theta}])^2 \quad \square
 \end{aligned}$$

Also, if  $\hat{\theta}$  is unbiased, then  $\text{MSE}[\theta] = \text{Var}[\hat{\theta}]$ . So, for unbiased estimators, one is more efficient than a second if and only if its MSE is smaller.

### Mean Square Error

MSE takes into account the variability of the estimator as well as the bias.

In general, when comparing two estimators  $\hat{\theta}_1$  and  $\hat{\theta}_2$  of  $\theta$ , we are often faced with a trade-off between variability and bias.

**Example 6.15** Let  $X \sim \text{Binom}(n, p)$ ,  $n$  known and  $p$  unknown. The sample proportion  $\hat{p}_1 = X/n$  is an unbiased estimator of  $p$  (see Exercises 6.1 and 6.24) with  $\text{E}[\hat{p}_1] = p$  and  $\text{Var}[\hat{p}_1] = p(1-p)/n$ , and  $\text{MSE}[\hat{p}_1] = p(1-p)/n$ .

An alternative estimator of  $p$  is  $\hat{p}_2 = (X+1)/(n+2)$ ; this adds one artificial success and one failure to the real data. Then

$$\text{E}[\hat{p}_2] = \frac{1}{n+2}(\text{E}[X] + \text{E}[1]) = \frac{np+1}{n+2}.$$

This is a case where we will not be able to “unbias”  $\hat{p}_2$ . The bias is

$$\text{Bias}[\hat{p}_2] = \frac{np+1}{n+2} - p = \frac{1-2p}{n+2}$$

and the variance is

$$\begin{aligned}
 \text{Var}[\hat{p}_2] &= \text{Var}\left[\frac{X+1}{n+2}\right] = \frac{1}{(n+2)^2}(\text{Var}[X] + \text{Var}[1]) \\
 &= \frac{1}{(n+2)^2}np(1-p) \\
 &= \frac{np(1-p)}{(n+2)^2}
 \end{aligned}$$

resulting in MSE

$$\text{MSE}[\hat{p}_2] = \frac{np(1-p)}{(n+2)^2} + \left(\frac{1-2p}{n+2}\right)^2 = \frac{np(1-p) + (1-2p)^2}{(n+2)^2}.$$

**Figure 6.6** Mean square error against  $p$ ,  $n = 16$ .

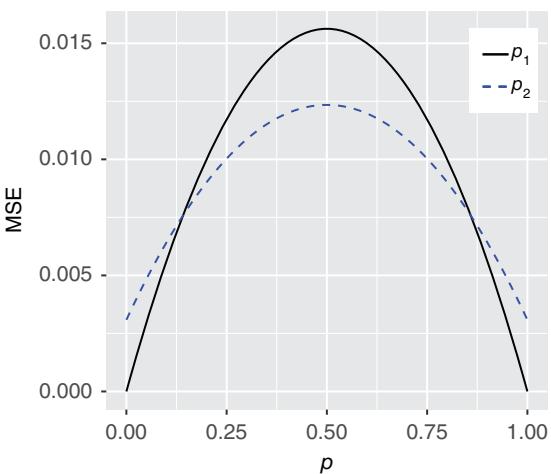


Figure 6.6 shows the MSE for both estimators as a function of  $p$ . While  $\hat{p}_2$  is biased (except for  $p = 0.5$ ), it has smaller MSE than  $\hat{p}_1$  except for  $p$  near 0 or 1.

### R Note

```
funMSE1 <- function(x, n)x*(1-x)/n
funMSE2 <- function(x, n)n*(1-x)*x/(n+2)^2 + (1-2*x)^2/(n+2)^2
ggplot(data.frame(x = c(0,1)), aes(x = x)) +
  stat_function(fun = funMSE1, args = list(n = 16), lty = 1) +
  stat_function(fun = funMSE2, args = list(n = 16), lty = 2) +
  labs(x = "p", y = "MSE")
```

In Exercise 6.37, you will investigate what happens to the mean square error when you adjust the sample size  $n$ .  $\square$

In practice, we may be unable to compute bias, variance, and MSE analytically, in particular, in situations in which the underlying distribution is unknown. In such situations, we may bootstrap to estimate these quantities.

### 6.3.4 Consistency

As we get more data, we expect an estimator to become more accurate. Our next criterion, *consistency*, says roughly that an estimator gives the right answer in the long run, as the sample size goes to infinity. This is a “sniff test” criterion – it is rare to use an estimator that fails this test.

**Definition 6.6** For a random sample of size  $n$ , let  $\hat{\theta}_n$  denote an estimator of  $\theta$  and let  $\{\hat{\theta}_n\}_{n=0}^{\infty}$  be a sequence of estimators. The estimators are *consistent* for

$\theta$  if and only if

$$\lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta| < \epsilon) = 1 \quad (6.18)$$

for every  $\epsilon > 0$ . ||

Since Equation (6.18) can also be expressed as  $\lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta| \geq \epsilon) = 0$ , this says that for any acceptable amount of error  $\epsilon > 0$ , the probability of an actual error worse than  $\epsilon$  goes to zero.

### Remark

- The limit in Equation (6.18) is referred to as *convergence in probability*.
- The “sequence of estimators” referred to in the definition generally refers to a single estimation procedure, such as  $\bar{X}$ ,  $S^2$ , or  $((n+1)/n)X_{\max}$ , being recomputed every time an observation is added to the sample. ||

**Example 6.16** Let  $X_1, X_2, \dots, X_n$  denote a random sample from  $N(\mu, 1)$  and let  $\bar{X}_n$  denote the sample mean.

From Theorem A.7,  $\bar{X}_n$  is normal with mean  $\mu$  and variance  $1/n$ . Thus,

$$\begin{aligned} \lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| < \epsilon) &= \lim_{n \rightarrow \infty} P(\mu - \epsilon < \bar{X}_n < \mu + \epsilon) \\ &= \lim_{n \rightarrow \infty} \frac{1}{\sqrt{2\pi/n}} \int_{\mu-\epsilon}^{\mu+\epsilon} e^{-(t-\mu)^2/(2/n)} dt \\ &= \lim_{n \rightarrow \infty} \frac{1}{\sqrt{2\pi}} \int_{-\epsilon/\sqrt{n}}^{\epsilon/\sqrt{n}} e^{-z^2/2} dz, \quad \text{where } z = \sqrt{n}(t - \mu). \end{aligned}$$

Since we are integrating the standard normal density over an interval approaching  $(-\infty, \infty)$ , the limit is 1.

Thus, the sample means are a consistent sequence of estimators of  $\mu$ . □

Rather than integrating pdfs, we can use the mean square error of an estimator to determine consistency: An estimator is consistent if its MSE goes to zero. The converse is not true, e.g. if  $P(|\hat{\theta} - \theta| < n^{-1/3}) = 1 - 1/n$  and  $P(|\hat{\theta} - \theta| = n) = 1/n$ , the estimator is consistent even though the MSE explodes.

**Proposition 6.5** Let  $\{\hat{\theta}_n\}$  be a sequence of estimators for  $\theta$ . If

$$\lim_{n \rightarrow \infty} E[\hat{\theta}_n] = \theta \quad \text{and} \quad \lim_{n \rightarrow \infty} \text{Var}[\hat{\theta}_n] = 0,$$

then  $\theta_n$  is consistent for  $\theta$ .

*Proof.* Recall from Proposition 6.4 that

$$E[(\hat{\theta}_n - \theta)^2] = \text{Var}[\hat{\theta}_n] + \text{Bias}[\hat{\theta}_n]^2.$$

In addition, Chebyshev's Inequality (Theorem A.11) gives

$$P(|\hat{\theta}_n - \theta| \geq \epsilon) \leq \frac{E[(\hat{\theta}_n - \theta)^2]}{\epsilon^2}.$$

Putting these two results together gives

$$P(|\hat{\theta}_n - \theta| \geq \epsilon) \leq \frac{\text{Var}[\hat{\theta}_n] + \text{Bias}[\hat{\theta}_n]^2}{\epsilon^2}.$$

Thus,  $\lim_{n \rightarrow \infty} E[\hat{\theta}_n] = \theta$  and  $\lim_{n \rightarrow \infty} \text{Var}[\hat{\theta}_n] = 0$  give  $\lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta| \geq \epsilon) = 0$ . This is equivalent to Equation (6.18).  $\square$

**Example 6.17** Returning to Example 6.16, we have that  $E[\bar{X}_n] = \mu$  and  $\text{Var}[\bar{X}_n] = 1/n$ . So by Proposition 6.5, the sample means are a consistent sequence of estimators of  $\mu$ .

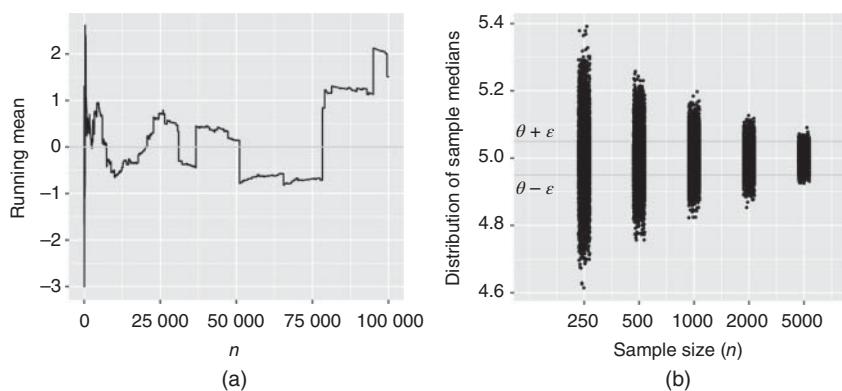
More generally, let  $X_1, X_2, \dots, X_n$  be i.i.d. from a distribution with mean  $\mu$  and variance  $\sigma^2$ . If  $\bar{X}_n$  denotes the sample mean, then  $E[\bar{X}] = \mu$  and  $\text{Var}[\bar{X}_n] = \sigma^2/n$ , so again by Proposition 6.5, the sample means are a consistent sequence of estimators for  $\mu$ .  $\square$

**Example 6.18** Let  $X_1, X_2, \dots, X_n$  be a random sample from the Cauchy distribution (refer to Example 6.7 and Exercise 2.11). Neither the mean nor any of its moments exist, although we can still compute the sample mean  $\bar{X}_n$ . Is this sample mean a consistent estimator of  $\theta$ ? It turns out this sample mean has the same distribution as a single observation (Stuart and Ord, 2009)! Hence,  $\bar{X}_n$  is not a consistent estimator of  $\theta$ . Figure 6.7 show a running mean for the Cauchy; there are enough occasional huge outliers that cause large enough jumps in the mean that the mean never settles down.

On the other hand, the sample medians are a consistent estimator of  $\theta$ . We illustrate this with a simulation: We draw random samples of sizes  $n$  from a Cauchy distribution with  $\theta = 5$ , compute the sample medians, and plot the distribution. Taking  $\epsilon = 0.05$ , we note the proportion of sample medians that fall within  $\theta \pm \epsilon$ . Figure 6.7 gives the results for runs of this simulation for different sample sizes. We can see that as sample sizes increase, a larger proportion of the sample medians fall within  $\theta \pm \epsilon$ . For  $n = 250$ , only 38.1% of the sample medians fall within this interval. On the other hand, for  $n = 5000$ , 97.8% of the sample medians are within  $\epsilon$  of  $\theta$ .  $\square$

### 6.3.5 Transformation Invariance\*

In discussing limitations of unbiasedness (Section 6.3.1), we declared that it seems wrong to use  $S^2$  when estimating  $\sigma^2$ , but to use  $S/0.973$  when estimating  $\sigma$ . This discrepancy is ugly. Our next criterion is one that, if satisfied, prevents such discrepancies.



**Figure 6.7** (a) Running mean for Cauchy. The  $y$  value is the running mean from a single sample,  $y_n = n^{-1} \sum_{i=1}^n x_i$ . There are enough huge outliers that the mean never settles down. (b) Comparison of distributions of sample medians from a Cauchy distribution with  $\theta = 5$ . For  $\epsilon = 0.05$ , the fraction of medians that fall in  $5 \pm 0.05$  are 0.381, 0.521, 0.683, 0.847, and 0.978 for sample sizes 250, 500, 1000, 2000, and 5000, respectively.

An estimation procedure is *transformation invariant* if it yields equivalent estimates for transformations of parameters. That is, if  $\zeta = h(\theta)$  for some invertible function  $h$ , then  $\hat{\zeta} = h(\hat{\theta})$ .

For example, the exponential distribution (Section B.8) can be given in terms of a rate parameter  $\lambda$  or a scale parameter  $\zeta = 1/\lambda$ ,

$$f(x) = \lambda e^{-\lambda x} = \frac{1}{\zeta} e^{-x/\zeta}$$

for  $x \geq 0$ . Both maximum likelihood and method of moments are invariant under this parameter transformation, with  $\hat{\zeta} = 1/\hat{\lambda}$ .

**Proposition 6.6** Let  $f(x; \theta) = g(x; \zeta)$ , where  $f$  and  $g$  are discrete or continuous densities with parameters  $\theta$  and  $\zeta$  related by  $\zeta = h(\theta)$  for some invertible function  $h$ . If  $\hat{\theta}$  and  $\hat{\zeta}$  are the maximum likelihood estimators for  $\theta$  and  $\zeta$ , respectively, then  $\hat{\zeta} = h(\hat{\theta})$ .

*Proof.* Let  $L_f$  and  $L_g$  denote the likelihood functions for the  $\theta$  and  $\zeta$  parameterizations, respectively. We have

$$L_f(\theta) = \prod f(x_i; \theta) = \prod g(x_i; \zeta) = L_g(\zeta),$$

when  $\zeta = h(\theta)$ . Since  $L_f$  is maximized at  $\hat{\theta}$ , this maximum value  $L_f(\hat{\theta})$  must be the maximum value of  $L_g$  also. Thus,

$$L_f(\hat{\theta}) = L_g(h(\hat{\theta})) = L_g(\hat{\zeta}).$$

□

The invariance of the MLE actually holds for arbitrary invertible functions  $h$ . Thus, for instance, if  $\hat{\theta}$  is the MLE for a parameter  $\theta$ , then  $\cos(\hat{\theta})$  is the MLE for the parameter  $\cos(\theta)$  (provided that the parameter is constrained to a region where the function is invertible).

The invariance property is useful in a wide variety of applications. For example, with exponential distributions, sometimes it is convenient to work with a rate parameter  $\lambda$ , other times with a scale parameter  $\varsigma = 1/\lambda$  (see Section B.8).

For normal distributions, we may work with  $\sigma$  or  $\sigma^2$ . This case is subtle – it involves thinking of  $\sigma^2$  as a parameter distinct from  $\sigma$ . Let  $\theta = h(\sigma) = \sigma^2$ , then we may rewrite the normal density using  $\theta$  and  $\sqrt{\theta}$  in place of  $\sigma^2$  and  $\sigma$ . If we maximize the likelihood with respect to  $\theta$ , this gives the equivalent answer as maximizing with respect to  $\sigma$ , with  $\hat{\theta} = \hat{\sigma}^2$ .

Similarly, if  $p$  is the probability for a binomial distribution and  $\hat{p}$  is its MLE (see Exercise 6.1), and we have the constraint that  $p \geq 0.5$ , then the MLE for  $\sqrt{p(1-p)}$  is  $\sqrt{\hat{p}(1-\hat{p})}$ . Without the constraint, the likelihood is undefined.

Methods of moment estimators are also invariant under invertible parameter transformations.

### 6.3.6 Asymptotic Normality of MLE\*

One nice property of maximum likelihood estimates is that they are asymptotically normal with the best possible variance (based on the Fisher information) under fairly general conditions. Caveat – in finite samples, the variance may be larger, and the distribution may be nonnormal.

**Theorem 6.3** Let  $X_1, X_2, \dots, X_n$  denote a random sample from the pdf  $f(x; \theta)$ . Let  $\hat{\theta}_n$  denote the MLE of  $\theta$ . Under certain regularity conditions on  $f$ , for large  $n$ ,  $\hat{\theta}_n$  is approximately normal with mean  $\theta$  and variance  $1/(nI(\theta))$ , where  $I(\theta)$  denotes the Fisher information (Equation (6.17)).

The proof is beyond the scope of this text. We refer the interested reader to Bickel and Doksum (2001).

**Example 6.19** Let  $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} f$ , where  $f(x; \theta) = \theta x^{\theta-1}$ , for  $\theta > 0$ ,  $0 < x < 1$ . Then  $\ln(f(x; \theta)) = \ln(\theta) + (\theta - 1)\ln(x)$  and  $\partial/\partial\theta(\ln(f(x; \theta))) = 1/\theta + \ln(x)$ . Hence, brushing up on our calculus skills, we find

$$I(\theta) = E \left[ \left( \frac{1}{\theta} + \ln(x) \right)^2 \right] = \frac{1}{\theta^2}.$$

The MLE of  $\theta$  is  $\hat{\theta} = -n / \sum \ln(x_i)$ . Thus, for large  $n$ ,  $\hat{\theta}$  is approximately normal with mean  $\theta$  and variance  $1/(nI(\theta)) = \theta^2/n$ .  $\square$

## 6.4 Statistical Practice

There are some additional considerations that matter in statistical practice. Here we discuss three closely related topics – what to measure, transformations, and robustness – and an additional topic, weighted estimation.

In practice, we should measure what matters. For example, here is the beginning of a problem at Google:

I have two data sets of travel prediction errors, in the form  $((\text{actualTime} - \text{predictedTime})/\text{actualTime})^2$  and I'm trying to determine if the difference in the means of these two data sets is statistically discernible.

That's a good start – to scale by the actual time. Comparing raw errors would imply that a prediction of 5 min when the actual time is 20 min would be no worse than a prediction of 45 min when the actual time is an hour. However, that scaling is not ideal – it implies that predicting 1 h when the actual time is 2 h is no worse than predicting 3 m when the actual time is 6 m. It would probably be better to divide by the square root of the actual time.

Similarly, for the Verizon example, the New York Public Utilities Commission mandates comparing the mean repair times between two groups. That is based on the total time for a group, so that an extra 5 min on any single repair has the same effect on the measurement, implying that an extra 5 min on a 1-min repair is no more annoying than an extra 5 min on a 48-h repair.

A common procedure in practice is to transform the data before computing statistics like means. In particular, for data like repair times that are positive with a long tail, we may take transformations such as the log or square root that are concave down. This has two effects. One is that it sharply reduces the size of the largest observations, so they have less effect on means or other estimates. Transforming reduces the effect of outliers. The second effect is that we can use a transformation to measure what matters. In both the Verizon and Google travel time prediction examples, there are diminishing effects of an extra minute as the times get larger. An ideal transformation would be  $h(x) = \text{cost}(x)$  so that comparing means of the transformed data is like comparing the average cost or average annoyance.

A log transformation is common, but is undefined at zero and has infinite slope, so it inappropriately magnifies small differences in  $x$  values near zero. Transformations like  $\log(x + 1)$  and  $\sqrt{x + 1}$  avoid the infinite derivatives.

Finally, real data are never normally distributed and are often long-tailed. In addition to using appropriate transformations, we can use statistics that are

*robust* – less sensitive to outliers – like medians or trimmed means. R has many robust functions for many applications.<sup>2</sup>

### 6.4.1 Are You Asking the Right Question?

It is part of the statistical consultant's role to not only provide effective estimation techniques, that measure what matters, but also to take a broader view – is the client even asking the right questions?

In the travel time prediction problem, the clients asked which prediction method provides smaller errors, and whether that difference is statistically discernible. But a more fundamental question is how to produce predictions with smaller errors? That may involve not just choosing between two prediction methods, but combining them. Perhaps there are certain situations when one method is more accurate than the other, or perhaps a linear combination of them would be more effective. If the variances of the errors of the two methods are  $\sigma_i^2$ ,  $i = 1, 2$ , and the correlation (a measure of how closely they are related, see Section 9.2) between them is  $\rho$ , then the variance of the error when using the combination  $\lambda\hat{\theta}_1 + (1 - \lambda)\hat{\theta}_2$  is  $\lambda^2\sigma_1^2 + (1 - \lambda)^2\sigma_2^2 + 2\lambda(1 - \lambda)\rho\sigma_1\sigma_2$ , and we can optimize this with respect to  $\lambda$ . (See Exercise 6.48).

### 6.4.2 Weights

In practice, estimates should often give more weight to some observations than to others. For example, in survey sampling, we may over-sample some groups of particular interest, then give those observations less weight when computing estimates. Sometimes duplicate observations are combined, with the count of duplicates recorded – we should use those counts as weights. Other times we can simply obtain better estimates by using weights. For example, in the Google mobile ads case study in Section 1.12, the observations correspond to campaigns of very different sizes – number of ad impressions, clicks, and conversions – and quantities like cost per click are naturally more variable for small campaigns, and result in outliers. We should weight the observations by some measure of size, to take advantage of the more accurate measurements. See Section 8.7, where we do this for the Google mobile ads case study of Section 1.12.

Here we derive the optimal weights, for the case of a weighted mean; the result is useful in other situations. Suppose that  $Y_1, \dots, Y_n$  are independent, with common mean  $\mu$  and individual variances  $\sigma_i^2$ . We will compute a weighted mean

$$\hat{\mu} = \sum_{i=1}^n w_i Y_i,$$

---

<sup>2</sup> <https://cran.r-project.org/web/views/Robust.html>.

with  $\sum_{i=1}^n w_i = 1$ . The variance is

$$\text{Var}[\hat{\mu}] = \sum_{i=1}^n w_i^2 \sigma_i^2.$$

Using Lagrange multipliers from multivariate calculus, we find that the optimal weights are  $w_i = c/\sigma_i^2$ , with  $c$  chosen so the weights sum to 1. The smaller that  $\sigma_i$  is, the more accurate the observation, and the larger the weight it receives.

Now consider that result another way. Suppose, for example that each  $Y_i$  is the average of some number of observations  $X_{ij}$  for  $j = 1, \dots, m_i$ , independent with common variance  $\sigma_X^2$ . Then  $\sigma_i^2 = \sigma_X^2/m_i$ . When  $m_i$  is large,  $Y_i$  is more accurate. The weights  $w_i = cm_i/\sigma_X^2$  are proportional to  $m_i$ ; in other words, we weight the  $Y_i$  proportional to the number of observations that they represent, and the weighted average of  $Y$ 's corresponds to a simple average of  $X$ 's.

Computing weighted estimates – weighted means, medians, Variances, etc., – is typically straightforward common sense, and many R functions provide weighted estimates, e.g. `weighted.mean`. However, computing standard errors for weighted estimates is complicated, it depends on how the weights arise. When an R function computes standard errors, before using them you should verify that the assumptions the function makes agree with the random sampling method and weights calculation that produced your data.

## Exercises

- 6.1** Let  $X$  be a binomial random variable,  $X \sim \text{Binom}(n, p)$ . Show that the MLE of  $p$  is  $\hat{p} = X/n$ .
- 6.2** Prove Proposition 6.2
- 6.3** Suppose a random sample with  $X_1 = 5, X_2 = 9, X_3 = 9, X_4 = 10$  is drawn from a distribution with pdf  $f(x; \theta) = \theta/(2\sqrt{x})e^{-\theta\sqrt{x}}$ , where  $x > 0$ . Use maximum likelihood to find an estimate for  $\theta$ .
- 6.4** Let  $X_1, X_2, \dots, X_n$  be a random sample from the distribution with pdf  $f(x; \theta) = (x^3 e^{-x/\theta}) / (6\theta^4)$  for  $x \geq 0$ . Calculate the maximum likelihood estimate of  $\theta$ .
- 6.5** Recall Theorem 6.1 where we found the maximum likelihood estimates for  $\mu$  and  $\sigma$  for a random sample  $X_1, X_2, \dots, X_n \sim N(\mu, \sigma^2)$ .
  - (a) Suppose instead,  $\mu$  is Unknown, but  $\sigma$  is known. Find the maximum likelihood estimate of  $\mu$ .
  - (b) Now, suppose  $\sigma$  is unknown and  $\mu$  is known. Find the MLE of  $\sigma$ .

- 6.6** Let  $X_1, X_2, \dots, X_n$  be a random sample from a distribution with pdf  $f(x; \theta) = e^{\theta-x}$  for  $x > \theta > 0$ .
- Show that the MLE of  $\theta$  does not exist.
  - What if we change the domain of  $X$  to  $X \geq \theta > 0$ ?
- 6.7** Let  $X_1, X_2, \dots, X_n$  be a random sample from a distribution with pdf  $f(x; \theta) = e^{-x}/(1 - e^{-\theta})$  for  $0 < x < \theta$ . Find the MLE of  $\theta$ .
- 6.8** The Maxwell–Boltzmann distribution is used to model the speed of particles in gases. The pdf is  $f(x; \theta) = \sqrt{2/\pi} x^2 e^{-x^2/(2\theta^2)}/\theta^3$ . If  $X_1, X_2, \dots, X_n$  are a random sample from this distribution, find the MLE of  $\theta$ .
- 6.9** Katie has  $N$  keys in her purse of which one opens the door to her house. When she arrives home at the end of a long day, she puts her hand into her purse, randomly grabs a key and tries to unlock the door. If she fails, she puts the key back into her purse and then randomly draws another key. Suppose over the course of 5 days, she manages to unlock her front door on the 8th, 12th, 7th, 6th, and 12th attempts. Find the MLE of  $N$ , the number of keys in her purse. (Assume the keys are similar, so each has the same chance to be picked.)
- 6.10** Suppose the weight  $X$  (in pounds) of a girl in a certain town has distribution  $N(\mu, 15^2)$ , while the weight  $Y$  of a boy in this town has distribution  $N(1.3\mu, 20^2)$ . The weights of a randomly chosen girl and boy are  $x = 95$ ,  $y = 130$  lb, respectively. Find the MLE of  $\mu$ .
- 6.11** Let  $f(x; \theta) = \theta x^{\theta-1}$ ,  $0 \leq x \leq 1, \theta > 0$ . Suppose  $X_1, X_2, X_3 \stackrel{\text{i.i.d.}}{\sim} f(x; \theta)$  and  $Y_1, Y_2 \stackrel{\text{i.i.d.}}{\sim} f(x; 2\theta)$ , where the  $X$ 's and  $Y$ 's are independent. If  $X_1 = 1, X_2 = 1/2, X_3 = 1/2$  and  $Y_1 = 1/3, Y_2 = 1/4$ , find the MLE of  $\theta$ .
- 6.12** Let  $X_1, X_2$  be random sample from  $N(\mu, 3^2)$  and  $Y_1, Y_2$  be a random sample from  $N(2\mu, 3^2)$ , and assume the  $X'_i$ 's are independent of the  $Y'_j$ 's. Suppose  $X_1 = 3, X_2 = -1, Y_1 = 3, Y_2 = 2$ . Find the MLE of  $\mu$ .
- 6.13** Let  $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Exp}(\lambda)$  and  $Y_1, Y_2, \dots, Y_m \stackrel{\text{i.i.d.}}{\sim} \text{Exp}(2\lambda)$ , and assume the  $X'_i$ 's are independent of the  $Y'_j$ 's. Find the MLE of  $\lambda$ .
- 6.14** Let  $X_1, X_2, \dots, X_n$  denote a random sample from the gamma distribution  $\text{Gamma}(r, \lambda)$ . Assume  $r$  and  $\lambda$  are unknown. Write down the equations that would be solved simultaneously to find the maximum likelihood estimators of  $r$  and  $\lambda$ .

- 6.15** Suppose the random variable  $X$  has pdf  $f(x; \alpha, \beta) = \alpha\beta x^{\beta-1}e^{-\alpha x^\beta}$  for  $x \geq 0; \alpha, \beta > 0$ .
- Find the maximum likelihood estimator for  $\alpha$ , assuming that  $\beta$  is known.
  - Suppose  $\alpha$  and  $\beta$  are both unknown. Write down the equations that would be solved simultaneously to find the maximum likelihood estimators of  $\alpha$  and  $\beta$ .
- 6.16** Let the five numbers 2, 3, 5, 9, 10 come from the uniform distribution on  $[\alpha, \beta]$ . Find the method of moments estimates of  $\alpha$  and  $\beta$ .
- 6.17** Let  $x_1 = 7.13, x_2 = 5.26, x_3 = 9.93, x_4 = 6.62, x_5 = 7.52$  be five random numbers from the gamma distribution  $\text{Gamma}(r, \lambda)$ . Use the method of moments to find estimates for  $r$  and  $\lambda$ .
- 6.18** Nobody likes to stand in line to order a hamburger or buy groceries, so understanding how to model service times (time to be served) is one area of research in queuing theory. Four students collected data on customers waiting in line at a college snack bar (B. Haynor et al. private communication, 2010). The data set `Service` contains service times (in minutes) for 174 customers. We will model the distribution of service times using the gamma distribution.
- Use the method of moments to estimate the parameters of the gamma distribution.
  - Create a histogram and ecdf of the data with the theoretical gamma distribution superimposed.
- 6.19** The Weibull distribution has been used to model the time between successive earthquakes (Hasumi et al., 2009; Tiampo et al., 2008). The data set `Quakes` contains the time between earthquakes (in days) for all earthquakes of magnitude 6 or greater from 1970 to 2009.<sup>3</sup> Modify the R scripts from the wind energy case study to model the times between earthquakes with the Weibull distribution. Include graphs like those in Figure 6.4. Hint: When searching for the shape parameter with the `uniroot` function, use `lower=0.8, upper=1`.
- 6.20** Let  $X_1 = 2, X_2 = 2, X_3 = 8, X_4 = 12$  be a random sample from the distribution with pdf  $f(x) = \lambda e^{-\lambda(x-a)}$ , where  $x \geq a > 0$  and  $\lambda > 0$ . Use the method of moments to find estimates of  $\lambda$  and  $a$ .

---

<sup>3</sup> <http://earthquake.usgs.gov/earthquakes>.

- 6.21** Let  $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$  denote a random sample from a distribution with pdf  $f(x; \theta) = \theta 2^\theta / x^{\theta+1}$ ,  $x \geq 2, \theta > 1$ .
- Use the method of moments to estimate  $\theta$ .
  - Use maximum likelihood to estimate  $\theta$ .
- 6.22** Let  $X_1 = 0.4, X_2 = 0.5, X_3 = 0.25, X_4 = 0.9, X_5 = 0.92$  be a random sample from a distribution with pdf  $f(x; \theta) = \theta x^{\theta-1}$ ,  $0 \leq x \leq 1, \theta > 0$ .
- Find the MLE of  $\theta$ .
  - Find the method of moments estimate of  $\theta$ .
- 6.23** Let  $X_1, X_2, \dots, X_m \stackrel{\text{i.i.d.}}{\sim} \text{Binom}(n, p)$ , independent binomial random variables. Assume that both  $n$  and  $p$  are unknown parameters. Suppose for  $m = 5$ , we have  $x_1 = 6, x_2 = 8, x_3 = 6, x_4 = 9$ , and  $x_5 = 7$ . Use the method of moments to estimate  $n$  and  $p$ .
- 6.24** Let  $X$  be a binomial random variable  $X \sim \text{Binom}(n, p)$ . Show that  $\hat{p} = X/n$  is an unbiased estimator of  $p$ .
- 6.25** Verify Equation (6.16).
- 6.26** Return to the setting in Exercise 6.6. We will use  $\bar{X}$  as an estimator of  $\theta$ . Determine whether or not  $\bar{X}$  is an unbiased estimator. If it is biased, calculate the bias and determine whether or not it is possible to obtain an unbiased estimator from it.
- 6.27** Let  $X_1, X_2, \dots, X_n$  be i.i.d. from the negative binomial distribution with  $P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}$ ,  $x = r, r+1, \dots$ , (see Section B.4). Show that  $\hat{p} = (r-1)/(X-1)$  is an unbiased estimator of  $p$ . Hint:  $1/(1-w)^m = \sum_{k=0}^{\infty} \binom{k+m-1}{m-1} w^k$ .
- 6.28** Let  $X_1, X_2, \dots, X_n$  be random variables with  $E[X_i] = \mu, i = 1, 2, \dots, n$ . Under what condition on the constants  $a_1, a_2, \dots, a_n$  is  $X = a_1 X_1 + a_2 X_2 + \dots + a_n X_n$  an unbiased estimator of  $\mu$ ?
- 6.29** Let  $X_1, X_2, \dots, X_n$  be independent random variables from a distribution with mean  $\mu = 0$  and variance  $\sigma^2$ . Let  $W = (1/n) \sum_{i=1}^n X_i^2$ . Show that  $W$  is an unbiased estimator of  $\sigma^2$ .
- 6.30** Let  $\hat{\theta}$  be an unbiased estimator of  $\theta$  and  $h(x) = a + bx$ , where  $a, b$  are constants. Show that  $h(\hat{\theta})$  is an unbiased estimator of  $h(\theta)$ .

- 6.31** Recall that the sample mean  $\bar{X}$  is an unbiased estimator of the population mean  $\mu$  (Theorem A.7). Let  $h(x) = 5x^2$ . Show that  $h(\bar{X})$  a biased estimator of  $h(\mu)$ . Is  $h(\bar{X})$  asymptotically unbiased?
- 6.32** Let  $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$  and  $\hat{\sigma}^2$  denote the MLE of  $\sigma^2$  (Theorem 6.1).
- (a) Find the bias of  $\hat{\sigma}^2$ .
  - (b) Find the variance of  $\hat{\sigma}^2$ . Hint: Theorem B.16.
  - (c) Find the mean square error of  $\hat{\sigma}^2$ .
- 6.33** Let  $\hat{\theta}_1$  and  $\hat{\theta}_2$  be two estimators of  $\theta$  with  $E[\hat{\theta}_1] = 0.9\theta$  and  $E[\hat{\theta}_2] = 1.2\theta$ . Also, suppose  $\text{Var}[\hat{\theta}_1] = 3$  and  $\text{Var}[\hat{\theta}_2] = 2$ . Find two unbiased estimators of  $\theta$  and determine which one is more efficient.
- 6.34** Suppose  $X_1, X_2, X_3$  are a random sample of size 3 from a distribution with pdf  $f(x) = (1/\theta)e^{-x/\theta}$  for  $x > 0, \theta > 0$ . Let  $\hat{\theta}_1 = X_1$ ,  $\hat{\theta}_2 = (X_1 + X_2)/2$  and  $\hat{\theta}_3 = (X_1 + 2X_2)/3$ . Show that these estimators of  $\theta$  are all unbiased and determine the relative efficiencies between them.
- 6.35** Let  $X_1, X_2, \dots, X_n$  be independent random variables from a distribution with mean  $\mu$  and variance  $\sigma^2$ . Let  $\hat{\mu}_1 = (1/2)(X_1 + X_2)$  and  $\hat{\mu}_2 = (1/4)X_1 + (X_2 + X_3 + \dots + X_{n-1})/(2n - 4) + (1/4)X_n$ . Show that these two estimators of  $\mu$  are unbiased and determine which is more efficient.
- 6.36** Let  $\hat{\theta}_1$  and  $\hat{\theta}_2$  be two different estimators for a parameter  $\theta$ . Suppose  $\text{Var}[\hat{\theta}_1] = 25$  and  $\text{Var}[\hat{\theta}_2] = 4$ .
- (a) If  $E[\hat{\theta}_1] = \theta$  and  $E[\hat{\theta}_2] = \theta + 3$ , which estimator has the smaller mean square error?
  - (b) Suppose  $E[\hat{\theta}_1] = \theta$  and  $E[\hat{\theta}_2] = \theta + b$ , for some positive number  $b$ . For what values of  $b$  (if any) does  $\hat{\theta}_2$  have a smaller mean square error than  $\hat{\theta}_1$ ?
- 6.37** Consider the MSE of the two different estimators for the binomial proportion in Section 6.3.3. The R code shows how to graph the MSE as a function of  $p$ . Recreate these graphs but with sample sizes  $n = 30$ ,  $n = 50$ ,  $n = 100$ ,  $n = 200$ . What is the effect of increasing the sample size?

- 6.38** Let  $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Unif}[0, \beta]$ . Let  $\hat{\beta}_1 = (n+1)X_{\min}$ .
- Is  $\hat{\beta}_1$  an unbiased estimator of  $\beta$ ?
  - Let  $\hat{\beta}_2$  be  $((n+1)/n)X_{\max}$ . (This is an unbiased estimator that we considered earlier in this chapter.) Compute  $\text{Var}[\hat{\beta}_1]/\text{Var}[\hat{\beta}_2]$ . What do you conclude?
- 6.39** Let  $X_1, X_2, X_3$  be randomly drawn from a distribution with pdf  $f(x; \theta) = 2\theta^2 x$  for  $0 < x < 1/\theta$ .
- Find the expected value of  $X_i$ .
  - Let  $T = X_1/9 + X_2/9 + X_3/3$  be an estimator of  $1/\theta$ . Find the bias and mean square error.
  - If possible, use  $T$  to get an unbiased estimator for  $1/\theta$ .
- 6.40** Let  $X_1, X_2, \dots, X_n$  be a random sample from a distribution with pdf  $f(x) = (6/\theta^6)x^5$  for  $0 \leq x \leq \theta$ . We will use  $X_{\max}$  as an estimator of  $\theta$ .
- Find the pdf for  $X_{\max}$ , the maximum of the sample.
  - Compute  $E[X_{\max}]$ .
  - Compute the bias of  $X_{\max}$ .
  - Compute the mean square error of  $X_{\max}$ .
- 6.41** Let  $X_1, X_2, \dots, X_n$  be i.i.d. from a distribution with pdf  $f(x; \beta) = \alpha x^{\alpha-1}/\beta^\alpha$  for  $0 \leq x \leq \beta$ , where  $\alpha > 0$  is a known constant, but  $\beta$  is unknown. We will use  $X_{\max}$  as an estimator of  $\beta$ .
- Find the pdf for  $X_{\max}$ , the maximum of the sample.
  - Compute  $E[X_{\max}]$ .
  - Compute the bias of  $X_{\max}$ .
  - Compute the mean square error of  $X_{\max}$ .
- 6.42** Let  $X_1, X_2, \dots, X_n$  be independent random variables with mean  $\mu$  and variance  $\sigma_1^2$ . Let  $Y_1, Y_2, \dots, Y_m$  be independent random variables with mean  $\mu$  and variance  $\sigma_2^2$ . Let  $W = a\bar{X} + (1-a)\bar{Y}$ , where  $0 < a < 1$ .
- Compute the expected value of  $W$ .
  - For what value of  $a$  is the variance of  $W$  a minimum?
- 6.43** Let  $X_1, X_2$  be independent exponential random variables with parameter  $\lambda$ . Let  $\bar{X} = (X_1 + X_2)/2$  be an estimator of  $1/\lambda$ .
- Show that  $\bar{X}$  is an unbiased estimator of  $1/\lambda$ .
  - Show that  $\text{Var}[\bar{X}] = 1/(2\lambda^2)$ .
  - Show that  $E[\sqrt{X_1 X_2}] = \pi/(4\lambda)$ . Fact:  $E[\sqrt{X_i}] = \sqrt{\pi}/(2\sqrt{\lambda})$ .
  - Compute the bias of the estimator  $\sqrt{X_1 X_2}$  of  $1/\lambda$ .

- 6.44** In Chapter 5, we claimed that “the ratio of sample means is not generally an unbiased estimator of the ratio of population means.” Prove this under the additional conditions that all random variables are strictly positive, that the denominator has nonzero variance, and that the numerator and denominator are independent.
- Show that  $1/\bar{X}$  is not unbiased for  $1/\mu$ . Hint: Approximate  $f(x) = 1/x$  with a Taylor series expansion about  $\mu$  and then evaluate at  $x = \bar{X}$ .
  - Show that if the two samples are independent,  $\bar{Y}/\bar{X}$  is not unbiased for  $\mu_Y/\mu_X$ .
- 6.45** Let  $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Unif}[0, \beta]$  and let  $\hat{\beta}_n = X_{\max}$ . Show that the sequence  $\{\hat{\beta}_n\}$  is consistent for  $\beta$ .
- 6.46** Let  $X_1, X_2, \dots, X_n$  be i.i.d. from a normal distribution with mean  $\mu$  and variance  $\sigma^2$ . Show that  $\hat{\sigma}_n^2 = (1/n) \sum_{i=1}^n (X_i - \bar{X})^2$  is a consistent estimator of  $\sigma^2$ . Hint: Theorem B.16.
- 6.47** Let  $X_1, X_2, \dots, X_n$  be a random sample from the exponential distribution with parameter  $\lambda > 0$ . Determine whether or not  $\hat{\lambda}_n = n / \sum_{i=1}^n X_i$  is a consistent estimator of  $\lambda$ .
- 6.48** Refer to Section 6.4. Suppose we have two methods for making a prediction, and the variances of the errors of the two methods are  $\sigma_1$  and  $\sigma_2$ , let  $\rho$  denote the correlation between the errors. If we use the combination  $\lambda \text{prediction}_1 + (1 - \lambda)\text{prediction}_2$ , then the variance of the error is  $\lambda^2\sigma_1^2 + (1 - \lambda)^2\sigma_2^2 + 2\lambda(1 - \lambda)\rho\sigma_1\sigma_2$ . Find the  $\lambda$  that minimizes this variance.

# 7

## More Confidence Intervals

According to a Gallup Poll conducted in 2021, 41% of 1010 adults aged 18 years of age or older considered themselves an environmentalist.<sup>1</sup> We learned in Chapter 6 that  $\hat{p} = 0.41$  is an unbiased point estimate of the true proportion  $p$  of adults who consider themselves an environmentalist, but we do not have any indication of how far off  $\hat{p}$  is from the true  $p$ . In Chapter 5, we used bootstrap percentile confidence intervals to give a range of plausible values for a parameter.

With the bootstrap, we use the data itself, through resampling to get an estimate for the variability of the sampling distribution of the parameter. In this chapter, we learn more classical approaches to constructing interval estimates, using theoretical models for the sampling distribution. We will also encounter the bootstrap again.

### 7.1 Confidence Intervals for Means

We begin with confidence intervals for a mean, or a difference in means. These are important in their own right and illustrate techniques that are useful for other situations.

#### 7.1.1 Confidence Intervals for a Mean, Variance Known

**Example 7.1** The Centers for Disease Control and Prevention maintains growth charts for infants and children.<sup>2</sup> For 13-year-old girls, the mean weight is 101 lb with a standard deviation of 24.6 lb. We assume that weights are normally distributed. The public health officials in Sodor are interested in the weights of the teens in their town: They suspect that the mean weight of their girls might be different from the mean weight in the growth chart, but are

1 <https://news.gallup.com/poll/348227/one-four-americans-say-environmentalists.aspx>.

2 <http://www.cdc.gov/growthcharts/zscore.htm>.

willing to assume that the variation is the same. If they survey a random sample of 150 thirteen-year-old girls and find that their mean weight – an estimate of the population mean weight – is 95 lb, how accurate will this estimate be?

We assume the 150 sample values are from a normal distribution,  $N(\mu, 24.6^2)$ . Then the sampling distribution of mean weights is  $N(\mu, 24.6^2/150)$ , by Corollary A.2. Let  $\bar{X}$  denote the mean of the 150 weights, so standardizing gives  $Z = (\bar{X} - \mu)/(24.6/\sqrt{150}) \sim N(0, 1)$ . For the standard normal random variable  $Z$ , we have  $P(-1.96 < Z < 1.96) = 0.95$ . Thus, we compute

$$\begin{aligned} 0.95 &= P(-1.96 < \frac{\bar{X} - \mu}{24.6/\sqrt{150}} < 1.96) \\ &= P(-1.96(24.6/\sqrt{150}) < \bar{X} - \mu < 1.96(24.6/\sqrt{150})) \\ &= P(-\bar{X} - 1.96(24.6/\sqrt{150}) < -\mu < -\bar{X} + 1.96(24.6/\sqrt{150})) \\ &= P(\bar{X} + 1.96(24.6/\sqrt{150}) > \mu > \bar{X} - 1.96(24.6/\sqrt{150})) \\ &= P(\bar{X} - 3.937 < \mu < \bar{X} + 3.937). \end{aligned} \tag{7.1}$$

The random interval  $(\bar{X} - 3.937, \bar{X} + 3.937)$  has a probability of 0.95 of containing the mean  $\mu$ . Now, once you have drawn your sample, the random variable  $\bar{X}$  is replaced by the (observed) sample mean weight of  $\bar{x} = 95$ , and the interval  $(91.1, 98.9)$  is no longer a random interval. We interpret this interval by stating that we are 95% confident that the population mean weight of 13-year-old girls in Sodor is between 91.1 and 98.9 lb.  $\square$

### Remark

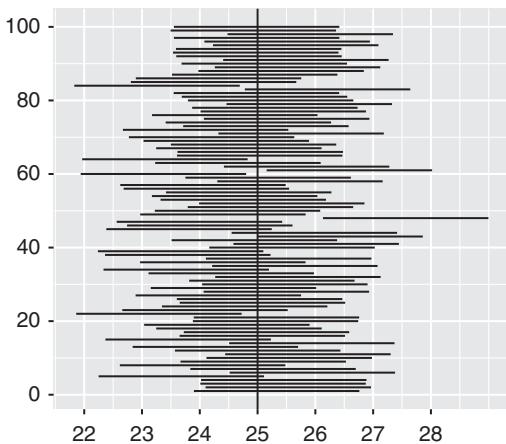
- The trick of doing algebra on equations that are inside a probability is often handy.
- Be careful when reading an equation such as  $0.95 = P(\bar{X} - 3.937 < \mu < \bar{X} + 3.937)$ . This does not mean that  $\mu$  is random, with a 95% probability of falling between two values. The parameter  $\mu$  is an unknown *constant*. Instead, it is the interval that is random, with a 95% probability of including  $\mu$ .

In the previous example, we computed a confidence interval of  $(91.1, 98.9)$ . We should not attribute a probability to this interval: the true mean is either in this interval or it is not! The statement “we are 95% confident” means that if we repeated the same process of drawing samples and computing intervals many times, then in the long run, 95% of the intervals would include  $\mu$ .  $\parallel$

More generally, for a sample of size  $n$  drawn from a normal distribution with unknown  $\mu$  and known  $\sigma^2$ , a 95% confidence interval for the mean  $\mu$  is

$$\left( \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right). \tag{7.2}$$

**Figure 7.1** Random confidence intervals,  $X_i \sim N(25, 4^2)$ ,  $n = 30$ . Notice that several miss the mean 25.



If we draw thousands of random samples from a normal distribution with parameters  $\mu, \sigma^2$  and compute a 95% confidence interval from each sample, then about 95% of the intervals would contain  $\mu$ .

We illustrate this with a simulation by drawing random samples of size 30 from  $N(25, 4^2)$ . For each sample, we construct the 95% confidence interval and check to see if it contains  $\mu = 25$ . We do this 1000 times and keep track of the number of times that the interval contains  $\mu$ . For good measure, we will graph some of the random intervals (Figure 7.1).

### R Note

```

counter <- 0                      # set counter to 0
df <- data.frame(x = c(22, 28), y = c(1,100))
p <- ggplot(df, aes(x = x, y = y)) + geom_vline(xintercept = 25)

for (i in 1:1000)
{
  x <- rnorm(30, 25, 4)           # draw a random sample of size 30
  L <- mean(x) - 1.96*4/sqrt(30) # lower limit
  U <- mean(x) + 1.96*4/sqrt(30) # upper limit
  if (L < 25 && 25 < U)          # check if 25 is in interval
    counter <- counter + 1        # if yes, increase counter by 1
  if (i <= 100)                   # plot first 100 intervals
    p <- p + annotate("segment", x = L, xend = U, y = i, yend = i)
}

p
counter/1000                      # proportion of times interval contains mu.

```

**Example 7.2** An engineer tests the gas mileage of a random sample of 30 of a certain car model that are ready to be sold. The 95% confidence interval for the mean mileage of all the cars is (29.5, 33.4) mpg. Critique the following statements:

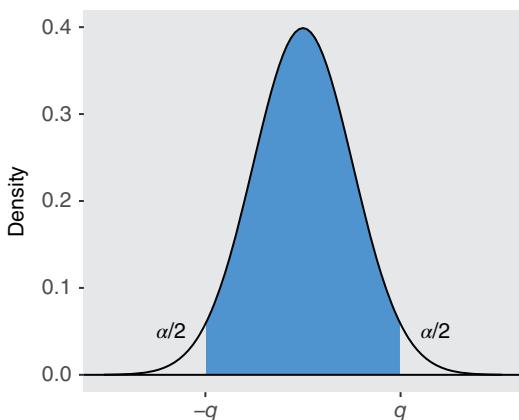
1. We are 95% confident that the gas mileage for cars of this model is between 29.5 and 33.4 mpg.
2. 95% of all samples will give an average mileage between 29.5 and 33.4 mpg.
3. There is a 95% chance that the true mean is between 29.5 and 33.4 mpg.

### Solution

1. *This is not correct:* A confidence interval is for a population parameter and in this case the mean, not for individuals.
2. *This is not correct:* Each sample will give rise to a *different* confidence interval and 95% of these intervals will contain the true mean.
3. *This is not correct:*  $\mu$  is not random. The probability that it is between 29.5 and 33.4 is 0 or 1.  $\square$

In our first example, we constructed a 95% confidence interval, but we can use other levels of confidence. More generally, let  $q$  denote the  $(1 - \alpha/2)$  quantile that satisfies  $P(Z < q) = 1 - \alpha/2$  (see Figure 7.2). Then, by symmetry  $P(-q < Z < q) = 1 - \alpha$ .

Let  $\bar{X}$  denote the mean of a random sample of size  $n$  from a normal distribution  $N(\mu, \sigma^2)$ . Then, since  $\bar{X} \sim N(\mu, \sigma^2/n)$ , by mimicking the algebra steps in



**Figure 7.2** Standard normal density with shaded area  $1 - \alpha$ .

Example 7.1 we have

$$\begin{aligned} 1 - \alpha &= P\left(-q < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < q\right) \\ &= P\left(\bar{X} - q\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + q\frac{\sigma}{\sqrt{n}}\right). \end{aligned}$$

In summary,

### Z Confidence Interval for a Normal Mean with Known Standard Deviation

If  $X_i \sim N(\mu, \sigma^2)$ ,  $i = 1 \dots n$  are a random sample, with  $\sigma$  known, then a  $(1 - \alpha) \times 100\%$  confidence interval for  $\mu$  is given by

$$\left(\bar{X} - q\frac{\sigma}{\sqrt{n}}, \bar{X} + q\frac{\sigma}{\sqrt{n}}\right), \quad (7.3)$$

where  $q$  denotes the  $(1 - \alpha/2)$  quantile of  $N(0, 1)$ .

It is common to write that quantile as  $z_{\alpha/2}$ , and write the interval as  $\bar{X} \pm z_{\alpha/2}\sigma/\sqrt{n}$ .

**Example 7.3** Suppose the random sample 3.4, 2.9, 2.8, 5.1, 6.3, 3.9 is drawn from the normal distribution with unknown mean  $\mu$  and known  $\sigma = 2.5$ . Find a 90% confidence interval for  $\mu$ .

#### Solution

The mean of the six numbers is 4.067. Here,  $1 - \alpha = 0.90$ , so  $\alpha/2 = 0.05$ . Thus, the 95th quantile is  $z_{0.05} = 1.645$  (i.e.  $P(Z < 1.645) = 0.95$ ). The 90% confidence interval is

$$\left(4.067 - 1.645\frac{2.5}{\sqrt{6}}, 4.067 + 1.645\frac{2.5}{\sqrt{6}}\right).$$

We are 90% confident that the population mean lies in the interval (2.389, 5.746).  $\square$

The term  $q(\sigma/\sqrt{n})$  (e.g.  $z_{\alpha/2}\sigma/\sqrt{n}$ ) is called the *margin of error*.

### Margin of Error

The *margin of error* (MOE) for a symmetric confidence interval is the distance from the estimate to either end. The confidence interval is of the form: estimate  $\pm$  MOE.

**Example 7.4** Suppose researchers want to estimate the mean weight of girls in Babb. They assume that the distribution of weights is normal with unknown mean  $\mu$ , but known standard deviation  $\sigma = 24.6$ . How many girls should they sample if they want, with 95% confidence, their margin of error to be at most 5 lb?

### Solution

Since  $q_{0.975} = 1.96$ , we set  $1.96(24.6/\sqrt{n}) \leq 5$ . This leads to  $n \geq 92.99$ , so there should be at least 93 girls in the sample.  $\square$

**Remark** Note that the width of the interval, given by  $q(\sigma/\sqrt{n})$ , depends on the level of confidence (which determines  $q$ ), the standard deviation, and the sample size. Analysts cannot control  $\sigma$ , but can adjust  $q$  or  $n$ .

To make the confidence interval narrower, they can either increase the sample size  $n$  or decrease the size of the quantile  $q$ , which amounts to decreasing the confidence level.  $\parallel$

### 7.1.2 Confidence Intervals for a Mean, Variance Unknown

In most real-life settings, a data analyst will not know either the mean or the standard deviation of the population of interest. How then would we get an interval estimate of the mean  $\mu$ ? We have used the sample mean  $\bar{X}$  as an estimate of  $\mu$ , so it seems natural to consider the sample standard deviation  $S$  as an estimate of the  $\sigma$ .

However, in deriving the confidence interval for  $\mu$ , we used the fact that  $(\bar{X} - \mu)/(\sigma/\sqrt{n})$  follows a standard normal distribution (if the population is normal). Does changing the  $\sigma$  to  $S$ , the sample standard deviation, change the distribution? We will use a simulation to investigate the distribution of  $(\bar{X} - \mu)/(S/\sqrt{n})$  for random samples drawn from  $N(\mu, \sigma^2)$ .

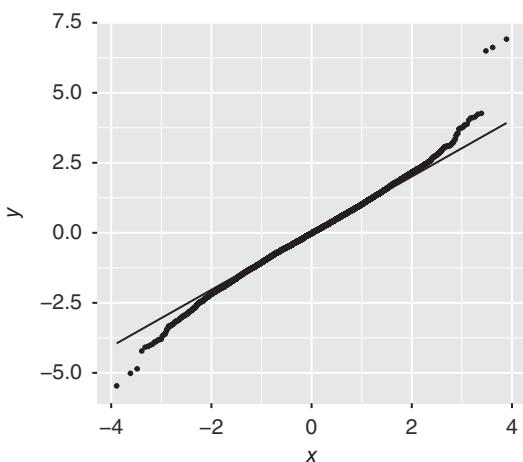
#### R Note

```

N <- 10^4
w <- numeric(N)
n <- 15           # sample size
for (i in 1:N)
{
  x <- rnorm(n, 25, 7)  # draw 15 from N(25, 7^2)
  xbar <- mean(x)
  s <- sd(x)
  w[i] <- (xbar-25) / (s/sqrt(n))
}
df <- data.frame(w)
ggplot(df, aes(sample = w)) + geom_qq(size = .8) +
  geom_qq_line()

```

**Figure 7.3** Normal quantile plot for sampling distribution of  $(\bar{X} - \mu)/(S/\sqrt{n})$ .

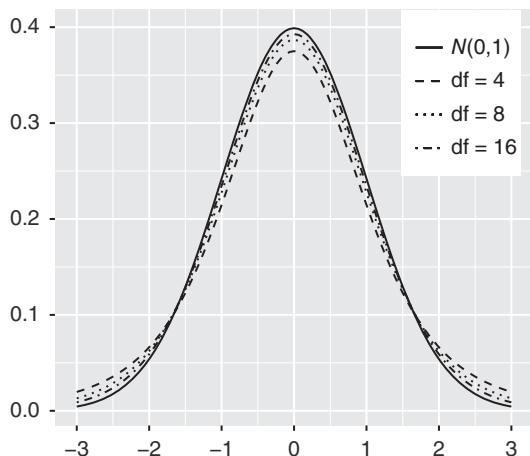


This distribution does have slightly longer tails than the normal distribution; you could not tell this from a histogram, but it is apparent in the normal quantile plot (Figure 7.3). Sometimes  $S$  is smaller than  $\sigma$ , and when the denominator is small, the ratio is large. In effect, having to estimate  $\sigma$  using  $S$  adds variability.

It turns out that  $T = (\bar{X} - \mu)/(S/\sqrt{n})$  has a Students  $t$  distribution with  $n - 1$  degrees of freedom (if the population is normal).

The density of a  $t$  distribution with  $k$  degrees of freedom is bell-shaped and symmetric about 0, with heavier (longer) tails than the standard normal. As  $k$  tends toward infinity, the density of the  $t$  distribution tends toward the density of the standard normal. For more details on this distribution, refer to Section B.11. Figure 7.4 shows densities for the standard normal and two  $t$  distributions; the shapes are similar, though the  $t$  densities are far greater in the tails.

**Figure 7.4** Density for standard normal and students  $t$  distributions with 4, 8, and 16 degrees of freedom.



We derive the confidence interval for  $\mu$  when  $\sigma$  is unknown in the same way as when  $\sigma$  is known. Let  $q = q_{1-\alpha/2}$  denote the  $(1 - \alpha/2)$  quantile of the  $t$  distribution with  $n - 1$  degrees of freedom,  $P(T < q) = 1 - \alpha/2$ ,  $0 < \alpha < 1$ . Then using symmetry of the  $t$  distribution, we have

$$\begin{aligned} 1 - \alpha &= P\left(-q < \frac{\bar{X} - \mu}{S/\sqrt{n}} < q\right) \\ &\vdots \\ &= P\left(\bar{X} - q\frac{S}{\sqrt{n}} < \mu < \bar{X} + q\frac{S}{\sqrt{n}}\right). \end{aligned}$$

### T Confidence Interval for a Normal Mean with Unknown Standard Deviation

If  $X_i \stackrel{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ ,  $i = 1 \dots n$  is a random sample, with  $\sigma$  unknown, then a  $(1 - \alpha) \times 100\%$  confidence interval for  $\mu$  is given by

$$\left(\bar{X} - q\frac{S}{\sqrt{n}}, \bar{X} + q\frac{S}{\sqrt{n}}\right), \quad (7.4)$$

where  $q$  denotes the  $(1 - \alpha/2)$  quantile of the  $t$  distribution with  $(n - 1)$  degrees of freedom.

It is common to write that quantile as  $t_{\alpha/2, n-1}$ , and write the interval as  $\bar{X} \pm t_{\alpha/2, n-1} s / \sqrt{n}$ .

**Example 7.5** The distribution of weights of boys in Babb is normal with unknown mean  $\mu$ . From a random sample of 28 boys, we find a sample mean of 110lb and a sample standard deviation of 7.5lb. To compute a 90% confidence interval, find the 0.95 quantile of the  $t$  distribution with 27 degrees of freedom, which is  $q = 1.7033$ . The interval is  $(110 - 1.7033(7.5/\sqrt{28}), 110 + 1.7033(7.5/\sqrt{28}))$ ; thus, we are 90% confident that the true mean weight is between 107.6 and 112.4 lb.  $\square$

### R Note

The functions `pt` or `qt` give probabilities or quantiles, respectively, for the  $t$  distribution.

For instance, to find  $P(T < 2.8)$  for the random variable  $T$  from a  $t$  distribution with 27 degrees of freedom,

```
> pt(2.8, 27)
[1] 0.9953376
```

To find the quantile  $q_{0.95}$  satisfying  $P(T < q_{0.95}) = 0.95$ ,

```
> qt(0.95, 27)
[1] 1.703288
```

Compare the 0.95 quantile for a  $t$  distribution with 27 degrees to that of the standard normal: 1.703 versus 1.645. Thus, the  $t$  interval is slightly wider than the  $z$  interval, reflecting, as we noted earlier, the extra uncertainty in not knowing the true  $\sigma$ .

**Example 7.6** Find a 99% confidence interval for the mean weight of baby girls born in North Carolina in 2004 (case study in Section 1.2).

### Solution

The mean and standard deviation of the weights of the  $n = 521$  girls is 3398.317 and 485.691 g, respectively. The normal quantile plot in Figure 7.5 shows that the weights are approximately normally distributed, so a  $t$  interval is reasonable.  $1 - \alpha = 0.99$ , so  $\alpha/2 = 0.005$ . The 0.995 quantile for the  $t$  distribution with 520 degrees of freedom is  $q_{0.995} = 2.585$ . Thus, the 99% confidence interval is  $3398.317 \pm 2.585(485.691/\sqrt{521}) = (3343.30, 3453.33)$  g.

### R Note

Use the `t.test` function to find confidence intervals.

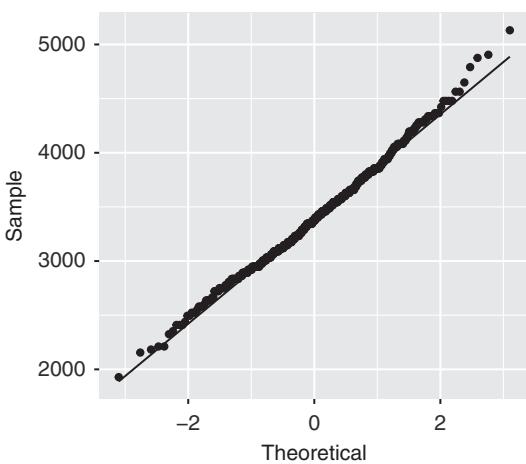
```
> girls <- NCBirths2004 %>% filter(Gender == "Female") %>%
  pull(Weight)
> t.test(girls, conf.level = .99)$conf
[1] 3343.305 3453.328
attr("conf.level")
[1] 0.99
```

□

#### 7.1.2.1 Check Conditions

With any statistical procedure, you should check whether the conditions underlying that procedure hold. This  $t$  confidence interval gives the desired coverage when observations are independent, identically distributed, and from a normal population.

Are the observations independent? Perhaps the data were collected by a clustered sampling procedure in which random regions were selected and then random observations selected from each region. If so the observations are not independent – data within a cluster tend to be more similar than data across clusters. Or were the data collected over time? If so there may be



**Figure 7.5** Normal quantile plot of weights of baby girls.

autocorrelation – that is, adjacent observations are correlated. You can check the correlation of  $(x_1, \dots, x_{n-1})$  and  $(x_2, \dots, x_n)$  (see Section 9.2). If there is autocorrelation, then the variance of  $\bar{X}$  is inflated, and the SE must reflect that or the interval is invalid.

**Remark** I (Tim) consulted for Verizon in a case for the Public Utilities Commission (PUC) of New York. I was a young guy and on the other side was an eminent statistician, who used ordinary two-sample  $t$  methods but neglected to take the autocorrelation of the observations into account. I showed that this completely invalidated the results, using simulation to help the PUC understand. An eyewitness reported that the other side was “furious, but of course they couldn’t refute any of it,” and Verizon won handily. ||

Are they identically distributed? Clusters may have different means and/or standard deviations. Data collected over time may have different distributions at different times.

If there is dependence or nonidentical distributions, you may still be able to use a  $t$  interval, but probably need to calculate the standard error and (less importantly) degrees of freedom differently.

Are there outliers? Since  $\bar{x}$  and  $s$  are sensitive to extreme values outliers can have a big impact on confidence intervals. You should investigate any outliers in your data: Are these recording errors or observations that are not representative of the population? If the former, correct them; and if the latter, remove them and re-do the analysis to see if they are influential. If the conclusion changes, you should report this along with the original results. But not every outlier should be removed; they may represent a small but important part of the population.

Better yet, even before you start computing estimates, consider using an estimate that is less sensitive to outliers than  $\bar{x}$ , for example, a trimmed mean (Section 5.5). Switching to a robust estimate after looking at the data is problematic; that corresponds to looking at the data, deciding you do not like the results, then throwing out some of the data to get a different result.

Finally, there is the question of normality. You can check whether the data appear normally distributed using a normal quantile plot. Suppose they do not appear normal; does that matter? How robust is the  $t$  interval to nonnormality? We turn to this question next.

### 7.1.2.2 Nonnormal Populations

$t$  confidence intervals were derived under the conditions that the underlying populations are normal. Then the interval is exact: A  $(1 - \alpha) \times 100\%$  interval covers  $\mu$  with probability  $1 - \alpha$  and misses  $\mu$  on either side with probability  $\alpha/2$ ; that is, the whole interval is above  $\mu$  with probability  $\alpha/2$  and below with probability  $\alpha/2$ .

In practice, they are often used for nonnormal populations. There are theoretical results that in the long run, as  $n \rightarrow \infty$ , they give the correct coverage. But how accurate are they in practice for finite  $n$ ? We will check this for a nonnormal population by running a simulation.

**Example 7.7** We draw random samples from the right-skewed gamma distribution with  $r = 5$  and  $\lambda = 2$  (see Figure 4.8) and count the number of times the 95% confidence interval misses the mean  $\mu = 5/2$  on each side.

#### R Note

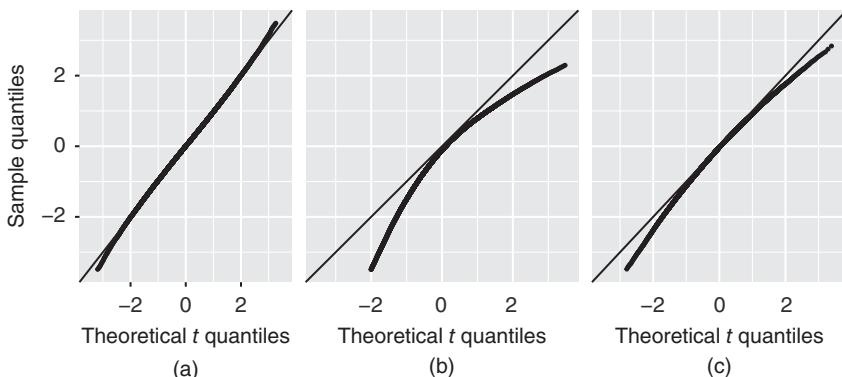
```
tooLow <- 0                      # set counter to 0
tooHigh <- 0                      # set counter to 0
n <- 20   # sample size
q <- qt(0.975, n-1)               # quantile
N <- 10^5
for (i in 1:N)
{
  x <- rgamma(n, shape = 5, rate = 2)
  xbar <- mean(x)
  s <- sd(x)
  L <- xbar - q*s/sqrt(n)
  U <- xbar + q*s/sqrt(n)
  if (U < 5/2)                   # Does right endpt miss 5/2?
    tooLow <- tooLow + 1          # If yes, increase counter
  if (5/2 < L)                   # Does left endpt miss 5/2?
    tooHigh <- tooHigh + 1        # If yes, increase counter
}
tooLow/N
tooHigh/N
```

What proportion of times did the confidence intervals miss the true mean  $5/2$ ? In one run of this simulation, about 4.5% of the time, the interval was too low and was below  $5/2$  and about 1.3% of the time, the interval was too high and was above the  $5/2$ . In Exercise 7.9, you will investigate the effect of changing the sample size.  $\square$

When the population is nonnormal but symmetric and the sample size is moderate or large, the  $t$  interval is very accurate. The main weakness of the  $t$  confidence interval occurs when the population is skewed. The simulation illustrates this problem.

To see this from another point of view, we will look at the distributions of the  $t$  statistics,  $T = (\bar{X} - \mu)/(S/\sqrt{n})$ , since the accuracy of  $t$  intervals depends on how close the  $t$  statistic is to having a  $t$  distribution.

Figure 7.6 compares the distribution of  $t$  statistics for samples of size  $n = 10$  from a uniform distribution, size  $n = 10$  from an exponential distribution, and size  $n = 100$  from an exponential distribution to the  $t$  distribution.<sup>3</sup> The range on all plots is truncated so that we can focus on the range of values important for confidence intervals. Notice that for the uniform population, the distribution of the  $t$  statistic is close to the  $t$  distribution, except in the tails. For exponential populations, the discrepancy is much larger, and the discrepancy decreases only slowly as the sample size increases. To reduce the discrepancy (the difference between actual and nominal probabilities) by a factor of 10 requires a sample size 100 times larger. For an exponential population, we must have  $n > 5000$ .



**Figure 7.6** Quantile–quantile plot of  $t$  statistics versus a  $t$  distribution with  $n - 1$  degrees of freedom for samples from (a) a uniform distribution with  $n = 10$ , (b) an exponential distribution with  $n = 10$ , and (c) an exponential distribution with  $n = 100$ .

<sup>3</sup> The principle here is the same as in the normal quantile plot. The quantiles of a sample are compared to the quantiles of the  $t$  distribution. If the distributions are the same, the points should roughly fall on the  $y = x$  line.

before the actual probabilities of a 95%  $t$  interval missing the true mean in either tail are within 10% of the desired probability of 2.5%; that is, the actual tail probabilities are between 2.25% and 2.75%.

Before using a  $t$  confidence interval, you should create a normal quantile plot to see whether the data are skewed. The larger the sample size, the more skewness can be tolerated. To see how much of an effect that skewness has for your size data, compare the interval to a bootstrap  $t$  interval, which handles skewness well; we cover this later (Section 7.5.2).

### 7.1.3 Confidence Intervals for a Difference in Means

Will new directed reading activities improve certain aspects of a child's reading ability? An educator conducted an experiment to test this (DASL). Twenty-one third graders took part in these directed reading activities for 8 weeks, while another class of 23 third graders did not. At the end of the study, all children took the degree of reading power (DRP) test, a standard test that measures various aspects of reading ability. The scores are shown in Figure 7.7.

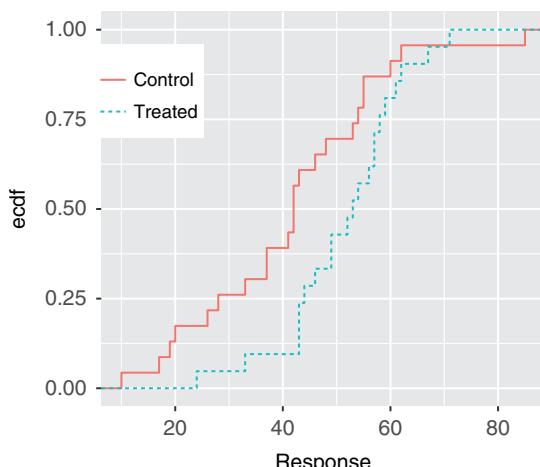
The mean DRP score for the students who received the treatment was 51.48, while the mean for the control group was 41.52. Though the difference seems large, it might be due to sampling variability. To check this, we need to understand the sampling distribution of  $\bar{X}_1 - \bar{X}_2$ .

Let  $X$  and  $Y$  be independent random variables with  $X \sim N(\mu_1, \sigma_1^2)$  and  $Y \sim N(\mu_2, \sigma_2^2)$ . Then from Theorem A.8,  $\bar{X} - \bar{Y} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$ .

For sample sizes  $n_1$  and  $n_2$

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right). \quad (7.5)$$

**Figure 7.7** Comparison of the empirical cumulative distribution functions (ecdfs) of reading scores for control and treatment groups. One student in the control group had an unusually high score (in the 80s) compared to everybody else in his/her group.



Of course, in practice we usually do not know the population variances, so we will plug in the sample variances. As in the single-sample case, we call this a  $t$  statistic,

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}. \quad (7.6)$$

The exact distribution of this statistic does not have a known simple solution. It does, however, have approximately a  $t$  distribution if the populations are normal. The difficult part is the degrees of freedom. A quick rule is to set the degrees of freedom equal to the smaller of  $n_1 - 1$  and  $n_2 - 1$ . A more accurate rule, which gives larger degrees of freedom (and hence shorter intervals), is Welch's approximation

$$\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}. \quad (7.7)$$

This is based on how accurately  $s_1^2/n_1 + s_2^2/n_2$  estimates  $\sigma_1^2/n_1 + \sigma_2^2/n_2$  (see Exercise 7.51).

In practice, the degrees of freedom is typically the least important part of the calculations, and the difference between the quick rule and Welch's approximation only matters if one of the sample sizes is below 30 (30 is about the point at which the difference between the rules may make a 10% difference in coverage probabilities for a two-sided 95% interval).

Let  $q$  denote the  $(1 - \alpha/2)$  quantile for the  $t$  distribution with  $\nu$  degrees of freedom:

$$P\left(-q < \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} < q\right) \approx 1 - \alpha$$

and solve both inequalities for  $\mu_1 - \mu_2$  to obtain a  $(1 - \alpha) \times 100\%$  confidence interval for  $\mu_1 - \mu_2$ .

### T Confidence Interval for a Difference in Means

If  $X_i \sim N(\mu_1, \sigma_1^2)$ ,  $i = 1, 2, \dots, n_1$  and  $Y_j \sim N(\mu_2, \sigma_2^2)$ ,  $j = 1, 2, \dots, n_2$ , then an approximate  $(1 - \alpha) \times 100\%$  confidence interval for  $\mu_1 - \mu_2$  is

$$(\bar{X} - \bar{Y}) \pm q \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, \quad (7.8)$$

where the degrees of freedom  $\nu$  is given either by Equation (7.7) or  $(\min(n_1, n_2) - 1)$  and  $q$  denotes the  $(1 - \alpha/2)$  quantile of the  $t$  distribution with  $\nu$  degrees of freedom.

**Example 7.8** (Reading scores in directed reading study.)

For the  $n_1 = 21$  third graders who participated in directed reading activities, the mean and the standard deviation of their DRP scores are  $\bar{x} = 51.48$ ,  $s_1 = 11.01$ , respectively, and for the  $n_2 = 23$  third-graders in the control group,  $\bar{y} = 41.52$ ,  $s_2 = 17.14$ . The mean difference is  $\bar{x} - \bar{y} = 9.96$  with the standard error of the difference and degrees of freedom

$$\sqrt{\frac{11.01^2}{21} + \frac{17.14^2}{23}} = 4.31 \quad \text{and} \quad \frac{(11.01^2/21 + 17.14^2/23)^2}{(11.01^2/21)^2/20 + (17.14^2/23)^2/22} \\ = 37.577,$$

respectively. The 0.975 quantile of the  $t$  distribution is 2.0251. Thus, the 95% confidence interval is  $9.96 \pm 2.0251(4.31) = (1.23, 18.69)$ . With 95% confidence, third-graders who participated in directed reading activities score, on average, from 1.23 to 18.69 points higher than third-graders who did not participate.

Now, one child in the control group received an unusually high score (85). If we remove this child, then the 95% confidence interval for the difference in mean scores becomes (3.98, 19.89). It does not appear that this child is influential: Removing this outlier does not change the confidence interval much.

The directed reading activities may improve DRP scores, on average, by as much as 19 points, or it may have little practical effect, improving by only about a few points.

**R Note**

```
t.test(Response ~ Treatment, data = Reading)$conf
[1] -18.67588 -1.23302
attr(, "conf.level"):
[1] 0.95
```

R computes the mean difference Control – Treatment.

□

**Remark** If the confidence interval for the difference in means contains 0, then we cannot rule out the possibility that the means might be the same,  $\mu_1 - \mu_2 = 0$  or, equivalently,  $\mu_1 = \mu_2$ .

If we were to construct separate 95% confidence intervals for the mean reading score for each of the treatment group and control group, then we would find an interval of (46.47, 56.49) for the treatment group and (34.11, 48.94) for the control group. The confidence intervals overlap, so we would not be able say whether or not there was a difference in the true mean scores. Looking for overlap in individual confidence intervals is not a good way to test for discernible differences. See Exercise 7.37. ||

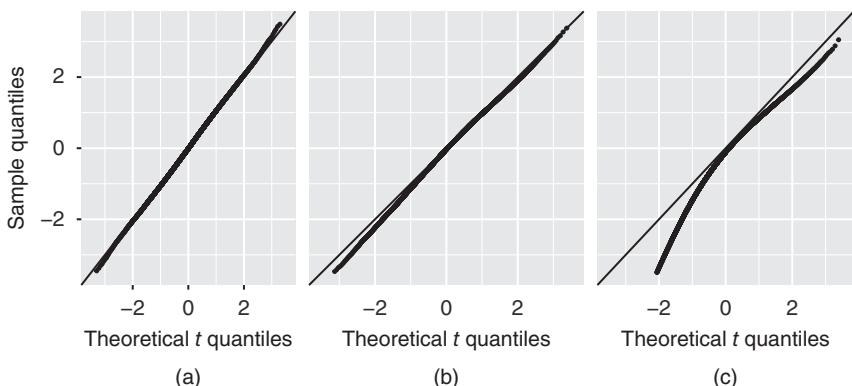
### 7.1.3.1 Check Conditions

As in the one-sample case, Section 7.1.2.1, you should check the conditions underlying these confidence intervals – that each sample is an i.i.d. sample from a normal population. Here there is one additional condition, that the two samples are independent. In particular, if the two samples are paired (e.g. before-and-after measurements on each person), then instead of a two-sample  $t$  interval you should do a paired  $t$  interval, see Section 7.1.4 below.

### 7.1.3.2 Nonnormal Populations, Revisited

Skewness is less of an issue for two-sample  $t$  confidence intervals than for one-sample intervals, as long as the samples are roughly the same size, because the skewness from the two samples tends to cancel out. In particular, if the populations have the same skewness and variance and the sample sizes are equal, then the skewness cancels out exactly, and the distribution of  $t$  statistics can be very close to a  $t$  distribution even for quite small samples. Otherwise, the bootstrap  $t$  interval (Section 7.5.2) handles skewness well, and also provides a way to check the effect of skewness on symmetric  $t$  intervals.

Figure 7.8 displays the results of three simulations in which two samples are drawn repeatedly from a right-skewed exponential distribution with  $\lambda = 1$ . In Figure 7.8a, the two samples are the same size, and we can see that the distribution of the two sample  $t$  statistic (Equation (7.6)) is very close to the  $t$  distribution. Figures 7.8b and c show the results when the two samples are different sizes. If the sample sizes are unbalanced, the skewness partially cancels out. The extreme case, as one sample size goes to infinity, reduces to a one-sample problem.



**Figure 7.8** Quantile–quantile plot of  $t$  statistics versus a  $t$  distribution when both samples are exponential and when the sample sizes are (a)  $n_1 = 10, n_2 = 10$ , (b)  $n_1 = 10, n_2 = 15$ , and (c)  $n_1 = 10, n_2 = 100$ . The degrees of freedom for the theoretical  $t$  quantiles are obtained using known variances in Welch's approximation.

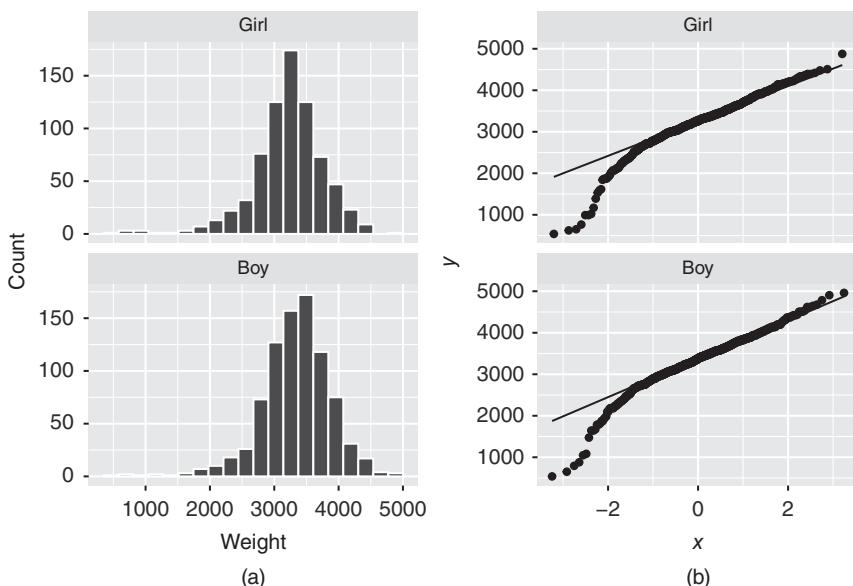
**Example 7.9** Consider the weights of boy and girl babies born in Texas in 2004 (case study in Section 1.2). Construct a 95%  $t$  confidence interval for the mean difference in weights (boys – girls).

### Solution

The weights are shown in Figure 7.9. The distribution of weights is left-skewed for both the boys and girls. While this may make  $t$  intervals for the individual means inaccurate, the skewness largely cancels out when computing a  $t$  confidence interval for the difference in means, when the sample sizes are similar.

The mean and standard deviation for the  $n_1 = 848$  boys are 3336.84 and 547.53 g, respectively, while for the  $n_2 = 739$  girls, they are 3220.94 and 550.82 g. Thus, the mean difference weights is 115.90 with standard error 27.64. The degrees of freedom is approximately 1552.91 so the corresponding 0.975 quantile is  $q_{0.975} = 1.96$ . The confidence interval is  $115.90 \pm 1.96 \cdot 27.64 = (61.68, 170.12)$  g. Thus, we are 95% confident that boy babies born in Texas in 2004 were, on average, from 61.68 to 170.12 g heavier than girl babies.

In this example, separate 95%  $t$  confidence intervals for the mean weights are (3299.94, 3373.75) for the boys and (3181.16, 3260.72) for the girls.  $\square$



**Figure 7.9** Weights of boy and girl babies born in Texas in 2004.

### 7.1.3.3 Pooling the Variances\*

If  $\sigma_1^2 = \sigma_2^2$ , then we can “pool the variances,” estimating the common variance using all the data. The *pooled sample variance* is

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

We have the following:

**Theorem 7.1** Let  $X_1, X_2, \dots, X_{n_1} \sim N(\mu_1, \sigma^2)$  and  $Y_1, Y_2, \dots, Y_{n_2} \sim N(\mu_2, \sigma^2)$  be two independent random samples with sample means and variances  $\bar{X}, S_1^2$  and  $\bar{Y}, S_2^2$ , respectively.

Then

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_p \sqrt{1/n_1 + 1/n_2}} \quad (7.9)$$

has a Student's  $t$ -distribution with  $(n_1 + n_2 - 2)$  degrees of freedom.  $\square$

This leads to the following:  $(1 - \alpha) \times 100\%$  *pooled variance two-sample t interval*, for  $\mu_1 - \mu_2$

$$\left( \bar{X} - \bar{Y} - q \cdot S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \bar{X} - \bar{Y} + q \cdot S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right), \quad (7.10)$$

where  $q$  is the  $(1 - \alpha/2)$  quantile of the  $t$  distribution with  $n_1 + n_2 - 2$  degrees of freedom.

In general, we advise against using this confidence interval. Pooling usually provides only a small gain when the variances are the same, but can be badly off when the variances are different. Nor can we test whether the variances are the same and decide whether to pool based on that because testing for equal variances is hard; the usual formula approach is extremely sensitive to nonnormality, and does *not* get better as sample sizes get large. An eminent statistician, George Box, wrote about testing to decide on pooling: “To make the preliminary test on variances is rather like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port.” (Box (1953)). You could use a permutation test to compare the variances, but why not just use the safer unpooled procedure?

An exception is when one sample size is very small so that the sample provides little information about the variance of its population. In that case, it may be better to assume equal variances, in spite of the bias this causes when the variances are unequal, to avoid the extra variability caused by a wild variance estimate in the small sample. If one sample is size 1 you have no choice, you must pool. For more discussion about pooling variances, see Moser and Stevens (1992), Miao and Chiou (2008), or Scheffe (1970).

### 7.1.4 Matched Pairs, Revisited

The two-sample  $t$  intervals above assume that the samples are independent for computing the standard errors. They are not independent when data are paired (Section 3.4). Typically,  $x$  and  $y$  are positively associated so assuming independence would overestimate standard errors for  $\bar{X} - \bar{Y}$ .

There is a simple remedy – we take the differences of each pair,  $D_i = X_i - Y_i$ , and treat this as a one-sample problem.  $\bar{D} = \bar{X} - \bar{Y}$ , so the middle of the interval is the same, only the standard error differs. We compute the standard error for  $\bar{D}$  which is just the standard deviation of  $D$  divided by the square root of the sample size. A *paired t confidence interval* is the one-sample  $t$  interval for the differences.

We again use the diving data (Diving2017) and consider these 12 divers as a representative sample of all elite female divers. If  $\mu_S$  and  $\mu_F$  represent the true mean diving scores in the semifinal and final rounds, respectively, we would like an interval estimate of the difference in mean scores. We take the difference between the final and semifinal scores to obtain a new variable, and calculate a one-sample  $t$ -confidence interval for the mean of this variable.

The mean and standard deviation of the difference are 11.975 and 34.849, respectively. The interval is  $11.975 \pm (2.201)(34.849 / \sqrt{12}) = (-10.167, 34.117)$ , where 2.201 is the 0.975-quantile of a  $t$ -distribution with 11 degrees of freedom. Since 0 lies in this interval, we cannot rule out the possibility that the mean scores are the same.

## 7.2 Confidence Intervals Using Pivots

In Section 7.1, we derived confidence intervals for means and difference of means by considering the sampling distribution of a statistic that depended on both  $\mu$  and  $\bar{x}$ . We were able to solve for  $\mu$  to obtain lower and upper limits that did not involve  $\mu$ . We can use this same idea for finding confidence intervals for other parameters.

Suppose  $X_1, X_2, \dots, X_n$  are a random sample from a distribution  $F$  with parameter  $\theta$ . Suppose for all  $\theta$ ,

$$P(L < \theta < U) = 1 - \alpha, \quad (7.11)$$

where  $L = g_1(X_1, X_2, \dots, X_n)$  and  $U = g_2(X_1, X_2, \dots, X_n)$  are functions of the  $n$  random variables, but not  $\theta$ . Then the interval  $(L, U)$  is a  $(1 - \alpha) \times 100\%$  confidence interval for  $\theta$ .

Note that confidence intervals are open at the endpoints,  $(L, U)$  always means  $L < \theta < U$ .

In practice, the interval should also satisfy

$$P(\theta \leq L) \leq \alpha/2, \quad (7.12)$$

$$P(\theta \geq U) \geq \alpha/2, \quad (7.13)$$

(with equality “=  $\alpha/2$ ” in continuous problems) otherwise (i) the interval is misleading – people will expect that an interval misses  $\theta$  with equal probabilities on both sides – and (ii) it raises ethical red flags, the possibility that unbalanced probabilities are chosen to obtain a desired answer. For example, a 95% interval should not miss 4% on one side and 1% on the other.

**Example 7.10** In the 95% confidence interval for a mean with  $\sigma$  known (Equation (7.2)), we have  $L = \bar{X} - 1.96\sigma/\sqrt{n}$  and  $U = \bar{X} + 1.96\sigma/\sqrt{n}$ .  $\square$

One approach to forming a confidence interval is to find a *pivotal quantity* or *pivot*, a random variable that depends on the sample  $X_1, X_2, \dots, X_n$  and the parameter  $\theta$ , say  $h(X_1, X_2, \dots, X_n, \theta)$ , but whose distribution *does not* depend on  $\theta$  or on any unknown parameters. For example, for a normal population with  $\sigma$  known,  $(\bar{X} - \mu)/\sigma$ , and  $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ , and  $(\bar{X} - \mu)$ , and are all pivotal quantities since their distributions,  $N(0, 1/n)$ ,  $N(0, 1)$ , and  $N(0, \sigma^2/n)$ , respectively, do not depend on  $\mu$ . For a normal population with  $\sigma$  unknown,  $(\bar{X} - \mu)/(s/\sqrt{n})$  is pivotal.

If  $q_1$  and  $q_2$  denote the  $\alpha/2$  and  $1 - \alpha/2$  quantiles of the distribution of the pivot, respectively, then we have

$$P(q_1 < h(X_1, X_2, \dots, X_n, \theta) < q_2) = 1 - \alpha.$$

Set  $q_1 = h(X_1, X_2, \dots, X_n, \theta)$  and  $q_2 = h(X_1, X_2, \dots, X_n, \theta)$  and solve for  $\theta$  to find the limits of the interval.

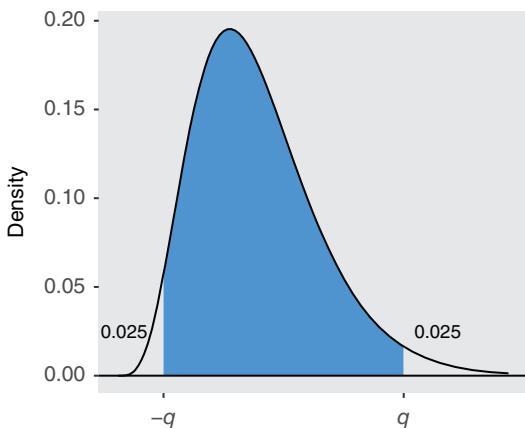
**Example 7.11** Let  $X_1, X_2, \dots, X_n$  be a random sample from an exponential distribution with unknown parameter  $\lambda$ . Then  $X_1 + X_2 + \dots + X_n$  has a gamma distribution with parameters  $n$  and  $\lambda$  by Theorem B.11, so  $h = \lambda(X_1 + X_2 + \dots + X_n)$  has a gamma distribution with parameters  $n$  and 1; this distribution does not depend on  $\lambda$  so  $h$  is a pivot.

Let  $q_1$  and  $q_2$  mark the 0.025 and 0.975 quantiles, respectively, for  $\text{Gamma}(n, 1)$  (see Figure 7.10). Then,

$$0.95 = P(q_1 < \lambda(X_1 + X_2 + \dots + X_n) < q_2)$$

$$= P\left(\frac{q_1}{\sum_{i=1}^n X_i} < \lambda < \frac{q_2}{\sum_{i=1}^n X_i}\right).$$

**Figure 7.10** Density for  $\text{Gamma}(5, 1)$ . Shaded region represents an area of 0.95.



Thus, the 95% confidence interval for  $\lambda$  is

$$\left( \frac{q_1}{\sum_{i=1}^n X_i}, \frac{q_2}{\sum_{i=1}^n X_i} \right).$$

For instance, suppose we observe 2, 2.5, 3, 4, 9. Then  $n = 5$ ,  $\sum_{i=1}^5 x_i = 20.5$ , and the 0.025 and 0.975 quantiles for  $\text{Gamma}(5, 1)$  are 1.6235 and 10.2415, respectively. Then we are 95% confident that the parameter  $\lambda$  lies in the interval  $(1.6235/20.5, 10.2415/20.5) = (0.0792, 0.4996)$ .

The usual estimator for  $\lambda$  is  $\hat{\lambda} = 1/\bar{X}$ , so we can rewrite the interval as

$$((q_1/n)\hat{\lambda}, (q_2/n)\hat{\lambda}).$$

In contrast to  $z$  and  $t$  confidence intervals that are of the form estimate  $\pm$  MOE, this is a multiplicative form, which makes sense for a quantity that must be positive. The interval is not symmetric about the estimate.  $\square$

For confidence intervals that must be strictly positive, a ratio pivot  $\hat{\theta}/\theta$  may be a good choice. If  $P(q_1 < \hat{\theta}/\theta < q_2) = 1 - \alpha$ , then  $P(q_2\hat{\theta} < \theta < q_1\hat{\theta}) = 1 - \alpha$  and  $(q_2\hat{\theta}, q_1\hat{\theta})$  is a  $1 - \alpha$  confidence interval.

**Example 7.12** Let  $X_1, X_2, \dots, X_n$  be a random sample from an exponential distribution with unknown mean  $\mu$  (yes, this is the same as the previous example, but using  $\mu = 1/\lambda$ .  $E(X) = \mu$ , so we let  $\hat{\mu} = \bar{X}$ , which has a gamma distribution with shape  $n$  and scale  $\mu/n$ , and mean  $\mu$ . Then  $\hat{\mu}/\mu = \bar{X}/\mu$  has a gamma distribution with shape  $n$  and scale  $1/n$ . This does not depend on  $\mu$  so  $\bar{X}/\mu$  is a pivot.

Let  $q_1$  and  $q_2$  be the  $\alpha/2$  and  $1 - \alpha/2$  quantiles of the gamma distribution with shape  $n$  and scale  $1/n$ , then the confidence interval is  $(q_2\bar{X}, q_1\bar{X})$ .

For instance, suppose we observe 2, 2.5, 3, 4, 9,  $\bar{x} = 4.1$ . The (0.025, 0.975) quantiles for  $\text{Gamma}(5, \text{scale} = 1/5)$  are (0.325, 2.05), and the confidence interval is (2.00, 12.6). This is equivalent to the interval for the rate parameter in the previous example.  $\square$

**Example 7.13** During World War II, the Western Allies estimated German tank production from the serial numbers on tanks, in particular on gearboxes. There is a nice discussion of this at [http://en.wikipedia.org/wiki/German\\_tank\\_problem](http://en.wikipedia.org/wiki/German_tank_problem). This statistical approach was sometimes much more accurate than other approaches, for example, (from Wikipedia):

The Allied conventional intelligence estimates believed the number of tanks the Germans were producing between June 1940 and September 1942 was around 1,400 a month. Using the above formula on the serial numbers of captured German tanks, (both serviceable and destroyed) the number was calculated to be 256 a month. After the war, captured German production figures from the ministry of Albert Speer show the actual number to be 255.

For a more in-depth analysis see Goodman (1952).

We will consider a simplified version of the problem – that the serial numbers began at 1 and continued up to  $\theta$ , where  $\theta$  is the total number produced, and that any tank is equally likely to be captured, independent of all others. (In practice, neither of these assumptions is true. For example, the newest tanks have had less opportunity to be captured, and tanks produced around the same time are more likely to be shipped to the same region, be in the same battles, and be captured together.)

We will simplify further by considering a continuous distribution. Suppose that the serial numbers are  $X_1, \dots, X_n \sim \text{Unif}[0, \theta]$ . In Section 6.1.2, we saw that the maximum of the observations  $X_{\max}$  is the maximum likelihood estimate, and we saw in Section 6.3.2 that a rescaled version  $((n+1)/n)X_{\max}$  was substantially more efficient than another unbiased estimate  $2\bar{X}$ . Since estimation based on  $X_{\max}$  worked so well, it should also work well for confidence intervals.

The pdf for  $X_{\max}$  is  $f_{\max}(x) = (n/\theta^n) x^{n-1}$  for  $0 \leq x \leq \theta$  (Equation (4.4)). By Proposition B.1, dividing every observation  $X$  by  $\theta$  makes every observation standard uniform,  $\text{Unif}[0, 1]$ , so  $Y = X_{\max}/\theta$  has pdf  $f_{\max}(y) = ny^{n-1}$  for  $0 \leq y \leq 1$ . This does not depend on  $\theta$ , so  $Y$  is a pivotal quantity. The corresponding cumulative distribution function (cdf) is  $F_{\max}(y) = y^n$  for  $0 \leq y \leq 1$ . We solve the two equations

$$F_{\max}(q_1) = \alpha/2$$

$$F_{\max}(q_2) = 1 - \alpha/2$$

to obtain  $q_1 = (\alpha/2)^{1/n}$  and  $q_2 = (1 - \alpha/2)^{1/n}$ .

We obtain the endpoints of the confidence interval by solving these equations for  $\theta$ :

$$\begin{aligned} 1 - \alpha &= P(q_1 < Y < q_2) \\ &= P\left(q_1 < \frac{X_{\max}}{\theta} < q_2\right) \\ &= P\left(\frac{X_{\max}}{q_2} < \theta < \frac{X_{\max}}{q_1}\right). \end{aligned}$$

For a 95% interval with  $n = 400$  captured tanks,  $q_1 = 0.025^{1/400} = 0.9908202$  and  $q_2 = 0.975^{1/400} = 0.9999367$ , so the confidence interval is  $(1.000063X_{\max}, 1.009265X_{\max})$ . This is a remarkably narrow confidence interval; if the largest serial number observed to date is  $10^4$ , we would be 95% confident that  $\theta$ , the true number of tanks produced, would be between 10 001 and 10 093. Note that this interval does not include the MLE  $X_{\max}$  – confidence intervals do not need to be symmetric about an estimate, or even include an estimate.  $\square$

### 7.2.1 Location and Scale Parameters\*

The examples above are all examples of either *location* or *scale* parameters. It is useful to recognize these situations because it governs what confidence intervals should look like.

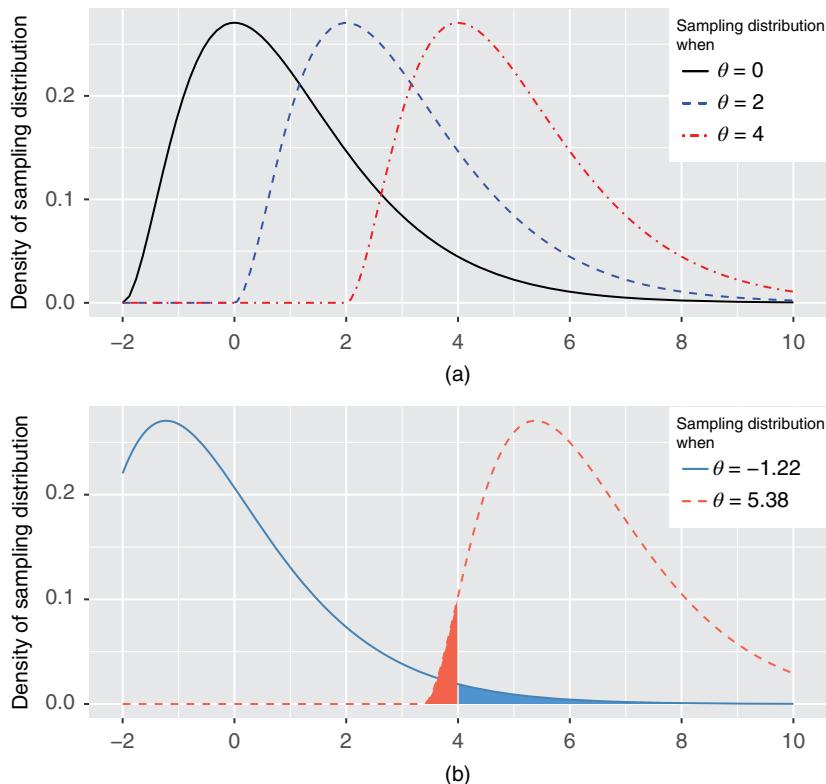
**Definition 7.1** Let  $F(\cdot; \theta)$  be a family of distributions, with  $-\infty < \theta < \infty$ . This is a *location family* if  $F(x; \theta) = H(x - \theta)$  for some function  $H$ . Equivalently, the distribution of  $X - \theta$  does not depend on  $\theta$ . Then we call  $\theta$  a *location parameter*.  $\parallel$

In other words, changing a location parameter shifts a distribution sideways. The standard deviation and shape do not change as the mean (or median, trimmed mean, etc.) changes. For example, the mean  $\mu$  of a normal distribution is a location parameter.

Good estimates for location parameters have the property that adding  $a$  to every observation increases  $\hat{\theta}$  by  $a$ . Then  $\hat{\theta} - \theta$  is pivotal.

We can use quantiles of this pivot to obtain confidence intervals.  $P(q_1 < \hat{\theta} - \theta) = 1 - \alpha/2 = P(\hat{\theta} - \theta < q_2)$ , and solving the inner inequalities  $q_1 < \hat{\theta} - \theta$  and  $\hat{\theta} - \theta < q_2$  for  $\theta$  yields confidence intervals ( $L = \hat{\theta} - q_2$ ,  $U = \hat{\theta} - q_1$ ).

This may seem backward, and indeed it is. With confidence intervals, we need to think backwards, to think like detectives – given the  $\hat{\theta}$  we observed, what values of  $\theta$  could have produced it? Figure 7.11 shows this; Figure 7.11a shows the sampling distribution of  $\hat{\theta}$ , for three different values of  $\theta$ . This estimator is positively biased, and skewed, so  $\hat{\theta}$  tends to be greater than  $\theta$ , sometimes much



**Figure 7.11** Sampling distributions for a location parameter. (a) Sampling distributions of  $\hat{\theta}$  for three values of  $\theta$ ; this estimator is positively biased and skewed. (b) The two sampling distributions with 2.5% probability of being above and below 4, respectively; the corresponding  $\theta$  values are the lower and upper endpoints of a confidence interval for  $\theta$ , when  $\hat{\theta} = 4$ .

greater. Hence, thinking backward, it is likely that  $\theta$  is less than  $\hat{\theta}$ . Figure 7.11b shows the two sampling distributions that have 2.5% probability of being above and below 4, respectively; if  $\hat{\theta} = 4$ , then (i) if  $\theta = -1.22$  the sampling distribution for  $\hat{\theta}$  has 2.5% probability above 4, and (ii) if  $\theta = 5.38$  the sampling distribution for  $\hat{\theta}$  has 2.5% probability below 4. These are the endpoints of an exact confidence interval. The interval  $(-1.22, 5.38) = (4 - 5.22, 4 + 1.38)$  reaches farther to the left to adjust for the positive bias and skewness of the estimator. Any  $\theta$  inside this range has the observed value of 4 in the middle of its sampling distribution, so these are the  $\theta$  values that are consistent with  $\hat{\theta} = 4$ .

Imagine that you are throwing a ball up in the air on a windy day.  $\theta$  is your position and  $\hat{\theta}$  is where it lands – random, plus some bias due to the wind. If all you observe is where the ball lands and you need to create a confidence

interval for where the ball was thrown from, you think backward – given the wind, where could the ball have been thrown from?

Bootstrap percentile intervals do the wrong thing for location parameters with bias – they correspond to throwing the ball 1000 times starting from where it landed, and taking the range of the middle 95% of the impacts as the confidence interval. This doubles the wind bias. If location families were common in practice, we would not use percentile intervals – but they’re not, I (Tim) cannot recall encountering a pure location family in practice. More common are scale families, or location-scale families.

**Definition 7.2** Let  $F(\cdot; \theta)$  be a family of distributions, with  $0 < \theta < \infty$ . This is a *scale family* if  $F(x; \theta) = H(x/\theta)$  for some cdf  $H$ . Equivalently, the distribution of  $X/\theta$  does not depend on  $\theta$ . Then we call  $\theta$  a *scale parameter*. For a continuous distribution  $f(x; \theta) = (1/\theta)h(x/\theta)$  for  $h = H'$ . ||

In other words, changing a scale parameter multiplies all values in a distribution, for example making the density wider and shorter, or narrower and taller (Figure 7.12). The standard deviation is proportional to the mean (or median, trimmed mean, etc.).

For example, for the continuous version of the German Tank problem  $\text{Unif}[0, \theta]$ ,  $\theta$  is a scale parameter. For the  $\text{Unif}[\alpha, \beta]$  family there is no scale parameter, but for  $\text{Unif}[0, \beta]$   $\beta$  is a scale parameter. For the  $N(0, \sigma^2)$  family,  $\sigma$  is a scale parameter, and also for  $N(a\sigma, \sigma^2)$  for a fixed constant  $a$ . For an exponential distribution, if we parameterize the distribution using  $\rho = 1/\lambda$ , then  $\rho$  is a scale parameter and similarly for gamma distributions with a fixed shape parameter  $r$ .

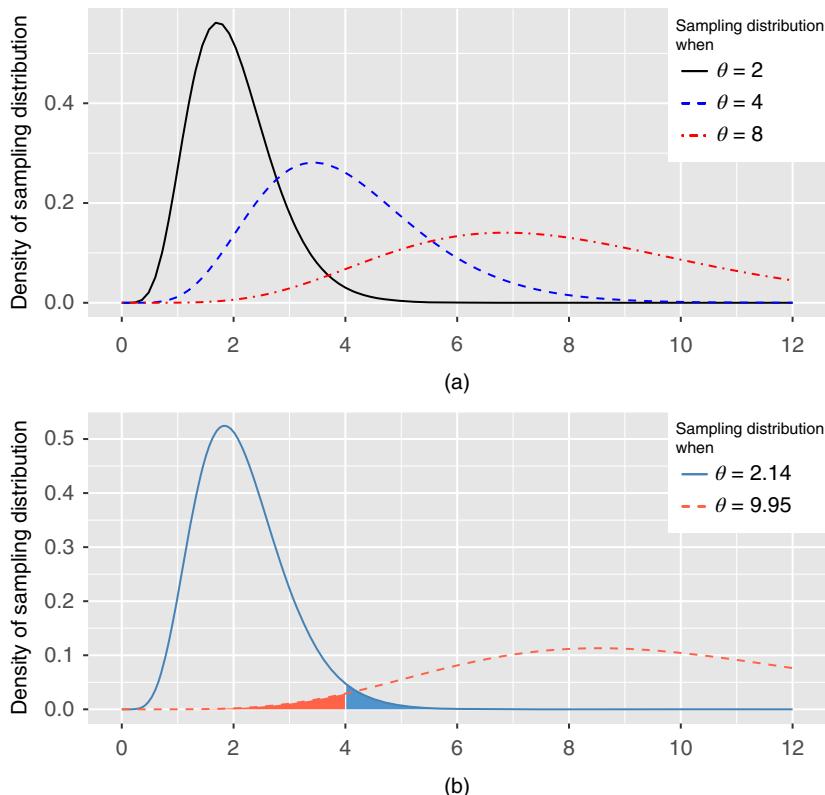
Data that must be positive, such as measurements of time, cannot come from a location family. They often come from a scale family.

For scale parameters, good estimates have the property that multiplying every observation by  $b$  increases  $\hat{\theta}$  by a factor  $b$ . Then,  $\hat{\theta}/\theta$  is pivotal.

We can use quantiles of this pivot to obtain confidence intervals.  $P(q_1 < \hat{\theta}/\theta) = 1 - \alpha/2 = P(\hat{\theta}/\theta < q_2)$ , and solving the inner inequalities  $q_1 < \hat{\theta}/\theta$  and  $\hat{\theta}/\theta < q_2$  for  $\theta$  yields confidence intervals ( $L = \hat{\theta}/q_2$ ,  $U = \hat{\theta}/q_1$ ).

Again, the confidence intervals seem backward. Given an observed  $\hat{\theta}$ , we need to consider what values of  $\theta$  could have produced that. Figure 7.12 shows sampling distributions for different values of the parameter. The striking characteristic is that here the sampling distributions corresponding to large values of  $\theta$  are much wider. Hence, when creating confidence intervals, the upper endpoints need to be farther away from  $\theta$  than do the lower endpoints.

**Remark** Indeed “thinking backward” yields a very general way to create confidence intervals, to *invert a hypothesis test* – to let the confidence interval



**Figure 7.12** Sampling distributions for a scale parameter. (a) Sampling distributions of  $\hat{\theta}$  for three values of  $\theta$ ; this estimator is unbiased and skewed. (b) The two sampling distributions with 2.5% probability of being above and below 4, respectively; the corresponding  $\theta$  values are the lower and upper endpoints of a confidence interval for  $\theta$ , when  $\hat{\theta} = 4$ .

include all values of  $\theta$  for which a two-sided hypothesis test would not reject the null hypothesis for that  $\theta$ . ||

We can combine the location and scale concepts to yield location-scale families, such as normal distributions.

**Definition 7.3** Let  $F(\cdot; \theta_L, \theta_S)$  be a family of distributions, with  $-\infty < \theta_L < \infty$  and with  $0 < \theta_S < \infty$ . This is a *location-scale family* if  $F(x; \theta_L, \theta_S) = H((x - \theta_L)/\theta_S)$  for some cdf  $H$ . Then  $\theta_L$  and  $\theta_S$  are location and scale parameters, respectively. Equivalently, the distribution of  $(X - \theta_L)/\theta_S$  does not depend on  $\theta_L$  or  $\theta_S$ . ||

For location-scale families, good estimators have the properties that adding  $\alpha$  to every observation increases  $\hat{\theta}_L$  by  $\alpha$  and does not change  $\hat{\theta}_S$ , while multiplying every observation by  $b$  increase  $\hat{\theta}_S$  by a factor  $b$  (and may change  $\hat{\theta}_L$ ). Then  $(\hat{\theta}_S/\theta_S)$  and  $(\hat{\theta}_L - \theta_L)/\hat{\theta}_S$  are pivotal, and we may obtain confidence intervals for  $\hat{\theta}_S$  the same way as for pure scale families. We may use quantiles of the second pivot to obtain confidence intervals for  $\theta_L$ .  $P(q_1 < (\hat{\theta}_L - \theta_L)/\hat{\theta}_S) = 1 - \alpha/2 = P((\hat{\theta}_L - \theta_L)/\hat{\theta}_S < q_2)$ , and solving the inner inequalities for  $\theta$  yields confidence intervals ( $L = \hat{\theta}_L - q_2 \hat{\theta}_S$ ,  $U = \hat{\theta}_L - q_1 \hat{\theta}_S$ ). Again, these are backward.

These look similar to the endpoints of a  $t$  interval  $\bar{x} \pm q \cdot s/\sqrt{n}$ . Note that if  $\hat{\theta}_S/\theta_S$  is a pivot, then so is  $d\hat{\theta}_S/\theta_S$  for any nonzero constant  $d$ ; similarly for  $(\hat{\theta}_L - \theta_L)/\hat{\theta}_S$ . The usual  $t$  interval is obtained using  $(\bar{x} - \mu)/(s/\sqrt{n})$  as a pivot.

### Common Pivots

Family	Pivot
Location	$\hat{\theta} - \theta$
Scale	$\hat{\theta}/\theta$
Location-scale	$(\hat{\theta}_L - \theta_L)/\hat{\theta}_S$
	$(\hat{\theta}_S/\theta_S)$

## 7.3 One-Sided Confidence Intervals

We have been discussing confidence intervals of the form  $(L, U)$  which gives both lower bound and upper bounds for the parameter  $\theta$  being estimated. In practice, we are often interested in only an upper bound  $U$  or only a lower bound  $L$ . For example, Google software engineers may want an upper bound on mean latency (time to load a web page), or financial analysts may want an upper bound on the proportion of mortgages in a portfolio that will default.

$[L, \infty)$  is a  $(1 - \alpha) \times 100\%$  one-sided confidence interval for  $\theta$  if the lower bound  $L$  satisfies  $P(L < \theta) = 1 - \alpha$  for all  $\theta$ . Similarly  $(-\infty, U]$  is a  $(1 - \alpha) \times 100\%$  one-sided confidence interval for  $\theta$  if  $P(\theta < U) = 1 - \alpha$  for all  $\theta$ .

For instance, let  $X_1, X_2, \dots, X_n$  be a random sample from a normal distribution with unknown  $\mu$  and let  $q$  denote the  $(1 - \alpha)$  quantile of the  $t$  distribution with  $n - 1$  degrees of freedom. Then

$$1 - \alpha = P\left(\frac{\bar{X} - \mu}{S/\sqrt{n}} < q\right) = P\left(\bar{X} - q\frac{S}{\sqrt{n}} < \mu\right),$$

so a one-sided  $(1 - \alpha) \times 100\%$  confidence interval for  $\mu$  is  $[\bar{X} - q(S/\sqrt{n}), \infty)$ . Similarly, a lower confidence interval for  $\mu$  is  $(-\infty, \bar{X} + q(S/\sqrt{n}))$ .

### One-Sided Confidence Interval for a Normal Mean

If  $X_1, X_2, \dots, X_n$  are a random sample from a normal distribution with unknown mean  $\mu$ , and  $q$  is the  $(1 - \alpha)$  quantile of the  $t$  distribution with  $n - 1$  degrees of freedom, then a  $(1 - \alpha) \times 100\%$  lower confidence bound is

$$\bar{X} - q \frac{S}{\sqrt{n}}$$

with corresponding upper confidence interval

$$\left[ \bar{X} - q \frac{S}{\sqrt{n}}, \infty \right)$$

and a  $(1 - \alpha) \times 100\%$  upper confidence bound is

$$\bar{X} + q \frac{S}{\sqrt{n}}$$

with corresponding lower confidence interval

$$\left( -\infty, \bar{X} + q \frac{S}{\sqrt{n}} \right].$$

**Example 7.14** Chemists at a state pollution control agency are concerned about lead levels in a certain lake. They take 15 samples of lake water and find an average lead level of 7  $\mu\text{g}/\text{dl}$  with standard deviation of 2  $\mu\text{g}/\text{dl}$ . Find a 95% lower confidence bound for the mean  $\mu$ .

#### Solution

We have  $\bar{x} = 7$ ,  $s = 2$  and  $q = 1.761$  is the 0.95 quantile of a  $t$  distribution with 14 degrees of freedom. Thus  $7 - 1.761(2/\sqrt{15}) \approx 6.091$ , so we are 95% confident that the mean lead level in this lake is at least 6.09  $\mu\text{g}/\text{dl}$  (the corresponding confidence interval is  $[6.09, \infty)$ ).  $\square$

#### R Note

To compute one-sided confidence intervals using `t.test`, add the argument `alt="less"` for a lower confidence interval and `alt="greater"` for an upper confidence interval.

```
> t.test(NCBirths2004$Weight, alt = "greater")$conf
[1] 3422.98   Inf
attr(, "conf.level"):
[1] 0.95
```

With 95% confidence, the mean weight of babies born in North Carolina in 2004 is at least 3422.98 g; the interval is  $[3422.98, \infty)$ .

**Example 7.15** Let  $X_1, X_2, \dots, X_n$  be a random sample from an exponential distribution with unknown mean  $\mu$ . Suppose  $n = 5$  and  $\bar{x} = 4.1$ . Example 7.12 contains a two-sided interval. This is similar, but we use the  $\alpha$  or  $1 - \alpha$  quantiles of the Gamma(5, scale = 1/5) distribution, 0.394 or 1.831, respectively.

A 95% upper confidence interval (a lower bound) for  $\bar{X}$  is  $(2.24, \infty)$ , where the lower endpoint is  $\bar{X}/1.831$ . A 95% lower confidence interval (upper bound) is  $(0, 10.4)$ . The lower endpoint of the latter is not  $-\infty$  because that would be silly; we know  $\mu > 0$ .  $\square$

**Example 7.16** (This example is for those who covered Section 7.2.1.) We revisit Example 7.14. Recall that  $n = 15$ ,  $\bar{x} = 7$ , and  $s = 2$ , and that measurements must be positive. This time we'll assume a scale family and use  $\bar{X}/\mu$  as a pivot, which implies that the standard deviation is proportional to  $\mu$ . This seems more reasonable than assuming that  $\sigma^2$  is a constant, particularly when  $\mu$  is near zero.

For a quick-and-dirty interval we'll approximate the sampling distribution of  $\bar{X}$  as  $N(\mu, \mu^2\sigma^2/n)$  for some  $\sigma^2$ . Setting  $\mu^2\sigma^2/n = s^2/n$  implies  $\sigma = s/\mu$  which we estimate as  $2/7$ . The distribution of  $\bar{X}/\mu$  is  $N(1, \sigma^2/n)$ , with 95% quantile  $1 + 1.64(2/(7\sqrt{15})) = 1.12$ . The interval is  $(7/1.12, \infty)$ , or  $(6.24, \infty)$ .

We use the  $t$  distribution quantiles with the scale approach to provide a bit more insurance due to parameters being estimated. Thus, the 95% quantile is  $1 + 1.761(2/(7\sqrt{15})) = 1.13$ , so the lower endpoint would be  $7/1.13 = 6.19$ .

Neither multiplicative endpoint is as low as the subtractive endpoint. Multiplicative is probably better. Note what happens in the extreme as the confidence level approaches 100% – the original approach would eventually make the lower endpoint negative, while a multiplicative endpoint would approach zero.  $\square$

## 7.4 Confidence Intervals for Proportions

We return to the poll results given at the beginning of this chapter: 41% of 1010 adults consider themselves environmentalists.

Let  $X$  denote the adults in a sample of size  $n$  who consider themselves environmentalists. We assume  $X$  is binomial,  $X \sim \text{Binom}(n, p)$ . From Chapter 6, we

know that the proportion of these adults,  $\hat{p} = X/n$ , is an unbiased estimator of  $p$  and from Corollary 4.2, for large  $n$ ,  $Z = (\hat{p} - p)/(\sqrt{p(1-p)/n})$  is approximately standard normal.

Thus,

$$P\left(-1.96 < \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} < 1.96\right) \approx 0.95. \quad (7.14)$$

Isolating the  $p$  in this expression requires a bit more algebra than the earlier problems. We set

$$-1.96 = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

and solve for  $p$  (we get the same answer if we had set the right-hand side of the above to 1.96). This leads to the quadratic equation

$$\left(\frac{n}{1.96^2} + 1\right)p^2 - \left(\frac{2n\hat{p}}{1.96^2} + 1\right)p + \frac{n\hat{p}^2}{1.96^2} = 0.$$

Using the quadratic formula to solve for  $p$  gives a 95% confidence interval  $(L, U)$ , where

$$L = \frac{\hat{p} + 1.96^2/(2n) - 1.96\sqrt{\hat{p}(1-\hat{p})/n + 1.96^2/(4n^2)}}{1 + 1.96^2/n}. \quad (7.15)$$

$$U = \frac{\hat{p} + 1.96^2/(2n) + 1.96\sqrt{\hat{p}(1-\hat{p})/n + 1.96^2/(4n^2)}}{1 + 1.96^2/n}. \quad (7.16)$$

Thus, for the example at the beginning of the section, using  $\hat{p} = 0.41$  and  $n = 1010$ , we have

$$\begin{aligned} & \frac{0.41 + (1.96^2)/(2 \cdot 1010) \pm 1.96\sqrt{(0.41 \cdot 0.59)/1010 + (1.96^2)/(4 \cdot 1010^2)}}{1 + (1.96^2/1010)} \\ &= 0.410 \pm 0.030 = (0.38, 0.44). \end{aligned}$$

Thus, we are 95% confident that between 38% and 44% of adults consider themselves environmentalists.

More generally, we have

### Score Confidence Interval for a Proportion

**Theorem 7.2** Consider a random sample of size  $n$  from a population, where the parameter  $p$  indicates the true proportion with a certain binary characteristic,  $0 < p < 1$ . Let  $X$  denote the number in the sample with this characteristic

and  $\hat{p} = X/n$  the sample proportion. Then an approximate  $(1 - \alpha) \times 100\%$  confidence interval  $(L, U)$  for  $p$  is  $(L, U)$  with

$$L = \frac{\hat{p} + q^2/(2n) - q\sqrt{\hat{p}(1-\hat{p})/n + q^2/(4n^2)}}{1 + q^2/n},$$

$$U = \frac{\hat{p} + q^2/(2n) + q\sqrt{\hat{p}(1-\hat{p})/n + q^2/(4n^2)}}{1 + q^2/n},$$

where  $q$  denotes the  $(1 - \alpha/2)$  quantile of  $N(0, 1)$ .

This confidence interval is called the  $(1 - \alpha) \times 100\%$  score confidence interval for the proportion  $p$ .

Please do not memorize this formula!

**Example 7.17** In the 2018 General Social Survey (case study in Section 1.7), 2193 participants (out of 2348) chose to answer the question about whether or not they favor the death penalty for murder. Of the respondents, 1385 favor the death penalty. Find a 90% confidence interval for the proportion of the population that favor the death penalty.

### Solution

With  $X = 1385$ ,  $n = 2193$ , we have  $\hat{p} = 0.632$  and  $q = 1.645$ . Thus, the 90% confidence interval is  $(0.614, 0.649)$ , so we are 90% confident that between 61.4% and 64.9% of the population favors the death penalty for murder.

#### R Note

The `prop.test` function computes score confidence intervals.

```
> prop.test(1385, 2193, conf.level = .9)$conf
[1] 0.6142292 0.6485521
...
```

If you omit the `conf.level` argument, then by default, a 95% confidence interval is calculated. `prop.test` also takes an argument of `alt = "greater"` or `alt = "less"` for one-sided confidence intervals.

```
> prop.test(1385, 2193, conf.level=.9, alt = "greater")$conf
[1] 0.6180305 1.0000000
...
```

We are 90% confident that at least 61.8% of the population favor the death penalty.



**Remark**

- This interval is also called a Wilson or Wilson score interval (Wilson (1927)).
- The center of the score interval is  $(\hat{p} + q^2/(2n))/(1 + q^2/n)$ . If we set  $\kappa = q^2$ , then this center can be written as  $\hat{p}(n/(n + \kappa)) + (1/2)(\kappa/(n + \kappa))$ , a weighted average of the observed proportion and 1/2. As  $n$  increases, more weight is given to  $\hat{p}$ . ||

**7.4.1 Agresti–Coull Intervals for a Proportion**

Now, the limits of the interval given by Theorem 7.2 are pretty messy so that in general, we would want to use software to do the calculations. However, in the event that we must resort to hand calculations, it would be nice to find a simpler expression for the confidence interval. Agresti and Coull (1998) considered the 95% confidence interval for which the 0.975 quantile is  $q \approx 1.96$ , and hence  $q^2 \approx 4$ .

**Agresti–Coull 95% Confidence Interval for a Proportion**

If  $X$  denotes the number of successes in a sample of size  $n$ , let  $\tilde{X} = X + 2$ ,  $\tilde{n} = n + 4$  and  $\tilde{p} = \tilde{X}/\tilde{n}$ . Then an approximate 95% confidence interval for  $p$  is

$$\left( \tilde{p} - 1.96 \sqrt{\frac{\tilde{p}(1 - \tilde{p})}{\tilde{n}}}, \tilde{p} + 1.96 \sqrt{\frac{\tilde{p}(1 - \tilde{p})}{\tilde{n}}} \right). \quad (7.17)$$

One way to remember this is “plus 4” – add four additional observations, split between success and failure, before calculating  $\tilde{p}$  and  $\tilde{n}$ .

**Example 7.18** Suppose the sample size is  $n = 210$  with  $x = 130$ . Then  $\tilde{x} = 132$ ,  $\tilde{n} = 214$ , and  $\tilde{p} = 132/214 = 0.6168$ . Thus, an approximate 95% confidence interval is given by

$$0.6168 \pm 1.96 \sqrt{\frac{0.6168(1 - 0.6168)}{214}} = 0.6168 \pm 0.0651 = (0.5517, 0.6819),$$

which gives almost the same result as Theorem 7.2. □

**Remark** The above interval is similar to an interval that is often taught in introductory statistics courses, sometimes called the *Wald confidence interval for a binomial proportion*. Suppose that in Equation (7.14) we ignore the fact that the standard error  $\sqrt{p(1 - p)/n}$  depends on  $p$  and simply substitute  $\hat{p}$  for  $p$ . We obtain the interval  $\hat{p} \pm 1.96 \sqrt{\hat{p}(1 - \hat{p})/n}$ . Many papers have been published

recently cautioning against the use of this interval: it is not very accurate, particularly for  $\hat{p}$  near zero or one (Agresti and Coull (1998); Brown et al. (2001); Newcombe (1998); Pan (2009)). ||

**Example 7.19** Senator Knudson prepares to conduct a survey to gauge voter support for her re-election campaign. She would like a confidence interval with an error of at most 4%, with 95% confidence. How large should the sample size be for the survey?

### Solution

Since the interval given by Equation (7.17) is symmetric, the margin of error is  $1.96\sqrt{\hat{p}(1 - \hat{p})/\tilde{n}}.$  Thus, we want to solve for  $\tilde{n}$  in

$$1.96\sqrt{\frac{\tilde{p}(1 - \tilde{p})}{\tilde{n}}} \leq 0.04.$$

Unfortunately, we do not know  $\tilde{p}$  – if we did, the Senator would not need to conduct the survey! We will use  $\tilde{p} = 0.5$  since this will maximize the expression under the radical sign (see Exercise 7.28).

$$\begin{aligned} 1.96\sqrt{\frac{0.5(1 - 0.5)}{\tilde{n}}} &\leq 0.04 \\ \left(\frac{1.96(0.5)}{0.04}\right)^2 &\leq \tilde{n} \\ 596.25 &\leq n. \end{aligned}$$

Thus, she should survey at least 597 people.

In some instances, based on prior knowledge, the analyst may substitute another estimate for  $\tilde{p}$  (e.g. the proportion from a previous poll). □

### 7.4.2 Confidence Interval for a Difference of Proportions

For the difference of two proportions,  $p_1 - p_2$ , an interval extending the score interval can be constructed, though there is not a closed form version (Wilson (1927)). It can be computed using statistical software. A closed form approximation (Agresti and Caffo (2000)) is

$$\tilde{p}_1 - \tilde{p}_2 \pm q\sqrt{\frac{\tilde{p}_1(1 - \tilde{p}_1)}{\tilde{n}_1} + \frac{\tilde{p}_2(1 - \tilde{p}_2)}{\tilde{n}_2}}, \quad (7.18)$$

where  $\tilde{p}_i = (X_i + 1)/(n_i + 2)$  and  $\tilde{n}_i = n_i + 2$  for  $i = 1, 2$ , and  $q$  is the  $(1 - \alpha/2)$  quantile of  $N(0, 1)$ . Like Equation (7.17), you can remember this as a “plus 4” interval, with the four added observations split among success and failure for the two samples before calculating  $\tilde{p}_*$  and  $\tilde{n}_*$ .

**Example 7.20** Researchers in the United Kingdom conducted a clinical trial to determine whether preoperative dexamethasone was effective in reducing nausea and vomiting after elective gastrointestinal surgery (Collaborators et al. (2017)). Of the 674 patients who received a single dose of 8 mg intravenous dexamethasone, 172 reported vomiting within 24 h of the surgery compared to the 223 out of 676 patients receiving standard care. Compute a 95% confidence interval for the true difference in proportions (dexamethasone – standard care).

### Solution

Let  $p_1$  and  $p_2$  denote the proportions of patients on dexamethasone and standard care, respectively, who reported vomiting. We want a confidence interval for  $p_1 - p_2$  based on our sample estimates of  $\hat{p}_1 = 172/674 = 0.255$  and  $\hat{p}_2 = 223/676 = 0.333$ . In R, the `prop.test` function computes confidence intervals for the difference in proportions.

#### R Note

```
> prop.test(c(172, 223), c(674, 676))$conf
[1] -0.12453865 -0.02483891
...
```

Thus, we are 95% confident that the percentage of patients on dexamethasone who vomited postsurgery is between 2.5% and 12.5% lower than the percentage of patients on standard care who vomited postsurgery. The “plus 4” interval (Equation (7.18)) is similar,  $(-0.1228, -0.2614)$ .  $\square$

## 7.5 Bootstrap Confidence Intervals

In this section, we will discuss two bootstrap intervals that are related to  $t$  intervals.

### 7.5.1 T Confidence Intervals Using Bootstrap Standard Errors

The first is just an ordinary  $t$  interval  $\hat{\theta} \pm t_{\alpha/2, n-1} \text{SE}$ , but with the standard error calculated using the bootstrap rather than a formula. This is particularly useful when we do not have a formula for the SE for a particular estimator – for example, a median, or trimmed mean.

One caution for small samples is that bootstrap standard errors tend to be too small, by a factor of about  $\sqrt{(n-1)/n}$  (see Section 5.8.2). An easy remedy is to multiply the bootstrap SE by  $\sqrt{n/(n-1)}$  before using it in the  $t$  interval.

This interval is a quick-and-dirty interval. It does no better at handling skewed distributions or other issues that come up in practice than a formula  $t$  interval does.

Please do not call this a “bootstrap  $t$  interval” – that name is reserved for the next interval.

### 7.5.2 Bootstrap $t$ Confidence Intervals

The next interval is the bootstrap  $t$  interval. The idea behind this interval is to use  $t$  statistics like  $(\hat{\theta} - \theta)/SE_{\hat{\theta}}$ , but to estimate the actual distribution of the  $t$  statistic by bootstrapping rather than just assuming that  $t$  statistics follow a Student’s  $t$  distribution. It turns out that this produces very good confidence intervals.

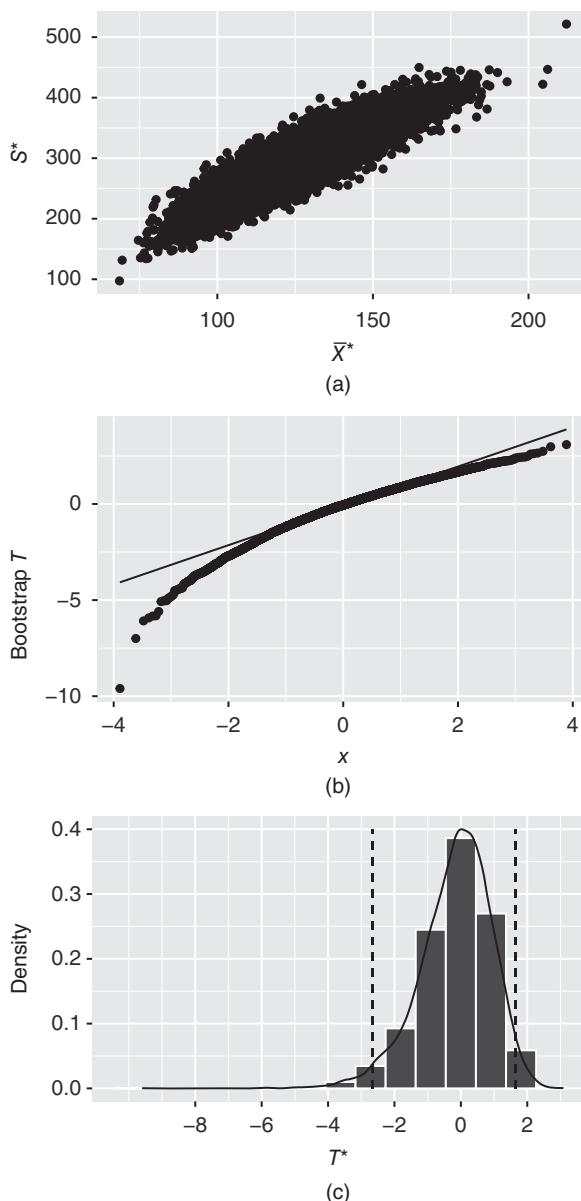
Recall the Bangladesh arsenic levels data in Section 5.2.2. The distribution of arsenic levels was right-skewed (Figure 5.6). The formula derived earlier (Section 7.1.2) for the confidence interval of the mean  $\mu$  assumes that the statistic  $T = (\bar{X} - \mu)/(S/\sqrt{n})$  follows a  $t$  distribution. That seems unlikely for data this skewed. Instead, we bootstrap the  $t$  statistic: for each of  $10^5$  resamples, we compute the resample mean  $\bar{X}^*$ , resample standard deviation  $S^*$ , and then compute the resample  $T$  statistic

$$T^* = \frac{\bar{X}^* - \bar{x}}{S^*/\sqrt{n}}.$$

In Figure 7.13, we can see that the bootstrap distribution for  $T^*$  is left-skewed(!); in fact, it is more left-skewed than the bootstrap distribution of the numerator (Figure 5.7) is right-skewed!! The reason for this is the denominator. There is a strong positive relationship between  $\bar{X}^*$  and  $S^*$  (Figure 7.13a). If a bootstrap resample contains relatively many of the large values from the right tail of the original data, then  $\bar{X}^*$  is large and  $S^*$  is especially large (standard deviations are computed by squaring distances from the mean, so they are affected even more by large observations than a mean is). The large denominator thus keeps  $T^*$  from being particularly large. Conversely, when there are relatively few of the large observations in the resample, then  $\bar{X}^* - \bar{x}$  is negative and the denominator can be especially small, thus resulting in a  $T$  ratio that is large negative (Figure 7.13c).

The 2.5% and 97.5% percentiles of the bootstrap  $t$  distribution are  $-2.66$  and  $1.63$ , compared to  $\pm 1.97$  for the Student’s  $t$  distribution. Then the bootstrap  $t$  interval is  $(\bar{x} - 1.63s/\sqrt{n}, \bar{x} + 2.66s/\sqrt{n})$ . Why swap and negative? Because we need to think backward, and use a pivot to obtain intervals. Suppose that  $Q_1$  and  $Q_2$  are the  $\alpha/2$  and  $(1 - \alpha/2)$  quantiles of the  $T$  distribution,

$$\begin{aligned} 1 - \alpha &= P(Q_1 < T < Q_2) \\ &= P\left(Q_1 < \frac{\bar{X} - \mu}{S/\sqrt{n}} < Q_2\right). \end{aligned}$$



**Figure 7.13** (a) Plot of  $S^*$  against  $\bar{X}^*$ . (b) A normal quantile plot of the bootstrap  $t$  statistics. (c) A histogram and density, of the bootstrap  $t$  statistics, with 2.5% and 97.5% quantiles shown.

Solving both inequalities for  $\mu$  gives the interval

$$\left( \bar{X} - Q_2 \frac{S}{\sqrt{n}}, \bar{X} - Q_1 \frac{S}{\sqrt{n}} \right).$$

The quantiles  $Q_1$  and  $Q_2$  are unknown, but we can estimate them using quantiles of  $T^*$ , the bootstrap  $t$  statistic. This works well if the distribution of  $T^*$  is good for approximating the distribution of  $T$ ; we saw in Section 5.8.4 that it is.

### Bootstrap $t$ Confidence Interval for $\mu$

For each of many resamples, calculate the bootstrap  $t$  statistic

$$T^* = \frac{\bar{X}^* - \bar{x}}{S^*/\sqrt{n}}.$$

Let  $\hat{Q}_1$  and  $\hat{Q}_2$  be the empirical  $\alpha/2$  and  $(1 - \alpha/2)$  quantiles of  $T^*$ , respectively (i.e. the interval between them contains the middle  $(1 - \alpha) \times 100\%$  of all the bootstrap  $t$  statistics). The bootstrap  $t$  confidence interval is

$$\left( \bar{x} - \hat{Q}_2 \frac{s}{\sqrt{n}}, \bar{x} - \hat{Q}_1 \frac{s}{\sqrt{n}} \right).$$

**Remark**  $\hat{Q}_1$  is negative and  $\hat{Q}_2$  is positive, and the interval is  $(\bar{x} - \text{positive}, \bar{x} - \text{negative})$ . ||

**Example 7.21** For the arsenic example,  $n = 271$ ,  $\bar{x} = 125.32$ ,  $s = 297.98$ ,  $\hat{Q}_1 = -2.66$ ,  $\hat{Q}_2 = 1.634$  (for a 95% interval), and the bootstrap  $t$  intervals is  $(\bar{x} - 1.632s/\sqrt{n}, \bar{x} + 2.644s/\sqrt{n}) = (95.7, 173.4)$ . For comparison, the symmetric  $t$  intervals is  $(89.7, 161.0)$ . The bootstrap  $t$  interval is further to the right, protecting against the possibility that the sample might not have enough observations from the long right tail of the population.

### R Note

The following script bootstraps the  $t$  statistics and stores it in the vector `Tstar`.

```
Arsenic <- Bangladesh$Arsenic
xbar <- mean(Arsenic)
N <- 10^4
n <- length(Arsenic)
Tstar <- numeric(N)
for (i in 1:N)
```

```
{
  x <- sample(Arsenic, size = n, replace = T)
  Tstar[i] <- (mean(x)-xbar) / (sd(x)/sqrt(n))
}
```

For bootstrap  $t$  confidence intervals, we need the empirical quantiles.

```
> quantile(Tstar, c(.025, .975))
-2.658719 1.634934
> xbar - quantile(Tstar, c(.975, .025)) *
+      sd(Arsenic)/sqrt(n)
95.72643 173.44466
```

□

**Example 7.22** In Section 5.3, we found the 95% bootstrap percentile interval for the mean weight of all babies born in North Carolina in 2004 to be (3419, 3478) g. The 95% bootstrap  $t$  confidence interval is (3418.8, 3478.4) g, and the 95%  $t$  formula confidence interval is (3419.3, 3478.4). All three intervals are quite close, so we may use any one of them. A check of a normal quantile plot shows that the distribution of weights is approximately normal. □

### 7.5.2.1 Bootstrap $t$ Confidence Intervals for Difference of Means

Bootstrap intervals for a difference in means follow the same idea. For matched pairs, we take the differences, then compute a one-sample bootstrap  $t$  interval for the mean of the differences, as above. For two independent samples, we do a two-sample bootstrap of the  $t$ -statistic for the difference of means, as follows.

#### Bootstrap $t$ Confidence Interval for $\mu_1 - \mu_2$

For each of many resamples, calculate the bootstrap  $t$  statistic

$$T^* = \frac{\bar{X}_1^* - \bar{X}_2^* - (\bar{x}_1 - \bar{x}_2)}{\sqrt{s_1^{*2}/n_1 + s_2^{*2}/n_2}}.$$

Let  $\hat{Q}_1$  and  $\hat{Q}_2$  be the empirical  $\alpha/2$  and  $(1 - \alpha/2)$  quantiles of the bootstrap  $t$  distribution, respectively. The bootstrap  $t$  confidence interval is

$$\left( \bar{x}_1 - \bar{x}_2 - \hat{Q}_2 \cdot \sqrt{s_1^2/n_1 + s_2^2/n_2}, \quad \bar{x}_1 - \bar{x}_2 - \hat{Q}_1 \cdot \sqrt{s_1^2/n_1 + s_2^2/n_2} \right).$$

Recall the Verizon example (Example 5.5), where we considered the difference in means of two very skewed distributions of repair times for two very unbalanced samples ( $n_1 = 23$  versus  $n_2 = 1664$ ). The 95% bootstrap  $t$  interval

for the difference in means is  $(-22.3, -2.0)$ . For comparison, the formula  $t$  interval is  $(-16.6, 0.4)$  and the bootstrap percentile interval is  $(-17.2, -1.6)$ . The more accurate bootstrap  $t$  interval stretches farther in the negative direction, even more than the bootstrap percentile interval.

### R Note

```
TimeILEC <- Verizon %>% filter(Group == "ILEC") %>% pull(Time)
TimeCLEC <- Verizon %>% filter(Group == "CLEC") %>% pull(Time)

thetahat <- mean(TimeILEC) - mean(TimeCLEC)
nx <- length(TimeILEC) # nx=1664
ny <- length(TimeCLEC) # ny=23
SE <- sqrt(var(TimeILEC)/nx + var(TimeCLEC)/ny)

N <- 10^4
Tstar <- numeric(N)
for (i in 1:N)
{
  bootx <- sample(TimeILEC, nx, replace = TRUE)
  booty <- sample(TimeCLEC, ny, replace = TRUE)
  Tstar[i] <- (mean(bootx) - mean(booty) - thetahat) /
    sqrt(var(bootx)/nx + var(booty)/ny)
}
```

To find the 95% bootstrap  $t$  confidence interval

```
> thetahat - quantile(Tstar, c(.975, .025)) * SE
  97.5%      2.5%
-22.23509  -2.03913
> t.test(TimeILEC, TimeCLEC)$conf    # compare to t-test

[1] -16.5568985  0.3618588
...
```

#### 7.5.2.2 Bootstrap $t$ Confidence Intervals for Other Statistics

The same basic procedure can also be used for confidence intervals for statistics other than one or two means – to compute a  $t$  statistic for each of many resamples, using the appropriate standard error for  $\hat{\theta}$ , find the quantiles of that bootstrap  $t$  distribution and create an estimate of the form  $(\hat{\theta} - \hat{Q}_2 \times \text{SE}, \hat{\theta} - \hat{Q}_1 \times \text{SE})$ .

There is one twist – that if we do not have a formula for the standard error, then we need to estimate it using the bootstrap. And to get the standard error based on each of the bootstrap samples, we need to resample from that bootstrap sample! We need two nested loops. This is known as *double bootstrap* or *iterated bootstrap*. This would get slow, if we use  $10^4$  bootstrap samples and  $10^4$  iterated resamples for each bootstrap sample. Fortunately, we

don't need that many; we can use fewer in the inner loop because all we are doing there is estimating the standard deviation of a bootstrap distribution, which is easier than estimating tail quantiles.

In Example 7.21, we computed the bootstrap  $t$  interval for the mean of arsenic levels; we will do the same for the midmean.

**Example 7.23** For the arsenic example,  $n = 271$ , the midmean (25% trimmed mean) is 35.96 (much smaller than the untrimmed mean). The bootstrap standard error is 5.56, and quantiles of the bootstrap  $t$  distribution are  $(-2.14, 1.82)$ , yielding an interval  $(25.8, 47.9)$ . For comparison, a  $t$  interval using the bootstrap standard error is  $(25.0, 46.9)$ .

### R Note

Bootstrap t with estimated standard errors iterated bootstrap.

```
Arsenic <- Bangladesh$Arsenic
estimate <- mean(Arsenic, trim = 0.25) # 35.95985

N <- 10^4 # outer loop
N2 <- 10^2 # inner loop
n <- length(Arsenic)
Tstar <- numeric(N)
estimateStar <- numeric(N)
seStar <- numeric(N)

for (i in 1:N)
{
  x <- sample(Arsenic, size = n, replace = T)

  # Inner loop to estimate standard error based on x
  estimate2 <- numeric(N2)
  for (j in 1:N2)
  {
    x2 <- sample(x, size = n, replace = T)
    estimate2[j] <- mean(x2, trim = 0.25)
  }

  estimateStar[i] <- mean(x, trim = 0.25)
  seStar[i] <- sd(estimate2)
  Tstar[i] <- (estimateStar[i] - estimate) / seStar[i]
}

> sd(estimateStar) # Standard error
[1] 5.565023
> quantile(Tstar, c(.025, .975))
-2.139846  1.817002
```

```
# Bootstrap t interval
> estimate - quantile(Tstar, c(.975, .025)) * sd(estimateStar)
25.84820 47.86815

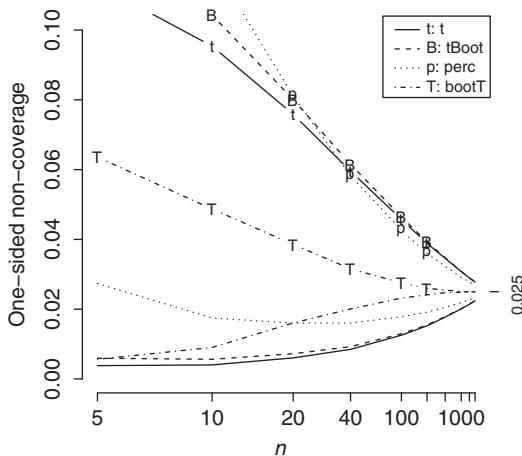
# Ordinary T interval with bootstrap standard error
> estimate + qt(c(.025, .975), n-1) * sd(estimateStar)
[1] 25.00350 46.91621
```

□

### 7.5.3 Comparing Bootstrap $t$ and Formula $t$ Confidence Intervals

We indicated earlier that symmetric  $t$  intervals are not accurate for skewed populations unless sample sizes are large. In this chapter, we investigate this more carefully, compare the performance of different intervals, and discuss why.

95% confidence intervals should miss the true parameter 2.5% of the time on each side. Figure 7.14 shows how often they actually miss. Focus first on the solid curves, corresponding to the ordinary  $t$  interval. The top curve solid is how often the  $t$  interval misses low (the upper endpoint of the interval falls below  $\mu$ ). The  $t$  interval is terrible for small samples, missing far too often,



**Figure 7.14** How often confidence intervals miss on each side. Confidence intervals for the mean of an exponential population. For 95% confidence intervals, the ideal non-coverage is 2.5% on each side. The actual non-coverage on each side is shown as a function of  $n$ .  $t$  is the ordinary  $t$  interval,  $B$  is a  $t$  interval with bootstrap standard error,  $p$  is a bootstrap percentile interval, and  $T$  is a bootstrap  $t$  interval. For each interval there are two curves; the top (bottom) curves, with (without) letters, show how often the interval misses by being too low (high).

until sample sizes are quite large. The bottom solid curve shows how often the interval misses by being too low; this occurs too rarely, less than the desired 2.5%. Together, they indicate that the  $t$  interval paints a biased picture, suggesting that the true mean is less than it actually is.

The  $t$  interval is not “10% accurate” (with actual miss probabilities within 10% of the desired 2.5%, i.e. between 2.25% and 2.75%) until  $n = 5000$ . While you may not need that level of accuracy in many applications, the overall poor performance suggests that we should avoid this interval for skewed data.

The  $t$  interval with bootstrap SE (Section 7.5.1) is worse, it misses even more often. For confidence intervals for a sample mean, there is no reason to use this. It is useful with estimators that do not have a good formula standard error.

The bootstrap percentile interval is actually worse than the ordinary  $t$  interval for small samples because it is too narrow. For larger samples, it is somewhat better because it does a better job of handling skewness.

The best interval by far is the bootstrap  $t$  interval. It does a much better job of handling skewness. For all of the other intervals, the actual noncoverage on each side is  $0.025 + c/\sqrt{n} + \text{smaller terms}$ , for some  $c$ , while for the bootstrap  $t$  it is  $0.025 + c/n + \text{smaller terms}$ .

It is worth thinking about why symmetric intervals do poorly for skewed data. For right-skewed data, when  $\bar{X} < \mu$ , then typically  $S < \sigma$  and consequently, the confidence intervals are narrow and the right endpoint of the interval falls below  $\mu$  too often. Conversely, when  $\bar{X} > \mu$ , then typically  $S > \sigma$ , so the intervals tend to be wide and do not miss  $\mu$  often enough. Overall, symmetric intervals tend to be to the left of where they should be and give a biased picture of where the mean is likely to be.

## 7.6 Confidence Interval Properties

Desirable properties of confidence intervals include accurate coverage, short length, transformation invariance, ease of use, and ease of interpretation.

### 7.6.1 Confidence Interval Accuracy

A confidence interval is accurate if it gives the correct coverage – a 95% interval actually includes the parameter 95% of the time, and misses 2.5% on each side.

The preceding discussion was about the mean for skewed distributions where symmetric  $t$  intervals are not very accurate, and bootstrap  $t$  intervals are. But the result holds more broadly – under fairly general assumptions in a wide variety of applications:

- $t$  confidence intervals are only *first-order accurate*, meaning they give the correct coverage in the long run but that the coverage errors converge to zero quite slowly with the difference (actual miss probability) – (desired miss probability) =  $c/\sqrt{n} + \text{smaller terms}$ , for some  $c$ , while

- bootstrap  $t$  intervals are *second-order accurate* with the difference  
(actual miss probability) – (desired miss probability) =  $c/n$  + smaller terms  
for some  $c$ .

There is also some behavior that is not captured by those asymptotic results. In particular, the bootstrap percentile interval is too short in small samples, and undercovers.

### 7.6.2 Confidence Interval Length

Obviously, it is nice if confidence intervals are relatively short. Typically, intervals associated with more efficient (smaller variance) estimators are shorter. In addition,  $t$  estimates pay a penalty for having to estimate the variance. For ordinary  $t$  intervals, this is reflected in the use of  $t$  quantiles, rather than  $z$  quantiles and in the fudge factor  $\sqrt{n/(n-1)}$  used in computing the sample standard deviation  $s$ . The bootstrap  $t$  pays a similar penalty and two additional penalties – one for handling skewness correctly instead of just assuming it is zero (this lengthens one side more than it shortens the other side), and one for estimating the skewness of the data; the latter penalty is relatively unimportant except for small samples, where it can be substantial; bootstrap  $t$  intervals can be quite wide in small samples.

### 7.6.3 Transformation Invariance

Suppose two parameters are related by an invertible transformation  $\psi = h(\theta)$ . A confidence interval is transformation invariant if  $L_\psi = h(L_\theta)$  and  $U_\psi = h(U_\theta)$ . For example, the endpoints of a confidence interval for  $\sigma^2$  would be the square of the endpoints for  $\sigma$ .

Bootstrap percentile intervals are transformation invariant. Bootstrap  $t$  intervals are not exactly invariant but they are close – close enough to be second-order accurate.

Symmetric  $t$  intervals are not transformation invariant. For example, for two binomial proportions, we can get very different results using  $t$  intervals for  $p_1/p_2$ ,  $p_2/p_1$ ,  $\ln(p_1/p_2)$ , odds ratio  $p_1(1-p_2)/((1-p_1)p_2)$ , or log-odds ratio (these are all used in practice).<sup>4</sup> This is disconcerting. In fact, using arbitrary transformations with symmetric  $t$  intervals, we can produce almost any confidence interval we want, and reach different conclusions. This raises ethical issues.

### 7.6.4 Ease of Use and Interpretation

Bootstrap  $t$  intervals are expensive computationally in cases where there is no formula for standard error, and we have to estimate the standard error by

---

<sup>4</sup> The delta method in Section 7.7 is a way to obtain standard errors for statistics like these.

bootstrapping, as in Section 7.5.2.2. Using  $10^4$  bootstrap samples and estimating SE using  $10^2$  iterated bootstrap samples requires computing the statistic  $1 + 10^4 + 10^6$  times.

Ordinary  $t$  intervals are easy to use in simple problems, where there is a known formula for standard error. For more general applications, the bootstrap intervals are easier.

One advantage of symmetric intervals is that they are easy to describe: estimate plus or minus something. Asymmetric intervals are more awkward.

### 7.6.5 Research Needed

Getting good confidence intervals is an area that needs more attention from statistical researchers. They should perform well for large and small samples, work in a wide variety of applications, be easy to implement in statistical software, and not be too computationally expensive.

## 7.7 The Delta Method\*

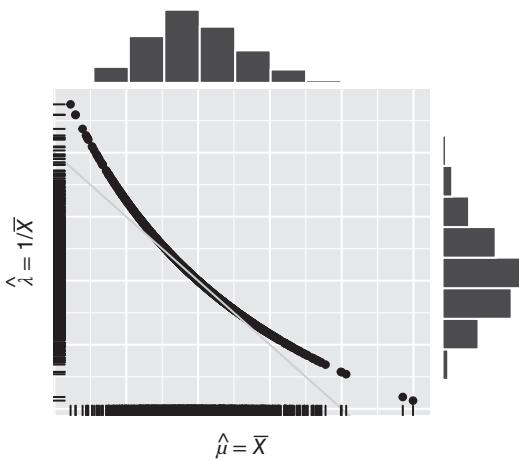
Throughout this chapter, we have seen that we need standard errors of some quantity to compute confidence intervals. In many instances, we have a formula to calculate the standard errors for a quantity, for example  $\bar{X}$ , but what we need is a standard error for a function of this quantity, such as  $1/\bar{X}$ .

- In Example 5.9, we are interested in the relative risk of cardiovascular disease, that is, the ratio of two proportions. We can compute standard errors for  $\hat{p}_1$  and  $\hat{p}_2$ , but for a  $t$  interval, we need a standard error for  $\hat{p}_1/\hat{p}_2$ .
- In the Verizon example in Section 1.3, we have standard errors for  $\bar{X}$  and  $\bar{Y}$  (estimates of  $\mu_X$  and  $\mu_Y$ , the means for the incumbent local exchange carrier (ILEC) and competing local exchange carrier (CLEC) distributions), but we are also interested in the mean repairs per hour  $1/\mu_X$  or  $1/\mu_Y$  and the ratio of the two average repair times  $\mu_Y/\mu_X$ . We need standard errors for  $1/\bar{X}$ ,  $1/\bar{Y}$ , and  $\bar{Y}/\bar{X}$ .
- In analyzing traffic deaths related to alcohol in Section 9.6, we may have estimates and standard errors for the logistic regression parameters  $\alpha$  and  $\beta$ , but when we want to compute standard errors for the predicted probability of alcohol involvement in the automobile death of a 20-year-old,  $\exp(\hat{\alpha} + 20\hat{\beta})/(1 + \exp(\hat{\alpha} + 20\hat{\beta}))$ .

We could use bootstrapping to estimate these standard errors. Here, we describe another method, the *delta method*.

Figure 7.15 show the basic idea; we have the standard error for one estimator ( $\bar{X}$  on the  $x$  axis), and we estimate the standard error for a related estimator

**Figure 7.15** Delta method.  $x$ : sample means  $\bar{X}$  from 200 samples of size  $n = 50$  from an exponential population with mean 1,  $y: \hat{\lambda} = 1/\bar{X}$ . The tangent line at  $\mu = 1$  is shown. The delta method approximate the standard error for  $y$  by the standard error for  $x$  times the absolute value of the slope of the tangent line.



( $\hat{\lambda} = 1/\bar{X}$ ) as the first standard error times (the absolute value of) the slope of the tangent line.

Write

$$\eta = g(\theta),$$

where  $\theta$  is a quantity whose standard error is known,  $\eta$  is the quantity of interest, and  $g$  is a differentiable function. The corresponding estimator is

$$\hat{\eta} = g(\hat{\theta}).$$

The core of the delta method is a linear Taylor series approximation of  $g(\hat{\theta})$  about  $\theta$ ,

$$\begin{aligned}\hat{\eta} &= g(\hat{\theta}) \\ &= g(\theta) + g'(\theta)(\hat{\theta} - \theta) + \dots.\end{aligned}$$

Now  $g(\theta)$  and  $g'(\theta)$  are constants, so

$$\text{Var}[\hat{\eta}] \approx g'(\theta)^2 \text{Var}[\hat{\theta}]$$

with corresponding standard deviation

$$\sigma_{\hat{\eta}} \approx |g'(\theta)|\sigma_{\hat{\theta}}.$$

Since  $g'(\theta)$  is unknown, we use the statistician's favorite trick of plugging in an estimator to obtain standard errors

$$s_{\hat{\eta}} \approx |g'(\hat{\theta})|s_{\hat{\theta}}. \tag{7.19}$$

For example, if  $g(\mu) = 1/\mu$ , then  $g'(\mu) = -1/\mu^2$  and the standard error for  $1/\bar{x}$  is  $s_{\bar{x}}/\bar{x}^2 = s/(\sqrt{nx^2})$ . The mean Verizon ILEC repair time is 8.4 h with standard deviation 14.7 and standard error for the mean of 0.36; the estimated mean number of repairs per hour is 0.12, with standard error  $0.12/8.4^2 = 0.005$ .

The name comes from writing  $\Delta = \hat{\theta} - \theta$  and then expanding  $\hat{\eta} = g(\theta + \Delta)$  as a function of  $\Delta$ .

In the case of a function of two parameters,

$$\eta = g(\theta_1, \theta_2)$$

$$\hat{\eta} = g(\hat{\theta}_1, \hat{\theta}_2),$$

with first-order Taylor series approximation

$$\begin{aligned}\hat{\eta} &\approx g(\theta_1, \theta_2) + \frac{\partial g(\theta_1, \theta_2)}{\partial \theta_1}(\hat{\theta}_1 - \theta_1) + \frac{\partial g(\theta_1, \theta_2)}{\partial \theta_2}(\hat{\theta}_2 - \theta_2) \\ &= \eta + g_1(\hat{\theta}_1 - \theta) + g_2(\hat{\theta}_2 - \theta),\end{aligned}$$

where  $g_1$  and  $g_2$  indicate the partial derivatives with respect to  $\theta_1$  and  $\theta_2$ . The only random terms are  $\hat{\theta}_1$  and  $\hat{\theta}_2$ , so

$$\text{Var}[\hat{\eta}] \approx \text{Var}[g_1\hat{\theta}_1 + g_2\hat{\theta}_2].$$

Now, we plug in estimates, then take the square root to obtain the standard error.

With more parameters, we have  $\eta = g(\theta_1, \theta_2, \dots, \theta_p)$  and  $\text{Var}[\hat{\eta}] \approx \text{Var}[g_1\hat{\theta}_1 + \dots + g_p\hat{\theta}_p]$ .

An important special case is when  $\bar{Y}/\bar{X}$  is an estimator for a ratio of means,  $\eta = \mu_Y/\mu_X$ . For simplicity, we assume here that  $\mu_X > 0$ . Let  $g(\mu_X, \mu_Y) = \mu_Y/\mu_X$ , then the partial derivatives are  $g_1 = -\mu_y/\mu_x^2$  and  $g_2 = 1/\mu_x$ .

$$\begin{aligned}g_1\bar{X} + g_2\bar{Y} &= (-\mu_y/\mu_x^2)\bar{X} + (1/\mu_x)\bar{Y} \\ &= (1/\mu_x)(\bar{Y} - \eta\bar{X}),\end{aligned}$$

so

$$\text{Var}[\bar{Y}/\bar{X}] \approx \frac{1}{\mu_x^2} \text{Var}[\bar{Y} - \eta\bar{X}]. \quad (7.20)$$

In fact,  $\bar{Y}/\bar{X}$  is asymptotically normal with mean  $\mu_Y/\mu_X$  and variance given by Equation (7.20) (see (Hesterberg (1991)) for the proof). There are two important special cases. For paired data, let  $r_i = y_i - \hat{\eta}x_i$ , and  $s_r$  be the corresponding sample standard deviation. The standard error is

$$\text{SE}(\bar{Y}/\bar{X}) = \frac{s_r}{\mu_x \sqrt{n}}.$$

For independent samples,

$$\text{SE}(\bar{Y}/\bar{X}) = \frac{1}{\mu_x} \sqrt{s_Y^2/n_2 + \hat{\eta}^2 s_X^2/n_1}.$$

**Example 7.24** In Example 5.9 (relative risk), 55 of 3338 smokers and 21 of 2676 nonsmokers died of cardiovascular disease, with a relative risk of  $\hat{r} = \hat{p}_1/\hat{p}_2 = 0.0165/0.0078 = 2.1$ . This is a ratio of means (a proportion is an average of Bernoulli variables) with independent numerator and denominator; we use Equation (7.7) with  $\hat{p}_1(1 - \hat{p}_1)$  in place of  $s_y^2$  and  $\hat{p}_2(1 - \hat{p}_2)$  for  $s_x^2$ .

$$\begin{aligned}s_{\hat{r}}^2 &= \frac{\hat{p}_1(1 - \hat{p}_1)/n_1 + \hat{r}^2 \hat{p}_2(1 - \hat{p}_2)/n_2}{\hat{p}_2^2} \\&= \frac{0.0162/3338 + 2.1^2(0.00779)/2676}{0.0078^2} \\&= 0.54^2.\end{aligned}$$

The estimated relative risk of 2.1 has a standard error of 0.54. A 95% confidence interval is  $\hat{r} \pm 1.96s_{\hat{r}} = (1.05, 3.15)$ . Hence, there is substantial uncertainty in the estimate.

If we flip the numerator and denominator,  $\hat{p}_2/\hat{p}_1$ , the resulting 95% confidence interval is  $(0.238, 0.714)$ . Inverting that gives a corresponding confidence interval for  $\hat{p}_1/\hat{p}_2$ ,  $(1/0.714, 1/0.0238) = (1.34, 4.20)$ . This is very different from  $(1.05, 3.15)$ .  $\square$

This is a weakness of standard intervals of the form  $\hat{\theta} \pm \text{SE}_{\hat{\theta}}$ : different analysts can get nonequivalent confidence intervals depending on how they express their problem.

To partially address this issue, analysts should use standard transformations when such transformations exist. For ratios, it is standard to use a log transformation; for example, work with

$$\hat{\eta} = \log(\bar{Y}/\bar{X}) = \log(\bar{Y}) - \log(\bar{X}).$$

**Example 7.25** Referring again to the relative risk example (Example 5.9), we continue with a log transformation,

$$\hat{\eta} = \log(\hat{p}_1/\hat{p}_2) = \log(\hat{p}_1) - \log(\hat{p}_2).$$

We again need the delta method. The partial derivatives are  $\partial\eta/\partial p_1 = 1/p_1$  and  $\partial\eta/\partial p_2 = 1/p_2$ , and

$$\hat{\eta} \approx \eta + (\hat{p}_1 - p_1)/p_1 - (\hat{p}_2 - p_2)/p_2$$

with variance (in this case, with independent samples)

$$\text{Var}[\hat{\eta}] \approx \frac{p_1(1 - p_1)/n_1}{p_1^2} + \frac{p_2(1 - p_2)/n_2}{p_2^2} = \frac{(1 - p_1)}{n_1 p_1} + \frac{(1 - p_2)}{n_2 p_2}.$$

The squared standard error is obtained by plugging in estimates. The resulting 95% confidence interval is  $(0.24, 1.24)$ . Transforming back to the original scale by exponentiating gives a 95% confidence interval for  $\hat{p}_1/\hat{p}_2$  of  $(1.27, 3.46)$ , intermediate between the intervals obtained using the two fractions, and probably more accurate than either.  $\square$

### Remark

- The delta method is associated with normal approximations. We usually apply it when the input estimator(s)  $\theta$  are asymptotically (multivariate) normally distributed, then under some conditions, the output  $\eta$  is also normally distributed with variance consistent with the standard error calculations above. Hence,  $t$  confidence intervals using delta method standard errors are asymptotically correct.
- A confidence interval procedure is *transformation invariant* if it yields equivalent confidence intervals for any monotone transformation of the parameter of interest. Bootstrap percentile intervals are transformation invariant,  $t$  intervals and  $z$  intervals are not transformation invariant. Bootstrap  $t$  intervals (Section 7.5.2) are approximately transformation invariant.  $\parallel$

## Exercises

- 7.1** A researcher hired by a farming organization obtains a random sample of 50 cows in a state and finds a 95% confidence interval for the mean milk production for cows  $\mu$  to be  $(22, 30)$  kg/day. Critique the following interpretations:
- There is a 95% chance that a cow produces, on average, between 22 and 30 kg of milk per day.
  - We are 95% confident that  $\bar{x}$  is between 22 and 30 kg of milk per day.
  - The mean milk production  $\mu$  for cows in the state will be 22–30 kg/day 95% of the time.
  - We are 95% confident cows in this state produce, on average, from 22 to 30 kg of milk per day.
  - In 95% of samples, the mean milk production will be between 22 and 30 kg/day.
- 7.2** For high school seniors in 2021 who took the SAT exam, the mean math SAT score was  $\mu = 528$  with a standard deviation of  $\sigma = 120$ . From a random sample of 34 students in your county, you find the average SAT score to be 538. You forgot to compute the standard deviation, so you decide to assume that the standard deviation of scores of high school seniors who took the SAT in your county is the same as the national standard deviation of 120. Compute a 95% confidence interval for the mean SAT score in your county.

- 7.3** Suppose that 20 years ago, the mean cholesterol level of adult men in a certain town was 185 mg/dl with a standard deviation of 50 mg/dl.
- Suppose you obtain a sample of size 100 and find the mean cholesterol level to be  $\bar{x} = 210$ . Assuming that  $\sigma$  hasn't changed, find a 90% confidence interval for the mean cholesterol level of the population (of adult men in this town now).
  - Suppose you decide to conduct a new study to determine the mean cholesterol levels of adult men in this town. Assuming that the standard deviation has not changed, how many people should you include in your sample if you want the margin of error to be at most 10 mg/dl, using 95% confidence?
  - If you want to be 99% confident with that MOE, then how large should your sample size be?
- 7.4** In R, the `qt` function computes quantiles of the  $t$  distribution.
- Find the quantile  $q$  used in a 90%  $t$  confidence interval for a sample of size  $n = 5$ ,  $n = 15$ ,  $n = 30$ ,  $n = 100$ . Compare to the corresponding quantile for the standard normal.
  - For a fixed sample size  $n$ , say,  $n = 15$ , find the quantile  $q$  used in a  $(1 - \alpha) \times 100\%$   $t$  confidence interval for 90% confidence, 95% confidence, 99% confidence. Which confidence level results in the narrowest interval, and which the widest interval (assuming a fixed  $n$ )?
- 7.5** Suppose you draw a random sample of size  $n$  from a normal distribution with unknown mean  $\mu$  and known standard deviation  $\sigma$  and construct a 95% confidence interval for  $\mu$ . If you want to halve the margin of error, how much larger would the sample size have to be?
- 7.6** Isabella is interested in the sugar content of vanilla ice cream. She obtains a random sample of  $n = 20$  brands and finds an average of 18.05 g with standard deviation 5 g (per half cup serving). Assuming that the data come from a normal distribution, find a 90% confidence interval for the mean amount of sugar in a half cup serving of vanilla ice cream.
- 7.7** Biologists “were interested in understanding the seasonal changes in body condition in order to investigate the effects of adaptive strategies on life histories of animals.” In their study of 26 striped skunks, they found that the total body mass ranged from 0.55 to 6.17 kg with a mean of 3.09 kg and standard deviation 1.35 kg (Ten Hwang et al. (2005)).
- Assume the distribution of total body mass of striped skunks is normal and compute a 95% confidence interval for the true mean total body mass of striped skunks.
  - Prior to your calculations, you create a histogram of your data. What should you be on the look-out for?

- 7.8** An engineer is studying the length of time his company's rechargeable batteries will work before needing to be recharged. He tests a random sample of 100 batteries and finds the average time that these batteries hold a charge is 120 h with standard deviation 12 h. Assume that the data exhibit only moderate skewness and find a 95% one-sided lower  $t$  confidence bound for the true mean length of battery life, and give an interpretation of the bound.
- 7.9** In Example 7.7, we drew random samples of size 20 from  $\text{Gamma}(5, 2)$  to see how often the  $t$  confidence interval for the mean misses the true mean. Repeat this simulation by changing the sample size, say  $n = 10$ ,  $n = 40$ ,  $n = 100$ , and  $n = 250$ . How does the sample size affect the frequency of missing  $\mu$ ?
- 7.10** Import the `FlightDelay` data set (see case study in Section 1.1). The variable `Delay` gives the lengths of flight delays. Since we have all flights departing from LaGuardia airport for American Airlines and United Airlines in May and June 2009, these values represent a population.
- Create a histogram of the delay times, describe the distribution and find the mean.
  - For random samples of size 30 from this population, will the 95% confidence intervals capture the true mean 95% of the time? We will simulate this question. Adapt the simulation in Section 7.1.1:

```

mu <- mean(FlightDelays$Delay)
counter <- counter + 1
p <- ggplot(data.frame(x = c(-20,100)), aes(x = x))
# Set up plot for CI's

for (i in 1:1000)
{
  x <- sample(FlightDelays$Delay, 30, replace = FALSE)
  L <- t.test(x)$conf.int[1]
  U <- t.test(x)$conf.int[2]
  if (L < mu && mu < U)
    counter <- counter + 1
  # plot first 100 CI's
  if (i <= 100)
    p <- p + annotate("segment", x = L, xend = U, y = i, yend = i)
}

# vertical line at true mu
p + geom_vline(xintercept = mu, lty = 2)

counter/1000  # fraction of times CI captures mu

```

Did 95% of confidence intervals capture the true mean? What is the issue here?

- (c) Notice the varying lengths of the confidence intervals. Why are some intervals very short while others very long?
- 7.11** Import the data set `Olympics2012` that has age, weight (pounds) and height (inches) information on a random sample of the over 10 000 athletes who participated in the 2012 Olympic Games in London. For the 26 female athletes in the sample, plot the distribution of their ages and then find a 95% confidence interval for the mean age.
- 7.12** Import the data set `Spruce` (case study in Section 1.10) into R.
- Create exploratory plots to check the distribution of the variable `Ht.change`.
  - Find a 95%  $t$  confidence interval for the mean height change over the 5-year period of the study and give a sentence interpreting your interval.
- 7.13** Researchers conducted a small study to determine the effects of diet on severely obese patients (Samaha et al. (2009)). After 6 months, the 43 patients who were randomly assigned to eat a low carbohydrate diet lost an average of 5.8 kg with standard deviation 8.6 kg. The 36 patients on the low fat diet lost an average of 1.9 kg with standard deviation 4.2 kg. Find a 95%  $t$  confidence interval for the mean difference in weight loss between the two groups (low-carb versus low-fat) and state the interpretation of the interval. (Use  $df = 63.185$ ).
- 7.14** Consider the data set `Girls2004` with birth weights of baby girls born in Wyoming or Alaska (case study in Section 1.2).
- Create exploratory plots and compare the distribution of weights between the babies born in the two states.
  - Find a 95%  $t$  confidence interval for the mean difference in weights for girls born in these two states. Give a sentence interpreting this interval.
- 7.15** Consider the data set `Girls2004` (see case study in Section 1.2).
- Create exploratory plots and compare the distribution of weights between babies born to nonsmokers and babies born to smokers.
  - Find a 95% one-sided lower  $t$  confidence bound for the mean difference in weights between babies born to nonsmokers and smokers. Give a sentence interpreting the interval.

- 7.16** Import the data set `Spruce` (case study in Section 1.10) into R. We will compare the mean height change of the seedlings planted in a fertilized plot to those planted in a nonfertilized plot.
- Create exploratory plots to compare the distributions of the variable `Ht.change` for the seedlings in the fertilized and nonfertilized plots.
  - Find a 95% one-sided lower  $t$  confidence bound for the mean difference in height change ( $F - NF$ ) over the 5 year period of the study, and give a sentence interpreting your interval.
- 7.17** Import the `FlightDelays` data set (case study in Section 1.1) into R. Although the data represent all flights for United Airlines and American Airlines in May and June 2009, assume for this exercise that these flights are a sample from all flights flown by the two airlines under similar conditions. We will compare the lengths of flight delays between the two airlines.
- Create exploratory plots of the lengths of delays for the two airlines.
  - Find a 95%  $t$  confidence interval for the difference in mean flight delays between the two airlines, and interpret this interval.
- 7.18** Import the data set `Olympics2012` (see Exercise 7.11). We will investigate the weights of the female athletes in the sample.
- Begin with an exploratory plot of the data. What do you observe?
  - Find a 95%  $t$  confidence interval for the mean population weight  $\mu$ .
  - Remove the outlier and find the 95%  $t$  confidence interval for the mean population weight  $\mu$ . Did this change much?
- 7.19** (Exercise 7.18 continued) We now consider the weights of the male athletes.
- Begin with an exploratory plot of this variable and describe the distribution.
  - Create a 95%  $t$ -confidence interval for the difference in mean weights between the males and females (use all the data in the sample). State your interpretation of this interval.
  - There are two outliers in the weights of the athletes. Remove these measurements and recreate the 95%  $t$  confidence interval for the difference in means. State your interpretation of this interval.
- 7.20** Is there a difference in the price of groceries sold by the two retailers Target and Walmart? The data set `Groceries` contain a sample of grocery items and their prices advertised on their respective websites on one specific day. (See Exercise 5.22.) Create a 95%  $t$ -confidence interval for the mean difference in prices. Is the outlier in the data influential?

- 7.21** Does chocolate ice cream have more calories than vanilla ice cream? The data set `IceCream` contains calorie information for a sample of brands of chocolate and vanilla ice cream. (See Exercise 3.20.) Create a 95% one-sided  $t$ -confidence interval to explore this question.
- 7.22** Run a simulation to see if the  $t$  ratio  $T = (\bar{X} - \mu)/(S/\sqrt{n})$  has a  $t$  distribution or even an approximate  $t$  distribution when the samples are drawn from a nonnormal distribution. Be sure to superimpose the appropriate  $t$  density curve onto your histograms. Try two different nonnormal distributions and remember to see if sample size makes a difference.
- 7.23** Following Theorem 7.1, we provide a confidence interval for the difference in means if we know that the population variances are the same. We also cautioned against using this result because it is difficult to determine whether the population variances are indeed the same. Run a simulation to compare the pooled and unpooled  $t$  confidence intervals for the difference in means, when the population variances are different.

The R code below will draw random samples of size  $m$  and  $n$  from  $N(8, 10^2)$  and  $N(3, 15^2)$ . We will count the number of times the two 95% confidence intervals capture the true difference in mean of 5.

```

pooled.count <- 0          # set counter to 0
unpooled.count <- 0        # set counter to 0

m <- 20                    # sample size
n <- 10                    # sample size

N <- 10000                 # number of runs
for (i in 1:N)
{
  x <- rnorm(m, 8, 10)      # Draw m from N(8,10^2)
  y <- rnorm(n, 3, 15)      # Draw n from N(3,15^2)
  # Conf ints, with pooled and unpooled variance
  CI.pooled <- t.test(x, y, var.equal = T)$conf
  CI.unpooled <- t.test(x, y)$conf

  # Is 5 in interval? If yes, increase counter.
  if (CI.pooled[1] <= 5 && 5 <= CI.pooled[2])
    pooled.count <- pooled.count + 1
  if (CI.unpooled[1] <= 5 && 5 <= CI.unpooled[2])
    unpooled.count <- unpooled.count + 1
}

pooled.count/N            # Proportion of times
unpooled.count/N          # CI covers 5

```

- (a) Compare the performance of the two versions of the confidence interval for the difference in means.
- (b) Repeat the simulation with different sample sizes, for example,  $m = 80, n = 40; m = 120, n = 80; m = 80, n = 80$ . Discuss.
- 7.24** Like many states, Tennessee conducts audits of stores to determine whether or not proper sales tax was assessed. The Department of Revenue obtains a random sample of transactions at the audited store and for each transaction looks at the tax error defined to be the amount of tax owed minus the amount of tax paid. The auditors examine the lower bound of a 75% one-sided upper  $t$  confidence interval and if it is larger than 0, the store owes the state money.<sup>5</sup> Suppose the Department of Revenue samples 500 transactions of a certain store and finds the average tax error to be US\$5.29 with a standard deviation of \$3.52. Compute a 75% one-sided upper  $t$  confidence interval for the true mean tax error.
- 7.25** In 2021, the Marist Poll surveyed 1037 American adults about hot dogs.<sup>6</sup>
- In the sample, 702 reported that they eat hot dogs. Find a 95% confidence interval for the true proportion of American adults who eat hot dogs.
  - Of the 630 hot dog eaters in the sample who add toppings, 47% claimed that mustard is the one topping they must have on the hot dog. Is it plausible that for hot dog eaters who add toppings, a majority claim that mustard is the one topping they must have on their hot dog?
- 7.26** Intestinal parasitic infections are a common problem in Countries, where access to clean water and sanitation may be limited. Researchers conducted a study to see if handwashing with soap was an effective method for reducing intestinal infection rates in school-aged children in northern Ethiopia (Mahmud et al. (2015)). Before the study, the participants either tested negative for the intestinal parasitic infection or were given a pretrial anti-parasitic treatment. Then they were randomly assigned to either a handwashing with soap treatment (before meals, after defecation, etc.) or to no intervention. After 6 months, 13 out of the 91 students assigned the handwashing treatment became reinfected with intestinal parasites compared to 33 out of the 87 students in the control group. Compute a 95% confidence interval for the difference in proportions and give a sentence interpreting your interval.

<sup>5</sup> <https://www.tn.gov/content/dam/tn/revenue/documents/taxes/sales/statisticalsampling.pdf>.

<sup>6</sup> <https://maristpoll.marist.edu/polls/life-marist-poll-results-analysis-hot-dog-toppings>.

- 7.27** A retail store wishes to conduct a marketing survey of its customers to see if customers would favor longer store hours. How many people should be in their sample if the marketers want their margin of error to be at most 3% with 95% confidence, assuming
- They have no preconceived idea of how customers will respond.
  - A previous survey indicated that about 65% of customers favor longer store hours.
- 7.28** Verify that  $\sqrt{\hat{p}(1 - \hat{p})}$  is a maximum when  $\hat{p} = 0.5$ .
- 7.29** Suppose researchers wish to study the effectiveness of a new drug to alleviate hives due to test anxiety. Seven hundred students are randomly assigned to take either this drug or a placebo. Suppose 34 of the 350 students who took the drug break out in hives compared to 56 of the 350 students who took the placebo.
- Compute a 95% confidence interval for the proportion of students taking the drug who break out in hives.
  - Compute a 95% confidence interval for the proportion of students on the placebo who break out in hives.
  - Do the intervals constructed in (a) and (b) overlap? What, if anything, can you conclude about the effectiveness of the drug?
  - Compute a 95% confidence interval for the difference in proportions of students who break out in hives by using or not using this drug and give a sentence interpreting this interval.
  - The results in (c) and (d) above seem to contradict each other. If you want to compare the effectiveness of the drug versus the placebo, which is the correct approach? Why?
- 7.30** An article in the March 2003 *New England Journal of Medicine* describes a study to see if aspirin is effective in reducing the incidence of colorectal adenomas, a precursor to most colorectal cancers (Sandler et al. (2003)). Of 517 patients in the study, 259 were randomly assigned to receive aspirin and the remaining 258 received a placebo. One or more adenomas were found in 44 of the aspirin group and 70 in the placebo group. Find a 95% one-sided upper bound for the difference in proportions ( $p_A - p_P$ ) and interpret your interval.
- 7.31** The data set Bangladesh has measurements on water quality from 271 wells in Bangladesh (Example 5.3). There are two missing values in the chlorine variable. Use the `drop_na` function in the `plyr` package to remove them.

```
chlorine <- drop_na(Bangladesh, Chlorine) %>% pull(Chlorine)
```

- (a) Compute the numeric summaries of the chlorine levels and create a plot and comment on the distribution.
  - (b) Find a 95%  $t$  confidence interval for the mean  $\mu$  of chlorine levels in Bangladesh wells.
  - (c) Find the 95% bootstrap percentile and bootstrap  $t$  confidence intervals for the mean chlorine level and compare results. Which confidence interval will you report?
- 7.32** The data set `MnGroundwater` has measurements on water quality of 895 randomly selected wells in Minnesota.
- (a) Create a histogram or normal quantile plot of the alkalinity and comment on the distribution.
  - (b) Find a 95%  $t$  confidence interval for the mean  $\mu$  of alkalinity levels in Minnesota wells.
  - (c) Find the 95% bootstrap percentile and bootstrap  $t$  confidence intervals for the mean alkalinity level and compare results. Which confidence interval will you report?
- 7.33** Consider the babies born in Texas in 2004 (`TXBirths2004`, case study in Section 1.2). We will compare the weights of babies born to nonsmokers and smokers.
- (a) How many nonsmokers and smokers are there in this data set?
  - (b) Create exploratory plots of the weights for the two groups and comment on the distributions.
  - (c) Compute the 95% confidence interval for the difference in means using the formula  $t$ , bootstrap percentile, and bootstrap  $t$  methods and compare your results. Which interval would you report?
  - (d) Modify your result from above to obtain a one-sided 95%  $t$  confidence interval (hypothesizing that babies born to nonsmokers weigh more than babies born to smokers).
- 7.34** Import the `FlightDelays` data set (case study in Section 1.1) into R. We will assume these data represent a sample of all flights flown by United Airlines and American Airlines in May and June 2009.
- (a) Obtain summary statistics and a quantile normal plot of the distribution of the lengths of flight delays that occurred on Fridays. Describe the distribution.
  - (b) Compute the 90% bootstrap  $t$ -confidence interval for the trimmed mean length of flight delays by trimming by 10% on either end of the distribution (see Example 7.23).
- 7.35** In the Google mobile ads case study (Section 1.12), Google was interested in comparing the amount that advertisers paid (`m.cost_pre`)

before and after Google's recommendation (`m.cost_post`). We will restrict our attention to the mobile platform.

- (a) Create quantile plots of each variable and describe the distribution.
  - (b) Explain why a two sample interval is not appropriate here.
  - (c) Compute the difference between these two variables and then create a quantile normal plot of the difference.
  - (d) Find a 95% formula  $t$  confidence interval for the true mean change in cost and give a sentence interpreting this interval.
  - (e) Find a 95% bootstrap  $t$  confidence interval for the true mean change. Does it differ much from the formula  $t$  interval? Why or why not?
- 7.36** As we have seen, we can create confidence intervals for many different parameters besides a mean or proportion. One statistic common in epidemiology or medical research is relative risk,  $p_1/p_2$ , where  $p_1$  and  $p_2$  are the proportions of people in two groups who have a disease or a condition. (See Example 5.9.) Suppose in a study of college students, 23% of sophomores reported binge drinking during homecoming compared to 12% of seniors. The sample relative risk is  $0.23/0.12 = 1.9$ , so sophomores are 1.9 times more likely to binge drink than seniors.
- (a) If in the study, the researchers compute a 95% confidence interval for the true relative risk to be  $(1.6, 2.3)$ , give a sentence interpreting this interval.
  - (b) Under what circumstance would a researcher conclude, based on a confidence interval for the relative risk, that there is no difference between the two groups?
- 7.37** In this exercise, we compare two different ways to judge whether the parameters  $\theta_1$  and  $\theta_2$  for two populations differ. We assume here that methods based on the CLT are reasonable. We calculate estimates  $\hat{\theta}_1$  and  $\hat{\theta}_2$  and the corresponding standard errors  $\hat{SE}_1$  and  $\hat{SE}_2$ , and calculate intervals:

$$\hat{\theta}_1 \pm 1.96\hat{SE}_1 \quad (7.21)$$

$$\hat{\theta}_2 \pm 1.96\hat{SE}_2. \quad (7.22)$$

One approach to determining whether or not  $\theta_1 = \theta_2$  is to note whether or not these confidence intervals overlap. Alternately, we may compute the confidence interval for the difference  $\theta_1 - \theta_2$

$$(\hat{\theta}_1 - \hat{\theta}_2) \pm 1.96\sqrt{\hat{SE}_1^2 + \hat{SE}_2^2} \quad (7.23)$$

and note to see whether or not the interval contains 0.

- (a) Explain why intervals (7.21) and (7.22) overlap if and only if  $(\hat{\theta}_1 - \hat{\theta}_2) \pm 1.96(\hat{SE}_1 + \hat{SE}_2)$  contains 0.
- (b) Explain why the ratio of the width of the interval in (a) to the width of interval (7.23),  
 $(\hat{SE}_1 + \hat{SE}_2)/\sqrt{\hat{SE}_1^2 + \hat{SE}_2^2}$ , is greater than 1.
- (c) What does this imply about the two methods for gauging whether or not  $\theta_1 = \theta_2$ ?  
See Schenker and Gentleman (2001) for more discussion of this issue.

**7.38** Let  $X_i \sim N(\mu_1, \sigma^2)$ ,  $i = 1, 2, \dots, n$  and  $Y_j \sim N(\mu_2, 2\sigma^2)$ ,  $j = 1, 2, \dots, m$ , be independent samples, and suppose  $\sigma^2$  is known. Find a 95% confidence interval for  $3\mu_1 + \mu_2$ .

**7.39** Let  $X_i \sim N(\mu_1, \sigma^2)$ ,  $i = 1, 2, \dots, n_1$  and  $Y_j \sim N(\mu_2, \sigma^2)$ ,  $j = 1, 2, \dots, n_2$  be independent samples with sample means and variances  $\bar{X}, S_1^2$  and  $\bar{Y}, S_2^2$ , respectively. Use Theorem 7.1 to verify that the confidence interval for  $\bar{X} - \bar{Y}$  is given by Equation (7.10).

**7.40** Suppose  $X \sim N(0, \sigma^2)$ . Then  $X^2/\sigma^2$  has a chi-square distribution with 1 degree of freedom (see Theorem B.14). Use this fact to find a 95% confidence interval for  $\sigma^2$ . (In R, the qchisq function computes quantiles for the chi-square distribution.)

**7.41** Let  $X_1, X_2, \dots, X_n$  be a random sample from a Poisson distribution with  $\lambda > 0$ . From Proposition 6.2 and Theorem A.7, we know that  $\bar{X}$  is an unbiased estimator of  $\lambda$ .

(a) Use the CLT approximation to find a 95% confidence interval for  $\lambda$ .  
(b) Compute the 95% confidence interval for the sample 4, 6, 7, 9, 10, 13.

**7.42** Let  $X_1, X_2, \dots, X_n$  be a random sample from  $N(\mu, \sigma^2)$  and let  $\bar{X}$  and  $S^2$  denote the sample mean and sample variance, respectively. Is the random variable  $(\bar{X} - \mu)/(S/\sqrt{n})$  a pivotal quantity? If yes, find a 95% confidence interval for  $\mu$ .

**7.43** Let  $X \sim \text{Gamma}(2, \lambda)$ . Then  $2\lambda X$  has a chi-square distribution with 4 degrees of freedom (see Exercise B.11). Use this fact to find a 95% confidence interval for  $\lambda$ . (In R, the qchisq function computes quantiles for the chi-square distribution.)

**7.44** Let  $X$  have the Pareto distribution  $\text{Pareto}(5, \beta)$  with pdf  $f(x) = 5\beta^5/x^6$  for  $x \geq \beta > 0$ .

- (a) Let  $Y = \log(X/\beta)$ . Show that  $Y \sim \text{Exp}(5)$ , the exponential distribution with parameter  $\lambda = 5$ , so  $Y$  is a pivotal statistic.
- (b) Suppose a single random value  $X$  is drawn from  $\text{Pareto}(5, \beta)$  and you observe  $x = 10$ . Use part (a) to find a 95% confidence interval for  $\beta$ .
- 7.45** For the German tank example (7.13), derive a confidence interval using  $\bar{X}/\theta$  as a pivotal statistic. Assume that  $\bar{X}$  has an approximate normal distribution (by the CLT). Find the 95% confidence interval, for  $n = 400$  when  $\bar{X} = 5000$ . Compare the width of this interval with the interval based on  $X_{\max}$ .
- 7.46** In this exercise, you will derive and use a “bootstrap  $Z$ ” interval.
- Following the steps in the derivation of the bootstrap  $t$  interval in Section 7.5.2, derive a bootstrap  $Z$  interval for  $\mu$ , for cases when  $\sigma$  is known.
  - Calculate this interval for the Verizon CLEC data; for  $\sigma$ , use the sample variance of the Verizon ILEC data. (In practice, we sometimes use methods for known  $\sigma$  when we can estimate  $\sigma$  from a large related data set.)
  - Compare that interval with a formula  $z$  interval. How does the bootstrap  $Z$  interval adjust for skewness?
- 7.47** Let  $X_1, X_2, \dots, X_n$  denote a random sample from  $N(\mu, \sigma^2)$  with sample mean  $\bar{X}$  and variance  $S^2$ . Let  $q_1$  and  $q_2$  denote the  $\alpha/2$  and  $(1 - \alpha/2)$  quantiles, respectively, of the chi-square distribution with  $n - 1$  degrees of freedom. Show that  $(1 - \alpha) \times 100\%$  confidence interval for  $\sigma^2$  is given by  $((n - 1)S^2/q_2, (n - 1)S^2/q_1)$ . (Hint: Theorem B.16.)
- 7.48** Robin wonders how much variation there is in a box of his favorite cereal. He buys eight boxes from eight different stores and finds the weights of the contents to be (in grams)

560	568	580	550	581	581	562	550
-----	-----	-----	-----	-----	-----	-----	-----

Assuming the data are from a normal distribution, find a 90% confidence interval for the variance  $\sigma^2$ . (In R, the `qchisq` function computes quantiles for the chi-square distribution.)

- 7.49** Let  $X$  be a random variable from a distribution with  $\text{pdf } f(x) = 2x/\theta^2$  for  $0 < x < \theta, \theta > 0$ . Show that  $Y = X/\theta$  is a pivot and then find a  $(1 - \alpha) \times 100\%$  confidence interval for  $\theta$

- 7.50** Let  $f$  be the pdf for some random variable and suppose  $\mu$  and  $\sigma > 0$  are two constants. Let  $h(x) = \frac{1}{\sigma}f((x - \mu)/\sigma)$ . Show that  $h(x)$  is also a pdf.

- 7.51** Welch's approximation is based on the following idea: If (a)  $Y \sim N(0, \sigma^2)$ , and  $\sigma^2$  is estimated by  $\hat{\sigma}^2$ , with  $\hat{\sigma}^2$  independent of  $Y$ , (b)  $E[\hat{\sigma}^2/\sigma^2] = 1$ , and (c)  $\text{Var}[\hat{\sigma}^2/\sigma^2] = 2/c$ , then  $Y/\hat{\sigma}$  has approximately a  $t$  distribution with  $v = c$ . Hence, an accurate estimate for  $\sigma$  gives high degrees of freedom and a poor estimate gives lower degrees of freedom (with a correspondingly wide confidence interval).

In the one sample case, where the sample is from a normal distribution,  $S^2/\sigma^2$  is a gamma distribution with mean 1 and variance  $2/(n - 1)$ , and is independent of the sample mean. Hence, the degrees of freedom is  $v = n - 1$  in the one-sample case.

In the two-sample case, we use  $S_1^2/n_1 + S_2^2/n_2$  as an estimate for  $\sigma_1^2/n_1 + \sigma_2^2/n_2$ . Show that if both populations are normal, then this estimate satisfies the three conditions for the idea. In particular, show that the value of  $c$  equals Welch's approximation. *Hint:* See Exercise B.12.

- 7.52** If  $(L, U)$  is a 95% CI for  $\theta$  and  $h$  is a transformation, then is  $(h(L), h(U))$  a 95% CI for  $h(\theta)$ ? We will run a simulation to check this. In particular, if  $(L, U)$  is a 95% confidence interval for  $\mu$ , is  $(L^2, U^2)$  a 95% confidence interval for  $\mu^2$ ? We will draw random samples from the normal distribution  $N(2, 7)$ .

```

N <- 10^4
counter <- 0
counter2 <- 0

n <- 30
mu <- 2      # true mean
sigma <- 7    # true st dev
for (i in 1:N)
{
  w <- rnorm(n, mu, sigma)          # sample
  se <- sd(w)/sqrt(n)              # standard error
  L <- mean(w) - qt(.975, n-1)*se # lower bound
  U <- mean(w) + qt(.975, n-1)*se # upper bound
  if (L < mu & mu < U)           # (L, U) captures true mu?
    counter <- counter + 1
  if (L^2 < mu^2 & mu^2 < U^2)     # (L^2, U^2) captures mu^2?
    counter2 <- counter2 + 1
}

counter/N      # capture rate for mu
counter2/N     # capture rate for mu^2

```

- (a) Does the transformed interval capture  $\mu^2$  95% of the time?
- (b) Make a conjecture about when a transformation  $h$  would result in  $(h(L), h(U))$  including  $h(\theta)$  95% of the time. Run simulations to explore this.
- 7.53** Suppose that  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Gamma}(r, \lambda)$ .
- Find the variance of  $\bar{X}$ .
  - Use the delta method to find the approximate variance of  $\bar{X}^2$ .
  - Use the delta method to find the approximate variance of  $\log(\bar{X})$ .
  - Use the delta method to find the approximate variance of  $\sqrt{\bar{X}}$ .
  - Check your answers using simulation.
- 7.54** Compute standard errors for the ratio of means  $\bar{Y}/\bar{X}$  for the Verizon data, where  $\bar{Y}$  and  $\bar{X}$  are the mean of CLEC and ILEC repair times, respectively, using
- the delta method for  $\bar{Y}/\bar{X}$
  - the delta method for  $\log(\bar{Y}/\bar{X})$
  - the ordinary bootstrap for  $\bar{Y}/\bar{X}$
  - the ordinary bootstrap for  $\log(\bar{Y}/\bar{X})$
- 7.55** (Exercise 7.54 continued) Compute confidence intervals,
- $t$  interval for  $\bar{Y}/\bar{X}$  using the delta method SE,
  - $t$  interval for  $\log(\bar{Y}/\bar{X})$  using the delta method SE,
  - bootstrap percentile interval for  $\bar{Y}/\bar{X}$ , and
  - bootstrap percentile interval for  $\log(\bar{Y}/\bar{X})$ . For degrees of freedom for the  $t$  intervals, use the smaller sample size minus 1.
- 7.56** (Exercises 7.54 and 7.55 continued)
- transform the two intervals Exercise 7.55 (b) and (d) on the log scale to the scale  $\bar{Y}/\bar{X}$ , and
  - compare the four intervals for  $\mu_Y/\mu_X$ .



# 8

## More Hypothesis Testing

In Section 3.3, we introduced hypothesis testing, a formal procedure to evaluate a statement about a population or populations. In particular, we tested hypotheses involving two population means using a permutation test. This test makes no assumptions about the distributions underlying the two populations. We now consider hypothesis testing in situations where we can make some assumptions about the distribution of a population or populations, or where we can use formulas or bootstrap to obtain approximate  $P$ -values.

The underlying procedure for analyzing a hypothesis test remains the same as before: we compute a test statistic from the data, and then compute a  $P$ -value, measuring how often chance alone would give a test statistic as extreme as the observed statistic, assuming the null hypothesis is true. If the  $P$ -value is small, then the results cannot easily be explained by chance alone, and we reject the null hypothesis.

To compute the  $P$ -value, we need to find a reference distribution, the null distribution – the distribution that the test statistic follows if the null hypothesis is true. In this chapter, we use mathematical derivations and approximations to find reference distributions from common families of distributions.

### 8.1 Hypothesis Tests for Means and Proportions: One Population

We begin by comparing with hypothesis tests for one population, either a single mean or proportion, and then consider the difference of two means or proportions.

#### 8.1.1 A Single Mean

**Example 8.1** For college-bound seniors in 2019, SAT math scores are normally distributed with a mean of 528 and a standard deviation of 117.

You suspect that seniors in the town of Sodor are much brighter than the country as a whole, so you decide to conduct a test. In a random sample of 56 seniors, you find the mean SAT math score to be 555. Is this sufficient evidence to conclude that Sodor seniors are smarter, or could a mean score of 555 be attributable to random variability?

The parameter of interest is  $\mu$ , the mean math SAT score in Sodor. You will assume that the standard deviation of math scores in Sodor are the same as the national standard deviation of  $\sigma = 117$ . The hypotheses are as follows:

$$H_0: \mu = 528 \quad \text{versus} \quad H_A: \mu > 528.$$

As in Section 3.3, to determine the null distribution, we assume that the null hypothesis is true; here we assume that the data are from a normal distribution with mean 528 and standard deviation 117. Thus, the sample mean  $\bar{x}$  comes from  $N(528, 117^2/56)$ . We standardize

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{555 - 528}{117/\sqrt{56}} = 1.7269 \sim N(0, 1).$$

How likely is this? We compute the  $P$ -value,  $P(Z \geq 1.7269) = 0.042$ . So if the null hypothesis is true, then only 4.2% of samples of size 56 give rise to a mean as or more extreme as 555. This is (mild) evidence against the null hypothesis, so we conclude that Sodor seniors are brighter than the national pool of seniors (at least as measured by the SAT!)  $\square$

Now, suppose that we are not willing to assume that the variability of scores in Sodor is the same as the variability in the national scores. We test the same hypotheses, but estimate the standard error from the data, using  $s/\sqrt{n}$  in place of  $\sigma/\sqrt{n}$ . Suppose  $s = 128$ . Then the test statistic is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{555 - 528}{128/\sqrt{56}} = 1.5785.$$

We saw previously that  $t = (\bar{x} - \mu)/(s/\sqrt{n})$  has a Student's  $t$ -distribution with  $(n - 1)$  degrees of freedom when the population is normal (Section 7.1.2), so the null distribution for the  $t$  statistic is a  $t$ -distribution. The one-sided  $P$ -value is  $P(T \geq 1.5785) = 0.0572$ , so about 5.7% of samples of size 56 from  $N(528, \sigma^2)$  would give a  $t$  statistic this large or larger, a slightly larger probability than when we assumed we knew the true  $\sigma$ .

Thus, in these two examples, we see a common theme: we calculate a test statistic from the data and then find or estimate the distribution for this test statistic assuming the null hypothesis is true. In permutation testing, the reference distribution is the permutation distribution obtained by permuting the data. Here the reference distributions are parametric, normal, or  $t$ -distributions.

### T-Test for a Normal Mean

Let  $X_1, X_2, \dots, X_n$  be a random sample from a normal population with unknown  $\mu$  and  $\sigma$ . Let  $\bar{X}$  and  $S$  denote the sample mean and standard deviation. To test

$$H_0: \mu = \mu_0 \quad \text{versus} \quad H_A: \mu \neq \mu_0,$$

we form the  $t$ -test statistic

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}.$$

Under the null hypothesis,  $T$  has a  $t$ -distribution with  $(n - 1)$  degrees of freedom. The  $P$ -value is the probability that chance alone would produce a test statistic as extreme as or more extreme than the observed value  $t = (\bar{x} - \mu_0)/(s/\sqrt{n})$ , if the null hypothesis is true.

**Example 8.2** The coffee vending machine at work dispenses 7 oz of coffee in paper cups. The staff suspects that the machine is under-filling the cups. From a sample of  $n = 15$  cups, they compute a mean of 6.6 oz and a standard deviation of 0.8 oz. Does this evidence support their suspicions? Assume that the coffee amounts dispensed by the machine are normally distributed.

### Solution

Let  $\mu$  denote the true amount of coffee dispensed by the machine. We test  $H_0: \mu = 7$  versus  $H_A: \mu < 7$ . If the null hypothesis is true, the amount of coffee dispensed follows a normal distribution with mean 7 and an unknown standard deviation. The test statistic is  $t = (6.6 - 7.0)/(0.8/\sqrt{15}) = -1.9365$ . We compare this to a  $t$ -distribution with 14 degrees of freedom and find a  $P$ -value of  $P(t \leq -1.9365) = 0.0366$ . Thus, less than 4% of samples would result in test statistics as extreme as the observed  $-1.9365$ . This suggests that the machine is under-filling the cups.  $\square$

#### 8.1.1.1 Check Conditions

The one-sample  $t$ -test is exact under the same conditions as a one-sample  $t$  confidence interval – independent and identically distributed observations from a normal distribution. See Section 7.1.2.1.

If the population distribution is not normal but has finite variance, then as  $n \rightarrow \infty$  the distribution of  $T$  approaches a standard normal distribution –  $\sqrt{n}$  times the numerator approaches normality by the central limit theorem, and  $\sqrt{n}$  times the denominator approaches the correct variance by an application of the strong law of large numbers. But that doesn't say how close the distribution is to a  $t$ -distribution or a normal distribution.

In practice we may use the  $t$ -distribution to compute approximate  $P$ -values if the sample size is large and the sample distribution is not too skewed.

What do we mean by “is large and not too skewed”? Frankly, the common practice of statistics is negligent in this area – it is rare in practice to actually answer that question in any meaningful way before proceeding with  $t$ -tests or intervals. Fortunately, there are relatively easy ways to check that: perform tests that do not require symmetric populations such as a permutation test, a bootstrap  $t$ -test (Section 8.2), or Johnson’s skewness-adjusted  $t$  test (Johnson, 1978) (see <https://github.com/Ichihara/MathStatsResamplingR>). If these alternative approaches give a  $P$ -value that is substantially different from the ordinary  $t$ -test, then we should not use the ordinary test.

Outliers are always a concern and should be inspected: do they represent an extreme value from the population, or a recording mistake? Conduct the test with and without the outlier to determine if the outlier is influential. If the conclusion changes, then you should report both outcomes: it is not acceptable to report the results without the outlier unless there is a clear identifiable reason why that observation does not belong in the sample.

### 8.1.2 One Proportion

In this section, we use normal approximations to obtain  $P$ -values for testing a single proportion. In some ways, this is unnecessary – with modern computers as it is easy to compute  $P$ -values exactly using binomial probabilities, even for very large  $n$ . But it is useful in other ways. The techniques we learn here can apply to other situations,  $Z$  values are often easier to interpret than  $P$ -values are, and  $Z$  values are useful in their own right to combine results from multiple tests on different data sets.

**Example 8.3** About 13% of the population is left-handed. A biologist suspects that the scientific community is not like the general population in terms of handedness. He will conduct a study, and if the  $P$ -value from his test is less than 5%, he will conclude that the evidence supports his theory. He queries 200 scientists and finds that 36, or 18%, are left-handed. Does this data support the biologist’s theory?

#### Solution

Let  $p$  denote the true proportion of left-handed scientists. Then

$$H_0: p = 0.13 \quad \text{versus} \quad H_A: p \neq 0.13.$$

To calculate the  $P$ -value, we assume the null hypothesis is true. Then  $X$ , the number of left-handed scientists is a binomial random variable,  $X \sim \text{Binom}(200, 0.13)$ . How unusual is it to get 36 or more left-handers in a sample of size 200? The one-sided  $P$ -value is

$$\begin{aligned}
 P(X \geq 36) &= P(X = 36) + P(X = 37) + \cdots + P(X = 200) \\
 &= \sum_{i=36}^{200} \binom{200}{i} 0.13^i (1 - 0.13)^{200-i} \\
 &= 0.0267.
 \end{aligned}$$

We multiply by 2 for a two-sided alternative hypothesis, so the  $P$ -value is 0.0533. Thus, this biologist would conclude that the data does not support the hypothesis that scientists are different from the general population in terms of handedness.  $\square$

### Exact and Z Test for a Proportion

Let  $X$  be a binomial random variable,  $X \sim \text{Binom}(n, p)$ . Suppose  $X = x$  is observed. To test

$$H_0: p = p_0 \quad \text{versus} \quad H_A: p \neq p_0,$$

we compute  $P(X \geq x)$  if  $x \geq np_0$  or  $P(X \leq x)$  if  $x \leq np_0$ .

For instance,

$$P(X \geq x) = \sum_{i=x}^n \binom{n}{i} p_0^i (1 - p_0)^{n-i}.$$

A Z-test uses a normal approximation, preferably with a continuity correction. If  $x \geq np_0$ , then  $z = (x - 1/2 - np_0)/\sqrt{np_0(1 - p_0)}$  and the one-sided  $P$ -value is  $P(Z > z)$ , where  $Z$  is standard normal. If  $x \leq np_0$ , then  $z = (x + 1/2 - np_0)/\sqrt{np_0(1 - p_0)}$  and the one-sided  $P$ -value is  $P(Z < z)$ .

For a two-sided test, we double the one-sided probability to obtain the  $P$ -value.

#### 8.1.2.1 Normal Approximation with Continuity Correction

Instead of the exact calculation, we may do quick-and-dirty calculations using the CLT. Under the null hypothesis, the distribution of  $\hat{p}$  is approximately normal with mean 0.13 and variance  $0.13(1 - 0.13)/200 = (0.02378)^2$ . Using the continuity correction, the one-sided  $P$ -value is

$$P(X \geq 36) = P(X > 35.5) = P(\hat{p} > 35.5/200) = P(\hat{p} > 0.1775).$$

Standardizing, we obtain

$$z = (0.1775 - 0.13)/0.02378 = 1.997.$$

Thus, the one-sided  $P$ -value is  $P(z \geq 1.997) = 0.022887$ , and the two-sided  $P$ -value is 0.0458. This indicates mild evidence against the null hypothesis.

The exact calculation and the CLT yielded dramatically different results, reaching opposite conclusions when the biologist decides to use 5% as his threshold, with  $P$ -values differing by 0.8%, a fraction of about 1/6 of the true

value. Here,  $np = 200 \times 0.13 = 26$ , well above a common rule of thumb that the CLT approximation is appropriate if  $np > 10$  and  $n(1 - p) > 10$ . Indeed, we saw in Section 4.3.3 that requiring  $np > 384$  and  $n(1 - p) > 384$  was a much better condition when accuracy matters. So, in the case of hypothesis testing, when the CLT yields a borderline answer, it is best to use exact calculations.

In contrast to confidence intervals, we do not add anything to the numerator or the denominator (recall Section 7.4.1). This is because in hypothesis testing, we calculate as if the null hypothesis is true, whereas in confidence intervals, we allow for uncertainty in the true parameter.

## 8.2 Bootstrap $t$ Tests

The idea behind the bootstrap  $t$  test is to use the same statistic as for an ordinary  $t$ -test, but to estimate the null distribution using the bootstrap instead of assuming that it has a  $t$ -distribution. We compute one-sided  $P$ -values using the fraction of bootstrap  $t$  statistics that exceed the observed statistic, just like we did for permutation tests in Section 3.3.

To test a single mean, the  $t$ -statistic is  $T = (\bar{X} - \mu)/(S/\sqrt{n})$ , and bootstrap  $t$ -statistic is  $T^* = (\bar{X}^* - \bar{X})/(S^*/\sqrt{n})$ , where  $*$  indicates a value from a bootstrap sample. The one-sided  $P$ -value is the fraction of the  $T^*$  that exceed  $T$ .

More generally, when the  $t$ -statistic is  $T = (\hat{\theta} - \theta)/SE$ , the bootstrap  $t$ -statistic is  $T^* = (\hat{\theta}^* - \hat{\theta})/SE^*$ . This applies in one-sample, two-sample, and more general problems.

### Bootstrap $t$ -Test

Let  $\hat{\theta}$  be an estimate of  $\theta$ , and  $SE$  the corresponding standard error.

To test

$$H_0: \theta = \theta_0 \quad \text{versus} \quad H_A: \theta \neq \theta_0,$$

we calculate the  $t$  statistic

$$T = \frac{\hat{\theta} - \theta_0}{SE}.$$

We draw bootstrap samples, and for each, compute the bootstrap  $t$ -statistic

$$T^* = \frac{\hat{\theta}^* - \hat{\theta}}{SE^*}.$$

The  $P$ -value is the fraction of times the bootstrap statistics exceed the original statistic. Multiply by 2 for a two-sided test.

**Example 8.4** A study recommends that the maximum concentration of arsenic in irrigation water be  $100 \mu\text{g/l}$  to prevent a build-up of the chemical

that might harm future crop production. How does the water in Bangladesh measure up to this recommendation? Let  $\mu$  denote the true mean level of arsenic in Bangladesh wells. We wish to test  $H_0: \mu = 100$  versus  $H_A: \mu > 100$ .

A one-sided  $t$ -test yields  $t = 1.3988$  with a  $P$ -value of 0.0815 that suggest that perhaps, the arsenic levels are, on average, acceptable for irrigation purposes. However, we have already seen that the distribution of arsenic levels is strongly right-skewed (Figures 5.6) that makes the use of the  $t$  confidence interval suspect.

If we use the bootstrap  $t$ -test instead, we find a  $P$ -value of 0.0511, a smaller probability, which makes it more plausible that something besides chance variability explains the arsenic levels.

### R Note

```
> t.test(Bangladesh$Arsenic, mu = 100, alt = "greater")
t = 1.3988, df = 270, p-value = 0.08151
...
To obtain the P-value from the bootstrap t approach:

Arsenic <- Bangladesh$Arsenic
observedT <- t.test(Arsenic, mu = 100)$statistic
xbar <- mean(Arsenic)
n <- length(Arsenic)
N <- 10^5
Tstar <- numeric(N)

for (i in 1:N)
{
  bootx <- sample(Arsenic, n, replace = TRUE)
  Tstar[i] <- (mean(bootx) - xbar) / (sd(bootx) / sqrt(n))
}

(sum(Tstar >= observedT) + 1) / (N + 1)
```

□

Bootstrap  $t$ -tests are usually quite accurate. Under fairly general assumptions, the difference between the actual type I error rate and the nominal rate is  $\leq c/n$  for some  $c$ , where  $n$  is the sample size (or smaller of two sample sizes in a two-sample problem). In contrast, for  $t$ -tests the errors are  $\leq c/\sqrt{n}$  (or  $\leq c/n$  if there is zero skewness).

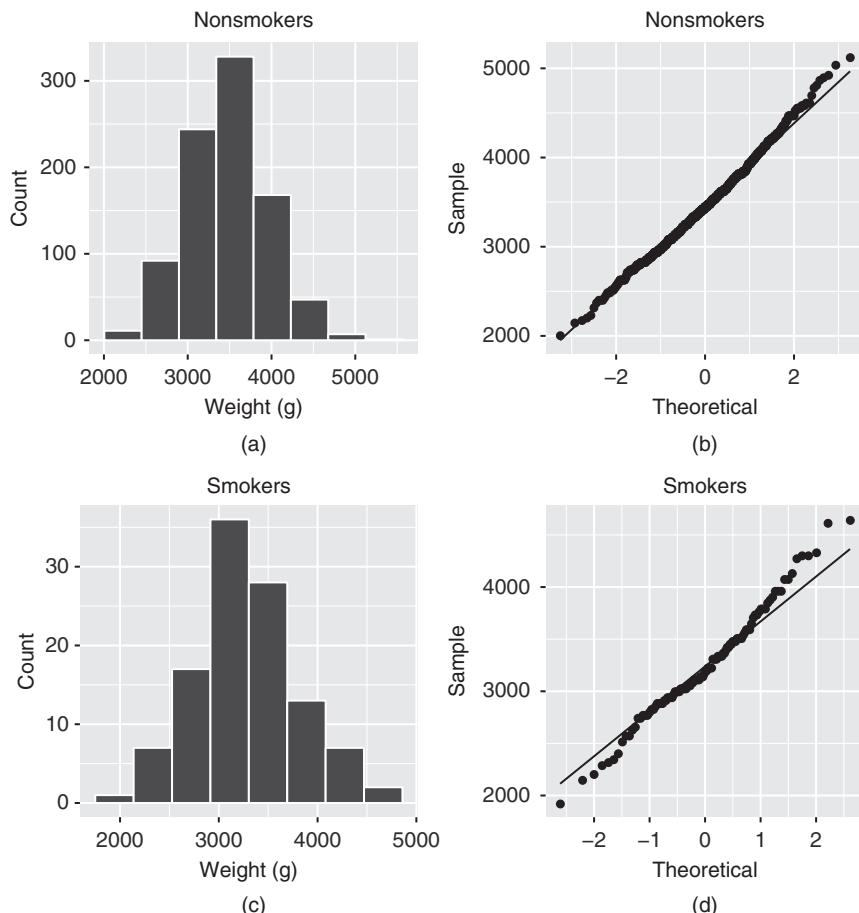
Permutation tests for comparing two samples are more accurate yet. Permutation tests are preferred, when it is possible to do them. But bootstrap  $t$ -tests can be used in a wider variety of applications, such as one-sample tests for means or proportions, or comparing two means when we believe the population variances are different (so that it is not appropriate to pool the data for a permutation test).

### 8.3 Hypothesis Tests for Means and Proportions: Two Populations

In Section 8.1, we discussed hypothesis testing when there was one population of interest. The same ideas apply when comparing two populations.

#### 8.3.1 Comparing Two Means

**Example 8.5** We return to the North Carolina babies case study. The mean and standard deviation of the weights of the  $n_1 = 898$  babies born to nonsmoking mothers are  $\bar{x}_1 = 3472$ ,  $s_1 = 479$  g, whereas the mean and standard deviation of the weights of the  $n_2 = 111$  babies born to smoking



**Figure 8.1** Distribution of weights of babies born to nonsmoking and smoking mothers.

mothers is  $\bar{x}_2 = 3257$ ,  $s_2 = 520$  g. Is the observed mean difference in weights of  $\bar{x}_1 - \bar{x}_2 = 215$  g easily explained by chance, or is there a real difference in the mean weights of North Carolina babies born to nonsmoking and smoking mothers in 2004? See Figure 8.1.

Let  $\mu_1$  and  $\mu_2$  denote the true mean weight of babies born to nonsmoking and smoking mothers, respectively. We consider the hypotheses

$$H_0: \mu_1 = \mu_2 \quad \text{versus} \quad H_A: \mu_1 > \mu_2.$$

As in Section 7.1.3, if we assume that the distribution of weights is normal for babies born to both nonsmoking and smoking mothers, then the statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

has approximately a  $t$ -distribution with degrees of freedom given either by Equation (7.7) (or the quick rule of  $(\min(n_1, n_2) - 1)$ ).

For these data, the statistic is

$$t = \frac{3471.912 - 3256.910}{\sqrt{478.5524^2/898 + 520.4788^2/111}} = 4.14 \quad \text{with}$$

$$v = \frac{\left( \frac{479^2}{898} + \frac{520^2}{111} \right)^2}{\left( \frac{479^2}{898} \right)^2/(898 - 1) + \left( \frac{520^2}{111} \right)^2/(111 - 1)} = 134.011$$

degrees of freedom (or 110 degrees of freedom). If the null hypothesis is true ( $\mu_1 - \mu_2 = 0$ ), then the chance of obtaining a statistic as extreme as 4.14 is  $P(t \geq 4.1411) = 0.00003$ , indicating the samples we obtained are rare – random chance alone would give a test statistic that large fewer than 3 out of 100 000 times. Thus, we conclude that babies born to mothers who smoke weigh, on average, less than babies born to mothers who do not smoke.

### R Note

The `t.test` function calculates a two sample  $t$  test.

```
> t.test(Weight ~ Smoker, data = NCBirths2004, alt = "greater")
...
t = 4.1411, df = 134.011, p-value = 3.04e-05
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
129.0090      Inf
sample estimates:
mean of x mean of y
3471.912   3256.910
```

The default is to test  $H_0: \mu_1 - \mu_2 = 0$ ; to specify a different value, for example,  $H_0: \mu_1 - \mu_2 = 50$ , add the argument `mu=50` to the `t.test` command.

From the R output, we can also see that with 95% confidence, babies born to mothers who do not smoke weigh at least 129 g more on average than babies born to mothers who smoke.  $\square$

### Two Sample $T$ -Test for Difference of Normal Means

Let  $X_1, X_2, \dots, X_{n_1} \sim N(\mu_1, \sigma_1^2)$  and  $Y_1, Y_2, \dots, Y_{n_2} \sim N(\mu_2, \sigma_2^2)$  be two independent normal random samples with sample means and standard deviations  $\bar{X}, S_1$  and  $\bar{Y}, S_2$ , respectively. To test

$$H_0: \mu_1 = \mu_2 \quad \text{versus} \quad H_A: \mu_1 \neq \mu_2,$$

we form the test statistic

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}.$$

If the null hypothesis is true, then  $T$  has approximately a  $t$  distribution with degrees of freedom given by Equation (7.7) (or the quick rule).

The  $P$ -value is the probability that chance alone would produce a test statistic as or more extreme than the observed value if the null hypothesis is true.

#### 8.3.1.1 Check Conditions

The conditions for a two-sample  $t$ -test are the same as for a two-sample  $t$  interval – two samples of independent, identically distributed observations, with independence between the samples, each normally distributed – and you can perform the same checks discussed in Section 7.1.2.1 with an additional check that the two samples are independent, in particular that they are not paired.

If the normality assumption is violated, in particular, if the distributions are skewed and the sample sizes are different, then the actual distribution of the  $t$ -statistic may be very different from the  $t$  distribution; in this case, a permutation test or bootstrap  $t$ -test is better.

In Example 3.4, we tested the equality of two means when both samples were strongly skewed, and the sizes were unbalanced: one sample had 1164 observations and the other only 23. Using a permutation test, we obtained a  $P$ -value of 0.0165, strongly suggesting that the mean repair times are not the same. Using the two sample  $t$ -test for the Verizon example results in a  $P$ -value of 0.02987, which is incorrect by a factor of roughly 1.8.

#### 8.3.1.2 Matched Pairs

The two-sample test requires that the data be from two independent populations. If the two samples are paired (not independent), then we can use the same technique that we used when constructing confidence intervals

(Section 7.1.4): take the differences between the two samples, turning the two sample problem with dependent data into a one sample problem. We then do a one-sample test of the differences, that is, an ordinary  $t$ -test or bootstrap  $t$  test. This is a *paired t test*.

### 8.3.1.3 Pooling the Variances\*

In Section 7.1.3.3, we mentioned a version of the confidence interval for the difference between two-sample means when we could assume that the population variances were equal. Similarly, for the two sample  $t$ -test for means, if we can assume the population variances are equal, then we can use the test statistic  $T = (\bar{X} - \bar{Y}) / (S_p \sqrt{1/n_1 + 1/n_2})$ , where  $S_p^2$  is the pooled sample variance. If the null hypothesis is true, then  $T$  follows a  $t$ -distribution with  $n_1 + n_2 - 2$  degrees of freedom.

We generally advise against pooling the variances, because it typically gains little in the ideal case that the variances are the same, but can be badly biased when they are not; see earlier remarks in Section 7.1.3.3.

## 8.3.2 Comparing Two Proportions

Here we focus on testing whether two proportions are the same. In this case, when the null hypothesis is true, then the two variances are the same, and we are able to pool. This leads to a simple closed-form test statistic which is much simpler than the confidence intervals for two proportions discussed in Section 7.4.2.

**Example 8.6** Nonresponse in surveys may lead to biased results and researchers often put great effort into reducing nonresponse. (Pejtersen, 2020) conducted a randomized experiment to see if a monetary incentive would improve the response rate for surveys sent by postal mail to vulnerable children and youth in Denmark. The participants aged 8–23 years came from families with severe social problems. Using random assignment, 143 participants received the survey with a supermarket voucher (worth 15 euros), while 119 participants received just the survey. Of those in the treatment group, 75.5% responded to the survey compared to 42.9% in the control group.

Let  $p_1$  and  $p_2$  denote the response proportions in the treatment and control group, respectively. We wish to test

$$H_0: p_1 = p_2 \quad \text{versus} \quad H_A: p_1 > p_2.$$

Thus, we need to find a test statistic based on  $\hat{p}_1$  and  $\hat{p}_2$ , the sample proportions, and an appropriate reference distribution.  $\square$

Before continuing with this example, recall that the sampling distribution of a proportion  $\hat{p}$  has mean  $p$  and variance  $p(1-p)/n$ , and is approximately normal if both  $np$  and  $n(1-p)$  are large. Hence, if  $\hat{p}_1$  and  $\hat{p}_2$  are independent,

then  $\hat{p}_1 - \hat{p}_2$  has mean  $p_1 - p_2$  and variance  $p_1(1 - p_1)/n_1 + p_2(1 - p_2)/n_2$ , and is approximately normal if all expected successes and failures are large. Standardizing gives

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{p_1(1 - p_1)/n_1 + p_2(1 - p_2)/n_2}}$$

and this has an approximate standard normal distribution for large samples.

Now, let us return to hypothesis testing. Under the null hypothesis,  $p_1 = p_2 = p$ . The test statistic becomes

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{p(1 - p)/n_1 + p(1 - p)/n_2}}.$$

This still involves the unknown parameter  $p$  that we need to estimate. Since we are assuming that the proportion of “successes” is the same in both populations, we pool the samples to estimate  $p$ . If  $X_1$  and  $X_2$  denote the number of “successes” in each sample, then a pooled estimate of  $p$  is

$$\hat{p}_p = \frac{X_1 + X_2}{n_1 + n_2}$$

and the standard error for  $\hat{p}_1 - \hat{p}_2$  is

$$\text{SE}_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_p(1 - \hat{p}_p)}{n_1} + \frac{\hat{p}_p(1 - \hat{p}_p)}{n_2}}. \quad (8.1)$$

Thus, if the null hypothesis  $p_1 = p_2$  is true, then

$$\begin{aligned} Z &= \frac{(\hat{p}_1 - \hat{p}_2) - E[\hat{p}_1 - \hat{p}_2]}{\text{SE}_{\hat{p}_1 - \hat{p}_2}} \\ &= \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_p(1 - \hat{p}_p)/n_1 + \hat{p}_p(1 - \hat{p}_p)/n_2}} \\ &= \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_p(1 - \hat{p}_p)(1/n_1 + 1/n_2)}} \end{aligned}$$

follows approximately a standard normal distribution.

### 8.3.2.1 Monetary Incentives in Surveys, Cont.

Going back to the experiment on monetary incentives in surveys, we have  $X_1 = 108$ ,  $n_1 = 143$  and  $X_2 = 51$ ,  $n_2 = 119$ , so the pooled estimate and standard

error are

$$\hat{p} = \frac{108 + 51}{143 + 119} = 0.6069 \quad \text{and}$$

$$\text{SE}_{\hat{p}_1 - \hat{p}_2} = \sqrt{0.6069(1 - 0.6069) \left( \frac{1}{143} + \frac{1}{119} \right)} = 0.0606.$$

Thus, the test statistic is  $z = (0.755 - 0.429)/0.6069 = 5.3795$  and the probability  $P(Z \geq 5.3795) = 3.7 \times 10^{-8}$ . If the survey response rate of youth who received the incentive is the same as the rate of youth who do not receive the incentive, then the chance of obtaining a test statistic as or more extreme than what we observed is almost 0. This study supports the hypothesis that monetary incentives improve the response rate of surveys sent to vulnerable children and youth.  $\square$

### Z Test for Difference of Two Proportions

Let  $X_1 \sim \text{Binom}(n_1, p_1)$  and  $X_2 \sim \text{Binom}(n_2, p_2)$  be independent. To test

$$H_0: p_1 = p_2 \quad \text{versus} \quad H_A: p_1 \neq p_2,$$

let  $\hat{p}_p = (X_1 + X_2)/(n_1 + n_2)$  and form the test statistic

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_p(1 - \hat{p}_p)(1/n_1 + 1/n_2)}}.$$

Then, under the null hypothesis,  $Z$  has an approximate standard normal distribution.

The  $P$ -value is the probability of obtaining a random test statistic as or more extreme than the observed  $Z$  if the null hypothesis is true.

This normal approximation tends to be reasonably accurate (more accurate than the normal approximation for a single proportion), though it is not accurate for some combinations of sample sizes, and for  $p$  near 0 or 1. You will investigate this in Exercise 8.25. For better accuracy we can use a continuity correction, adding or subtracting 0.5 from each success count in the direction that gives a larger  $P$ -value; there are messy details that are best handled by software. In Section 10.3.2, we introduce *Fisher's exact test* which is a more accurate test for comparing two proportions.

### R Note

The `prop.test` function performs the two sample proportion test. By default, a continuity correction is applied in the calculations.

```
> prop.test(c(108,51), c(143,119))
2-sample test for equality of proportions
with continuity correction

data: c(108, 51) out of c(143, 119)
X-squared = 27.699, df = 1, p-value = 1.417e-07
alternative hypothesis: two.sided
95 percent confidence interval:
 0.2055235 0.4478232
sample estimates:
 prop 1    prop 2
 0.7552448 0.4285714
```

We are 95% confident that the percentage of youth receiving a monetary incentive who respond to a survey is from 20.6% to 44.8% higher than the percentage of youth not receiving a monetary incentive.

In the `prop.test` output, the test statistic given is the  $\chi^2$  statistic of 27.699. The chi-square distribution with one degree of freedom is the same as the square of the standard normal distribution (Theorem B.14). Thus, squaring the  $z$  statistic computed in the example above yields  $5.3795^2 = 28.939$ , which is close to the R output labeled `X-squared`.

**Example 8.7** Do younger people tend to be more receptive to usage of gender-neutral pronouns? Another survey by The Pew Research Center for the People and the Press published in 2019 considered generational differences in social and political issues.<sup>1</sup> A random sample of teens and adults were surveyed on whether forms or online profiles asking for a person's gender should include options other than "man" or "woman." Of the 1178 Generation Z (born between 1997 and 2005), 695 responded yes, compared to 1337 out of 2674 Millennials (born between 1981 and 1996). Does this indicate that a higher proportion of Gen Z support inclusion of additional gender options than Millennials?

### Solution

Let  $p_1$  and  $p_2$  denote the proportion of Generation Z and Millennials, respectively, who support the inclusion of additional gender options on forms. We wish to test  $H_0: p_1 = p_2$  versus  $H_A: p_1 > p_2$ .

---

<sup>1</sup> <https://www.pewsocialtrends.org/2019/01/17/generation-z-looks-a-lot-like-millennials-on-key-social-and-political-issues/>.

The pooled estimate of  $p$  and the standard error are

$$\hat{p}_p = \frac{695 + 1337}{1178 + 2674} = 0.5275 \quad \text{and}$$

$$\text{SE}_{\hat{p}_1 - \hat{p}_2} = \sqrt{0.5275(1 - 0.5275) \left( \frac{1}{1178} + \frac{1}{2674} \right)} = 0.0175.$$

Thus, the test statistic is  $z = (0.59 - 0.50)/0.0175 = 5.1429$ . If the proportions are the same, then the one-sided  $P$ -value is  $P(Z \geq 5.1429) = 1.35 \times 10^{-7}$ . It is unlikely that the observed difference in proportions between the two generations is due to chance. Thus, we conclude that a higher proportion of Generation Z support the inclusion of additional gender options than Millennials.  $\square$

### 8.3.3 Matched Pairs for Proportions

The two-sample test of proportions requires that the data be from independent populations. If the data are paired, we need to do something different. As usual, we begin by taking differences, but now there is a twist.

**Example 8.8** A postdoc studying stress conducts a survey at her university and finds that of 250 first-year students, 55 of them said they drink at least two cups of coffee a day. The postdoc does a follow-up survey of the following year and finds that of these 250 now sophomores, 71 said they drink at least two cups of coffee a day. She wants to use the two sample proportions test to see if the difference is statistically discernible, but she decides to consult you first. What would you advise?

#### Solution

Since the postdoc is comparing the two proportions for the same sample of students, her data are paired, not independent. You must advise her that she cannot use a two-sample proportions test.

We proceed as with the paired  $t$  test and take differences. Here, we need to know how many students drank at least two cups of coffee in both their first and sophomore years. Suppose that 50 drank coffee both years: 5 quit, and 21 started.

		Sophomores	
		No	Yes
First-years	Coffee?	174	21
	Yes	5	50

So there are 5 students with a difference of  $-1$  (decreased coffee consumption), 224 with a difference of  $0$  (that is, no change in their behavior), and 21 with a difference of  $+1$  (increased coffee consumption).

Now there is a twist – we will ignore the people who did not change. They are not important for our analysis! That leaves us with 26 people, 21 of whom increased their consumption and 5 cut. Now, we just do a one-sample test of a proportion with  $H_0: p = 0.5$  and  $H_A: p \neq 0.5$ , where  $p$  is the proportion who increase their consumption. The exact two-sided  $P$ -value is  $2P(X \geq 21)$  when  $X \sim \text{Binom}(26, 0.5)$ , or 0.0025. An imbalance this large between those who increase and decrease consumption would rarely occur by chance.

If the postdoc had performed a two sample proportions test, she would have found a  $P$ -value of 0.1223 and would most likely have concluded that there is no difference in coffee consumption.

This test for paired proportions – taking differences, ignoring the cases that do not change, and then computing a one-sample proportions test of  $H_0: p = 0.5$  – is known as McNemar's test for paired proportions (see Agresti, 2012).  $\square$

**Example 8.9** Two economists conducted a study to determine if there was discrimination against homosexuals in the housing market (Ahmed and Hammarstedt, 2009). The researchers created two fictitious couples, one heterosexual and one homosexual (males), who apply for rental housing in an apartment advertised by landlords on Blocket.se [www.blocket.se](http://www.blocket.se), a large buy-and-sell website in Sweden. Each of the 408 landlords in the sample received an application from both couples. (The application letters had similar content.) One hundred and seventy-seven of the landlords contacted both couples (that is, responded to the email), 177 did not reply to either couple, 50 replied to just the heterosexual couple, and 4 replied to just the homosexual couple. Determine whether or not this supports the hypothesis that there is discrimination against homosexuals in the housing (rental) market in Sweden.

### Solution

Comparing the proportion of landlords who responded to the application from the heterosexual couple ( $227/408 = 0.556$ ) to the proportion of landlords who responded to the application from the homosexual couple ( $181/408 = 0.444$ ), we ask, is this difference statistically discernible?

		Heterosexual couple	
		Landlords	Responded
Homosexual couple	Responded	177	4
	Did not respond	50	177

Again, we are comparing two proportions for a matched pair since the application for each couple was similar. Putting the data in a table, we proceed as in the previous example: we ignore the counts for which the landlords behaved in the same manner (responded to both couples or did not respond to both couples). Instead, we look at the 54 cases when one couple received a response and the other did not.

Let  $p$  denote the proportion of landlords who responded only to the heterosexual couple so we are testing  $H_0: p = 0.5$  versus  $H_A: p \neq 0.5$ . If  $X \sim \text{Binom}(54, 0.5)$ , then the exact two-sided  $P$ -value is  $2P(X \geq 50) \approx 0$ .

We conclude that there is evidence of discrimination against homosexuals in the rental housing market in Sweden.  $\square$

## 8.4 Type I and Type II Errors

There are two possible errors in a hypothesis test: rejecting  $H_0$  when it is true or failing to reject it when it isn't.

**Definition 8.1** A *type I error* occurs if we reject the null hypothesis when it is true. This is also known as a false positive. The type I error rate (false positive rate) is the probability of rejecting  $H_0$  when it is true.

A *type II error* occurs if we do not reject the null hypothesis when the alternative is true. This is a false negative. The type II error rate (false negative rate) is the probability of failing to reject  $H_0$  when it is false.  $\parallel$

In classical hypothesis testing, we do not treat the hypotheses the same. We assume the null hypothesis is true unless the data provide strong evidence otherwise.

Similarly, in US criminal trials, a defendant is considered “innocent until proven guilty,” and the proof must be “beyond a reasonable doubt.” Innocence corresponds to  $H_0$ , and guilt to  $H_A$ .

Which error is more serious, convicting an innocent person (type I) or freeing a guilty person (type II)? Our justice system sidesteps this question; it holds that convicting an innocent person is bad, and the probability of a wrongful conviction must be small. The severity of a type II error does not really enter the picture. Neither does the probability of a false acquittal – we do not accept less evidence to convict when there is not much evidence either way (Table 8.1).

Similarly, in the classical approach to hypothesis testing, we do not adjust thresholds to balance the two kinds of errors, taking into account their relative size and severity. Instead, we set thresholds to limit the probability of a type I error to a prespecified value, say 5%.

**Table 8.1** Decisions by a jury in a murder trial.

		Truth	
		Innocent	Guilty
Jury decision	Guilty	Type I error	Correct
	Not guilty	Correct	Type II error

**Example 8.10** A pharmaceutical company is testing a new drug that it hopes will lower cholesterol levels in patients. They conduct a study to test  $H_0$ : the drug is not effective in lowering cholesterol levels versus  $H_A$ : the drug is effective in lowering cholesterol levels. Describe the type I and type II errors in this case and the practical consequences of making these errors.

### Solution

A type I error occurs if the researchers conclude that the drug is effective when in fact it is not. Practical consequences of this include endangering the health of patients who take the drug in the mistaken belief that it will lower their cholesterol, wasted costs, and potential liability claims filed against the company. A type II error occurs if the researchers conclude the drug is not effective when in fact it is. Practical consequences of this include the failure to save lives, or otherwise, improve patient's health, and lost revenue for the company.  $\square$

#### 8.4.1 Type I Errors

In the classical approach to hypothesis testing, we specify the type I error and set decision thresholds accordingly. Alternately, for a given threshold, we may calculate the type I error rate.

**Example 8.11** We return to math SAT scores for 2019 college-bound seniors (scores are distributed  $N(528, 117^2)$ ). Suppose local educators in a certain city wants to know how their students compare to the national average. The educators have also decided that if their students average much lower than the national average of 528 points, then they will request more money from the city council for new teaching reforms. The educators will obtain a random sample of scores and test  $H_0: \mu = 528$  versus  $H_A: \mu < 528$ , where  $\mu$  denotes the mean math SAT score in their city. If their sample size is 100 and they decide that a sample average of 516 or less is their criterion for making the funding request, what is the probability they make a type I error? Assume that the standard deviation of scores in the city is  $\sigma = 117$ .

**Solution**

$$\begin{aligned}
 P(\text{Type I error}) &= P(\text{Reject } H_0 \mid H_0 \text{ true}) \\
 &= P(\bar{X} \leq 516 \mid \mu = 528) \\
 &= P\left(\frac{\bar{X} - 516}{117/\sqrt{100}} \leq \frac{516 - 528}{117/\sqrt{100}}\right) \\
 &= P(Z \leq -1.0256) = 0.153.
 \end{aligned}$$

Thus, there is a 15.3% chance of the educators unnecessarily requesting funds from the city when, in fact, their student's performance is in line with the national student body.  $\square$

**Example 8.12** A company claims that only 3% of people who use their facial lotion develops an allergic reaction (rash). You are a bit suspicious of their claims since you think a higher proportion of people at your college are allergic to this lotion. You query a random sample of 50 people and ask them to try the lotion. If more than three people develop the rash, you will send a nasty e-mail to the company CEO. What is the probability that you make a type I error?

**Solution**

If  $p$  denotes the true proportion of people at your college who are susceptible to the rash, you want to test  $H_0: p = 0.03$  versus  $H_A: p > 0.03$ . Let  $X$  denote the number of people who develop the rash. If  $H_0$  is true, then  $X \sim \text{Binom}(50, 0.03)$ , and

$$\begin{aligned}
 P(\text{Type I error}) &= P(X \geq 4 \mid H_0 \text{ true}) \\
 &= P(X \geq 4 \mid X \sim \text{Binom}(50, 0.03)) \\
 &= \sum_{i=4}^{50} \binom{50}{i} (0.03)^i (1 - 0.03)^{50-i} \\
 &= 0.0627.
 \end{aligned}$$

Thus, there is a 6.3% chance of a type I error.  $\square$

So what is an acceptable probability for making a type I error? Obviously, it depends on the stakes: in the murder trial analogy, we would want the probability of sending an innocent person to prison (or worse) to be very small (preferably close to 0!) On the other hand, in the funding request example, perhaps a type I error probability of 5–10% would be acceptable.

In practice, there are many situations, where we do consider the severity of type II errors. For example, in the early days of the AIDS/HIV epidemic, the Food and Drug Administration (FDA) allowed the first lifesaving drug, AZT, to be approved before all the planned data was collected.

**Definition 8.2** In a hypothesis test, we specify a probability threshold in advance, often called  $\alpha$ . If the  $P$ -value is less than or equal to this threshold, we declare the outcome statistically discernible, otherwise, not discernible. This threshold is called the *alpha level* of the test, and is the maximum type I error rate that we allow for the test. ||

If the test statistic has a discrete distribution the actual type I error rate might be smaller than  $\alpha$ . For example if  $\alpha = 0.05$  and two successive values of the test statistic give  $P$ -values of 0.04 and 0.06, respectively, then the first value is statistically discernible, but the second is not, and the type I error rate is 0.04.

What should  $\alpha$  be? In 1925, Ronald Fisher, a British statistician, published an influential book, *Statistical Methods for Research Workers* (Fisher, 1925), in which he proposed an alpha level of 0.05. Since that time, 0.05 and to a lesser extent 0.10 and 0.01 have become common thresholds for declaring an outcome as statistically discernible (recall Definition 8.2).

**Example 8.13** In the funding example, suppose the educators decide that they want the probability of a type I error to be 10%. For what value of  $C$  would sample means  $\bar{X} \leq C$  result in rejecting the null hypothesis? (Keep  $n = 100$ ,  $\sigma = 117$ .)

### Solution

$$\begin{aligned} 0.10 &= P(\text{Reject } H_0 \mid H_0 \text{ true}) \\ &= P(\bar{X} \leq C \mid \mu = 515) \\ &= P\left(\frac{\bar{X} - 528}{117/\sqrt{100}} \leq \frac{C - 528}{117/\sqrt{100}}\right) \\ &= P\left(Z \leq \frac{C - 528}{11.7}\right). \end{aligned}$$

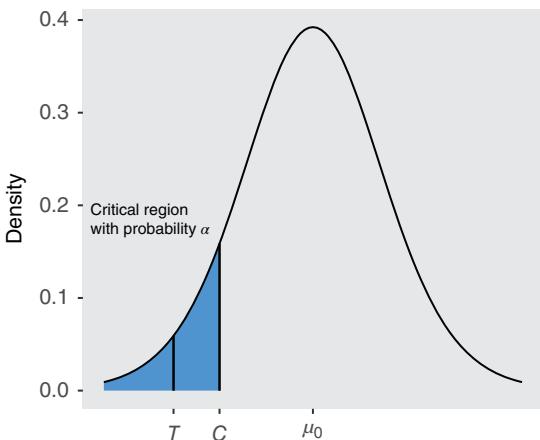
For the normal distribution, the 0.1 quantile is  $-1.2816$ . Thus, setting

$$-1.2816 = \frac{C - 528}{11.7}$$

and solving for  $C$  yields  $C = 513.01$ .

Thus, sample means of 513.01 or less would result in the educators concluding their students' score, on average, less than the national average in math SAT scores. □

**Figure 8.2** Critical region, critical value  $C$ , and a test statistic  $T$  for which  $P(T \leq C) \leq \alpha$  for a test where the alternative is  $\mu < \mu_0$ .



**Definition 8.3** Suppose we conduct a hypothesis test of  $H_0$  versus  $H_A$  at the  $\alpha$  alpha level. Let  $\mathcal{R}$  denote the set of all values of the test statistic for which we reject the null hypothesis. Then  $\mathcal{R}$  is called the *rejection region* or *critical region* and values at the endpoints (or boundary) of  $\mathcal{R}$  are called *critical values*. See Figure 8.2. ||

In the previous example,  $\mathcal{R} = (-\infty, 513.01]$  is the critical region and  $C = 513.01$  is the critical value.

**Example 8.14** Suppose you are losing badly at craps, so you begin to suspect that one die is loaded. You decide to test this theory by throwing the die 60 times and recording the number of times 1 appears. If you decide to use a 5% alpha level, for what number of occurrences of 1 would you conclude that the die is not fair?

### Solution

Let  $p$  denote the proportion of times the 1 appears. You test  $H_0: p = 1/6$  versus  $H_A: p \neq 1/6$ .

Let  $X$  denote the number of times 1 appears in 60 tosses of the die. If the die is fair, then  $X \sim \text{Binom}(60, 1/6)$  and you would not reject  $H_0$  if the number of times 1 appears is near 10. So the probability of a type I error is

$$\begin{aligned} 0.05 &= P(\text{Reject } H_0 \mid H_0 \text{ true}) \\ &= P(X \text{ not close to } 10 \mid X \sim \text{Binom}(60, 1/6)). \end{aligned}$$

Since the die is probably biased if  $X$  is too low or too high, let us set these tail probabilities to be  $0.05/2 = 0.025$ :

$$0.025 = P(X \leq C_1 \mid X \sim \text{Binom}(60, 1/6)) = \sum_{i=0}^{C_1} \binom{60}{i} (1/6)^i (5/6)^{60-i}$$

**Table 8.2** Cumulative probabilities for  $\text{Binom}(60, 1/6)$ .

$C_1$	0	1	2	3	4	5
$P(X \leq C_1)$	$1.77 \times 10^{-5}$	$2.307 \times 10^{-4}$	0.0015	0.0063	0.0202	0.0512
$C_2$	16	17	18	19	20	21
$P(X \geq C_2)$	0.0338	0.0164	0.0074	0.0031	0.0012	0.0005

and

$$0.025 = P(X \geq C_2 \mid X \sim \text{Binom}(60, 1/6)) = \sum_{i=C_2}^{60} \binom{60}{i} (1/6)^i (5/6)^{60-i}.$$

The critical region is  $\mathcal{R} = \{0, 1, 2, 3, 4\} \cup \{17, 18, \dots, 60\}$ , and  $P(X \leq 4) + P(X \geq 17) = 0.0366$  (see Table 8.2).  $\square$

In this example, by the discrete nature of  $X$ , we cannot get an exact type I error probability, so we are conservative, choosing critical values so that the actual type I error rate is at most equal to the nominal alpha level. Furthermore, for a two-sided test, the presumption is that both one-sided type I error rates are  $\leq \alpha/2$ . In general, we accept the null hypothesis unless the data provide a strong signal otherwise, and here a value of  $C_1 = 5$  or  $C_2 = 16$  would not provide that.

Some authors would allow the two critical values to be adjusted up or down to get the sum of the one-sided type I error rates as close to  $\alpha$  as possible. We feel this is misleading – if one of the one-sided type I error rates is greater than  $\alpha/2$ , then it is easier to reject on that side than is implied by the nominal alpha level. It opens the door for outright abuse, where one looks at the data before deciding which way to adjust the critical values.

**Example 8.15** An analyst draws a random sample of size 8,  $X_1, X_2, \dots, X_8$ , from a distribution with pdf  $f(x; \theta) = (\theta + 1)x^\theta$ ,  $0 \leq x \leq 1$ ,  $\theta > 0$ . She wants to test  $H_0: \theta = 2$  versus  $H_A: \theta > 2$ . As a decision rule, she records  $Y$ , the number of observations greater than or equal to 0.88. She rejects  $H_0$  if  $Y \geq 5$ . What is the probability of a type I error?

### Solution

She wants

$$P(\text{Reject } H_0 \mid H_0 \text{ true}) = P(Y \geq 5 \mid \theta = 2).$$

If  $\theta = 2$  is true, then the pdf is  $f(x; 2) = 3x^2$ . First, we compute the probability of any observation drawn from this distribution to be greater than 0.88:

$$p = P(X_i \geq 0.88 \mid \theta = 2) = \int_{0.88}^1 3x^2 \, dx = 0.3185.$$

Thus, the number of observations greater than or equal to 0.88 is  $Y \sim \text{Binom}(8, 0.3185)$ . We then compute

$$P(Y \geq 5 \mid \theta = 2) = \sum_{k=5}^8 \binom{8}{k} (0.3185)^k (1 - 0.3185)^{8-k} = 0.0736.$$

### R Note

```
sum(dbinom(5:8, 8, 0.3185)) # alternatively,
1 - pbinom(4, 8, 0.3185)
```

□

## 8.4.2 Type II Errors and Power

We have seen that we can control the probability of type I errors by setting the alpha level  $\alpha$ . We now consider type II errors. Let  $\beta$  denote the probability of a type II error:  $\beta = P(\text{Do not reject } H_0 \mid H_A \text{ true})$ . The complement of this,  $1 - \beta = P(\text{Reject } H_0 \mid H_A \text{ true})$ , is known as *power*.

**Definition 8.4** *Power* is the probability of correctly rejecting a false null hypothesis.

$$1 - \beta = P(\text{Reject } H_0 \mid H_A \text{ true}).$$

||

High power is desirable. You may read things like “the study was under-powered”; that usually means that not enough data were collected for the study to have a reasonable chance of detecting an effect. Particularly for underpowered studies, failure to reject  $H_0$  doesn’t mean that  $H_0$  is true; there just was not enough data collected to know.

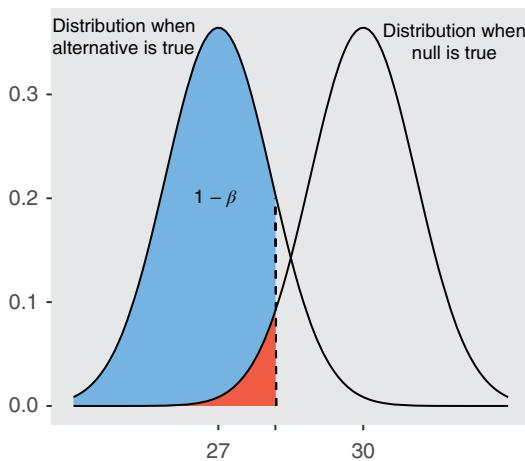
**Example 8.16** Suppose the heights of 2-year-old girls are normally distributed with a mean of 30 in. and a standard deviation of 6 in. Let  $\mu$  denote the true mean heights of these girls. A researcher wishes to test

$$H_0: \mu = 30 \quad \text{versus} \quad H_A: \mu < 30.$$

She plans to obtain a random sample of 30 girls and measure their heights, using  $\alpha = 0.05$  for an alpha level. What is the probability of correctly rejecting the null hypothesis if the mean height of 2-year-old girls in this community is actually 27 in.?

### Solution

If the null hypothesis holds, then the sampling distribution of  $\bar{X}$  is normal with mean 30 and standard error  $6/\sqrt{30} = 1.095$ . Using a one-sided test at



**Figure 8.3** Distributions of the test statistic under the null and alternative hypotheses.

$\alpha = 0.05$ , she rejects the null hypothesis if the  $z$ -score of her test statistic is  $Z \leq -1.645$  (the 0.05 quantile of the standard normal). This corresponds to  $Z = (\bar{X} - 30)/1.095 \leq -1.645$  or  $\bar{X} \leq 28.1987$ . In other words, the critical region is  $\mathcal{R} = (-\infty, 28.1987]$  and the critical value is  $C = 28.1987$ . See Figure 8.3.

So, if the true mean height is 27 in., what is the probability of correctly rejecting the null hypothesis of  $\mu = 30$ ?

$$\begin{aligned} 1 - \beta &= P(\text{Reject } H_0 \mid H_A \text{ true}) \\ &= P(\bar{X} \leq 28.1987 \mid \mu = 27) \\ &= P\left(\frac{\bar{X} - 27}{1.095} \leq \frac{28.1987 - 27}{1.095}\right) \\ &= P(Z \leq 1.0947) \\ &= 0.8632. \end{aligned}$$

Thus, she has about an 86% chance of correctly rejecting the null hypothesis if indeed the true mean height of 2-year-old girls is 27 in.  $\square$

**Example 8.17** A team of researchers plans a study to see if a certain drug can increase the speed at which mice move through a maze. An average decrease of 2 s through the maze would be considered effective, so the researchers would like to have a good chance of detecting a change this large or larger. Would 20 mice be a large enough sample? Assume the standard deviation is  $\sigma = 3$  s and that the researchers will use an alpha level of  $\alpha = 0.05$ .

### Solution

Let  $\mu$  denote the true mean decrease in time through the maze. Then the researchers are testing  $H_0: \mu = 0$  versus  $H_A: \mu > 0$ . If the null hypothesis holds, then the sampling distribution of  $\bar{X}$  is normal with mean 0 and standard error  $3/\sqrt{20} = 0.6708$ . Using a one-sided test at  $\alpha = 0.05$ , the researchers will reject the null hypothesis if the z-score of the test statistic satisfies  $Z \geq 1.645$ . This corresponds to  $Z = (\bar{X} - 0)/0.6708 \geq 1.645$  or  $\bar{X} \geq 1.1035$ . So, if the true mean decrease in time is 2 s, what is the probability of correctly rejecting the null hypothesis of  $\mu = 0$ ?

$$\begin{aligned} 1 - \beta &= P(\text{Reject } H_0 \mid H_A \text{ true}) \\ &= P(\bar{X} \geq 1.1035 \mid \mu = 2) \\ &= P\left(\frac{\bar{X} - 2}{0.6708} \geq \frac{1.1035 - 2}{0.6708}\right) \\ &= P(Z \geq -1.3365) \\ &= 0.9093. \end{aligned}$$

Thus, the researchers have a 91% chance of correctly concluding the drug is effective if the true average decrease in time is 2 s.  $\square$

In many instances, researchers decide on the power they want in a test and then determine the sample size needed to obtain that power.

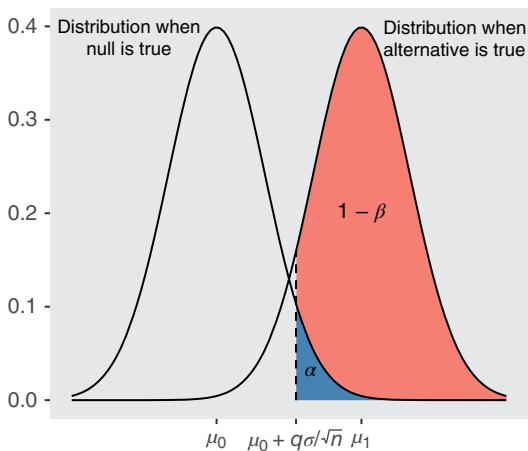
**Example 8.18** Suppose the researchers in the previous example want a 95% chance of rejecting  $H_0: \mu = 0$  at  $\alpha = 0.01$  if the true change is a 1.5 s decrease in time. What is the smallest number of mice that should be included in the study?

### Solution

On the standard normal curve,  $q = 2.3264$  is the cutoff value for the upper 0.01 tail (i.e. the 0.99 quantile). Thus, we need  $(\bar{X} - 0)/(3/\sqrt{n}) \geq 2.3264$ , or  $\bar{X} \geq 6.9792/\sqrt{n}$ .

Thus,

$$\begin{aligned} 0.95 &= P\left(\frac{\bar{X} - 0}{3/\sqrt{n}} \geq \frac{6.9792}{\sqrt{n}} \mid \mu = 1.5\right) \\ &= P\left(\frac{\bar{X} - 1.5}{3/\sqrt{n}} \geq \frac{6.9792/\sqrt{n} - 1.5}{3/\sqrt{n}}\right) \\ &= P\left(Z \geq 2.3264 - \frac{1.5}{3/\sqrt{n}}\right). \end{aligned}$$



**Figure 8.4** Distributions under the null and alternative hypotheses. Shaded regions represent power ( $1 - \beta$ ) and alpha level ( $\alpha$ ). Moving the critical value ( $\mu_0 + q\sigma/\sqrt{n}$ ) to the left increases power.

Using the 0.05 quantile for the standard normal,

$$-1.645 = 2.3264 - \frac{1.5}{3/\sqrt{n}}.$$

Thus,  $n = 64$  is the smallest number of mice that the researchers should use.  $\square$

More generally, suppose we test  $H_0: \mu = \mu_0$  versus  $H_A: \mu > \mu_0$ , where  $\sigma$  is known, at the  $\alpha$  alpha level. What is the power if the alternative is  $\mu = \mu_1$  (Figure 8.4)?

If  $H_0$  is true, then the sampling distribution of  $\bar{X}$  is normal with mean  $\mu_0$  and standard error  $\sigma/\sqrt{n}$ . Let  $q$  denote the  $1 - \alpha$  quantile for the standard normal. The corresponding critical value  $C$  for the sampling distribution of  $\bar{X}$  is found by

$$\begin{aligned} \frac{C - \mu_0}{\sigma/\sqrt{n}} &= q, \\ C - \mu_0 &= q \frac{\sigma}{\sqrt{n}}, \\ C &= \mu_0 + q \frac{\sigma}{\sqrt{n}}. \end{aligned}$$

Thus,

$$1 - \beta = P(\text{Reject } H_0 \mid H_A \text{ true})$$

$$\begin{aligned}
&= P\left(\bar{X} > \mu_0 + q \frac{\sigma}{\sqrt{n}} \mid \mu = \mu_1\right) \\
&= P\left(\frac{\bar{X} - \mu_1}{\sigma/\sqrt{n}} > \frac{\mu_0 - \mu_1 + q(\sigma/\sqrt{n})}{\sigma/\sqrt{n}}\right) \\
&= P\left(Z > \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} + q\right) \\
&= P\left(Z > q - \frac{(\mu_1 - \mu_0)}{\sigma/\sqrt{n}}\right). \tag{8.2}
\end{aligned}$$

Thus, we see that the power of a test,  $1 - \beta$ , is determined by the size of the lower bound of  $Z$  in Equation (8.2),

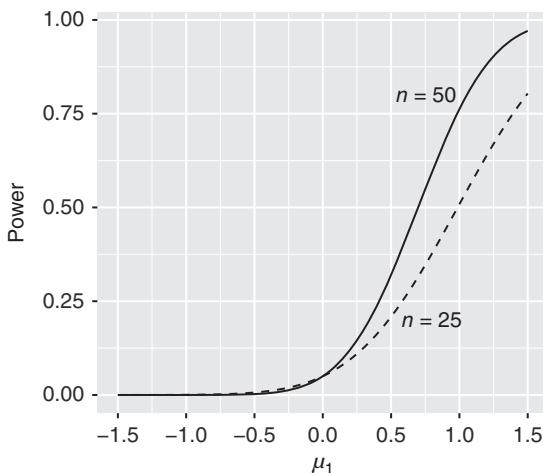
$$Z > q - \frac{(\mu_1 - \mu_0)}{\sigma/\sqrt{n}}.$$

The smaller that lower bound, the larger the power. The population standard deviation  $\sigma$  is not controllable by the analyst. The other factors that determine the power are the following:

- *Effect size*: The difference between the hypothesized mean  $\mu_0$  and the actual mean  $\mu_1$ . The larger the difference, the more likely we would detect the difference.
- *Denominator  $\sigma/\sqrt{n}$* : The larger the sample size, the smaller the denominator, and hence, the larger the amount being subtracted. This too increases power.
- *Alpha level  $\alpha$* : Increasing  $\alpha$  decreases the quantile  $q$  and pushes that lower bound to the left, increasing power.

Figure 8.5 displays two plots of power  $1 - \beta$  against different alternatives  $\mu_1$  for a hypothesis test  $H_0: \mu = 0$  versus  $H_A: \mu > 0$ . Note that in both the  $n = 50$  and the  $n = 25$  cases, as the effect size increases (i.e. as the alternative value increases), the power increases. Also, for a fixed alternative value, the power is higher for the  $n = 50$  sample.

In conducting a hypothesis test, a researcher will want to minimize all errors. We can set the type I error by declaring the alpha level before collecting the data. Minimizing the probability of a type II error ( $\beta$ ) means maximizing power. But as we can see from Figure 8.5, for a fixed sample size and standard deviation, increasing power results in increasing  $\alpha$ , the probability of a type I error. Thus, the only way to simultaneously decrease the probabilities of a type I and type II error is to increase the sample size.



**Figure 8.5** Plot of power against different alternatives  $\mu_1$  for  $n = 50$  and  $n = 25$  when hypothesized  $\mu = 0$ .

Similar calculations as above can be done for the alternatives  $H_A: \mu < \mu_0$  or  $H_A: \mu \neq \mu_0$  (see Exercises 8.41 and 8.42).

### Remark

- For many studies, power is set to at least 80%.
- For hypothesis tests of a mean where the population standard deviation is not known, the sampling distribution of the test statistic  $T = (\bar{X} - \mu_0)/(S/\sqrt{n})$  is a *t* distribution (approximately, or exactly if the population is normal). If the alternative hypothesis is true,  $T$  no longer has a *t* distribution, but has a *noncentral t distribution* (see Exercise 8.45).
- In general, power calculations are more difficult for hypothesis tests that involve parameters other than the mean, and researchers often use specialized statistical software to do these calculations.

||

### 8.4.3 P-Values Versus Critical Regions

In Chapter 3 and in most of this chapter, we conducted tests by computing a test statistic from the data and then finding the probability of obtaining a test statistic as extreme as the one observed, given that the null hypothesis is true. We used this probability, the *P*-value, to decide how likely it was that chance variation could explain our results. Small *P*-values supported the alternative hypothesis.

Another approach is to set the type I error and then determine a corresponding critical region (Definition 8.3). Typically, we specify a range of values of the test statistic for which we would reject  $H_0$  at a specified alpha level. We can compute the critical region without any data.

Conversely, computing a  $P$ -value requires having the data; it also does not require us to specify the alpha level. Thus, in reporting a  $P$ -value, we allow the reader to determine whether the outcome is statistically discernible.

For instance, consider Example 8.2. For a one-sided test with  $\alpha = 0.05$ , to find the critical region, we set

$$\begin{aligned} 0.05 &= P(\text{Reject } H_0 \mid H_0 \text{ true}) \\ &= P(\bar{X} \leq C \mid \mu = 7) \\ &= P\left(\frac{\bar{X} - 7}{0.8/\sqrt{15}} \leq \frac{C - 7}{0.8/\sqrt{15}}\right) \\ &= P\left(T \leq \frac{C - 7}{0.8/\sqrt{15}}\right). \end{aligned}$$

For the  $t$  distribution with 14 degrees of freedom, the 0.05 quantile is  $-1.76131$ . Thus, setting

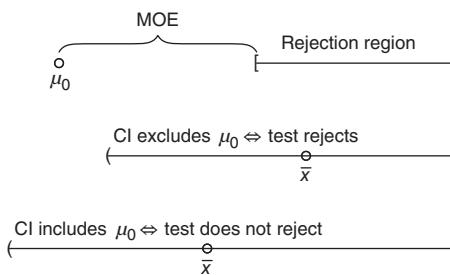
$$-1.7613 = \frac{C - 7}{0.8/\sqrt{15}}$$

yields  $C = 6.6362$ . Thus, a sample mean less than 6.6362 would result in rejecting the null hypothesis.

The sample mean for the coffee example was 6.6, which is (barely) in the critical region of  $(-\infty, 6.6362]$ , so we would reject the null hypothesis.

#### 8.4.4 Relationship Between Confidence Intervals and Hypothesis Tests

There is a close relationship between confidence intervals and hypothesis tests. They measure the same thing – the consistency of the data with parameter values. Typically, a test rejects  $H_0: \theta = \theta_0$  if and only if the confidence interval excludes  $\theta_0$ . Figure 8.6 shows this relationship.



**Figure 8.6** Correspondence between hypothesis tests and confidence intervals. The top shows a one-sided hypothesis test – the null hypothesis, margin of error (MOE), critical value, and rejection region  $[\mu_0 + \text{MOE}, \infty)$ . The middle and bottom show confidence intervals of the form  $(\bar{x} - \text{MOE}, \infty)$  for two values of  $\bar{x}$ . The interval excludes  $\mu_0$  if and only if the test rejects the null hypothesis.

We demonstrate this for the one-sample mean from a normal distribution. Let  $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$  where  $\mu$  is unknown and  $\sigma^2$  is known. Suppose we are testing

$$H_0: \mu = \mu_0 \quad \text{versus} \quad H_A: \mu \neq \mu_0,$$

using an alpha level of  $\alpha$ . Using the  $z$ -score as the test statistic

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}},$$

we find the critical region to be the set of  $(X_1, \dots, X_n)$  for which  $\bar{X}$  satisfies:

$$\left| \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right| \geq q_{1-\alpha/2},$$

where  $q_{1-\alpha/2} = q$  denotes the  $1 - \alpha/2$  quantile of the standard normal distribution.

The complement of this critical region is  $\bar{X} - q \frac{\sigma}{\sqrt{n}} < \mu_0 < \bar{X} + q \frac{\sigma}{\sqrt{n}}$ , called the *acceptance region* for the test. In particular, if the hypothesized value  $\mu_0$  for the parameter satisfies the above, then we would not have sufficient evidence to support the alternative hypothesis: that is, we “accept” the null hypothesis.

On the other hand, we have seen (Equation (7.2)) that this interval is also the  $(1 - \alpha/2) \times 100\%$  confidence interval for  $\mu$ .

This relationship between a hypothesis test at an  $\alpha$  alpha level and the  $(1 - \alpha) \times 100\%$  confidence interval holds more generally. That is, in a test of the parameter  $\theta$ , say  $H_0: \theta = \theta_0$  versus  $H_A: \theta \neq \theta_0$ , the  $(1 - \alpha/2) \times 100\%$  confidence interval is the set of values for which we do not reject the null hypothesis. Equivalently, if the confidence interval does not contain  $\theta_0$ , then we would reject the null hypothesis. See Casella and Berger (2001) for the general case.

**Example 8.19** In Example 8.5, we tested the hypothesis  $H_0: \mu_1 - \mu_2 = 0$  versus  $H_A: \mu_1 - \mu_2 > 0$ , where  $\mu_1, \mu_2$  denote the true mean weights of babies born to nonsmoking and smoking mothers, respectively. From the R output of the `t.ttest` command (see the **R Note** for this example), we see that the null hypothesis would be rejected at a  $\alpha = 0.05$  alpha level. The one-sided 95% confidence interval is  $(129.00, \infty)$ . The hypothesized value of the parameter 0 is not contained in this interval.  $\square$

We claim that this relationship – a confidence interval is the set of parameter values for which a hypothesis test doesn’t reject – holds more generally. There are two meanings to that. First, it holds for many of the confidence interval and test procedures we have discussed, including one and two-sample  $t$  or  $z$  tests/confidence intervals (CIs) for means, bootstrap  $t$  tests/CIs,  $t$  with bootstrap SE tests/CIs. It doesn’t quite hold for proportions methods defined earlier.

That's a bit problematic – it means we could get incompatible results, say a test rejecting  $\theta_0$  but a confidence interval including it.

Second, we could define one in terms of the other. Given a confidence interval procedure (that works for any level of  $\alpha$ ), we could (a) perform a hypotheses test at a fixed alpha level by rejecting if the  $1 - \alpha$  confidence interval excludes  $\theta_0$  or (b) calculate a  $P$ -value as the smallest  $\alpha$  for which a  $1 - \alpha$  confidence interval excludes  $\theta_0$ . Conversely, given a test procedure (that works for any parameter value), we can define a  $(1 - \alpha)$  confidence interval as the set of parameter values for which the test does not reject.

For example, we can turn the bootstrap percentile interval into a hypothesis test: Reject if the interval excludes  $\theta_0$ . The one-sided  $P$ -value is the fraction of the bootstrap statistics that are as or more extreme than the observed value.

We can also create compatible versions of proportions tests and intervals. In particular, we can create an exact confidence interval for a single binomial proportion based on the exact test for a binomial proportion, by setting  $\alpha/2$  equal to each one-sided probability and solving for  $p$ :

### Exact Confidence Interval for a Single Proportion

Let  $X$  be a binomial random variable,  $X \sim \text{Binom}(n, p)$ . Suppose  $X = x$  is observed. The endpoints of an exact  $(1 - \alpha)$  confidence interval  $(L, U)$  are obtained by solving

$$\alpha/2 = P(X \geq x | L = p) = \sum_{i=x}^n \binom{n}{i} p^i (1-p)^{n-i}$$

and

$$\alpha/2 = P(X \geq x | U = p) = \sum_{i=0}^x \binom{n}{i} p^i (1-p)^{n-i}$$

for  $p$ .

### R Note

The `R binom.test` function performs exact tests and confidence intervals for a single proportion.

```
> binom.test(7, 21, 0.5)
Exact binomial test
...
95 percent confidence interval:
 0.1458769 0.5696755
...
> pbinom(7, 21, 0.5696755) # Verify
```

```
[1] 0.02499999  
> 1 - pbinom(6, 21, 0.1458769)  
[1] 0.02499996
```

Unfortunately, we cannot turn a two-sample permutation test into a confidence interval, because that test only works for the null hypothesis that  $\theta_1 - \theta_2 = 0$ , not other values. Similarly, we also cannot do this for the two-sample test of proportions in Section 8.3 that uses a pooled proportion estimate.

## 8.5 Interpreting Test Results

Many fields such as psychology, health, and the natural sciences are going through what is being called a *replication* or *reproducibility* crisis: the results of a published study cannot be replicated independently by another researcher using the same or comparable study methods. For example, The Reproducibility Project: Psychology,<sup>2</sup> a 4-year open project started in 2011, attempted to replicate the findings of 100 research papers published in three major psychology journals during 2008. The researchers on this project were able to replicate less than half of the results, especially for studies with marginal  $P$ -values. Only 24% (8 of 34) studies with  $0.02 < P < 0.05$  achieved  $P < 0.05$  in the replication, compared to 63% (20 of 32) with  $P < 0.001$ . Although other factors such as study design and context (the time and location of the study compared to where and when the study was replicated) might explain some of the differences in outcome, much of the attention on the replication crisis, and on studies in general, has focused on the use and abuse of  $P$ -values and hypothesis testing. The American Statistical Association, the largest professional organization of statisticians, felt compelled to issue a statement on this topic (Wasserstein and Lazar, 2016). One journal, *Basic and Applied Social Psychology*, has even banned  $P$ -values and inferential statistics from articles, instead favoring descriptive statistics (Trafimow and Marks, 2015).

In this section, we will look at some of the issues that have raised concerns, including the temptation to focus on  $P$ -values rather than practical significance, terminology that exacerbates that temptation, arbitrary cutoffs, over-interpretation of negative results, and inflated type I error rates due to multiple testing.

---

<sup>2</sup> <https://osf.io/ezcuj/wiki/home/>.

### 8.5.1 Terminology

Statisticians have used the terminology “statistically significant” to indicate an outcome unlikely to be due to chance going as far back as 1885 (Edgeworth, 1885) and 1925 (Fisher, 1925). Unfortunately, it is easy to misinterpret a result that is “statistically significant”—does this indicate a difference that is significant in the ordinary sense or large enough to matter in practice?

Furthermore, it is common to abbreviate “statistically significant” to “significant,” which could mean anything. For instance, an earlier edition of this book reported in Example 1.9: “those in the tai chi group reported a significant decrease in knee pain”—it was unclear whether this meant statistical or practical significance (the former was intended).

We use “statistically discernible” in place of “statistically significant” for outcomes that are unlikely to be due to chance; “alpha level” in place of “significance level” for the maximum type I error in a test (Definition 3.4); and “hypothesis tests” in place of “significance tests.” However, you will still see the use of “significant” in the statistics literature. Data scientists who work at Instacart use an abbreviation “StatSig”; this at least avoids the temptation to shorten “statistically significant” to “significant.” See Witmer (2019) and Wasserstein et al. (2019) for more on this issue.

### 8.5.2 Arbitrary Thresholds

As we noted in Section 8.4.1, an alpha level of 0.05 is a common threshold used to determine statistical discernibility. Fisher’s rationale was “... $P = .05$ , or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not” (Fisher, 1925). Since that time, 0.05 has taken on an almost mythical prominence, where the quest for a  $P$ -value less than 0.05 overshadows other issues such as study design or appropriateness of a test. However, there is no real scientific reason why deviations greater than 2 standard deviations as opposed to, say, 1.5 or 3 standard deviations, should be considered statistically discernible.

The  $P$  value is a probability measuring the likelihood of observing an outcome *given that the null hypothesis is true*. As such, with an alpha level of  $\alpha = 0.05$ , calculating a  $P$ -value of 0.049 does not mean that the null hypothesis is false nor does a  $P$ -value of 0.051 mean the null hypothesis is true. The context of the question as well as the real consequences of making a wrong decision should guide a researcher in determining how to proceed.

### 8.5.3 Statistical Discernibility Versus Practical Importance

Researchers designing a study typically have a theory that they are trying to confirm: Does this new drug reduce heart attacks? Will this new teaching

method improve test scores? Did the percentage of adults who have graduated from college change from 10 years ago? Thus, a statistically discernible result – a small  $P$ -value – that supports the researchers theory is a desired outcome: the efforts they put into their study have paid off. However, a small  $P$ -value that indicates that their result is not likely to be due to chance does not necessarily indicate that their result is of practical importance.

**Example 8.20** A drug company is researching a new drug to help severely obese patients lose weight. They conduct a clinical trial to test  $H_0: \mu = 0$  versus  $H_A: \mu > 0$ , where  $\mu$  is the mean weight lost. Suppose after 1 year, the researchers measure an average weight loss of 4 lb with standard deviation of 3 lb for the 50 patients in the study who took the drug. The test statistic is  $t = 4/(3/\sqrt{50}) = 9.4281$  so comparing this with a  $t$  distribution with 49 degrees of freedom, we find a  $P$ -value of  $6.85 \times 10^{-13}$ , essentially 0! Thus, there is strong evidence that obese patients will lose weight on this drug. On the other hand, a 95% confidence interval shows that, on average, the weight loss will be between 3.1 and 4.9 lb.

The results are statistically discernible – the positive weight lost by the patients is unlikely to be due to chance, so the drug does work – but it seems unlikely that the company will be able to successfully market a drug that promises severely obese patients that they can expect to lose, on average, at most 5 lb! □

In practice, the null hypothesis is almost never exactly true, and with enough data, even very small, practically unimportant differences are statistically discernible. A confidence interval for the parameter of interest is often more informative than the  $P$ -value and should be reported along with the  $P$ -value of the hypothesis test.

### 8.5.4 Negative Results

**Lack of Discernibility** In a criminal trial, the jury may decide on a “not guilty” verdict. Does this verdict mean the defendant is innocent? Not necessarily! In many instances, the prosecutor failed to present a strong enough case to convict the defendant, so a “not guilty” decision is not the same as a declaration of innocence!

The same is true in hypothesis testing. If you conduct a hypothesis test at a specified alpha level and do not reject the null hypothesis, this does not necessarily mean that the null hypothesis is true. It could be true, or it could be that you just failed to gather enough evidence to support the alternative hypothesis. For instance, the power of your test may have been too low: perhaps your sample size was too small for the effect size you were trying to detect.

## 8.5.5 Inflated False Positive Rate

Actual false positive rates are often much larger than nominal values. Partly this is due to using inaccurate tests, e.g.  $t$ -tests for skewed data. Another factor is multiple testing, including data snooping.

### 8.5.5.1 Data Snooping

As we noted earlier, researchers often have a vested interest in finding a statistically discernible result in their studies. However, a finding in which one does not reject the null hypothesis does not mean that the researcher should perform more tests on their data. Data dredging, data snooping, or “p-hacking” – that is, cherry-picking data to get a statistically discernible outcome – can result in spurious or biased results.

For example, consider a pharmaceutical company testing a new drug for efficacy. The researchers may find that the benefit of the drug is not statistically discernible in their test as a whole. But, if they then look at enough subpopulations, say smokers, elderly, teenagers, farmers, Hispanic children, and so on, they have a high probability of finding some subpopulation in which the benefits of the drug are statistically discernible, purely by chance. The comic xkcd has a humorous take on this issue.<sup>3</sup>

Even if the null hypothesis is true, with a 0.05 alpha level, there is a 5% chance of incorrectly rejecting  $H_0$ . Suppose for a given study, all null hypotheses are true. If we run two independent tests, each at the 5% level, there is nearly a 10% chance of getting at least one positive result (“reject the null hypothesis”): the probability of no positive result is  $0.95^2 = 0.9025$ , so the probability of at least one positive result is  $1 - 0.9025$ . Then for 20 independent tests, the probability of not rejecting any null hypotheses is  $0.95^{20} = 0.3585$ , so the chance of rejecting the null hypothesis at least once is  $1 - 0.3585 = 0.6415$  – in other words, there is a 64% chance that we will declare something statistically discernible even though there is nothing going on.

A common issue in tech (Google, Instacart, and others) is multiple testing over time. There is a natural tendency to monitor an experiment, and stop if a hypothesis test yields a discernible result. But doing tests multiple times inflates the type I error rate.

### 8.5.5.2 Adjustments for Multiple Testing

One way to adjust for such *multiple testing* is by using a *Sidak correction*. Suppose we wish to run  $k$  tests on a data set. Let  $\alpha^*$  denote the nominal type I error for each of these tests. Let  $\alpha$  denote the desired overall Type I rate (the error in conducting all  $k$  tests). If each null hypothesis is true, and the tests are independent, then the probability of no rejections of the null hypothesis is

---

<sup>3</sup> <https://xkcd.com/882>.

$(1 - \alpha^*)^k$ , so that the probability of at least one rejection of the null hypothesis is  $1 - (1 - \alpha^*)^k$ . Thus, to obtain an overall type I error rate of  $\alpha$ , we set

$$\alpha = 1 - (1 - \alpha^*)^k$$

and solve for  $\alpha^*$ , yielding

$$\alpha^* = 1 - (1 - \alpha)^{1/k}.$$

For example, if we conduct five hypothesis tests and we wish to keep the overall Type I error at 0.05, then we should set the alpha level for each individual test at  $\alpha^* = 1 - (1 - 0.05)^{1/5} = 0.01021$ . Note that  $\alpha^* \approx 0.01 = \alpha/5$ . A quick approximation to the Sidak correction is the *Bonferroni correction*, which stipulates that if we want the overall type I error to be  $\alpha$  and we run  $k$  tests, then the type I error for each of these tests should be set to  $\alpha/k$ .

The Bonferroni correction is a bit conservative – that is, a smaller  $P$ -value (stronger evidence) is required to reject a null hypothesis. It can also be used when the tests are not independent, as the overall type I error rate does not exceed  $\alpha$ , even if the rejections are mutually exclusive. In practice, mutually exclusive tests are rare, while positively correlated tests are common; for example, tests for the effectiveness of a drug for the population as a whole, and for any subpopulation, are positively correlated because they share some of the same data. In this case, Bonferroni is even more conservative than necessary. However, in practice, that may not matter.

**Example 8.21** *Google Analytics Content Experiments*<sup>TM</sup> allows website owners to experiment with different versions of their website. The website owner wants site visitors to do something, for example, buy something, download software, donate, or send a letter to Congress. She may create different versions of the home page, with different text and/or images. She adds some JavaScript provided by Google to the default page; that JavaScript randomly leaves visitors on the default page, or sends them to one of the alternate versions. Then more Google JavaScript records whether visitors *convert* (take the desired action). Google reports the results to the website owner.

Google tested an earlier version of GACE itself and got 30% more downloads of *Picasa*<sup>TM</sup> (a precursor of Google Photos<sup>TM</sup>) using a page that (i) did not include a screen shot, (ii) had a “Try Picasa Now” button rather than “Free Download,” and other minor changes.

Let  $n_0, n_1, \dots, n_k$  be the number of visitors allocated to the default page and  $k$  alternate pages, and  $X_0, X_1, \dots, X_k$  be the number of conversions on each page. Let  $p_j$  denote the true conversion rate for the  $j$ th arm of this trial, and  $\hat{p}_j = X_j/n_j$  the estimated proportion. Google reports when an alternate page is discernibly better than the default page. This corresponds to running  $k$  hypothesis tests, with  $H_0: p_j = p_0$  versus  $H_A: p_j > p_0$ .

The tests are positively correlated because they share some of the same data; if  $\hat{p}_0$  happens to be low, then all tests are more likely to reject the null hypothesis. Hence, it may appear that Bonferroni would be too conservative. Some simulations showed that while it was conservative, using a more accurate correction made little difference. We omit details here because they have not been approved for public release, but you may try your own simulations, see Exercise 8.46.

One hint – when  $k > 1$ , it is best to let  $n_0$  be larger than the other  $n_j$  values; since the default page is used in every comparison, it is worth using extra data to make  $\hat{p}_0$  more accurate. This also makes the tests less correlated.  $\square$

## 8.6 Likelihood Ratio Tests

In this section, we discuss a class of tests, based on likelihood ratios. Recall that in Chapter 6, we introduced the likelihood function  $L(\theta | x_1, x_2, \dots, x_n)$  that measures the consistency of the parameter  $\theta$  and the observed data. In a likelihood ratio test (LRT), we decide whether data are more consistent with the alternative or null hypothesis by comparing  $L$  for values of  $\theta$  from the two hypotheses. In certain situations, that of simple hypotheses, likelihood ratio tests are the best possible tests, obtaining the smallest possible type II error rate for a given Type I error rate. In other settings, LRTs are useful both for deriving problem-specific methods, and as a test method in their own right.

### 8.6.1 Simple Hypotheses and the Neyman–Pearson Lemma

**Definition 8.5** A hypothesis is *simple* if it completely specifies the distribution of the population. Otherwise, it is *composite*.  $\parallel$

For example, in the case of a sample drawn from a normal population with unknown mean but known standard deviation  $\sigma_0$ , the null hypothesis of  $H_0: \mu = 4$  is a simple hypothesis – the parameters are  $(4, \sigma_0)$ . On the other hand, the alternative hypothesis is composite:  $(\mu, \sigma_0)$ , where  $\mu$  has many possible values. Similarly, for the case of an unknown standard deviation, both the null and alternative hypotheses are composite.

In the case that both hypotheses are simple,

$$H_0: \theta = \theta_0 \quad \text{versus} \quad H_A: \theta = \theta_A, \tag{8.3}$$

we create the *LRT statistic*

$$T = \frac{L(\theta_0)}{L(\theta_A)} = \frac{L(\theta_0 | X_1, X_2, \dots, X_n)}{L(\theta_A | X_1, X_2, \dots, X_n)}, \tag{8.4}$$

as the ratio of the two likelihood values. The *LRT* rejects the null hypothesis for small values of  $T$  since small values of this ratio indicate that the alternative

hypothesis is more likely. In particular, the test rejects  $H_0$  if  $T \leq c$  for some critical value  $c$ , which is practice is chosen based on the desired type I error rate. The  $P$ -value is the probability under the null hypothesis of a random value of the LRT  $\leq$  the observed value.

**Example 8.22** Suppose  $X_1, X_2, \dots, X_9$  is drawn from the exponential distribution with pdf  $f(X; \lambda) = \lambda e^{-\lambda X}$ . To test

$$H_0: \lambda = 8 \quad \text{versus} \quad H_A: \lambda = 10,$$

we compute the LRT statistic

$$\begin{aligned} T &= \frac{L(\lambda = 8)}{L(\lambda = 10)} = \frac{\prod_{i=1}^9 f(X_i; 8)}{\prod_{i=1}^9 f(X_i; 10)} \\ &= \frac{8^9 e^{-8 \sum_{i=1}^9 X_i}}{10^9 e^{-10 \sum_{i=1}^9 X_i}} \\ &= (0.8)^9 e^{2 \sum_{i=1}^9 X_i}. \end{aligned}$$

We reject the null hypothesis if

$$(0.8)^9 e^{2 \sum_{i=1}^9 X_i} < c$$

or, equivalently,

$$\sum_{i=1}^9 X_i < \frac{1}{2} \ln(1.25^9 \times c).$$

Call the right-side of  $c_2$ ; thus, the critical region is of the form

$$\mathcal{R} = \left\{ (X_1, X_2, \dots, X_9) \in \mathbf{R}^9 \mid \sum_{i=1}^9 X_i < c_2, X_i \geq 0 \right\}.$$

Suppose we wish to specify an alpha level of  $\alpha = 0.05$ . Since  $X_i \sim \text{Exp}(8) = \text{Gamma}(1, 8)$ , by Theorem B.11, we have  $\sum_{i=1}^9 X_i \sim \text{Gamma}(9, 8)$ . We set

$$0.05 = P((X_1, X_2, \dots, X_9) \in \mathcal{R}; H_0) = P\left(\sum_{i=1}^9 X_i < c_2 \mid H_0\right)$$

and then find the 0.05-quantile of  $\text{Gamma}(9, 8)$ , which is  $q = 0.5869 = c_2$ .  $\square$

The Neyman–Pearson lemma indicates that this test is the best possible test, for comparing simple null and alternative hypotheses.

**Theorem 8.1 (The Neyman–Pearson Lemma)** Let  $X_1, X_2, \dots, X_n$  be a random sample from a distribution with parameter  $\theta$ . Suppose we wish to test two

simple hypotheses

$$H_0: \theta = \theta_0 \quad \text{versus} \quad H_A: \theta = \theta_A.$$

The LRT rejects the null hypothesis if the test statistic  $T = L(\theta_0)/L(\theta_A)$  satisfies  $T < c$  at an alpha level  $\alpha$ . Then any test with alpha level  $\leq \alpha$  has power less than or equal to the power for this LRT. That is, for a fixed  $\alpha$ , the LRT minimizes the probability of a type II error.  $\square$

See, for instance, Casella and Berger (2001) for a proof.

Here is an analogy that may help you understand the LRT and Neyman–Pearson lemma. Suppose that you are farming some land and can plant potatoes and berries. You have to grow enough potatoes for food to survive, but beyond that you want to grow as many berries as possible because you like them. You have different plots with different yields for each plant. Which crop should you plant in each plot? Should you plant potatoes in the most fertile land? Or berries there? Actually, you should decide based on the relative yields – plant potatoes in those plots where the ratio (potato yield)/(berry yield) is the highest, and berries where the ratio (berry yield)/(potato yield) is highest. Plant just enough potatoes to meet your food quota, starting with those plots with the highest (potato/berry) yield ratio, then plant berries in the remaining plots.

Here, potatoes correspond to the null hypothesis, potato yields to  $L(\theta_0|x)$ , and where you plant potatoes to the acceptance region  $\mathcal{A}$ : this must have probability  $\int_{\mathcal{A}} L(\theta_0|x)dx$  equal to at least  $(1 - \alpha)$  to meet the quota. Berry yields correspond to likelihood under the alternative hypothesis, berries grown to power, and where we plant berries to the rejection region  $\mathcal{R}$ ; we want to maximize  $\int_{\mathcal{R}} L(\theta_A|x)dx$  subject to the quota. The test statistic  $T$  corresponds to (potato yield)/(berry yield); the best rejection region is where this is small. We accept  $H_0$ , where the ratio  $T = L(\theta_0|x)/L(\theta_A|x)$  is high, and reject where  $T$  is low.

**Example 8.23** Let  $X_1, X_2, \dots, X_n$  be  $n$  independent Bernoulli trials with parameter  $p$ . Suppose we wish to test  $H_0: p = 0.4$  versus  $H_A: p = 0.5$ . Let  $X = \sum_{i=1}^n X_i$ . Then the LRT statistic is

$$T = \frac{L(p = 0.4)}{L(p = 0.5)} = \frac{(0.4)^X (0.6)^{n-X}}{(0.5)^X (0.5)^{n-X}} = \left(\frac{2}{3}\right)^X (1.2)^n.$$

We reject  $H_0$  if

$$\left(\frac{2}{3}\right)^X (1.2)^n < c$$

or, equivalently,

$$X > \frac{\ln(c) - n \ln(1.2)}{\ln(2/3)}.$$

Let  $c_2 = (\ln(c) - n \ln(1.2)) / \ln(2/3)$ . Then the critical region is

$$\mathcal{R} = \left\{ (X_1, X_2, \dots, X_n) \in \{0, 1\}^n \mid \sum_{i=1}^n X_i > c_2 \right\}.$$

For a test with  $\alpha = 0.05$ , we want

$$0.05 = P((X_1, X_2, \dots, X_n) \in \mathcal{R} \mid H_0) = P\left(\sum_{i=1}^n X_i > c_2\right).$$

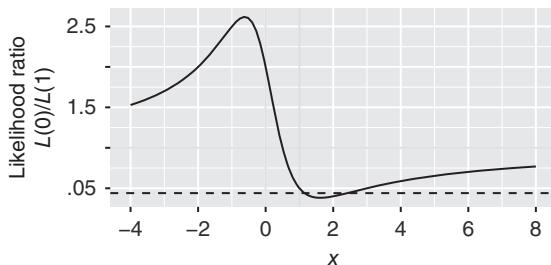
Then  $c_2$  will be the 0.05 quantile of the binomial distribution with parameters  $p = 0.4$  and  $n$ .

If  $n = 10$ , then  $P(\sum_{i=1}^{10} X_i > 7) = 0.012$ , while  $P(\sum_{i=1}^{10} X_i > 6) = 0.055$ , so in this case, we cannot achieve a test with alpha level  $\alpha = 0.05$  exactly. On the other hand, if we specify  $\alpha = 0.012$ , then any other hypothesis test with  $\alpha < 0.012$  would have less power than this LRT (i.e. the test that rejects  $H_0$  when  $\sum_{i=1}^{10} X_i > 7$ ).  $\square$

**Example 8.24** Let the random variable  $X$  have a Cauchy distribution with pdf  $f(x) = 1/(\pi(1 + (x - \theta)^2))$  for  $-\infty < x < \infty$ . The likelihood ratio for testing  $H_0: \theta = \theta_0$  versus  $H_A: \theta = \theta_A$  is  $T(x) = (1 + (x - \theta_A)^2)/(1 + (x - \theta_0)^2)$ . If  $\theta_A > \theta_0$ , then  $T$  is decreasing between  $\theta_0$  and  $\theta_A$  and somewhat beyond, but eventually increases again, with  $\lim_{x \rightarrow \infty} T = 1$ . Hence, the rejection region is a finite interval and does not include extremely large values of  $x$ . In other words, larger values of  $x$  provide stronger evidence against the null hypothesis, but only so far; eventually, a larger value of  $x$  just means an outlier (which are common with this distribution) rather than evidence. See Figure 8.7.

If there are  $n$  observations, then the log-likelihood ratio is  $\sum_{i=1}^n \ln(1 + (x_i - \theta_A)^2) - \ln(1 + (x_i - \theta_0)^2)$ . Again, values of  $x_i$  that are large but not too large provide the most evidence against  $H_0$ .  $\square$

In practice, we never know the true underlying distribution, but deriving tests based on normal, Cauchy, or other distributions is a way to develop tests that have desirable properties in practice. Tests that are theoretically optimal for



**Figure 8.7** Likelihood ratio for Cauchy,  $\theta_0 = 0$ ,  $\theta_A = 1$ , for a single observation. The dotted line is at 0.44 – this gives a rejection region  $(1.15, 2.42)$  with  $\alpha \approx 10\%$ .

normal distributions work well for data that are approximately normal. Tests for long-tailed distributions like Cauchy or  $t$  distributions are more robust against outliers.

### 8.6.2 Likelihood Ratio Tests for Composite Hypotheses

Simple alternative hypotheses are rare, but the LRT is also useful when one or both hypotheses are composite. We cannot compute the simple ratio  $T = L(\theta_0)/L(\theta_A)$  because there is no single  $\theta_0$  and/or single  $\theta_A$ . Instead, we use the most likely  $\theta$  values, that maximize  $L$  subject to the relevant hypotheses. For the numerator, we maximize subject to the null hypothesis, and for the denominator, we maximize over the combined parameter space – this is simpler than limiting to the alternative hypothesis, and ultimately makes little difference in the tests.

Let  $\Omega$  denote the set of possible values for  $\theta$ , and  $\Omega_0$  the subset that satisfies the null hypothesis. We test

$$H_0: \theta \in \Omega_0 \quad \text{versus} \quad H_A: \theta \in \Omega \setminus \Omega_0. \quad (8.5)$$

**Definition 8.6** The *LRT statistic* for testing the hypothesis given by Equation (8.5) is

$$T(X) = \frac{\max_{\Omega_0} L(\hat{\theta} | x)}{\max_{\Omega} L(\hat{\theta} | x)}. \quad (8.6)$$

A LRT rejects the null hypothesis when  $T(X) \leq c$  for some number  $c$ ,  $0 \leq c \leq 1$ . Given an observed value  $T(x)$ , the *P-value* is  $P(T(X) \leq T(x))$ . ||

Note that  $T(X) \leq 1$ , and is smaller when the data are less consistent with the null hypothesis.

We illustrate this test by first considering a simpler example with one parameter before moving on to examples with two parameters, which is where the real power of the LRT comes in.

**Example 8.25** Let  $X_1, X_2, \dots, X_n$  be a random sample from  $N(\mu, 1)$ . Suppose we wish to test

$$H_0: \mu = 8 \quad \text{versus} \quad H_A: \mu < 8.$$

The likelihood of  $\mu$  is (see Equation (6.5))

$$L(\mu) = \frac{1}{(\sqrt{2\pi})^n} e^{-(1/2)\sum_{i=1}^n (X_i - \mu)^2}.$$

In particular, suppose we have  $X_1 = 7, X_2 = 7, X_3 = 6.6$ , and  $X_4 = 6$ . Under the null hypothesis, the likelihood is

$$\begin{aligned} L(\mu = 8) &= \frac{1}{(\sqrt{2\pi})^4} e^{-(1/2)((7-8)^2 + (7-8)^2 + (6.6-8)^2 + (6-8)^2)} \\ &= 0.000473. \end{aligned}$$

In this case, the null hypothesis is simple, so we do not need to maximize across a set of possible  $\theta$  values.

The maximum likelihood estimate of  $\mu$  is the sample mean,  $\hat{\mu} = \bar{X} = 6.65$  (see Exercise 6.5), so the likelihood is

$$\begin{aligned} L(\hat{\mu}) &= \frac{1}{(\sqrt{2\pi})^4} e^{-(1/2)((7-6.65)^2 + (7-6.65)^2 + (6.6-6.65)^2 + (6-6.65)^2)} \\ &= 0.0181. \end{aligned}$$

The LRT statistic is  $T = 0.00473/0.181 = 0.259$ . In rough terms, this means that the null hypothesis is only one-fourth as consistent with the data as is the alternative hypothesis. But is that beyond normal chance variation – enough to reject the null hypothesis? We need to know more about the distribution of the test statistic when the null hypothesis is true. To do this, we will do more general calculations and avoid plugging in our specific numbers too early. The hypotheses may be written as follows:

$$H_0: \mu = \mu_0 \quad \text{versus} \quad H_A: \mu < \mu_0.$$

There is a simple null hypothesis, so the numerator of the test statistic is

$$L(\mu = \mu_0) = \frac{1}{(\sqrt{2\pi})^n} e^{-(1/2) \sum_i (X_i - \mu_0)^2}.$$

For the denominator, we assume that  $\bar{X} < \mu_0$  in the following calculations; otherwise,  $T = 1$  and we would just accept  $H_0$ . The likelihood evaluated at the maximum likelihood estimate of  $\hat{\mu} = \bar{X}$  is

$$L(\hat{\mu} = \bar{X}) = \frac{1}{(\sqrt{2\pi})^n} e^{-(1/2) \sum_i (X_i - \bar{X})^2}.$$

Thus, the ratio of the two likelihoods is

$$\begin{aligned} T(X) &= \frac{\exp(-(1/2) \sum_{i=1}^n (X_i - \mu_0)^2)}{\exp(-(1/2) \sum_{i=1}^n (X_i - \bar{X})^2)} \\ &= \exp \left[ \frac{1}{2} \left( - \sum_{i=1}^n (X_i - \mu_0)^2 + \sum_{i=1}^n (X_i - \bar{X})^2 \right) \right] \\ &= \exp[-n(\bar{X} - \mu_0)^2 / 2], \end{aligned}$$

where the last equality comes from the calculation

$$\sum_{i=1}^n (X_i - \mu_0)^2 = \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \mu_0)^2.$$

A LRT rejects  $H_0$  when  $T(X)$  is sufficiently small. Thus, we can specify a critical region  $\mathcal{R}$  to be sample means  $\bar{X}$  such that  $\bar{X} < \mu_0$  and

$$T(x) = \exp\left[-\frac{n}{2}(\bar{X} - \mu_0)^2\right] \leq c$$

or  $\bar{X} < \mu_0$  and

$$|\bar{X} - \mu_0| \geq \sqrt{-\frac{2}{n} \ln(c)}.$$

In other words, the critical region is when

$$\bar{X} \leq \mu_0 - \sqrt{-\frac{2}{n} \ln(c)},$$

for some  $0 < c < 1$  or, equivalently, when

$$\bar{X} \leq \mu_0 - d,$$

for some  $d > 0$ . In this example, the LRT reduces to a one-sided test based on  $\bar{X}$ . Since  $\bar{X} \sim N(\mu_0, 1/n)$  when  $H_0$  is true, a suitable critical value for a size  $\alpha$  test would be  $d = z_\alpha/\sqrt{n}$ . For this example, we would reject at the 5% level when  $\bar{X} \leq 8 - 1.644/2 = 7.18$ . Since  $\bar{x} = 6.65$ , we conclude that the mean  $\mu$  is less than 8.  $\square$

**Example 8.26** Let  $X_1, X_2, \dots, X_n$  be a random sample from a normal distribution with unknown mean and variance  $N(\mu, \sigma^2)$ . Suppose we wish to test

$$H_0: \mu = \mu_0 \quad \text{versus} \quad H_A: \mu \neq \mu_0.$$

The likelihood is (see Equation (6.5))

$$L(\mu, \sigma^2) = \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-(1/2)\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2}.$$

We need to maximize the likelihood under each hypothesis. It is simpler to maximize the log-likelihood

$$l(\mu, \sigma^2) = \ln(L) = -(n/2) \ln(2\pi) - n \ln(\sigma) - (1/2) \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2.$$

Under  $H_0$ ,  $\mu = \mu_0$ , and we maximize the log likelihood with respect to  $\sigma$ ; we set the derivative equal to zero and solve

$$\frac{dl}{d\sigma} = -n/\sigma + \sigma^{-3} \sum_{i=1}^n (X_i - \mu_0)^2 = 0$$

to find the maximum at  $\hat{\sigma}_0^2 = (1/n) \sum_{i=1}^n (X_i - \mu_0)^2$ .

Under  $H_A$ , we set derivatives with respect to  $\mu$  and  $\sigma$  equal to zero and solve

$$\frac{dl}{d\mu} = \frac{-1}{2\sigma^2} \sum_{i=1}^n 2(X_i - \mu)(-1) = 0,$$

which gives  $\hat{\mu}_A = \bar{X}$  and

$$\frac{dl}{d\sigma} = -n/\sigma + \sigma^{-3} \sum_{i=1}^n (X_i - \mu)^2 = 0,$$

which gives  $\hat{\sigma}_A^2 = (1/n) \sum_{i=1}^n (X_i - \bar{X})^2$ .

As a side note,  $\hat{\sigma}_0^2$  is an odd estimate – it is inflated when  $\bar{X}$  is far from  $\mu_0$ . In contrast,  $\hat{\sigma}_A^2$  is a reasonable estimate, similar to the usual sample variance  $s^2$  but with a denominator of  $n$  rather than  $n - 1$ .

We now plug these estimates into the likelihood. Let  $a = (2\pi)^{-n/2}$ , then

$$L(\mu_0, \hat{\sigma}_0^2) = a \hat{\sigma}_0^{-n} e^{-(1/2) \sum_{i=1}^n (X_i - \mu_0)^2 / \hat{\sigma}_0^2} = a \hat{\sigma}_0^{-n} e^{-n/2}$$

and

$$L(\mu_A, \hat{\sigma}_A^2) = a \hat{\sigma}_A^{-n} e^{-(1/2) \sum_{i=1}^n (X_i - \bar{X})^2 / \hat{\sigma}_A^2} = a \hat{\sigma}_A^{-n} e^{-n/2}$$

and the likelihood ratio simplifies to

$$T(X) = L(\mu_0, \hat{\sigma}_0^2) / L(\bar{X}, \hat{\sigma}_A^2) = (\hat{\sigma}_A / \hat{\sigma}_0)^n.$$

This is small when

$$\hat{\sigma}_A^2 / \hat{\sigma}_0^2 = \frac{(1/n) \sum_{i=1}^n (X_i - \bar{X})^2}{(1/n) \sum_{i=1}^n (X_i - \mu_0)^2} = \frac{(1/n) \sum_{i=1}^n (X_i - \bar{X})^2}{(1/n) \sum_{i=1}^n (X_i - \bar{X})^2 + (\bar{X} - \mu_0)^2}$$

is small; that is when

$$\frac{(\bar{X} - \mu_0)^2}{(1/n) \sum_{i=1}^n (X_i - \bar{X})^2}$$

is large. This looks familiar; it is nothing more than  $T^2 n^2 / (n - 1)$ , where  $T = (\bar{X} - \mu_0) / (S/\sqrt{n})$  is the usual  $T$  statistic for testing the mean. In other words, we can use the LRT to derive the usual  $t$  test for two-sided alternative hypotheses.  $\square$

This demonstrates one use for the LRT – to derive tests that are useful in their own right. It is also useful for finding the shape of test regions in an impartial way. Without this, there is the risk of  $P$ -value hacking – of an analyst choosing a test statistic and/or rejection region to produce a low  $P$ -value.

## 8.7 Statistical Practice

Recall Section 6.4, where we discussed a number of practical issues – measuring what matters, using transformations, robust estimates, and weighted estimation. We return to those points here in the context of hypothesis testing.

Please review the description of the Google Mobile Ads Optimization data in Section 1.12. The experiment is designed to reduce discrepancies in the return on investment (ROI) (value/cost) between desktop and mobile platforms; such discrepancies suggest suboptimal bidding by advertisers. That is, if the return on investment were lower on mobile (in the pre period), Google would recommend lower bids on mobile (for the post period). Advertisers specify a mobile bid multiplier, for example, a multiplier of 2.0 indicates the bid on mobile is double that on desktop. The effect of the experiment is governed by `mult.change`, the change in the multiplier between the pre-experiment period and the experimental period. If the multiplier was previously 2.0 and `mult.change = -0.5`, the multiplier would be reduced to 1.5.

The analysis involved a substantial amount of exploratory data analysis (EDA). We omit most of that here, but you can follow the script `MobileAds.R`, available from <https://github.com/lchihara/MathStatsResamplingR>.

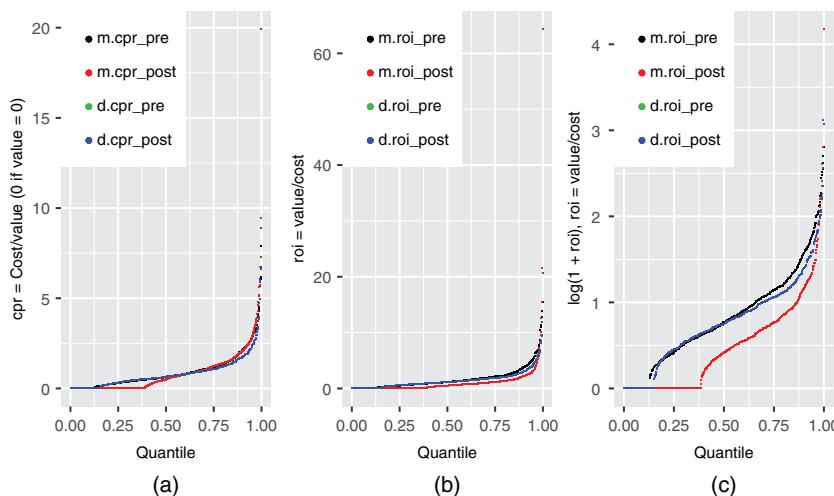
The original analysis looked for mobile/desktop discrepancies in `cpr`, which is `cost/value`, except that `cpr` is recorded as 0 rather than  $\infty$  when the denominator is 0. That could lead to misleading comparisons; say that `cpr` is 1 on desktop, both pre and post, and that `cpr` on mobile is 10 pre and 0 post. The difference  $|1 - 10|$  appears worse than the difference  $|1 - 0|$ ; but in fact the latter really means  $|1 - \infty|$ . In addition, `cpr` has some very large values due to small denominators, particularly for rows with few clicks.

Instead of `cpr`, we will work with the inverse, return-on-investment, which has the zeroes in the numerator: `roi=value/cost`. Even this has a very skewed distribution, due to some small denominators. This is typical of variables obtained by ratios.

A common remedy for ratio variables is a log transformation. However, some `roi` values are zero. And log doesn't measure what is important; it has an infinite derivative at 0, so it magnifies any differences near zero, even if not important. For example, the difference between `roi = 0.01` and `roi = 0.0001` may not matter in practice, but a log transformation implies that the difference matters as much as the difference between `roi = 1` and `roi = 100`.

Instead, we add 1 before taking log:  $\ln(1 + \text{roi})$ . Figure 8.8 shows the distribution for (a) `cpr`, (b) `roi`, and (c)  $\ln(1 + \text{roi})$ . Note that (i) the cases with `value = 0` are in their proper place in (b) and (c), (they should be at  $y = \infty$  instead of  $y = 0$  in (a)), and (ii) the distribution in (c) is not long-tailed.

Next, we take differences between the mobile and desktop platforms.  $\text{error} = \ln(1 + m.\text{roi}) - \ln(1 + d.\text{roi})$ . A successful experiment makes these



**Figure 8.8** Google Mobile Ads Optimization, distribution of return on investment and related quantities. (a)  $cpr = \text{cost}/\text{value}$  (except  $cpr$  is recorded as 0 when  $\text{value} = 0$ ). (b)  $roi = \text{value}/\text{cost} = \text{return on investment}$ . (c)  $\ln(1 + roi)$ . (Don't worry if you cannot distinguish the groups; what matters is the shape.)

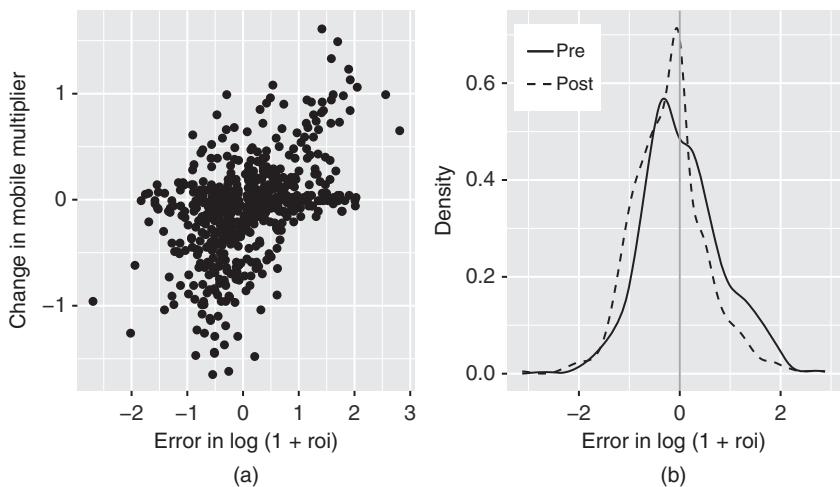
differences small. Positive values in the pre-experiment period indicate greater ROI for mobile, suggesting that the mobile bid multiplier should be increased. Figure 8.9 shows a positive correlation between these differences and the change in mobile multiplier, with a correlation of 0.47 (see Section 9.2), indicating that advertisers did tend to increase the mobile bid multiplier when the ROI was greater for mobile. The correlation for the after period was weak (0.05), indicating that advertisers had largely taken advantage of the opportunity to balance out mobile and desktop.

Figure 8.9 suggests that the experimental effect is positive; the errors are smaller for the postperiod (with the experiment) than the preperiod (before the experiment).

Now, we turn to testing. Let  $X$  and  $Y$  represent  $cpr$  errors pre and post. The analyst originally asked how to test for a discernible difference in mean absolute errors for

$$\frac{1}{n} \sum_{i=1}^n |Y_i| - |X_i|,$$

when both  $|X|$  and  $|Y|$  are highly skewed. Such skewness makes one-sample  $t$ -tests invalid, with a type I error rate far from the desired value. However, here the data are paired, and the differences  $|Y_i| - |X_i|$  actually have a fairly symmetric distribution, even for the errors on the original scale. Hence, it would be permissible to do a paired  $t$ -test. That does not mean that we should. Means



**Figure 8.9** Google Mobile Ads Optimization, (a) Positive values on the x-axis indicate a greater return on investment for mobile than desktop in the preperiod. Positive values on the y-axis indicate that the mobile bid multiplier was increased for the experiment. (b) Differences between desktop and mobile return on investment.

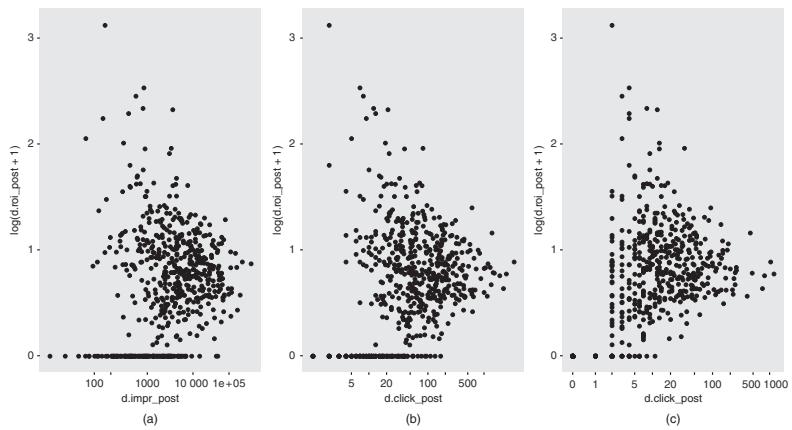
are very sensitive to outliers, so the test results would be dominated by a few outliers, and it would be difficult to detect real differences; the power would be low. In other words, type I error rate would be fine, but type II error rate would not. We could get better power by testing differences in trimmed means or other robust statistics, or take the differences in means of transformations of  $X$  and  $Y$ .

That becomes unnecessary when defining errors as differences in  $\ln(1 + \text{roi})$ . In fact, rather than working with the mean difference in absolute error, we can work with the difference in mean squared error

$$\frac{1}{n} \sum_{i=1}^n Y_i^2 - X_i^2.$$

Statisticians have long found that working with mean squared errors is more effective than working with the mean absolute error, in most situations where the errors do not have long-tailed distributions. This is better at measuring what matters (doubling an error is typically four times as bad, in practice; small errors are within the range of random variation that people expect, while large errors cause problems). It also has nicer mathematical properties, that we see in Chapter 9.

There is an additional consideration here: the errors are not identically distributed. ROI is more variable in smaller campaigns, whether size is measured in impression, clicks, or conversions (see Figure 8.10). We will give



**Figure 8.10** Return on investment (ROI) is more variable when the size of the campaign is small, measure by any of (a) impressions, (b) clicks, and (c) conversions.

more weight to the more accurate estimates, from the larger campaigns, as indicated in Section 6.4. Here we weight by a measure of size: the smaller of `m.click_pre`, `m.click_post`, `d.click_pre`, and `d.click_post`. We could use impressions or conversions instead of clicks. We use the smaller of the four values because accuracy tends to be determined by the least accurate observation. We let

$$w_i = \min_{\text{4 groups}} \text{click}_i$$

and let the test statistic be

$$T = \sum_{i=1}^n w_i(Y_i^2 - X_i^2).$$

We can evaluate the  $P$ -value using a paired permutation test, in which we randomly switch the  $X_i$  and  $Y_i$  within each row. In this case, we will use a normal approximation to the permutation distribution. Let  $D_i = Y_i^2 - X_i^2$ , then in every permutation sample  $E(D_i) = 0$  and  $\text{Var}[D_i] = (Y_i^2 - X_i^2)^2$ , hence,  $T$  has mean 0 and

$$\text{Var}[T] = \sum_{i=1}^n w_i^2(Y_i^2 - X_i^2)^2.$$

The sample size is reasonably large (628 observations), the skewness of  $T$  is zero, and there are no outliers, so a normal approximation should be very accurate. We find that  $T = -649$  (negative numbers indicate that the experiment is doing well), the standard deviation of the null distribution is 510, and  $Z$ -value is  $-1.27$ . A one-sided  $P$ -value is 0.10, indicating that the difference is not statistically discernible.

Note that our test is only for a subset of the actual data. With more data, the difference may be discernible.

### 8.7.1 More Campaigns with No Clicks and No Conversions

One thing we observe during EDA is that there are more campaigns with no clicks in the post period, for desktop and especially for mobile, similarly for conversions. For example, here are desktop conversions, before and after.

d.conv.post		
d.conv.pre	Zero	Nonzero
Zero	51	37
Nonzero	63	504

There are more ad groups with conversions pre (567) than post (531). To test whether that difference is statistically discernible, we cannot do a

two-sample test of proportions, because pre and post are not independent (they represent the same campaigns). As in Section 8.3.3, we take differences, ignore cases that are same, and only work with the changes: 63 cases with conversions before but not after, and 37 with the converse. A one-sample proportions test with  $H_0: p = 0.5$  gives a two-sided  $P$ -value of 0.012; that imbalance would occur infrequently by chance. Pre/post differences for mobile conversions, and pre/post differences for clicks on both platforms, are even more extreme than this, so it seems clear that there are reductions in both clicks and conversions. As this occurs for both platforms, it may not be due to the campaign, but may have some other cause, such as a large TV campaign starting or stopping, or the preperiod including a time when people do more shopping.

## Exercises

For all hypothesis tests, state your answer in the context of the problem (i.e. do not just state “reject null” or “accept null”).

- 8.1** Calcium levels in healthy adults are normally distributed with mean 9.5 mg/dl and unknown standard deviation. A physician in a rural community suspects that the mean calcium levels are different for women in this community. He collects measurements from 20 healthy women and finds  $\bar{x} = 9.2$  with sample standard deviation  $s = 1.1$ . Does this support his hypothesis?
- 8.2** The mean cholesterol level for vegetarians in the United States is 161 mg/dl, and the distribution is normal. You suspect that vegetarians from Sodor are different from vegetarians in the United States. You will conduct a study. For a random sample of 24 vegetarians from Sodor, you find a sample mean cholesterol level of 164 mg/dl with a standard deviation of 5 mg/dl. Do these data support the hypothesis that cholesterol levels for vegetarians in Sodor are different from the United States?
- 8.3** What is normal body temperature? The standard has been 98.6 °F. Suppose a medical worker suspects that body temperatures in children in Sodor are higher than the norm. She obtains measurements from a random sample of 18 children and finds the following:

98.0	98.9	99.0	98.9	98.8	98.6	99.1	98.9	98.5
98.9	98.9	98.4	99.0	99.2	98.6	98.8	98.9	98.7

- (a) What are the hypotheses to test?  
(b) Carry out the test and state a conclusion in a complete sentence.
- 8.4** The average of the opening price of stocks on NASDAQ on January 2, 2017, was US\$30.29. Suppose we want to know if the average of the opening price of stocks on December 1, 2017, was the same or different. The data set Nasdaq contains a random sample of 50 stock funds from NASDAQ from December 1, 2017, with mean US\$23.29 (standard deviation US\$22.73), which is lower than US\$30.29, but could this difference be due to chance variability?  
(a) Create a histogram and quantile normal plot of the data and describe the distribution.  
(b) Use the formula  $t$  test to carry out the test at a 5% alpha level. What conclusion do you draw?  
(c) If you had used the bootstrap  $t$  test instead, would you have reached the same conclusion?
- 8.5** The data set Walleye from the Minnesota Pollution Control Agency contains data on the lengths (inches) and weight (pounds) measurements for a sample of 60 walleye caught in Minnesota lakes during the 1990s (B. Monson, private communication). There is some suspicion that on average, walleye are smaller now than in the past (due to overfishing). Suppose historical records from the early 1900s indicate that the average weight of walleye caught by fishermen was 2.5 lb. Assume the data are representative of caught walleye; do they indicate that walleye from the 1990s weigh less?  
(a) Create a histogram and quantile normal plot of the data and describe the distribution.  
(b) Use the  $t$  test to carry out a test at a 5% alpha level. What conclusion do you draw?  
(c) If instead, you use the bootstrap  $t$  test, would you reach the same conclusion?
- 8.6** Suppose the number of births per month in Sodor can be modeled by a Poisson random variable with parameter  $\lambda > 0$ . The town elder claims that, on average, there are 15 births per month, but you suspect it is more. If in 1 month, there are 20 births, does this support your hypothesis?
- 8.7** Researchers at the Minnesota Pollution Control Agency collected samples of fish in different lakes around the state to analyze mercury concentrations (B. Monson, private communication). Below are mercury levels (in parts per million) in the fish for two different lakes.

Whitewater	0.26	0.263	0.267	0.281	0.288	0.297	0.315	0.315	0.380
Wirth	0.226	0.232	0.246	0.246	0.249	0.256	0.275	0.283	0.302

Do the data suggest that the true mean mercury levels differ for the two lakes? Set up and carry out a hypothesis test and state your conclusion in a complete sentence.

- 8.8** One of the factors of interest in the black spruce case study in Section 1.10 was whether or not competition would affect the growth rate of the seedlings. The biologist removed weeds and other growth from some of the plots (“No Competition”), but not others (“Competition”).
- (a) Conduct a test to see if the average change in diameter of the seedlings over the 5-year period of the study is different between the seedlings in the two groups (data set *Spruce*).
  - (b) Recall that in the study design, the seedlings were randomly assigned to the “No Competition” or “Competition” plots. What does this imply for the conclusion reached in part (a)?
- 8.9** The data set *AleLager* contains calories and alcohol content (by volume) for a sample of domestic and international ales and lager beers (per 12 oz). Investigate the hypothesis that ales have more calories than lagers.
- 8.10** The resplendent quetzal (*Pharomachrus mocinno*), a bird found in Central America, is known for its colorful plumage; it is also the national bird of Guatemala. This species typically nests in abandoned woodpecker nests in dead tree trunks (snags). Biologists were interested in the heights of these nests and the snags for the resplendent quetzal found in Guatemala and Costa Rica (Siegfried et al., 2010). Import the data set *Quetzal* that has data for 21 nests and the snags in which the nests were located.
- (a) Compute summary statistics and create exploratory plots of the nest heights (in meters) grouped by country. What do you observe?
  - (b) Conduct a hypothesis test to see if there is a difference in mean heights of the nests between the two countries.
  - (c) There is an outlier in one of the countries. Remove this nest and repeat the test. Does your conclusion change?
- 8.11** The data set *Salaries* contains salary information on a random sample of major league baseball players from 1985 and 2015. The salary data are in millions, and the 1985 numbers are in 2015 dollars.

- (a) Compute the mean and standard deviation of the salaries and create quantile normal plots for each year.
- (b) Perform a  $t$  test to compare the means. At a 5% alpha level, what conclusion would you draw?
- (c) Perform a permutation test to compare the means. What conclusion would you draw?
- (d) Which result would you report? Why?
- 8.12** In Section 8.2, we provided the code for a one sample bootstrap  $t$  test, but indicated that there is also a two sample versions. Determine the algorithm for this procedure and then use it to test the mean salaries in Exercise 8.11. *Hint:* Adapt the code in Example 7.21 and Section 7.5.2.
- 8.13** When you get fitted for glasses, the optometrist will measure the pupillary distance (PD) of each eye, the distance of the pupil to the middle of the bridge of your nose. Are people symmetric with respect to their eyes? The data set `Eyes` contains PD measurements (in mm) for a sample of volunteers. The variables `hand` and `eye` indicate which hand or eye is the dominant one.
- (a) Inspect the data set. Are the PD measurements an example of matched pairs or independent data? Explain.
  - (b) Compute summary statistics of the PD measures for the left eye and the right eye.
  - (c) Conduct a hypothesis test to determine whether or not the mean PD measurements are the same for each eye.
  - (d) Does eye symmetry depends on eye dominance? Is this a matched pairs or independent sample setting? Perform the appropriate test and state your conclusion.
- 8.14** In this simulation, we will explore effects of sample size and the variance assumption in the two-sample  $t$  test (pooled and unpooled). In the R code below, random samples of size  $m$  and  $n$  are drawn from a normal distribution  $N(30, 5^2)$  and then the  $P$ -values for the pooled and unpooled two sample  $t$ -test are extracted. We compute the proportion of times the null hypothesis of equal means is rejected (false positives).
- (a) Run the simulation with  $m = n = 30$  and compare the outcomes. Now, what if the samples are unbalanced, say,  $m = 30, n = 300$ ?
  - (b) Now, change the variances, say, to  $12^2$  for the second population (`rnorm(n, 30, 12)`). Run the simulation with  $m = n$  for various values. Then consider unequal sample sizes, for example,  $m = 30, n = 300$  and then  $m = 300, n = 30$ . Discuss.

```

m <- 30                                # Set sample sizes
n <- 30
sigma1 <- 5
sigma2 <- 5

pooled.count <- 0                      # Set counters
unpooled.count <- 0
for (i in 1:10^5)
{
  x <- rnorm(m, 30, 5)
  y <- rnorm(n, 30, 5)

  p.pooled <- t.test(x, y, var.equal=TRUE)$p. value
  p.unpooled <- t.test(x, y)$p. value

  # Increase count by one if null is rejected
  pooled.count <- pooled.count + (p.pooled < 0.05)
  unpooled.count <- unpooled.count + (p.unpooled < 0.05)
}
pooled.count / 10^5    # fraction of times discernible
unpooled.count / 10^5

```

- 8.15** In 1987, 39.1% of sex workers in Bamako, Mali, were HIV-positive. A sociologist suspects that this proportion is higher today. To check this, she surveys a random sample of  $n = 130$  sex workers from Bamako and finds  $\hat{p} = 0.48$ . Is this sufficient evidence to conclude that the proportion of sex workers in Bamako who are HIV-positive is greater than 0.391?
- 8.16** According to the 2016 American Community Survey, 26% of residents of Illinois have completed high school. A college professor suspects this percentage is lower in a certain county in the state. He surveys a random sample of adults in this county and finds that of the 310 in his sample, 69 have completed high school. Are the data consistent with the professor's hypothesis?
- 8.17** Are there regional differences in support for same-sex marriage? A 2017 survey conducted by the Pew Center for the People & the Press and the Pew Forum on Religion & Public Life found that 62% of 552 people from the Midwest supported same-sex marriage compared to 68% of 577 people from the West.<sup>4</sup> Conduct a test to see if this difference is statistically discernible.

<sup>4</sup> <http://www.people-press.org/2017/06/26/support-for-same-sex-marriage-grows-even-among-groups-that-had-been-skeptical>.

- 8.18** Infections following surgery are a serious concern that can have a major impact on a patient's road to recovery. One approach to counter infection is to kill surgical pathogens by oxidation. In one study (Greif et al., 2000), researchers randomly assigned 250 patients to receive 30% inspired oxygen and 250 patients to receive 80% inspired oxygen. All patients were undergoing surgery for colorectal resection. Of the patients receiving 30% inspired oxygen, 28 had a surgical wound infection compared to 13 patients who received the 80% inspired oxygen treatment.
- (a) Conduct a test to see if this difference is statistically discernible.
  - (b) Compute a 95% confidence interval for the difference and give a sentence interpreting this interval.
  - (c) There was no control group – a group that received no inspired oxygen – in this study. What is the implication of this?
- 8.19** Body lice infestations commonly occur in populations without access to bathing facilities, such as the homeless. In addition to causing severe itching and rashes in the infested person, body lice can spread diseases such as typhus and trench fever. Researchers in France conducted a randomized double-blind study to determine if underwear treated with permethrin, a type of insecticide, would protect against body lice infestation (Benkouiten et al., 2014). The homeless people who were recruited from two homeless shelters in Marseille were randomly assigned to wear either the permethrin-treated or untreated underwear. After 2 weeks, 11 out of the 32 homeless who wore the permethrin-treated underwear were free of lice compared to 3 out of the 28 who wore the untreated underwear.
- (a) Conduct a test to see if this difference could be due to chance variability.
  - (b) Compute a two-sided confidence interval for the difference in proportions and give a sentence interpreting this interval.
  - (c) In the study, there were originally 40 homeless persons assigned to the permethrin-treated underwear group and 33 homeless to the control group, but only 32 and 28 from each group, respectively finished the trial. How might this bias the results of this study?
- 8.20** Suppose scientists are interested in knowing whether a certain disease is associated with smoking while in high school. From a medical database, they find 200 twins in which one of the twins developed the disease (cases), while the other one did not (controls). They contact these twins and ask them if they smoked while in high school. The table below shows the responses.

		Cases	
Smoked?		No	Yes
Controls	No	100	25
	Yes	14	61

- (a) What proportion of the cases smoked while in high school? What proportion of the controls smoked while in high school?  
 (b) Determine whether or not the difference in proportions is statistically discernible.

- 8.21** Obesity in children is an international public health concern. Researchers in Spain conducted a study to see if educational interventions (classroom activities related to healthy food habits and exercise, materials promoting wellness distributed to families of school children, etc.) could help alleviate this problem (Llargues et al., 1979). At the start of the study, the children were asked whether or not they ate vegetables daily, and then asked the same question 2 years later. Of the 190 children, 28 children who had initially replied “No” changed to “Yes,” 23 students switched from “Yes” to “No,” and the remaining students had no change in their initial responses. Determine whether or not there is evidence that this educational intervention was effective in changing the children’s eating habits with regards to vegetables.
- 8.22** The School Breakfast Program (SBP) is a national program that provides breakfast to schoolchildren in low-income areas. Researchers conducted a study to determine whether parents of fourth-grade children in one county in Georgia were aware of their child’s participation in the program (Guinn et al., 2002). The parents of 337 children were asked “Does this child usually eat school breakfast?” and their response of “yes” or “no” was obtained. Data collectors observed whether or not these children participated in the SBP on a set of randomly selected days over a 24 week period; the participation rate was deemed “usual” if the child participated at least 50% of the time. The researchers found that there were 69 instances in which the parent responded “yes,” while the child was observed to not participate in the SBP, and 11 instances in which the parent responded “no,” but the child was observed participating in the SBP. In the remaining 147 cases, the parent’s assessment was consistent with the child’s behavior. Is there a statistically discernible difference in the proportion of parents who responded “yes” and the proportion of children who usually participate in the SBP?

- 8.23** In the Google mobile ads case study (Section 1.12), one of the variables recorded was the cost per click cpc.
- During the experiment (post), what proportion of the campaigns reported a cpc of 0 on the mobile platform? On the desktop platform?
  - Determine if this difference between the mobile and desktop platforms could be due to chance variability.
- 8.24** In the Google mobile ads case study (Section 1.12), one of the variables recorded was the number of conversions (purchases) conv. We will focus on the desktop platform.
- What proportion of the campaigns before the experiment had 0 conversions? Afterwards?
  - Determine if this difference before and after the experiment could be due to chance variability.
- 8.25** We indicated in Section 8.3.2 that the normal approximation for the two-proportion  $Z$  statistic is accurate in many cases, though not always. In this exercise, you will explore this.
- Use simulation to explore the sampling distribution of the  $Z$  statistic, for  $p = 0.1$  and  $p = 0.5$ , for small and large sample sizes, with equal, slightly different, and unbalanced sample sizes, e.g.  $(n_1, n_2)$  equal to  $(100, 100)$ ,  $(100, 95)$ ,  $(100, 10)$   $(10, 10)$ ,  $(10, 13)$ . In all cases, use at least  $10^4$  replications. Describe the distributions.
  - Investigate the discreteness of the distribution. Simulate the numerator  $X_1/n_1 - X_2/n_2$ . For what values of  $(n_1, n_2)$  is this approximately continuous or not? What effect does dividing by the denominator have on the discreteness of  $Z$ ?

```

n1 <- 100      # Size of sample 1
n2 <- 100      # Size of sample 2
N <- 10^4      # Size of sampling distribution
p <- 0.1       # Probability
x1 <- rbinom(N, size = n1, p)
x2 <- rbinom(N, size = n2, p)
phat <- (x1+x2) / (n1+n2)
propDiff <- x1/n1 - x2/n2
SE <- sqrt(phat * (1-phat) * (1/n1+1/n2))

df <- data.frame(x = propDiff/SE)
ggplot(df, aes(sample = x)) + geom_qq() +
  geom_abline(slope = 1, intercept = 0)

```

- 8.26** A pharmaceutical company is conducting a clinical trial to determine the effectiveness of a new drug to lower cholesterol levels. If  $\mu$  denotes the mean change (before–after) in cholesterol levels, they will test  $H_0: \mu = 0$  versus  $H_A: \mu > 0$ . Describe the type I and type II errors that could be made and the practical consequences of making these errors.
- 8.27** The Environmental Protection Agency (EPA) sets a standard for arsenic levels in water to be no more than 10 ppb (parts per billion). You suspect that, on average, arsenic levels in your community's water are much higher than this standard. You will study this by measuring arsenic levels in water samples selected from 15 households and testing  $H_0: \mu = 10$  ppb versus  $H_A: \mu > 10$  ppb, where  $\mu$  denotes the mean arsenic level (ppb) in your community.
- Describe the type I and type II errors that could be made in this study and the practical consequences of making these types of errors.
  - You gather your sample, compute the sample mean and standard deviation, then have your intern carry out the one-sample test for a mean. He uses the standard normal distribution instead of the  $t$  distribution and claims to reject the null hypothesis. If you had used the  $t$  distribution, would you have necessarily come to the same conclusion? Explain.
- 8.28** Researchers conducted a double-blind clinical trial to see if lorcaserin is effective in reducing weight (Smith et al., 2010). Participants were obese or overweight adults with a mean body-mass index of 36.2; 1595 patients were randomly assigned to take a 10 mg dose of lorcaserin while 1587 patients received a placebo.
- “At 1 year, 47.5% of patients in the lorcaserin group and 20.3% in the placebo group had lost 5% or more of their body weight ( $P < 0.001$ ), corresponding to an average loss of  $5.8 \pm 0.2$  kg with lorcaserin and  $2.2 \pm 0.1$  kg with placebo during year 1 ( $P < 0.001$ ).”
- In the quote from the study, two  $P$ -values are given. What hypotheses are being tested? What would the researchers conclude?
  - What type of errors (Type I or II?) could be made and what would be the practical consequences?
  - Can the researchers claim that lorcaserin *causes* weight loss? Why or why not.
- 8.29** Sagal is interested in the lengths of a certain species of fish in Lyman Lake. Assume the lengths (in centimeters) are normally distributed with unknown mean  $\mu$  but known standard deviation  $\sigma = 4$ . She decides to

test  $H_0: \mu = 25$  versus  $H_A: \mu > 25$  at  $\alpha = 0.05$  with a sample size of 30. What is the power of the test if, in fact,  $\mu = 27$ ?

- 8.30** The Food and Drug Administration sets an action level for mercury in fish at 1 ppm – that is, if mercury levels are higher than this value in commercial fish, the FDA will take action to remove the fish from stores. Suppose a public health official is worried about mercury levels in Lake Nyanza. She will obtain a random sample of  $n$  fish and find the average amount of mercury. If  $\mu$  denotes the mean amount of mercury in fish in the lake, she will test  $H_0: \mu \leq 1$  versus  $H_A: \mu > 1$  at  $\alpha = 0.01$ . Suppose the mercury levels in fish in this river are normally distributed with  $\sigma = 0.3$ . How large should the sample be if she wants a 90% chance of detecting a mean of  $\mu = 1.2$  (or higher)?
- 8.31** In the town of Sodor, the average weight of mice is 340 g with standard deviation 30 g. The mayor in Shetland suspects that mice in his town weigh more. If the true average weight of mice in Shetland is 354 g (or more), he will contact the public health officials. He plans to catch a sample of  $n$  mice to test this theory. If he conducts a test at the 0.1 alpha level, what must  $n$  be if he wants an 85% chance of detecting this weight?
- 8.32** Nick's favorite brand of cereal comes in boxes of 570 g. He suspects that the company is underfilling the boxes. If the true average weight is 561 g (or less), he will contact the Better Business Bureau. If he decides to test his hypothesis at a 0.05 alpha level, how many boxes of cereal would he need to sample if he wants an 80% chance of detecting a mean of 561 g? Assume for the null hypothesis that cereal weights are distributed normally with mean  $\mu = 570$  and standard deviation  $\sigma = 14$  g.
- 8.33** Let  $X_1, X_2, \dots, X_{12}$  be a random sample from a Bernoulli distribution with unknown success probability  $p$ . We will test  $H_0: p = 0.3$  versus  $H_A: p < 0.3$ , rejecting the null if the number of successes,  $Y = \sum_{i=1}^{12} X_i$ , is 0 or 1.
- Find the probability of a type I error.
  - If the alternative is true, find an expression for the power,  $1 - \beta$ , as a function of  $p$ .
  - Plot the power against  $p$ .
- 8.34** Let  $X_1, X_2, \dots, X_{50}$  be a random sample from a Bernoulli distribution with unknown success probability  $p$ . We will test  $H_0: p = 0.6$  versus  $H_A: p \neq 0.6$ , rejecting the null hypothesis at a  $\alpha = 0.05$  alpha level. Find the critical region for this test.

- 8.35** Let  $X_1, X_2, \dots, X_5$  be a random sample from the Poisson distribution with  $\lambda > 0$ . A researcher wishes to test  $H_0: \lambda = 2$  versus  $H_A: \lambda > 2$ . She will reject  $H_0$  if  $\sum_{i=1}^5 X_i \geq 16$ .
- Compute the probability of a type I error.
  - If in fact,  $\lambda = 4$ , what is the probability of a type II error? *Hints:* (a) Theorem B.6. (b) the R `ppois` function computes probabilities for the Poisson distribution.
- 8.36** Suppose a single measurement is taken from a distribution with pdf  $f(x) = \lambda e^{-\lambda x}$ ,  $x > 0$ . The hypotheses are  $H_0: \lambda = 1$  versus  $H_A: \lambda < 1$ , and the null hypothesis is rejected if  $x \geq 3.2$ .
- Calculate the probability of committing a type I error.
  - Calculate the probability of committing a type II error if in fact,  $\lambda = 1/5$ .
- 8.37** Miles draws a random sample  $X_1, X_2, \dots, X_{10}$  from an exponential distribution with  $\lambda > 0$ . He wishes to test  $H_0: \lambda = 0.25$  versus  $H_A: \lambda < 0.25$ . He decides he will reject the null hypothesis if at least three of the values are greater than 9.
- Compute the probability of a type I error.
  - If the true  $\lambda$  is 0.15, what is the power of his test?
- 8.38** Let  $X_1, X_2, \dots, X_{15}$  be a random sample from the exponential distribution with  $\lambda > 0$ . To test  $H_0: \lambda = 1/5$  versus  $H_A: \lambda < 1/5$ , you will use  $X_{\min}$  as a test statistic. If  $X_{\min} \geq 1$ , you will reject the null hypothesis. See Exercise 4.30.
- Compute the probability of a type I error.
  - Find the power of the test if in fact,  $\lambda = 1/25$ .
- 8.39** Suppose that a random sample of size 5 is drawn from a uniform distribution on  $[0, \theta]$ . To test  $H_0: \theta = 2$  versus  $H_A: \theta > 2$ , you will use  $X_{\max}$  as a test statistic. If you reject the null hypothesis when  $X_{\max} \geq k$ , what value must  $k$  be to make the probability of a type I error equal to 0.05?
- 8.40** A sample of size 1 from a distribution with pdf  $f(x) = (1 + \theta)x^\theta$ ,  $0 \leq x \leq 1$ ,  $\theta > 0$ , is drawn to test  $H_0: \theta = 2$  versus  $H_A: \theta > 2$ . The null hypothesis is rejected if  $x \geq 3/4$ .
- Find an expression for power,  $1 - \beta$ , as a function of  $\theta$ .
  - Graph the power against  $\theta$  in the interval  $(2, 10)$ .
- 8.41** Derive an expression for power like Equation (8.2) for the alternative hypothesis  $H_A: \mu < \mu_0$ . Include a graph similar to Figure 8.5.

- 8.42** Derive an expression for power like Equation (8.2) for the alternative hypothesis  $H_A: \mu = \mu_0$ . Include a graph similar to Figure 8.5.
- 8.43** Suppose you have a large data set and you plan to conduct 12 hypothesis tests. If you want the overall type I error to be 0.1, what should the type I error rate be for each individual test?
- 8.44** The data set for the General Social Survey (GSS) case study in Section 1.7 has 17 variables. Suppose an ambitious student wants to perform a certain hypothesis test for every possible pair of variables. If she wants the overall type I error rate to be 0.05, what alpha level should she use for each test?
- 8.45** Consider a random sample of size  $n$  drawn from a normal population. For a hypothesis test for the mean, if the null hypothesis is true, we noted that the  $t$  test statistic follows a  $t$  distribution with  $n - 1$  degrees of freedom. But what if the alternative hypothesis is true?

In this exercise, you will test the hypothesis  $H_0: \mu = 5$  versus  $H_A: \mu \neq 5$  when in fact, the samples are being drawn from  $N(7, 1)$ .

```

N <- 10^4
tstat <- numeric(N)
for (i in 1:N)
{
  w <- rnorm(30, 7, 1)
  tstat[i] <- (mean(w) - 5)*sqrt(30)
}

df <- data.frame(tstat)
ggplot(df, aes(x = tstat)) +
  geom_histogram(aes(y = stat(density)), bins = 20,
    color = "white") +
  xlim(c(-3,16)) +
  stat_function(fun = dt, args = list(df = 29), color = "red")

```

- Does the density of the  $t$  distribution approximate the sampling distribution of the  $t$  test statistic? In fact, the distribution of the  $t$  test statistic is a noncentral  $t$  distribution with noncentral parameter  $ncp = (\mu - \mu_0)\sqrt{n}/\sigma$ , where  $\mu$  is the true mean and  $\mu_0$  is the hypothesized mean.
- Check that the noncentral parameter for this example is 10.95 and then add the following curve to the graphs above

```
stat_function(fun = dt, args = list(df = 29, ncp = 10.95),
  color = "blue")
```

- 8.46** In Example 8.21, we indicated that when a single control group is compared to multiple treatment groups, the control group size should be larger than the treatment groups, and that the Bonferroni correction is conservative. Check this. Suppose there is one treatment group, with sample size  $n_0$ , and  $k = 10$  treatment groups, each with the same sample size  $n_1$ , with  $n_0 + kn_1 = n$ , for some fixed  $n$ .
- Find the optimal  $n_0$ , to minimize  $\text{Var}[\hat{p}_i - \hat{p}_0]$ . Hint 1: Minimize assuming the null hypothesis is true, that  $p_i = p_0$  for all  $i$ . Hint 2: Treat  $n_0$  as continuous; either solve for the constraint and differentiate with respect to  $n_0$ , or use a Lagrange multiplier.
  - How much larger is  $n_0$  than  $n_1$ ?
  - Say that  $n = 1100$ ; do simulations (with say  $10^4$  replications) with (1) the optimal  $n_0$  for this  $n$ , and (2) equal group sizes  $n_0 = n_i = n/(k+1)$ . (a) How do the variances of  $\text{Var}[\hat{p}_i - \hat{p}_0]$  compare between (1) and (2)? For both, also record (b) the correlations of the 10 differences  $\hat{p}_i - \hat{p}_0$ , (c) the fraction of times that each of the 10 tests comparing  $\hat{p}_i$  to  $\hat{p}_0$  gives a  $P$ -value below 0.05, (d) the fraction of times that at least test gives a  $P$ -value below 0.05.
  - How well would Bonferroni work here?
  - Find the correlation of the difference estimators analytically. How does the correlation change when  $n_0/n$  changes?
- 8.47** Let  $X$  be a random variable with  $\text{pdf}(x; \theta) = \theta x^{\theta-1}$ ,  $0 < x < 1$  and  $\theta > 0$ . Consider  $H_0: \theta = 1/2$  versus  $H_A: \theta = 1/4$ .
- Derive the most powerful test, using  $\alpha = 0.05$ .
  - Compute the power of this test.
- 8.48** Suppose  $X_1, X_2, \dots, X_n$  are a random sample from a population with an exponential distribution with  $\lambda > 0$ . Derive the most powerful test for  $H_0: \lambda = 7$  versus  $H_A: \lambda = 5$ .
- 8.49** Consider a sequence of  $n$  independent Bernoulli random variables  $X_1, X_2, \dots, X_n$ . Derive the most powerful test for  $H_0: p = 0.4$  versus  $H_A: p = 0.3$ .
- 8.50** Consider a sequence of  $n$  independent Bernoulli random variables  $X_1, X_2, \dots, X_n$ . Derive the most powerful test for  $H_0: p = p_0$  versus  $H_A: p = p_a$  for  $p_0 < p_a$ .
- 8.51** In genetics, certain characteristics of a plant (among other living things) are governed by three possible genotypes, denoted by AA, Aa, and aa. Suppose the probabilities of each type in a particular population is  $\theta^2, 2\theta(1-\theta), (1-\theta)^2$ , respectively, where  $0 < \theta < 1$ . In a random

sample of  $n$  plants, let  $X_1, X_2, X_3$  denote the numbers of each genotype. Derive the most powerful test to test  $H_0: \theta = 0.4$  versus  $H_A: \theta = 0.7$ . The critical region can be specified by an expression involving  $X_1$  and  $X_2$ .

- 8.52** Let  $X_1, X_2, \dots, X_n$  be a random sample from a normal distribution with unknown mean  $\mu$  and unknown variance  $\sigma^2$ . Suppose we wish to test  $H_0: \mu = \mu_0$  versus  $H_A: \mu < \mu_0$  at alpha level  $\alpha$ . Show that the likelihood ratio test reduces to a one-sided  $t$  test.
- 8.53** Let  $X_1, X_2, \dots, X_n$  be a random sample from the uniform distribution over  $[0, \theta]$ . Suppose we wish to test  $H_0: \theta = 5$  versus  $H_A: \theta < 5$  at alpha level  $\alpha = 0.05$ . Use the likelihood ratio test to find a test statistic and the critical region.
- 8.54** Let  $X_1, X_2, \dots, X_n \sim N(\mu_1, \sigma^2)$  and  $Y_1, Y_2, \dots, Y_m \sim N(\mu_2, \sigma^2)$  are random samples, where  $\mu_1, \mu_2$ , and  $\sigma^2$  are unknown. Suppose we wish to test  $H_0: \mu_1 = \mu_2$  versus  $H_A: \mu_1 < \mu_2$  at alpha level  $\alpha$ . Show that the likelihood ratio test reduces to a one-sided, two sample  $t$  test.



**9**

## Regression

In the black spruce case study in Section 1.10, the biologist was interested in how much the seedlings grew over the course of the study. Let  $(x_1, y_1), (x_2, y_2), \dots, (x_{72}, y_{72})$  denote the height and diameter change, respectively, for each of the 72 seedlings. In Figure 9.1, we see that there is a strong, positive, and roughly linear relationship between height and diameter changes.

In this chapter, we will describe a method to model this relationship – that is, find a mathematical equation that explains the relationship between change in height and the change in diameter.

### 9.1 Covariance

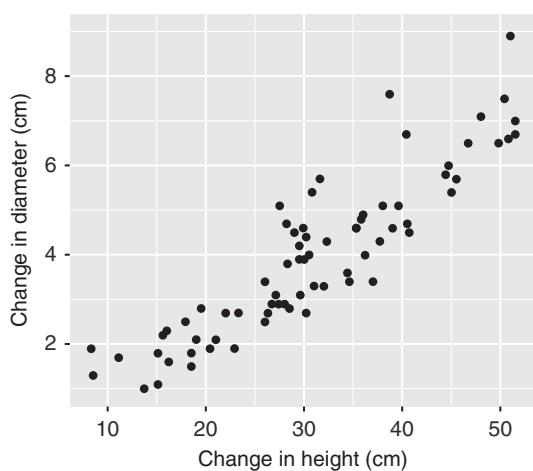
In Chapter 2, we introduced the scatter plot as a graphical tool to explore the relationship between two numeric variables. For example, referring to Figure 9.2a, we might describe the relationship here between the two variables as positive, linear and moderate to moderately strong.

Now, consider the graph in Figure 9.2b. How would you describe the relationship here? Perhaps linear, positive, and strong?

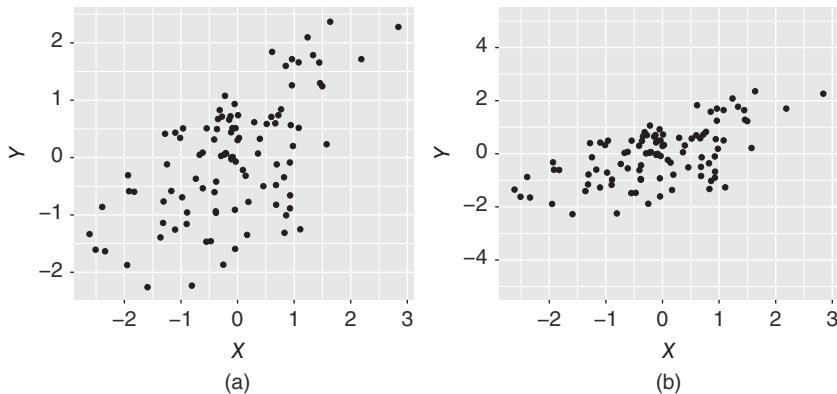
In fact, these two graphs are of the same two variables! The difference in the two impressions is due to the  $y$ -axis scaling. In the first graph, the range of the  $y$ -axis is roughly  $-2.5$  to  $2.5$ ; in the second graph, the  $y$ -axis range is roughly  $-4.5$  to  $4.5$ . Graphs are excellent tools for exploring data, but issues such as scaling can distort our perception of underlying properties and relationships. Thus, we will consider a numeric measure that indicates the strength of a linear relationship between the two variables.

We return to the Black Spruce data and recreate the scatter plot of diameter change against height change, add a vertical line to mark the mean of the height changes ( $\bar{x}$ ) and a horizontal line to mark the mean of the diameter changes ( $\bar{y}$ ) (Figure 9.3).

For points in quadrant I, both  $x_i - \bar{x}$  and  $y_i - \bar{y}$  are positive so  $(x_i - \bar{x})(y_i - \bar{y})$  is positive. Similarly, in quadrant III,  $(x_i - \bar{x})(y_i - \bar{y})$  is positive since both terms



**Figure 9.1** Change in diameter against change in height of seedlings over a 5-year period.



**Figure 9.2** Two scatter plots of two numeric variables.

are negative, and in quadrants II and IV,  $(x_i - \bar{x})(y_i - \bar{y})$  is negative since the terms have opposite signs. For the Spruce data, on average,  $(x_i - \bar{x})(y_i - \bar{y})$  is positive since most of the points are in quadrants I and III.

This motivates the following definition, a measure of how  $X$  and  $Y$  are related.

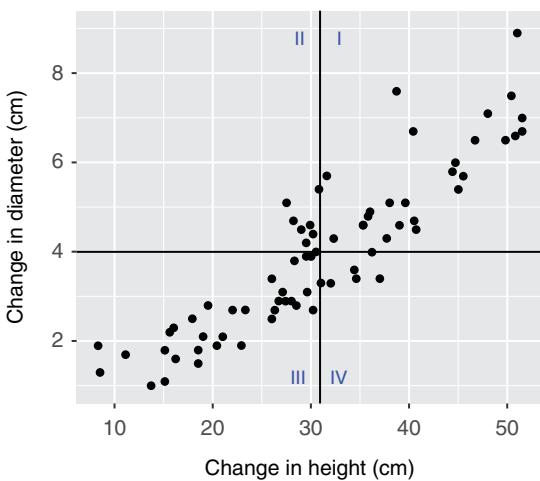
**Definition 9.1** The *covariance of  $X$  and  $Y$*  is

$$\text{Cov}[X, Y] = E[(X - \mu_X)(Y - \mu_Y)].$$

||

If  $\text{Cov}[X, Y] > 0$ , then we say that  $X$  and  $Y$  are *positively correlated*:  $X$  and  $Y$  tend to vary in the same direction, either both greater or less than their respective means. If  $\text{Cov}[X, Y] < 0$ , then we say that  $X$  and  $Y$  are *negatively correlated*:  $X$  or  $Y$  tend to vary in opposite directions, one greater than and one less than their mean.

**Figure 9.3** Change in diameter against change in height of seedlings over a 5-year period with vertical and horizontal lines at the mean values.



The covariance of a variable with itself is its variance:

$$\text{Cov}[X, X] = E[(X - \mu_X)^2] = \text{Var}[X].$$

**Proposition 9.1**  $\text{Cov}[X, Y] = E[XY] - \mu_X\mu_Y$ .

*Proof.*

$$\begin{aligned} \text{Cov}[X, Y] &= E[(X - \mu_X)(Y - \mu_Y)] \\ &= E[XY - \mu_YX - \mu_XY + \mu_X\mu_Y] \\ &= E[XY] - \mu_YE[X] - \mu_XE[Y] + \mu_X\mu_Y \\ &= E[XY] - \mu_Y\mu_X - \mu_X\mu_Y + \mu_X\mu_Y \\ &= E[XY] - E[X]E[Y]. \end{aligned}$$

□

**Example 9.1** Let  $X$  and  $Y$  have the joint distribution:

$$f(x, y) = \begin{cases} \frac{3}{2}(x^2 + y^2), & 0 < x < 1, 0 < y < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Find the covariance of  $X$  and  $Y$ .

### Solution

The marginal distributions of  $X$  and  $Y$  are:

$$\begin{aligned} f_X(x) &= \int_0^1 \frac{3}{2}(x^2 + y^2) dy = \frac{3}{2} \left( x^2 + \frac{1}{3} \right), \\ f_Y(y) &= \int_0^1 \frac{3}{2}(x^2 + y^2) dx = \frac{3}{2} \left( y^2 + \frac{1}{3} \right). \end{aligned}$$

Thus,

$$\mathbb{E}[X] = \int_0^1 x \cdot \frac{3}{2} \left( x^2 + \frac{1}{3} \right) dx = \frac{5}{8}$$

and, similarly,  $\mathbb{E}[Y] = 5/8$ . In addition

$$\mathbb{E}[XY] = \int_0^1 \int_0^1 xy \cdot \frac{3}{2}(x^2 + y^2) dxdy = \frac{3}{8}.$$

Thus,

$$\text{Cov}[X, Y] = \frac{3}{8} - \frac{5}{8} \cdot \frac{5}{8} = -\frac{1}{64}. \quad \square$$

**Corollary 9.1** If  $X$  and  $Y$  are independent, then  $\text{Cov}[X, Y] = 0$ .

**Remark** The converse is false:  $\text{Cov}[X, Y] = 0$  does not imply that  $X$  and  $Y$  are independent. For example, if  $X$  has a symmetric distribution and  $Y = (X - \mu_X)^2$ , then  $Y$  completely depends on  $X$ , but the covariance is zero.  $\parallel$

Covariances of sums of random variables add up; this yields a useful expression for the variance of a sum of random variables.

### Theorem 9.1

$$\text{Cov} \left[ \sum_{i=1}^n X_i, \sum_{j=1}^m Y_j \right] = \sum_{i=1}^n \sum_{j=1}^m \text{Cov}[X_i, Y_j].$$

*Proof.*

$$\begin{aligned} \text{Cov} \left[ \sum_{i=1}^n X_i, \sum_{j=1}^m Y_j \right] &= \mathbb{E} \left[ \left( \sum_{i=1}^n X_i - \mathbb{E} \left[ \sum_{i=1}^n X_i \right] \right) \left( \sum_{j=1}^m Y_j - \mathbb{E} \left[ \sum_{j=1}^m Y_j \right] \right) \right] \\ &= \mathbb{E} \left[ \sum_{i=1}^n (X_i - \mu_{X_i}) \sum_{j=1}^m (Y_j - \mu_{Y_j}) \right] \\ &= \sum_{i=1}^n \sum_{j=1}^m \mathbb{E}[(X_i - \mu_{X_i})(Y_j - \mu_{Y_j})] \\ &= \sum_{i=1}^n \sum_{j=1}^m \text{Cov}[X_i, Y_j]. \end{aligned} \quad \square$$

**Corollary 9.2** If  $X_1, X_2, \dots, X_n$  are random variables, then

$$\text{Var} \left[ \sum_{i=1}^n X_i \right] = \sum_{i=1}^n \sum_{j=1}^n \text{Cov}[X_i, X_j] = \sum_{i=1}^n \text{Var}[X_i] + 2 \sum_{1 \leq i < j \leq n} \text{Cov}[X_i, X_j].$$

In particular,  $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}[X, Y]$ .

An important special case is when  $X$  and  $Y$  are independent.

**Corollary 9.3** If  $X$  and  $Y$  are independent, then

$$\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y].$$

## 9.2 Correlation

In the black spruce case study, the biologist recorded measurements in centimeters. Now, suppose he decides to convert his measurements to inches.

Let  $X, Y$  denote the heights and diameters in centimeters, whereas  $X', Y'$  denote the measurements in inches. There are 0.3937 in. to the centimeter, so

$$\begin{aligned}\text{Cov}[X', Y'] &= \text{Cov}[0.3937X, 0.3937Y] \\ &= E[(0.3937X)(0.3937Y)] - E[0.3937X]E[0.3937Y] \\ &= 0.3937^2 \text{Cov}[X, Y].\end{aligned}$$

Thus, the covariance decreases by a factor of 0.155.

But changing the measurement units really does not affect how strongly the variables are related. We could even do a scatter plot that would look exactly the same, except for axis labels. We need a measure of the relationship that is unitless. This leads us to correlation.

**Definition 9.2** The *correlation coefficient* of random variables  $X$  and  $Y$  is

$$\rho(X, Y) = \frac{\text{Cov}[X, Y]}{\sigma_X \sigma_Y}. \quad \parallel$$

Correlation is not affected by adding constants or multiplying by positive constants.

**Proposition 9.2** Let  $X' = a + bX$  and  $Y' = c + dY$  for constants  $a, b \geq 0$ , and  $c, d \geq 0$ . Then

$$\rho(X', Y') = \rho(a + bX, c + dY) = \rho(X, Y).$$

□

*Proof.* Exercise.

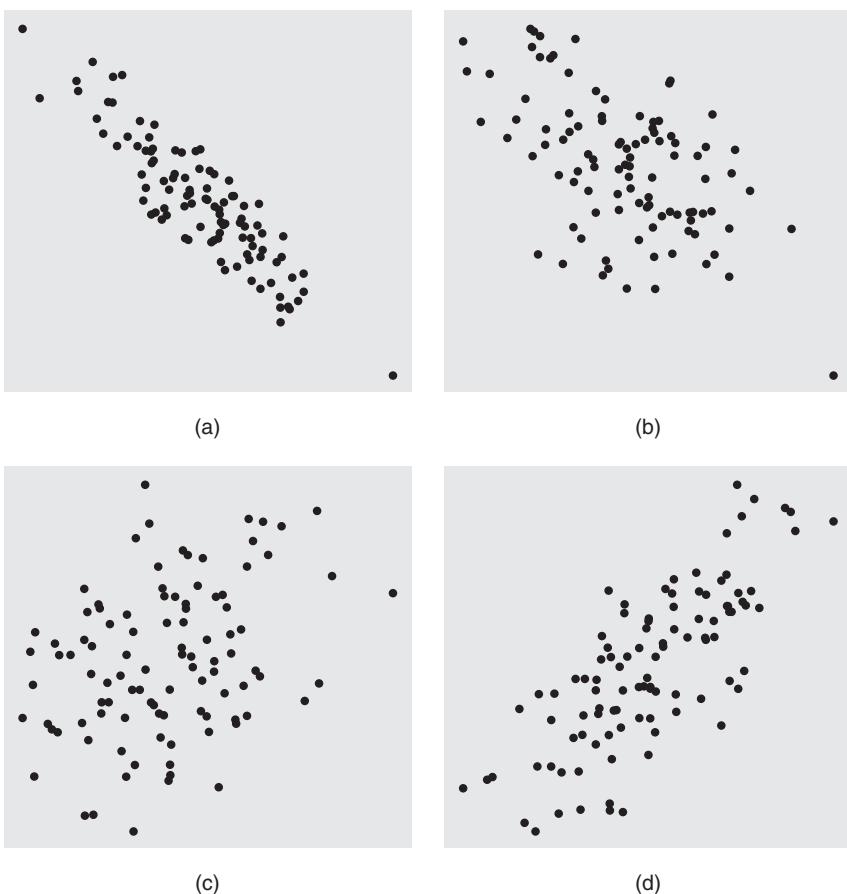
Since correlation is not affected by linear transformations, the correlation may be expressed in terms of the correlation of standardized variables, which in turn equals the covariance of the standardized variables.

**Corollary 9.4**

$$\rho(X, Y) = \rho\left(\frac{X - \mu_X}{\sigma_X}, \frac{Y - \mu_Y}{\sigma_Y}\right) = \text{Cov}\left[\frac{X - \mu_X}{\sigma_X}, \frac{Y - \mu_Y}{\sigma_Y}\right].$$

Correlation is bounded by  $-1$  and  $1$  (Figure 9.4).





**Figure 9.4** Examples of correlation. (a) Correlation  $-0.9$ , (b) Correlation  $-0.5$ , (c) Correlation  $0.3$ , (d) Correlation  $0.7$ .

**Proposition 9.3**  $|\rho(X, Y)| \leq 1$

*Proof.* Let  $Z_X = \frac{X - \mu_X}{\sigma_X}$  and  $Z_Y = \frac{Y - \mu_Y}{\sigma_Y}$ , so  $\rho(X, Y) = \text{Cov}[Z_X, Z_Y]$ .

$$\begin{aligned}\text{Var}[Z_X \pm Z_Y] &= \text{Var}[Z_X] + \text{Var}[Z_Y] \pm 2\text{Cov}[Z_X, Z_Y] \\ &= 2 \pm 2\rho(X, Y).\end{aligned}$$

But variances are always nonnegative, so

$$2 \pm 2\rho(X, Y) \geq 0,$$

$$\pm\rho(X, Y) \geq -1,$$

$$|\rho(X, Y)| \leq 1.$$

□

**Proposition 9.4**  $|\rho(X, Y)| = 1$  if and only if  $Y = a + bX$  for some real numbers  $a$  and  $b$ .

*Proof.* From the proof of the previous proposition, if  $\rho(X, Y) = 1$ , then  $\text{Var}[Z_X - Z_Y] = 0$ .

Thus,  $Z_X - Z_Y = C$  for some constant  $C$ .

$$\begin{aligned} Z_Y &= Z_X - C, \\ \frac{Y - \mu_Y}{\sigma_Y} &= \frac{X - \mu_X}{\sigma_X} - C, \\ Y &= \frac{\sigma_Y}{\sigma_X}X - \sigma_Y C + \mu_Y - \mu_X \frac{\sigma_Y}{\sigma_X}, \\ Y &= bX + a. \end{aligned}$$

Similarly for  $\rho(X, Y) = -1$ . □

We leave the converse as an exercise.

The *sample correlation* for data  $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$  is obtained by plugging in sample moments for population moments. The population correlation is

$$\rho(X, Y) = \frac{\text{Cov}[X, Y]}{\sigma_X \sigma_Y}.$$

The sample correlation can be written in different ways:

$$\begin{aligned} r &= \frac{(1/n) \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(1/n) \sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{(1/n) \sum_{i=1}^n (y_i - \bar{y})^2}} \\ &= \frac{(1/(n-1)) \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(1/(n-1)) \sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{(1/(n-1)) \sum_{i=1}^n (y_i - \bar{y})^2}} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}. \end{aligned} \tag{9.1}$$

It does not matter whether you use a divisor of  $n$  or  $n - 1$ , or no divisor, provided you are consistent.

**Remark** Many textbooks give the following algebraically equivalent form of the correlation as a calculation aid:

$$r = \frac{\left( (1/n) \sum_{i=1}^n x_i y_i \right) - \bar{x} \bar{y}}{\sqrt{\left( (1/n) \sum_{i=1}^n x_i^2 \right) - \bar{x}^2} \sqrt{\left( (1/n) \sum_{i=1}^n y_i^2 \right) - \bar{y}^2}}.$$

We advise against using this version since it is inaccurate due to round-off error; the variances in the denominator can even end up negative. Try looking up “software numerical accuracy” in your favorite web search engine for background, or see McCullough (2000). ||

### R Note

The `cor` function computes the correlation between two variables.

```
#base R
> cor(Spruce$Ht.change, Spruce$Di.change)
[1] 0.9021
# dplyr package
> Spruce %>% summarize(corr = cor(Ht.change, Di.change))
  corr
1 0.9020819
```

Recall the syntax for scatter plots (Section 2.6):

```
ggplot(Spruce, aes(x = Ht.change, y = Di.change)) + geom_point()
plot(Di.change ~ Ht.change, data = Spruce)      #base R
```

## 9.3 Least Squares Regression

We introduced correlation as a numeric measure of the strength of the linear relationship between two numeric variables. We now characterize a linear relationship between two variables by determining the best line that describes the relationship.

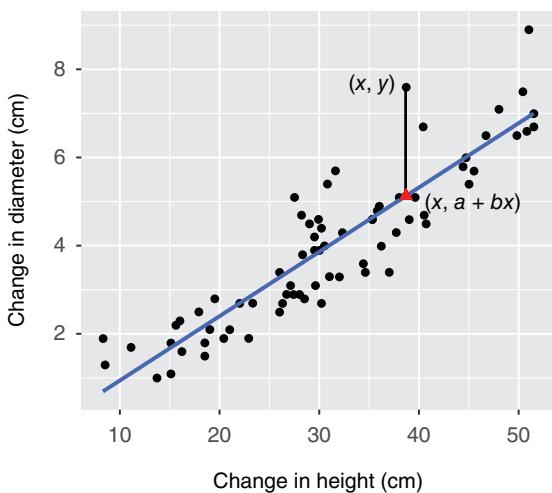
What do we mean by “best”? In most statistical applications, we pick the line  $y = a + bx$  to make the vertical distances from observations to the line small, as shown in Figure 9.5. The reason using vertical distances is that we typically use one variable, the  $x$  variable, to predict or explain the other ( $y$ ), and we try to make the prediction errors as small as possible.

Next, we need some way to measure the overall error, taking into account all vertical distances. The most natural choice would be to add the distances,

$$\sum_i |y_i - (a + bx_i)| \tag{9.2}$$

and in some applications this is a good choice. But more common is to choose  $a$  and  $b$  to minimize the *sum of squared distances*

$$g(a, b) = \sum_{i=1}^n (y_i - (a + bx_i))^2. \tag{9.3}$$

**Figure 9.5** “Best fit” line.

To minimize, we set the partial derivatives equal to zero:

$$\begin{aligned}\frac{\partial g}{\partial a} &= 2 \sum_{i=1}^n (y_i - a - bx_i)(-1) = 0, \\ \frac{\partial g}{\partial b} &= 2 \sum_{i=1}^n (y_i - a - bx_i)(-x_i) = 0,\end{aligned}$$

and solve for  $a$  and  $b$ ; this simplifies to

$$b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}. \quad (9.4)$$

$$a = \bar{y} - b\bar{x}. \quad (9.5)$$

The line  $\hat{y} = a + bx$  is called the *least-squares regression* line. In practice, we use statistical software to calculate the coefficients.

**Example 9.2** For the Spruce data, let  $x$  denote height change and  $y$  denote the diameter change. Then the least-squares line is

$$\hat{y} = -0.519 + 0.146x.$$

For every centimeter increase in the change in height, there is an associated increase of 0.146 cm in the change in diameter.

To predict the change in diameter for a seedling that grows 25 cm, we compute  $\hat{y} = -0.519 + 0.146 \times 25 = 3.13$ ; that is, we predict that a seedling that grows 25 cm in height would grow 3.13 cm in diameter.  $\square$

**Definition 9.3** For any  $x$ , let  $\hat{y} = a + bx$ , then  $\hat{y}$  is called a *predicted value* or a *fitted value*. ||

Note that Equation (9.5) can be written as  $\bar{y} = a + b\bar{x}$ , which implies that  $(\bar{x}, \bar{y})$  lies on the least-squares line  $y = a + bx$ .

**Proposition 9.5**  $(1/n) \sum_{i=1}^n \hat{y}_i = \bar{y}$ .

That is,  $\bar{\hat{y}} = \bar{y}$ . The mean of the predicted  $y$ 's is the mean of the observed  $y$ 's.

□

*Proof.* Exercise.

Next, we see that there is a relationship between correlation and least-squares regression – in particular, the slope of the least-squares line is proportional to the correlation.

Let

$$ss_x = \sum_{i=1}^n (x_i - \bar{x})^2. \quad (9.6)$$

$$ss_y = \sum_{i=1}^n (y_i - \bar{y})^2. \quad (9.7)$$

$$ss_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}). \quad (9.8)$$

Then, from Equation (9.4), we reexpress the estimated slope as

$$b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{ss_{xy}}{ss_x}.$$

We can also reexpress the correlation (Equation (9.1)) as follows:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{ss_{xy}}{\sqrt{ss_x ss_y}}.$$

Therefore,  $ss_{xy} = r \sqrt{ss_x} \sqrt{ss_y}$ , so we have

$$\begin{aligned} b &= r \frac{\sqrt{ss_x} \sqrt{ss_y}}{ss_x} \\ &= r \frac{\sqrt{ss_y}}{\sqrt{ss_x}} \end{aligned}$$

$$\begin{aligned}
 &= r \frac{\sqrt{(1/(n-1)) \sum_{i=1}^n (y_i - \bar{y})^2}}{\sqrt{1/(n-1) \sum_{i=1}^n (x_i - \bar{x})^2}} \\
 &= r \frac{s_y}{s_x},
 \end{aligned}$$

where  $s_x$  and  $s_y$  denote the sample standard deviations for  $x$  and  $y$ .

**Example 9.3** The correlation between the diameter change and height change is 0.9021. The standard deviations of the diameter changes and height changes are 1.7877 and 11.0495, respectively. Thus,  $b = 0.9021 \cdot 1.7877 / 11.0495 = 0.146$ , which agrees with what we obtained in Example 9.2.  $\square$

### Least-Squares Regression

Let  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  be  $n$  observations. The least-squares regression line is  $\hat{y} = a + bx$  where

$$b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}. \quad (9.9)$$

$$a = \bar{y} - b\bar{x}. \quad (9.10)$$

In addition,

$$b = r \frac{s_y}{s_x}, \quad (9.11)$$

where  $r$  is the sample correlation and  $s_x$  and  $s_y$  are the sample standard deviations of the  $x_i$ 's and  $y_i$ 's, respectively.

The variable being predicted,  $y$ , is called the "outcome," "response," or "fitted" variable. The variable used for predicting is called the "predictor" or "explanatory" variable.

Some authors refer to the  $y$  and  $x$  variables as "dependent" and "independent," respectively. However, this can be confused with independence and dependence of random variables.

### R Note

```
> spruce.lm <- lm(Di.change ~ Ht.change, data = Spruce)
> spruce.lm
...

```

```
Coefficients:
(Intercept) Ht.change
-0.5189      0.1459
> ggplot(Spruce, aes(x = Ht.change, y = Di.change)) + geom_point() +
+   geom_smooth(method = lm, se = FALSE)
```

To obtain the predicted values,

```
> fitted(spruce.lm) # or predict(spruce.lm)
  1        2        3        4        5        6
6.0488081 4.7644555 3.8595707 4.7060758 4.6331013 5.9612386
...
...
```

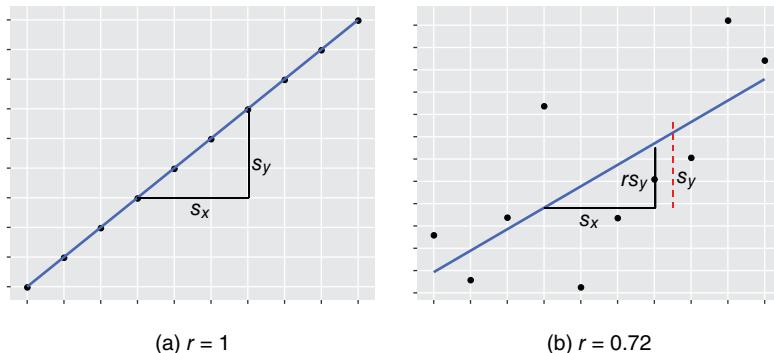
To calculate sums of squares in R, a shortcut is to use the relationship  $ss_x = (n - 1)s_x^2$ , where the sample variance  $s_x^2$  is calculated in R using var. For example, to find  $\sum_{i=1}^{72} (x_i - \bar{x})^2$  for the height change variance in the Spruce data set,

```
> (nrow(Spruce) - 1) * var(Spruce$Ht.change)
[1] 8668.38
```

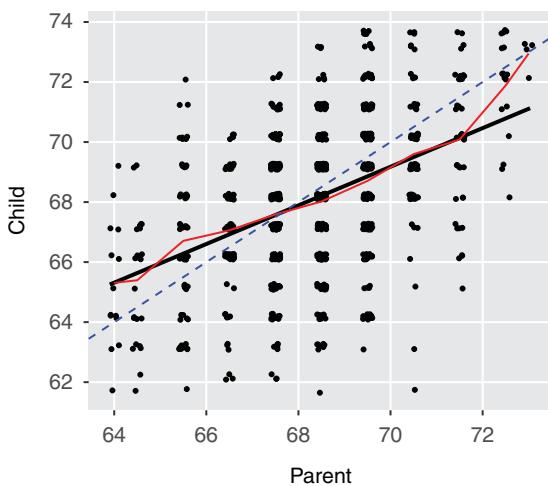
### 9.3.1 Regression toward the Mean

If  $x$  and  $y$  are perfectly correlated ( $r = 1$ ), then the slope is  $s_y/s_x$ , so every change of one standard deviation in  $x$  results in a change of one standard deviation of  $y$ . But if  $r \neq 1$ , then for a change of one standard deviation in  $x$ , the vertical change is less than one standard deviation of  $y$ ; so  $\hat{y}$  is less responsive to a change in  $x$ . If  $\rho = 0$ , then the regression line is flat. See Figure 9.6.

This phenomenon is the origin of the name “regression.” Sir Francis Galton studied the heights of parents and children. He found that although



**Figure 9.6** (a) The relationship between the regression slope,  $s_y$  and  $s_x$  with perfectly correlated data. (b) The relationship without perfect correlation. For every change in one standard deviation of  $x$ ,  $\hat{y}$  changes less than one standard deviation in  $y$ .



**Figure 9.7** Heights of parents and children. The data are *jittered* – a small amount of random noise is added so that multiple points with the same  $x$  and  $y$  are visible. The  $x$ -axis contains the “midparent” height – the average of the father’s height and 1.08 times the mother’s height. The  $y$ -axis contains the average adult child’s height, with female heights multiplied by 1.08. The dashed line is the  $45^\circ$  line. The solid line is the least-squares regression line and the zigzag line connects the mean child height with each midparent height.

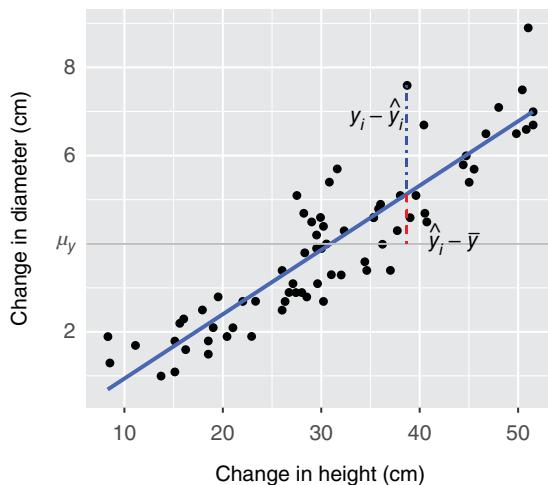
the children of tall parents were tall, on average, they were less so than their parents. Similarly, children of short parents averaged less short than their parents. The data, from Verzani (2010), are shown in Figure 9.7. Galton termed this “regression toward mediocrity,” with the implication that in the long run, everyone would become of average height. This is of course not true (just look at the coauthors of this book!)<sup>1</sup> The regression line does not give the whole picture. There is also substantial variability above and below the regression line. Some offspring of average-height parents end up tall or short, so that over time the variability above and below the mean remains roughly constant.

We now call a wide variety of models for the relationship between an outcome and predictors “regression models.” Another common term is “machine learning” models. These are conceptually the same, though we tend to use the latter term for more complicated models applied to larger data sets.

### 9.3.2 Variation

Let us look at the different components of variation in regression in more detail, to see where Galton went wrong. We will use the Spruce data.

<sup>1</sup> One of us is 5'2", the other 6'3"



**Figure 9.8** The partitioning of the variability. The horizontal line is the mean of the observed  $y$ 's.

We partition the difference between an observed  $y$  value and the mean of the observed  $y$  values into two parts (Figure 9.8),

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}).$$

By algebraic manipulation, we can show that

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2. \quad (9.12)$$

This is related to the Pythagorean theorem for the sides of triangles; the vectors  $(y_1 - \hat{y}_1, \dots, y_n - \hat{y}_n)$  and  $(\hat{y}_1 - \bar{y}, \dots, \hat{y}_n - \bar{y})$  are orthogonal. We say that the *total variation* (of the  $y$ 's) equals the *variation of the residuals* plus the *variation of the predicted values*.

Also,

$$\begin{aligned} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 &= \sum_{i=1}^n (a + bx_i - (a + b\bar{x}))^2 \\ &= \sum_{i=1}^n (b(x_i - \bar{x}))^2 \\ &= b^2 \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= b^2 ss_x \\ &= r^2 ss_y. \end{aligned} \quad (9.13)$$

Thus,

$$\begin{aligned} r^2 &= \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \frac{\frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \frac{\text{variance of predicted } y\text{'s}}{\text{variance of observed } y\text{'s}}. \end{aligned}$$

$r^2$  is the proportion of the variation of the observed  $y$ 's that is explained by the regression line.

### R-Squared = Proportion of Variance Explained

The  $R$ -squared coefficient is the square of the correlation,  $r^2$ . In other words,  $r^2 \times 100\%$  of the variance, or variation, of  $y$  is explained by the linear regression. We say that  $r^2$  is the proportion of the variance explained by the regression model.

**Example 9.4** For the black spruce case study,  $r^2 = 0.9021^2 = 0.8138$  so about 81% of the variability in the diameter changes is explained by this model.  $\square$

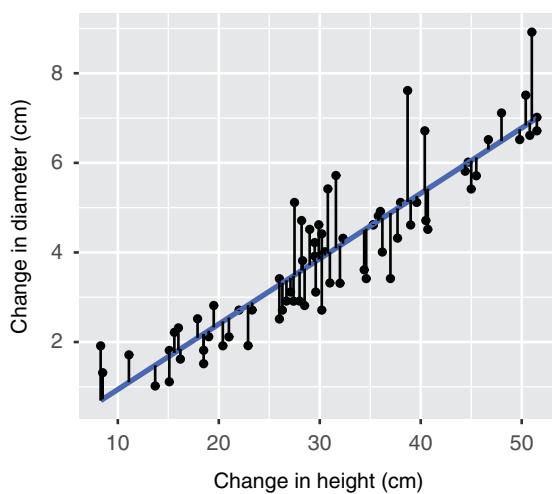
### 9.3.3 Diagnostics

We can fit a linear regression line to any two variables, whether or not there is a linear relationship. But for predictions from the line to be accurate, the relationship should be approximately linear. A linear relationship is also required (together with some additional conditions) for the standard errors and confidence intervals in Section 9.4 to be correct. So our next step is to check if it is appropriate to model the relationship between these two variables with a straight line.

**Definition 9.4** Let  $(x_i, y_i)$  be one of the data points. The number  $y_i - \hat{y}_i$  is called a *residual*.

A *residual plot* is a plot of  $y_i - \hat{y}_i$  against  $x_i$  for  $i = 1, 2, \dots, n$ .  $\parallel$

The residual is the difference between an observed  $y$  value and the corresponding fitted value; it is providing information of how far off the least-squares line is in predicting the  $y_i$  value at a particular data point  $x_i$ . If the residual is positive, then the predicted value is an underestimate, whereas if the residual is negative, then the predicted value is an overestimate (Figure 9.9).



**Figure 9.9** Residuals are the (signed) lengths of the line segments drawn from each observed  $y$  to the corresponding predicted  $\hat{y}$ .

**Example 9.5** In Example 9.2, we computed the least-squares line  $\hat{y} = -0.519 + 0.146x$ . Thus, for the first tree in the data set, the predicted diameter change is  $\hat{y} = -0.519 + 0.146 \times 5.416 = 6.049$ , so the corresponding residual is  $5.416 - 6.049 = -0.633$  (Table 9.1). The least-squares line overestimates the diameter change for this tree.  $\square$

The plot of residuals against the predictor variable  $(x_i, y_i - \hat{y}_i)$  provides visual information on the appropriateness of a straight-line model. Ideally, points should be scattered randomly about the reference line  $y = 0$ .

Residual plots are useful for the following:

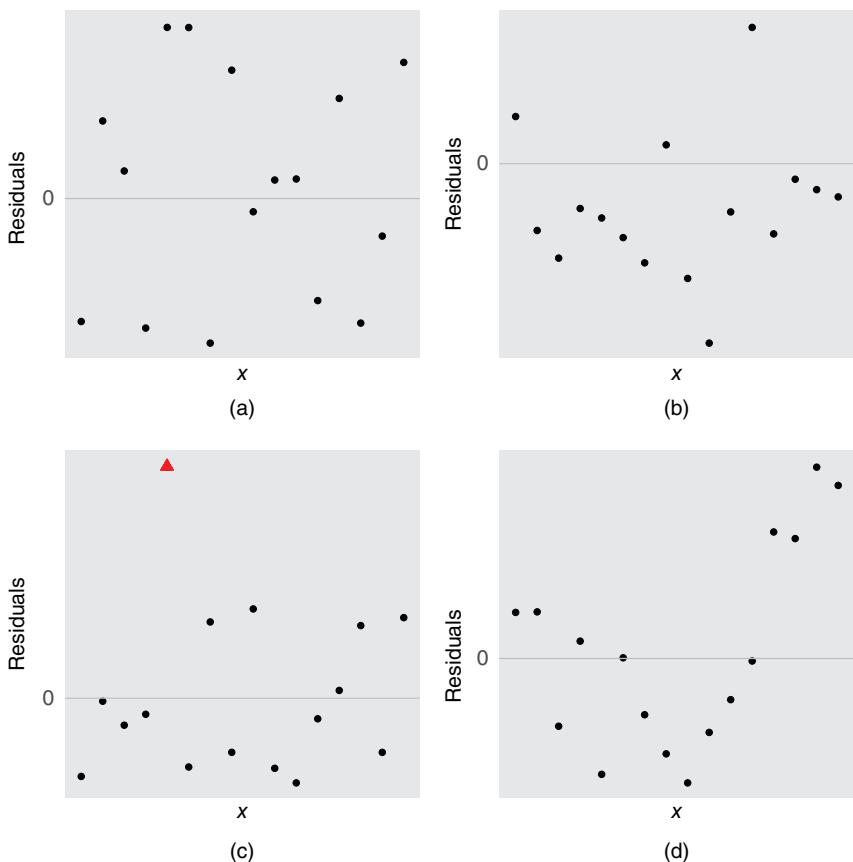
- *Revealing curvature:* That is, for indicating that the relationship between the two variables is not linear.
- *Spotting outliers.*

If outliers are noticed, then you should check to see if they are influential: does their removal dramatically change the model?

See Figure 9.10 for examples illustrating these points.

**Table 9.1** Partial view of Spruce data.

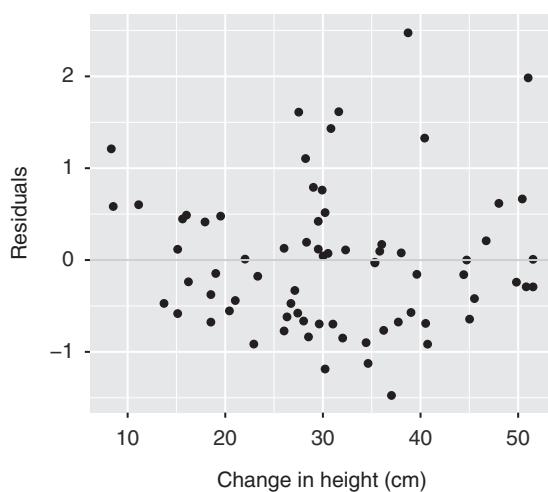
Tree	Height change	Diameter change	Predicted $\hat{y}$	Residual
1	45.0	5.416	6.049	-0.633
2	36.2	4.009	4.764	-0.755
3	30.0	3.914	3.859	0.054
4	35.8	4.813	4.706	0.106
5	35.3	4.6125	4.633	-0.021



**Figure 9.10** Examples of residual plots. (a) A good straight line fit. (b) The regression line is consistently overestimating the  $y$  values. (c) An outlier. (d) Curvature – a straight line is not an appropriate model for the relationship between the two variables.

For the Spruce data, the residual plot reveals that the distribution of the residuals is right-skewed – most residuals are negative but small, and there are a smaller number of positive residuals, but they are larger (Figure 9.11). This does not mean that a linear relationship is inappropriate, but it does cause problems for some methods that assume that residuals are normally distributed.

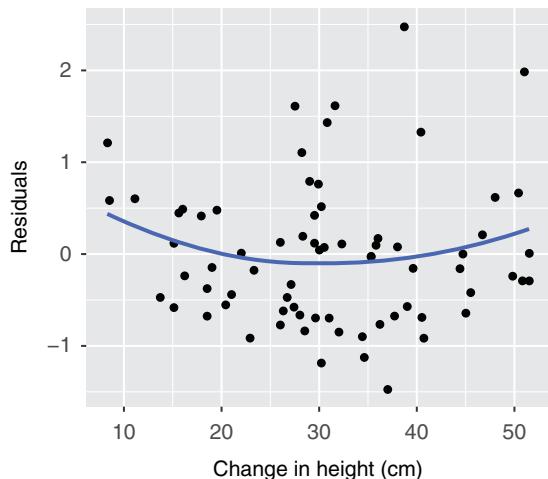
A bigger issue is whether there may be curvature. Here some caution is needed – the human eye can be good at creating patterns out of nothing (the ancient star constellations are one example). Here, ignoring the most negative residual, the residuals appear to have a curved bottom. This impression is reinforced by a second set of points slightly higher, also curved upward. But these may be purely random artifacts. There do appear to be a large



**Figure 9.11** Residual plot for the Spruce data.

number of negative residuals in the middle – but there are also a number of even bigger positive residuals in the middle. There do appear to be a lack of large negative residuals on both sides – but this may be simply because there are fewer observations on each side.

A more effective way to judge curvature is to add a *scatter plot smooth* to the plot, a statistical procedure that tries to find a possibly curved relationship in the data. There are many such procedures, for example, the “connect-the-dot” procedure shown in Figure 9.7. Figure 9.12 shows another procedure, “loess.” Calculating these is beyond the scope of this book, but most statistical software offers options to create these smoothers.



**Figure 9.12** Residual plot for the Spruce data, with a scatter plot smooth indicating slight curvature.

The smoother added to the residual plot (Figure 9.12) indicates slight curvature that should lead the researcher to some more investigation. Indeed, for these data (case study in Section 1.10), the observations do not come from one population since the seedlings were planted under different conditions.

#### R Note (Spruce Regression Example, Continued)

The `resid` function gives the residuals for a regression.

```
Spruce$Residuals <- resid(spruce.lm)
ggplot(Spruce, aes(x = Ht.change, y = Residuals)) +
  geom_point() + geom_hline(yintercept = 0) +
  geom_smooth(method = "loess", se = FALSE, span = 2)
```

The last layer adds a loess *smoother*.

**Example 9.6** Here are sugar and fat content (in grams per half cup serving) for a random sample of 20 brands of vanilla ice cream.

Sugar	15	13	20	23	11	21.5	12	23	23.0	19
Fat	8	8	16	14	7	15.5	8	21	15.5	16
Sugar	19.0	19	21.8	17	20	17	20	16	11	12
Fat	4.5	13	13.5	12	16	8	15	8	7	6

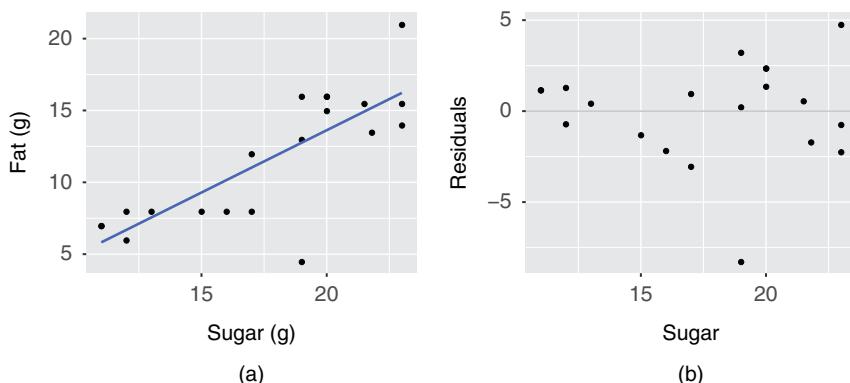
The mean and standard deviation of the fat values are 11.6 and 4.5236 g, respectively, while those for sugar are 17.665 and 4.1323 g. The correlation between fat and sugar is 0.792.

For a least-squares line of fat on sugar, the slope of the least-squares line is  $b = 0.792 \times (4.5236/4.1323) = 0.867$  while the intercept is  $a = 11.6 - 0.867 \times 17.655 = -3.707$ . Thus, the equation is  $\hat{fat} = -3.707 + 0.867 \cdot sugar$ .

For every gram increase in sugar, there is an associated increase of 0.867 g in fat content.

About 62.7% of the variability in fat content can be explained by this least-squares line.

The residual plot (Figure 9.13) reveals a large negative outlier at about 19 g of sugar. Removing this observation results in a least-squares line of  $\hat{fat} = -3.913 + 0.903 \cdot sugar$ . The slope of the regression line does not change very much though the proportion of the variability in fat that is explained by the model does (78.3%).  $\square$



**Figure 9.13** (a) Scatter plot of fat content against sugar content in ice cream. (b) Residual plot for the least-squares regression.

### 9.3.4 Multiple Regression

The ideas of linear regression can be applied in the case, where there are multiple predictors; then instead of  $\hat{y} = a + bx$ , the typical equation is of the form  $\hat{y} = a + b_1x_1 + \dots + b_px_p$ , where  $p$  is the number of predictors. In some cases, such as when Google uses regression to improve web search answers, the number of predictors can be in the billions.

One special case is when there are multiple groups in the data. In the Spruce example, there are two additional predictors – whether the tree was fertilized or not, and whether the tree faced competition. One relatively simple model in this case is

$$\hat{y} = 0.51 + 0.104 \cdot \text{Ht . change} + 1.03 \cdot \text{Fertilizer} - 0.49 \cdot \text{Competition},$$

where we convert the categorical predictors to *dummy variables* – 1 if Fertilizer = “F” and 0 for “NF”; 1 if Competition = “C” and 0 for “NC.” This equation suggests that trees that grew taller tended to grow thicker, that for a given change in height those trees that were fertilized tended to grow thicker, and that for a given change in Height, trees that were in competition did not grow as thick – it seems they spend more energy growing taller rather than thicker.

This model has a single slope and different intercepts for the four groups defined by Fertilizer and Competition. Other models can be fit, for example we may allow for different slopes in the different groups.

The formulas for calculating multiple regression coefficients are beyond the scope of this course, but can be performed using statistical software; for example the R command for the model above is `lm(Di.change ~ Ht.change+Fertilizer+Competition, data=Spruce)`.

For more about multiple regression, see Kutner et al. (2005), Weisberg (2005), or Draper and Smith (1998).

## 9.4 The Simple Linear Model

The least-squares regression line is the “best fit” line for a set of  $n$  ordered pairs,  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  whether this set represents a population or is a sample from a larger population. If this set is just a sample from a larger population, then the least-squares line is an estimate of a “true” least-squares line fit to the entire population.

Thus, in the case of a sample, after we calculate sample estimates such as the sample correlation or regression slope, we often want to quantify how accurate these estimates are, for example using standard errors, confidence intervals, and hypothesis tests. We will do so here using our usual bag of tricks – permutation tests, bootstrapping, and formulas based on certain conditions.

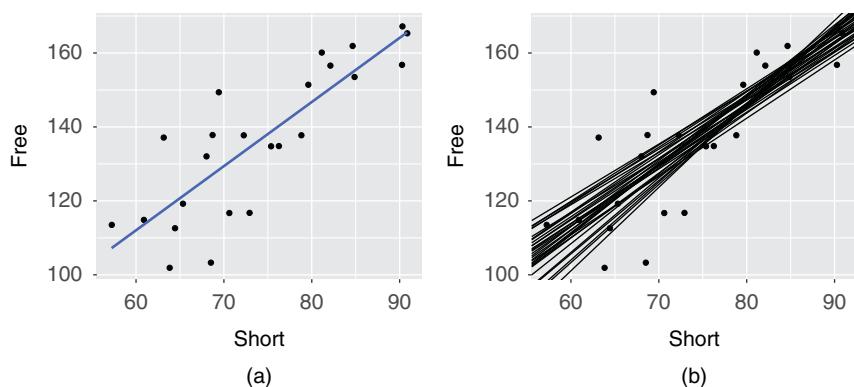
**Example 9.7** In figure skating competitions, skaters perform twice: a 2 min short program and a 4 min free skate program. The scores from the two segments are combined to determine the winner. What is the relationship, if any, between the score on the short program and the score on the free skate portion? We will investigate this by looking at the scores of 24 male skaters who competed in the 2010 Olympics in Vancouver. We will consider these observations as a sample taken from a larger population of all Olympic-level male figure skaters.

Figure 9.14a displays the scores, together with the least-squares regression line. The scores are highly correlated, with a correlation of 0.84. The regression line for predicting the score on the free skate program, based on the short program scores, is  $\hat{Y}_{\text{Free}} = 7.97 + 1.735 \cdot \text{Short}$ .

But how accurate are these numerical results? Are the correlation and regression slope discernibly different from zero? What are standard errors or confidence intervals for the correlation, slope, or the prediction for  $\hat{Y}$  at a particular value of  $x$ ?

Figure 9.14b shows regression lines from 30 bootstrap samples. This gives a useful impression of the variability of the regression predictions. The predictions are most accurate for values of  $x$  near the center, and become less accurate as we extrapolate in either direction. □

In Sections (9.4.1 and 9.5), we will obtain standard errors and confidence intervals two different ways, first using formulas and then using resampling. These are complementary – the bootstrap is better at visual impressions, while



**Figure 9.14** Scores of the 24 finalists in the 2010 Olympics men's figure skating contest for the short program and free program: (a) The least-squares regression line. (b) The regression lines from 30 bootstrap samples.

the formula approach gives mathematical expressions that quantify the visual impressions.

The least-squares regression line is derived without making any assumption about the data. That is, we do not require one or both variables to be an independent random sample drawn from any particular distribution or even for the relationship to be linear. However, in order to draw inferences or calculate confidence intervals, we require some conditions.

### Conditions for the Simple Linear Model

Let  $(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)$  be points with fixed  $x$  values and random  $Y$  values, independent of other  $Y$ 's, where the distribution of  $Y$  given  $x$  is normal,  $Y_i \sim N(\mu_i, \sigma^2)$ , with

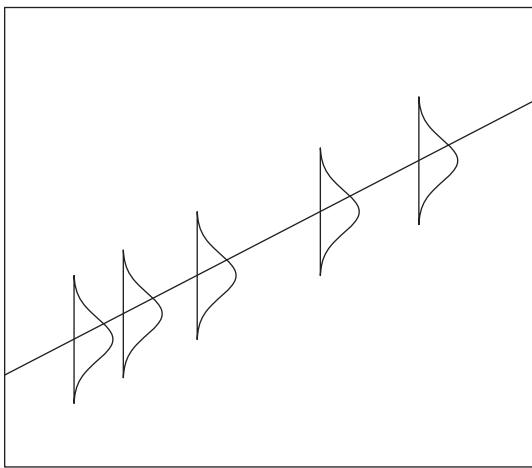
$$\mu_i = E[Y_i] = \alpha + \beta x_i,$$

for constants  $\alpha$  and  $\beta$ .

In other words, the conditions are

- The  $x$  values are fixed, not random.
- The relationship between the  $x$  values and the means  $\mu_i$  is linear,  $E[Y_i | x_i] = \alpha + \beta x_i$ .
- The residuals  $\epsilon_i = Y_i - \mu_i$  are independent.
- The residuals have constant variance.
- The residuals are normally distributed.

See Figure 9.15. In practice, the linear and independence conditions are very important, the others less so – we explain why in Section 9.4.3.



**Figure 9.15** Linear regression conditions: each  $Y_i$  is normal with mean  $\alpha + \beta x_i$  and constant variance. Conditions not shown are that the  $x$  values are fixed and observations are independent.

**Theorem 9.2** Let  $(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)$  satisfy the conditions for a linear model. Then the maximum likelihood estimates are

$$\hat{\beta} = \frac{\sum(x_i - \bar{x})(Y_i - \bar{Y})}{\sum(x_i - \bar{x})^2}. \quad (9.14)$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}. \quad (9.15)$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum(Y_i - \hat{Y})^2. \quad (9.16)$$

*Proof.* Since the  $Y_i$ 's are normally distributed, we can form the likelihood function:

$$\begin{aligned} L(\alpha, \beta, \sigma) &= \prod_1^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(Y_i - \mu_i)^2}{2\sigma^2}} \\ &= \prod_1^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(Y_i - \alpha - \beta x_i)^2}{2\sigma^2}} \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{\frac{-1}{2\sigma^2} \sum_1^n (Y_i - \alpha - \beta x_i)^2}. \end{aligned}$$

Thus, the log-likelihood is

$$\ln(L) = -n \ln(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2} \sum_1^n (Y_i - \alpha - \beta x_i)^2.$$

We take the partial derivatives with respect to  $\alpha$ ,  $\beta$ , and  $\sigma$ , respectively,

$$\begin{aligned}\frac{\partial \ln(L)}{\partial \alpha} &= \frac{-1}{\sigma^2} \sum_1^n (Y_i - \alpha - \beta x_i)(-1). \\ \frac{\partial \ln(L)}{\partial \beta} &= \frac{-1}{\sigma^2} \sum_1^n (Y_i - \alpha - \beta x_i)(-x_i). \\ \frac{\partial \ln(L)}{\partial \sigma} &= \frac{-n}{\sigma} + \frac{1}{\sigma^3 \sum_1^n (Y_i - \alpha - \beta x_i)^2}.\end{aligned}$$

Equating each partial derivative to 0 and doing some algebra yields the maximum likelihood estimates

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{x}.$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2.$$

□

Note that these maximum likelihood estimates for  $\beta$  and  $\alpha$  are exactly the same as the least-squares estimates (Equations (9.9) and (9.10)).

We state without proof:

**Theorem 9.3** Suppose  $(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)$  satisfy the conditions for a linear model. Then

1.  $\hat{\sigma}^2$ ,  $\hat{\beta}$  and  $\bar{Y}$  are mutually independent.
2.  $n(\hat{\sigma}^2 / \sigma^2)$  has a chi-square distribution with  $n - 2$  degrees of freedom.

We would not actually use  $\hat{\sigma}^2$  much; instead, we will use an unbiased version:

### Corollary 9.5

$$S^2 = \frac{n}{n-2} \hat{\sigma}^2 = \frac{1}{n-2} \sum (Y_i - \hat{Y}_i)^2$$

is an unbiased estimator of  $\sigma^2$ .

We call  $S$  the *residual standard deviation* or *residual standard error*. Note that it is computed with a divisor of  $n - 2$ , corresponding to the degrees of freedom (this is because the means are affected by two estimated parameters  $\hat{\alpha}$  and  $\hat{\beta}$ ).

*Proof.* From Theorem B.12, the expected value of a chi-square distribution with  $n - 2$  degrees of freedom is  $n - 2$ . Thus, from Theorem 9.3 (2),

$E[n(\hat{\sigma}^2/\sigma^2)] = n - 2$ , or upon rearranging,

$$E[S^2] = E\left[\frac{n}{n-2}\hat{\sigma}^2\right] = \sigma^2.$$

□

### 9.4.1 Inference for $\alpha$ and $\beta$

We now consider some properties of the maximum likelihood estimators  $\hat{\alpha}$  and  $\hat{\beta}$  of the intercept and slope for the linear model  $E[Y] = \alpha + \beta x$ .

**Theorem 9.4** Let  $(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)$  satisfy the conditions for a simple linear model, and let  $\hat{\alpha}$  and  $\hat{\beta}$  denote the estimators of  $\alpha$  and  $\beta$ , respectively. Then,

1.  $\hat{\alpha}$  and  $\hat{\beta}$  are normal random variables;
2.  $E[\hat{\alpha}] = \alpha$  and  $E[\hat{\beta}] = \beta$ ;
3.  $\text{Var}[\hat{\beta}] = \sigma^2/\text{ss}_x$ ;
4.  $\text{Var}[\hat{\alpha}] = \sigma^2[1/n + \bar{x}^2/\text{ss}_x]$ ,

where  $\text{ss}_x$  is given in Equation (9.6).

*Proof.*

$$\begin{aligned}\hat{\beta} &= \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{1}{\text{ss}_x} \sum_{i=1}^n ((x_i - \bar{x})Y_i - (x_i - \bar{x})\bar{Y}) \\ &= \frac{1}{\text{ss}_x} \left( \sum_{i=1}^n (x_i - \bar{x})Y_i - \sum_{i=1}^n (x_i - \bar{x})\bar{Y} \right) \\ &= \frac{1}{\text{ss}_x} \sum_{i=1}^n (x_i - \bar{x})Y_i.\end{aligned}$$

Note that  $\hat{\beta}$  is a linear combination of independent normal random variables, so is also a normal random variable (Theorem A.9).

$$\begin{aligned}E[\hat{\beta}] &= E\left[\frac{1}{\text{ss}_x} \sum_{i=1}^n (x_i - \bar{x})Y_i\right] \\ &= \frac{1}{\text{ss}_x} \sum_{i=1}^n (x_i - \bar{x})E[Y_i] \\ &= \frac{1}{\text{ss}_x} \sum_{i=1}^n (x_i - \bar{x})(\alpha + \beta x_i) \\ &= \alpha \frac{\sum_{i=1}^n (x_i - \bar{x})}{\text{ss}_x} + \beta \frac{\sum_{i=1}^n (x_i - \bar{x})x_i}{\text{ss}_x} \\ &= \beta,\end{aligned}$$

where the last equality follows from  $\sum_{i=1}^n (x_i - \bar{x}) = 0$  and  $ss_x = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (x_i - \bar{x})x_i$ .

Thus,  $\hat{\beta}$  is an unbiased estimator of  $\beta$ .

The variance is

$$\begin{aligned}\text{Var}[\hat{\beta}] &= \text{Var} \left[ \frac{1}{ss_x} \sum_{i=1}^n (x_i - \bar{x})Y_i \right] \\ &= \frac{1}{(ss_x)^2} \sum_{i=1}^n \text{Var}[(x_i - \bar{x})Y_i] \\ &= \frac{1}{(ss_x)^2} \sum_{i=1}^n (x_i - \bar{x})^2 \text{Var}[Y_i] \\ &= \frac{1}{(ss_x)^2} \sum_{i=1}^n (x_i - \bar{x})^2 \sigma^2 \\ &= \frac{\sigma^2}{ss_x}.\end{aligned}$$

Thus, the sampling distribution of  $\hat{\beta}$  is normal with mean  $\beta$  and variance  $\sigma^2/ss_x$ .

The proof for  $\hat{\alpha}$  is similar.  $\square$

Now, since  $\hat{\beta}$  follows a normal distribution, we can form the  $z$  statistic,

$$Z = \frac{\hat{\beta} - \beta}{\sqrt{\sigma^2/ss_x}} = \frac{\hat{\beta} - \beta}{\sigma/\sqrt{ss_x}},$$

which follows a standard normal distribution.

In practice,  $\sigma$  is unknown, so we plug in the estimate  $S$  to obtain

$$\frac{\hat{\beta} - \beta}{S/\sqrt{ss_x}}.$$

As in earlier chapters (e.g. Section 7.1.1), replacing the population standard deviation with an estimate results in a  $t$  rather than standard normal distribution.

Let  $SE_{\hat{\beta}} = S/\sqrt{ss_x}$ , the estimate of the standard error of  $\hat{\beta}$ ; then, we have the following theorem:

**Theorem 9.5** Let  $(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)$  satisfy the conditions for the simple linear model. Then,

$$T = \frac{\hat{\beta} - \beta}{SE_{\hat{\beta}}}$$

follows a  $t$  distribution with  $n - 2$  degrees of freedom.

*Proof.* From Theorem 9.4

$$Z = \frac{\hat{\beta} - \beta}{\sigma / \sqrt{ss_x}}$$

follows a standard normal distribution. Also, from Theorem 9.3,

$$\frac{n\hat{\sigma}^2}{\sigma^2} = \frac{(n-2)S^2}{\sigma^2}$$

has a  $\chi^2$  distribution with  $n-2$  degrees of freedom, and  $Z$  and  $(n-2)S^2/\sigma^2$  are independent. Thus, from Theorem B.17, the ratio

$$\frac{Z}{\sqrt{\frac{(n-2)S^2/\sigma^2}{n-2}}} = \frac{\hat{\beta} - \beta}{S / \sqrt{ss_x}}$$

has a  $t$  distribution with  $n-2$  degrees of freedom.  $\square$

In practice, we often interested in testing to see if the slope  $\beta$  is zero or we will want to calculate a confidence interval for  $\beta$ .

### Inference for $\beta$

To test the hypothesis  $H_0: \beta = 0$  versus  $H_A: \beta \neq 0$ , form the test statistic

$$T = \frac{\hat{\beta}}{SE_{\hat{\beta}}}.$$

Under the null hypothesis and linear model conditions,  $T$  has a  $t$  distribution with  $n-2$  degrees of freedom.

A  $(1-\alpha) \times 100\%$  confidence interval for  $\beta$  is given by

$$\hat{\beta} \pm q SE_{\hat{\beta}},$$

where  $q$  is the  $1-\alpha/2$  quantile of the  $t$  distribution with  $n-2$  degrees of freedom and  $SE_{\hat{\beta}} = S / \sqrt{ss_x}$ .

**Example 9.8** The data set `Skating2010` contains the scores from the short program and free skate for men's figure skating in the 2010 Olympics.

### R Note

```
> skate.lm <- lm(Free ~ Short, data = Skating2010)
> summary(skate.lm)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) 7.9691    18.1175    0.440     0.664
Short       1.7347    0.2424    7.157 3.56e-07 ***
---
Signif. codes: 0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.36 on 22 degrees of freedom
Multiple R-squared: 0.6995,      Adjusted R-squared: 0.6859
F-statistic: 51.22 on 1 and 22 DF,  p-value: 3.562e-07
```

From the R output, we obtain  $S = 11.36$ , the estimate for  $\sigma$ .

To test  $H_0: \beta = 0$  versus  $H_A: \beta \neq 0$ , we use  $t = \hat{\beta}/SE_{\hat{\beta}} = 1.7347/0.2424 = 7.157$ . We compare this to a  $t$  distribution with 22 degrees of freedom to obtain a  $P$ -value of  $2 \times 1.780036 \times 10^{-7} = 3.56 \times 10^{-7}$ . Thus, we conclude that  $\beta \neq 0$ .

To compute a 95% confidence interval for the true  $\beta$ , we first find the 0.975 quantile for the  $t$  distribution with 22 degrees of freedom,  $q = 2.0738$ . Then,

$$1.7347 \pm 2.0738 \times 0.2424 = (1.232, 2.2374).$$

Thus, we are 95% confident that the true slope  $\beta$  is between 1.23 and 2.24.  $\square$

Similarly, we could give a standard error for  $\hat{\alpha}$  and calculate a  $t$  statistic for testing  $H_0: \alpha = 0$ . Statistical software routinely provides these. We caution, however, that it is rarely appropriate to do that test. It may be tempting to test whether one can simplify a regression model by omitting the intercept. But unless you have a physical model that omits the intercept, you should include the intercept in describing a linear relationship. And even when there is such a physical model, in practice including the intercept provides a useful fudge factor for adjusting the discrepancy between theory and reality.

#### 9.4.2 Inference for the Response

In many applications, we will be interested in estimating the mean response for a specific value of  $x$ , say  $x = x_s$ . If  $\hat{Y}_s = \hat{\alpha} + \hat{\beta}x_s$  denotes the point estimate of  $E[Y_s]$ , we need the sampling distribution of  $\hat{Y}_s$ .

We state results for both  $\bar{Y}$  and  $\hat{Y}$ .

**Theorem 9.6** Let  $(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)$  satisfy the conditions for a simple linear model. Then,

1.  $\bar{Y}$  is a normal random variable.
2.  $E[\bar{Y}] = \alpha + \beta\bar{x}$ .
3.  $\text{Var}[\bar{Y}] = \sigma^2/n$ .
4.  $\hat{Y}_s$  is a normal random variable.

5.  $E[\hat{Y}_s] = E[Y_s] = \alpha + \beta x_s$ .
6.  $\text{Var}[\hat{Y}_s] = \sigma^2 [1/n + (x_s - \bar{x})^2 / ss_x]$ .

*Proof.* We leave the proof for the normality, mean, and variance of  $\bar{Y}$  as an exercise.

From Theorem 9.4,

$$E[\hat{Y}_s] = E[\hat{\alpha} + \hat{\beta}x_s] = E[\hat{\alpha}] + E[\hat{\beta}]x_s = \alpha + \beta x_s.$$

Using Equation (9.15), we have  $\hat{Y}_s = \bar{Y} + (x_s - \bar{x})\hat{\beta}$ , which is a linear combination of two independent normal variables. Also, by Theorems 9.3 and 9.4,

$$\begin{aligned}\text{Var}[\hat{Y}_s] &= \text{Var}[\bar{Y}_s + (x_s - \bar{x})\hat{\beta}] \\ &= \text{Var}[\bar{Y}_s] + (x_s - \bar{x})^2 \text{Var}[\hat{\beta}] \\ &= \frac{\sigma^2}{n} + (x_s - \bar{x})^2 \frac{\sigma^2}{ss_x}.\end{aligned}$$

□

Again, using the residual standard error  $S$  as an estimate of  $\sigma$ , we have that

$$T = \frac{\hat{Y}_s - E[\hat{Y}_s]}{S \sqrt{1/n + (x_s - \bar{x})^2 / ss_x}}$$

follows a  $t$  distribution.

Let  $SE_{\hat{Y}_s} = S \sqrt{1/n + (x_s - \bar{x})^2 / ss_x}$ , the estimate of the standard error of  $\hat{Y}_s$ . We summarize without formal proof:

**Theorem 9.7** Let  $(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)$  satisfy the conditions for a simple linear model. Let  $x = x_s$  be a specific value of the predictor variable and  $\hat{Y}_s = \hat{\alpha} + \hat{\beta}x_s$ . Then,

$$T = \frac{\hat{Y}_s - E[\hat{Y}_s]}{SE_{\hat{Y}_s}}$$

follows a  $t$  distribution with  $n - 2$  degrees of freedom.

### Confidence Interval for $E[Y_s]$

A  $(1 - \alpha) \times 100\%$  confidence interval for  $E[Y_s]$  at  $x = x_s$  is given by

$$\hat{Y}_s \pm qSE_{\hat{Y}_s} = \hat{Y}_s \pm qS \sqrt{\frac{1}{n} + \frac{(x_s - \bar{x})^2}{ss_x}},$$

where  $q$  is the  $1 - \alpha/2$  quantile of the  $t$  distribution with  $n - 2$  degrees of freedom and  $S$  is the residual standard error.

We see that the variance of  $\hat{Y}_s$  is smallest at  $x_s = \bar{x}$  and increases as  $(x_s - \bar{x})^2$  increases. In other words, the farther the  $x_s$  is from  $\bar{x}$ , the less accurate the predictions.

**Example 9.9** In the Olympic skating Example 9.8, suppose we consider a short program score of 60. Then the estimate of the mean free skate score is  $\hat{E}[Y_s] = 7.969 + 1.735 \times 60 = 112.07$ . From the data set, we find the mean and standard deviation of the short score to be  $\bar{x} = 74.132$  and  $s_x = 9.771$ , respectively. Thus, with  $n = 24$ ,  $S = 11.36$ , and  $ss_x = (n - 1)s_x^2 = 2195.691$ , the standard error is

$$11.36 \sqrt{\frac{1}{24} + \frac{(60 - 74.132)^2}{2195.61}} = 4.137;$$

the 0.975 quantile for the  $t$  distribution with 22 degrees of freedom  $q = 2.074$ . Thus, the 95% confidence interval for the mean free skate score when the short score is 60 is

$$112.07 \pm 2.074 \times 4.137 = (103.5, 120.7).$$

We conclude that with 95% confidence, the expected free skate score is between 103.5 and 120.7 when the short program score is 60 points.  $\square$

What if instead of the mean free skate score corresponding to a short score of 60, we want an estimate of an individual free skate score? In this case, we need to take into account the uncertainty in the expected value as well as the random variability of a single observation. Thus, the variance of the prediction error is

$$\text{Var}[Y - \hat{Y}] = \text{Var}[Y] + \text{Var}[\hat{Y}] = \sigma^2 + \sigma^2 \left[ \frac{1}{n} + \frac{(x - \bar{x})^2}{ss_x} \right].$$

Thus, the estimate of the prediction standard error is

$$\text{SE}_{\text{prediction}} = S \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{ss_x}}. \quad (9.17)$$

### Prediction Interval for $Y_s$

A  $(1 - \alpha) \times 100\%$  prediction interval for  $Y_s$  at  $x = x_s$  is given by

$$\hat{Y}_s \pm q \text{SE}_{\text{prediction}} = \hat{Y}_s \pm q S \sqrt{1 + \frac{1}{n} + \frac{(x_s - \bar{x})^2}{ss_x}}, \quad (9.18)$$

where  $q$  is the  $1 - \alpha/2$  quantile of the  $t$  distribution with  $n - 2$  degrees of freedom and  $S$  is the residual standard error.

This interval is very sensitive to normality – if the residual distribution is not normal this should not be used, even if  $n$  is huge.

**Example 9.10** Suppose a male skater scores a 60 on his short program. Find a 95% prediction interval for his score on the free skate.

### Solution

Referring to Example 9.9, we have  $\hat{Y}_s = 112.07$ . The standard error of prediction is

$$11.36 \sqrt{1 + \frac{1}{24} + \frac{199.7134}{2195.61}} \approx 12.09.$$

Thus,

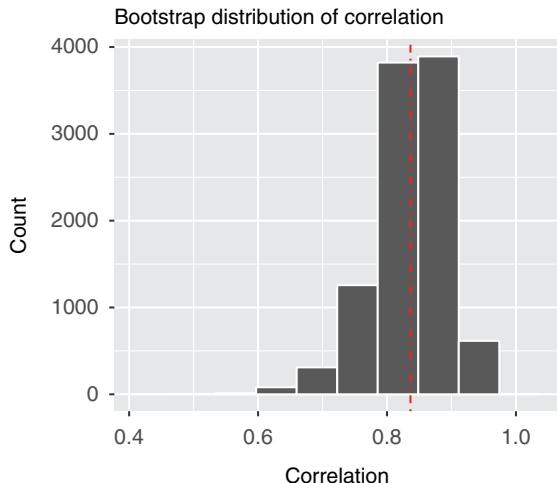
$$112.07 \pm 2.074 \times 12.09 = (86.995, 137.146).$$

Note that the prediction interval is much wider than the confidence interval; also see Figure 9.16.  $\square$

**Example 9.11** A wildlife biologist is interested in the relationship between the girth and length of grizzly bears. Suppose from a sample of 17 bears, she finds the following linear relationship:  $\text{Girth} = 32.67 + 0.55\text{Length}$ , where the measurements are in centimeters. Suppose the residual standard error and  $R$ -squared are 4.69 and 0.792, respectively. In addition, suppose the mean and standard deviation of the length measurements are 141.59 and 16.19 cm, respectively.

- (a) Find the standard deviation of the girth measurements.
- (b) Find a 95% confidence interval for the average girth of a bear whose length is 120 cm.

**Figure 9.16** Plot of pointwise confidence and prediction intervals for the Olympic skating data.



- (c) Find a 95% prediction interval for the girth of an individual bear whose length is 120 cm.

### Solution

We are given  $s = 4.69$ ,  $r = \sqrt{0.792} = 0.890$ ,  $\hat{\beta} = 0.55$ , and  $s_x = 16.19$ .

- (a) Thus, the standard deviation of the girth measurements is  $s_y = \hat{\beta}s_x/r = 0.55 \cdot 16.19/0.890 = 10.01$ .  
 (b)  $Y_s = 32.67 + 0.55 \times 120 = 98.67$ . Since  $ss_x = (n - 1)s_x^2$ , we have

$$\text{SE}_{\bar{Y}_s} = 4.69 \sqrt{\frac{1}{17} + \frac{(120 - 141.59)^2}{(16 \times 16.19^2)}} = 1.934.$$

The 0.975 quantile for a  $t$  distribution on 15 degrees of freedom is  $q = 2.131$ , so a 95% confidence interval for the mean girth of grizzlies that are 120 cm in length is  $98.67 \pm 2.131 \times 1.934 = (94.5, 102.8)$  cm.

- (c) The standard error of prediction is

$$\text{SE}_{\text{prediction}} = 4.69 \sqrt{1 + \frac{1}{17} + \frac{(120 - 141.59)^2}{(16 \times 16.19^2)}} = 5.07.$$

Thus, a 95% prediction interval for the girth of an individual grizzly that is 120 cm in length is  $98 \pm 2.131 \times 5.07 = (87.9, 108.5)$  cm.  $\square$

### 9.4.3 Comments About Conditions for the Linear Model

We began Section 9.4 with certain conditions. We now discuss how important those conditions are.

#### 9.4.3.1 The $x$ Values Are Fixed

In practice, this condition holds in some designed experiments, say when estimating the relationship between crop yield and fertilizer, where the amount of fertilizer applied to each plot or plant is specified in advance. It does not hold in the more common case that both the  $X$  and  $Y$  values are random, and the pairs  $(X_i, Y_i)$  are drawn at random from a joint distribution.

This condition played a key role in derivations of the properties of  $\hat{\beta}$  and other estimates. But in practice this condition is relatively unimportant. We routinely use regression predictions, tests, and intervals even when the  $X$  values are random.

In fact, it is typically better to do the analysis as if the  $X$  values are fixed, even if they are random. In doing so, we are *conditioning on the observed information*. Information is a concept that relates to how accurately we can make estimates. For example, a larger sample size corresponds to more information and more precise estimates. In simple linear regression, the information also depends on

how spread out the  $x$  values are. Recall that  $\text{Var}[\hat{\beta}] = \sigma^2/\text{ss}_x$  – the more spread out the  $x$  values, the more accurate the estimate of slope.

Now, what does it mean to condition on the observed information? Suppose you are planning a survey and your roommate agrees to help. Each of you will poll 100 people to end up with a total sample size of 200. However, she gets sick and cannot help. When analyzing the results of your survey and computing standard errors, should you take into account that the eventual sample size was random, with a high probability of being 200? No, you should not – you should just analyze the survey based on the amount of information you have, not what might have been. Similarly, when computing standard errors for the regression slope, it is generally best to compute them based on how spread out the  $X$  values actually are, rather than adjusting for the fact that they could have been more or less spread out.

#### 9.4.3.2 The Relationship Between the Variables Is Linear

This is critical. Suppose, for example, that the real relationship is quadratic. Then the predictions are wrong. Also, the residuals standard error is inflated, because  $\sum(Y_i - \hat{Y}_i)^2$  includes not only the random deviations but also the systematic error, the differences between the line and the curve.

In practice, this linearity condition is often violated. If the violation is small, we may proceed anyway, but if it is larger, then predictions, standard errors, confidence intervals, and  $P$  values are all incorrect.

#### 9.4.3.3 The Residuals Are Independent

This condition is also critical. This condition is often violated when the observations are collected over time. Often, data are collected over time, and successive residuals are positively correlated, in which case the actual variances are larger than indicated by the usual formulas.

#### 9.4.3.4 The Residuals Have Constant Variance

This is less important when doing inferences for  $\hat{\beta}$ . The condition is often violated: in particular, we often see the residual variance increase (or decrease) with  $x$ , with an average value in the middle. Then, the differences between reality and the conditions tend to cancel out in computing  $\text{Var}[\hat{\beta}]$ .

However, when computing  $\text{Var}[\hat{Y}]$  when  $x \neq \bar{x}$  or  $\text{Var}[\hat{\alpha}]$ , this condition does matter. We will see an example below.

#### 9.4.3.5 The Residuals Are Normally Distributed

Here we benefit from a version of the central limit theorem – if the sample size is large and the information contained in  $\text{ss}_x$  is not concentrated in a small number of observations, then  $\hat{\alpha}$ ,  $\hat{\beta}$ , and  $\hat{Y}$  are approximately normally distributed even when the residuals are not normal, and confidence intervals are approximately correct.

Prediction intervals are another story. They are a prediction for a single value, not an average, and a large sample size does not make these approximately correct if the residual distribution is nonnormal.

### Summary of Conditions for Linear Model

The critical conditions are that the relationship between the two variables is linear and that the observations are independent. The constant variance condition can be important, but the most common violations have little effect on inferences for  $\hat{\beta}$ . Normality and fixed  $X$  values are relatively unimportant for confidence intervals, but normality matters when computing a prediction interval.

## 9.5 Resampling Correlation and Regression

Another approach to obtaining inferences is to resample. We begin with the bootstrap, for standard errors and confidence intervals then use permutation tests for hypothesis testing.

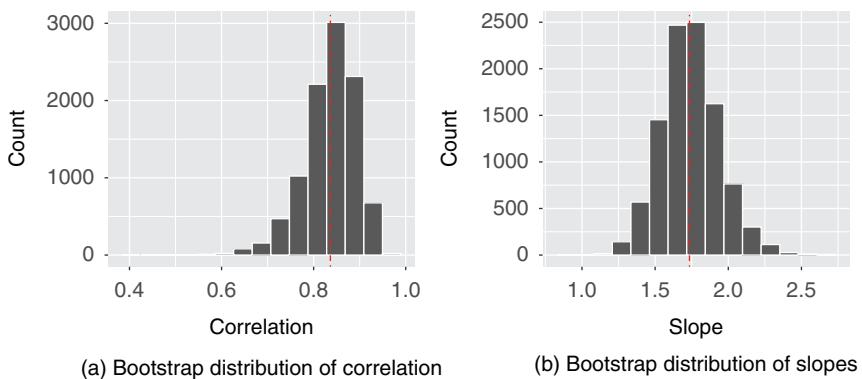
To bootstrap, we treat these the skaters as a random sample of the population of all Olympic-quality male skaters. Then, to create a bootstrap sample, we resample the skaters. For each bootstrap sample, we calculate the statistic(s) of interest.

Here is the general bootstrap procedure for two variables:

### Bootstrap for Two Variables

Given a sample of size  $n$  from a population with two variables,

1. Draw a resample of size  $n$  with replacement from the sample; in particular, draw  $n$  bivariate observations  $(x_i, y_i)$ . If the observations are rows and variables are columns, we resample whole rows.
2. Compute a statistic of interest, such as the correlation, slope, a prediction at a specific value of  $x$   $\hat{E}[\hat{Y}] = \hat{\alpha} + \hat{\beta}x$ , or a  $t$  statistic for any of these quantities.
3. Repeat this resampling process many times, say 10 000.
4. Construct the bootstrap distribution of the statistic. Inspect its spread, bias, and shape. Use it for confidence intervals or hypothesis tests.



**Figure 9.17** Bootstrap distributions of (a) correlations and (b) slopes for scores from the 2010 Olympics men's figure skating competition. The vertical lines are the corresponding statistics for the original data.

Figure 9.17a shows the bootstrap distribution for the correlation coefficient. This distribution is skewed and indicates bias – the mean of the bootstrap distribution is smaller than the correlation of the original data. This is typical for correlations – they are biased toward zero. This is reasonable – if the original correlation is near 1 (or -1), the correlation for a sample cannot get much larger (or smaller), but can get much smaller (larger).

The bootstrap standard errors are 0.57 for the correlation and 0.20 for the slope. As before, these are the standard deviations of the bootstrap distributions.

We can use the range of the middle 95% of the bootstrap values as a rough confidence interval. For instance, in one simulation for the skating data, we found a 95% percentile confidence interval for the correlation to be (0.70, 0.92) and for the slope  $\beta$  of the regression line to be (1.37, 2.17). In Example 9.8, using the  $t$  distribution, we found a 95% confidence interval for the slope to be (1.23, 2.34).

In this case, the classical interval is probably more accurate. In most of the regression problems you will encounter, the bootstrap does not offer much improvement in accuracy over classical intervals, except when the conditions behind classical intervals are violated (see the Bushmeat Case Study, Section 9.5.2). Still, the bootstrap offers a way to check your work and provides graphics that may help you understand confidence intervals and standard errors in regression problems.

### R Note

The script for bootstrapping the correlation, slope, and mean response is given below. Since we want to resample the observations  $(x_i, y_i)$ , we will resample the corresponding row numbers: that is, we will draw samples of size 24 (the number of skaters) with replacement from 1, 2, ..., 24 and store these in the vector `index`. The command `Skating2010[index, ]` creates a new data frame from the original with rows corresponding to the rows in `index` and keeping all the columns.

```
N <- 10^4
cor.boot <- numeric(N)
beta.boot <- numeric(N)
alpha.boot <- numeric(N)
yPred.boot <- numeric(N)
n <- nrow(Skating2010) # number of skaters = 24
for (i in 1:N)
{
  index <- sample(n, replace = TRUE) # sample from 1,2,...,n
  Skate.boot <- Skating2010[index, ] # resampled data

  cor.boot[i] <- cor(Skate.boot$Short, Skate.boot$Free)

  #recalculate linear model estimates
  skateBoot.lm <- lm(Free ~ Short, data = Skate.boot)
  alpha.boot[i] <- coef(skateBoot.lm)[1] # new intercept
  beta.boot[i] <- coef(skateBoot.lm)[2] # new slope
  yPred.boot[i] <- alpha.boot[i] + 60 * beta.boot[i]
}

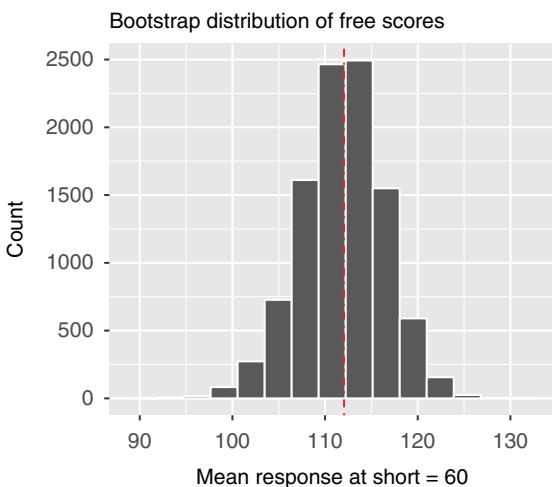
mean(cor.boot)
sd(cor.boot)
quantile(cor.boot, c(.025,.975))

observed <- cor(Skating2010$Short, Skating2010$Free)

df <- data.frame(cor.boot, beta.boot, alpha.boot, yPred.boot)
ggplot(df, aes(x = cor.boot)) +
  geom_histogram(bins = 20, color = "white") +
  geom_vline(xintercept = observed, color = "red", lty = 2)
```

The commands for the summaries and plots of the slope and response results are similar.

**Figure 9.18** Distribution of bootstrapped free program scores when the short program score is 60.



For a specific  $x$ , bootstrapping gives a percentile confidence interval for the expected value  $E[Y]$ .

Figure 9.18 shows the bootstrap distribution for the mean free skate program score corresponding to a short program score of 60. This distribution is centered at the original prediction and is roughly normally distributed, perhaps with a long left tail. The range of the middle 95% of the bootstrap lines (for a given  $x$ ) gives the percentile confidence interval for that  $x$ . For instance, for the skating data, a 95% bootstrap percentile confidence interval for  $E[Y]$  at  $x = 60$  is (102.5, 120.4): we are 95% confident that at  $x = 60$ , the corresponding mean  $Y$  value is between 102.5 and 120.4.

On the other hand, a prediction interval gives a range for an individual – for a male who scores 60 on the short program, a 95% prediction interval should have a 95% chance of containing the free program score for that individual. The algorithm for a prediction interval for a response at a given  $x$  is more involved since we need to take into account the variability of an individual, and the central limit theorem does not apply. See Davison and Hinkley (1997) for a way to compute prediction intervals.

### 9.5.1 Permutation Tests

To test whether there is a relationship between  $x$  and  $y$ , or whether they are independent, we turn to permutation tests. The procedure here is to create

a permutation sample by randomly permuting just one (not both) of the two variables and then computing a statistic such as correlation or slope.

### Permutation Test of Independence of Two Variables

Given a sample of size  $n$  from a population with two variables,

1. Draw a permutation resample of size  $n$  without replacement from one of the variables; keep the other variable in its original order.
2. Compute a statistic that measures the relationship, such as the correlation or slope.
3. Repeat this resampling process many times, say 9999.
4. Calculate the  $P$ -value.

For the skating scores, the  $P$ -values are essentially zero; the probability of random chance alone producing a correlation as strong as 0.84 is minuscule, so we conclude that the two scores are not independent.

### R Note

Script for testing to see whether the short program score and the free skate score are independent. We permute just one of the variables (`Short`) while leaving `Free` fixed.

```
N <- 9999
n <- nrow(Skating2010) # number of observations
result <- numeric(N)
observed <- cor(Skating2010$Short, Skating2010$Free)
for (i in 1:N)
{
  index <- sample(n, replace = FALSE)
  Short.permuted <- Skating2010$Short[index]
  result[i] <- cor(Short.permuted, Skating2010$Free)
}
(sum(observed <= result) + 1) / (N + 1) # P-value
```

### 9.5.2 Bootstrap Case Study: Bushmeat

Many species of wildlife are going extinct due to habitat loss, climate change, and hunting. Brashares et al. (2004) found evidence of a direct link between fish supply (in kg) and subsequent demand for bushmeat<sup>2</sup> in Ghana. Table 9.2 and

---

<sup>2</sup> From Wikipedia: bushmeat is meat from terrestrial wild animals, killed for subsistence or commercial purposes throughout the humid tropics of the Americas, Asia, and Africa.

**Table 9.2** Bushmeat: local supply of fish per capita and biomass of 41 species in nature preserves.

Year	Fish	Biomass	Year	Fish	Biomass	Year	Fish	Biomass
1970	28.6	942.54	1980	21.8	862.85	1990	25.9	529.41
1971	34.7	969.77	1981	20.8	815.67	1991	23.0	497.37
1972	39.3	999.45	1982	19.7	756.58	1992	27.1	476.86
1973	32.4	987.13	1983	20.8	725.27	1993	23.4	453.80
1974	31.8	976.31	1984	21.1	662.65	1994	18.9	402.70
1975	32.8	944.07	1985	21.3	625.97	1995	19.6	365.25
1976	38.4	979.37	1986	24.3	621.69	1996	25.3	326.02
1977	33.2	997.86	1987	27.4	589.83	1997	22.0	320.12
1978	29.7	994.85	1988	24.5	548.05	1998	21.0	296.49
1979	25.0	936.36	1989	25.2	524.88	1999	23.0	228.72

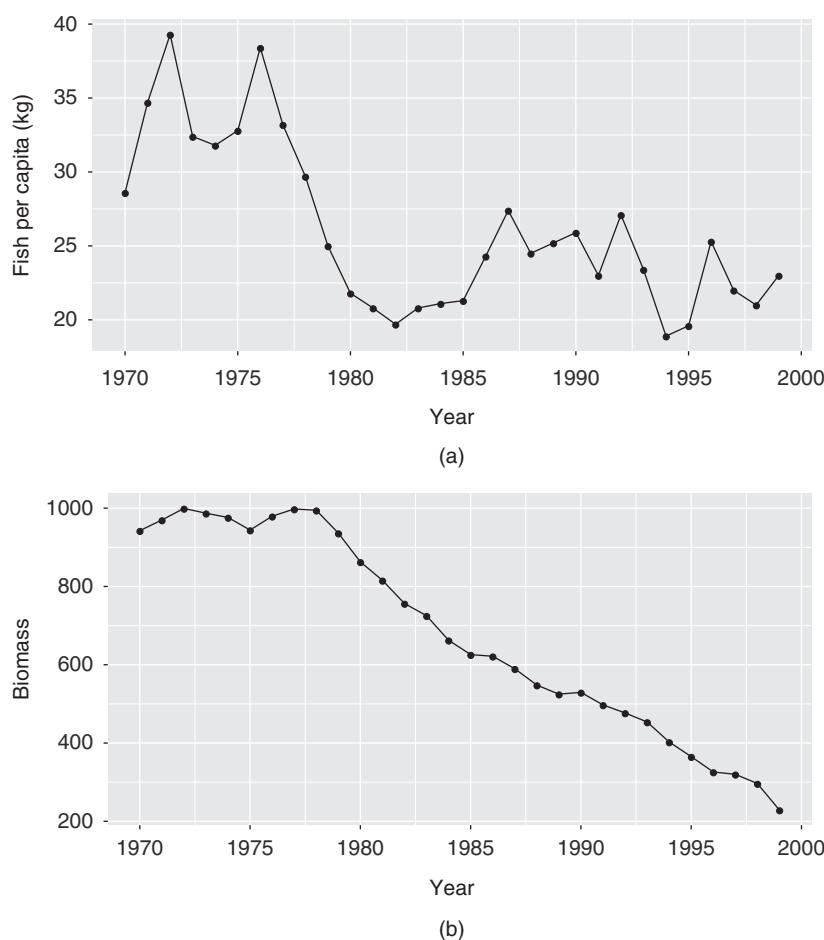
Figure 9.19 contain data from 30 years of local fish supply and biomass of 41 species in nature preserves.

There is a general decline in biomass over the study period, but a closer look suggests that the decline is steeper in years with a small supply of fish. Rather than looking at the biomass for each year, we look at the percentage change. In Figure 9.20, we observe a positive relationship between fish supply and percentage change in biomass. The correlation is 0.67 and a regression of percent change in biomass against fish supply gives a slope of 0.64, suggesting that each increase of 1 kg fish per capita results in 0.64% loss of biomass and that with sufficiently large fish supplies, estimated at 33.3 (the  $x$  intercept of the least-squares line), there would be no loss in biomass.

However, these are estimates based on a limited amount of data, so we turn to the bootstrap to assess variability.

Figure 9.21 shows two views of the bootstrap output. Figure 9.21a is a graphical bootstrap – for 40 bootstrap samples, we calculate the slopes and intercepts and draw the corresponding lines over the original data and original line. We notice that there appears to be a moderate amount of variability in the regression slopes, but not to the extent that any slopes are negative. There is also variability in the height of the regression lines, especially as we move to the right or the left. If we extrapolate to the left, all the way to zero fish, it would show the variability in the intercept  $\hat{\alpha}$ .

Figure 9.21a shows that the regression lines have the smallest variability in the middle; the farther one goes to either side, the less accurate the answers are. However, the smallest variance occurs not at  $\bar{x} = 26.1$ , as implied by Theorem 9.6, but farther right. This is because the condition of constant variance is violated. Looking back at the original scatter plot Figure 9.20,

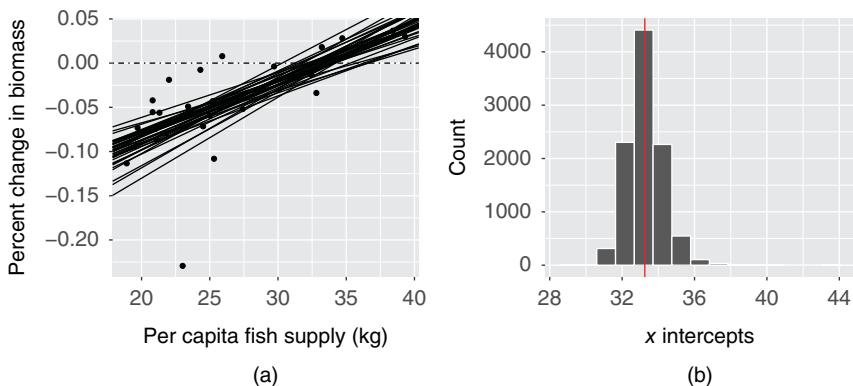
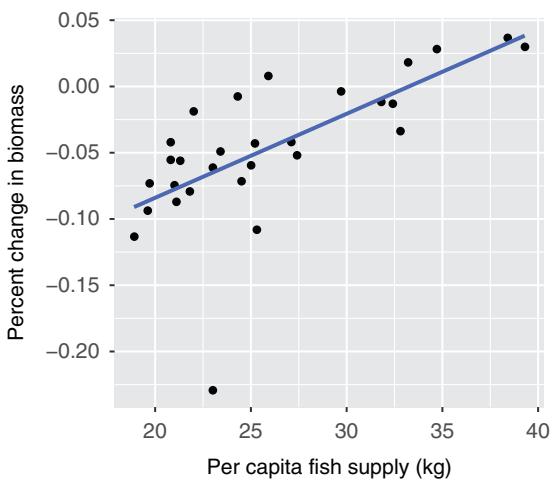


**Figure 9.19** Bushmeat data; (a) fish per capita and (b) biomass of 41 species of wildlife in nature preserves for 30 years.

we see that the residuals are smaller on the right. This is probably not just random variation, for two reasons, related to the numerator and denominator of the definition of relative change. When fish are more plentiful, there is less demand for bushmeat and correspondingly less variability in that demand; at the extreme, if there is zero demand, the variance is zero. Second, the observations on the left typically occur in later years when the denominator is smaller, hence, even constant variability in the numerator results in greater variability in the ratio.

The results suggest that increasing the fish supply would reduce bushmeat harvest. An important question is what level of fish would stop the loss of

**Figure 9.20** Scatter plot of percent change in biomass against fish supply with least-squares line imposed.



**Figure 9.21** (a) Regression lines from 40 bootstrap samples of the bushmeat data.  
(b) Bootstrap distribution of the  $x$  intercept.

wildlife? Based on the original regression line, the value would be 33.25 (the  $x$  intercept of the line). We can use the bootstrap to get an idea how accurate that number is.

We use more bootstrap samples to do this. We use only 40 samples for plotting the regression lines because otherwise the figure becomes a mass of black ink. But now, for better accuracy in estimating the intercept, we use  $10^4$  samples.

Figure 9.21b displays the bootstrap distribution of the  $x$  intercept, that is, the estimated supply of fish needed to stop the loss of wildlife. The original value of 33.25 falls in the middle of this distribution. The middle 95% range is 31.78 and 35.04, giving a rough idea of the reliability of the estimate. We are

95% confident that the supply of fish needed to forestall loss of biomass lies in that interval. Curiously, the interval  $(31.5, 35.43) = (33.25 - 1.75, 33.25 + 2.18)$  stretches farther to the right, which gives a pessimistic story – it takes a lot of fish to gain confidence on the positive side.

We must admit that the bootstrap we just did, sampling with replacement from the data, assumes that the original data are i.i.d. from a bivariate population. This condition is violated, because the data occur over time; they are neither independent nor identically distributed. The 22% drop in biomass in 1 year, for example, occurs in the final year, when the denominator is small, making a large change in either direction relatively easy. There are procedures intended for time series data, both bootstrap and formula based; these are more complicated and are beyond the scope of this book.

## 9.6 Logistic Regression

According to the Centers for Disease Control and Prevention, the leading cause of death for people under the age of 34 is motor vehicle-related injuries.<sup>3</sup> Since many of these accidents are due to impaired driving – driving while under the influence of alcohol or drugs – there is a lot of emphasis on educating young drivers on the dangers of combining drinking and driving. But is there evidence that in fatal accidents, drinking and age are linked?

The Fatality Analysis Reporting System (FARS) (<http://www.nhtsa.gov/FARS>) database contains data on all fatal traffic accidents in the United States, the District of Columbia and Puerto Rico since 1975. FARS is maintained by The National Center for Statistics and Analysis, part of the National Highway Traffic Safety Administration. We investigate the relationship between the involvement of alcohol and age of the driver in a random sample of 100 driver fatalities in 2009 in Pennsylvania; the drivers were driving a car, SUV, or light pickup truck (vehicles such as motor homes, convertibles, or commercial vehicles are excluded). One variable is a binary variable coded 1 if alcohol was involved and 0 otherwise; another variable is age of the driver, in years (Table 9.3).

Let  $Y_i$  denote the alcohol involvement variable and  $x_i$  the predictor variable age,  $i = 1, 2, \dots, 100$ . For these individuals, we assume that the  $Y_i$ 's are independent Bernoulli random variable with  $P(Y_i = 1) = p_i$ .

We want to understand the relationship between  $p_i = E[Y_i]$  and  $x_i$ . In Section 9.4, we considered linear regression of the form  $E[Y_i] = p_i = \alpha + \beta x_i$ .

---

<sup>3</sup> <http://www.cdc.gov/Motorvehiclesafety>.

**Table 9.3** Part of the data on driver fatalities in Pennsylvania.

ID	Alcohol	Age
1	0	86
2	0	38
3	0	40
4	0	20
5	1	27

But if  $\beta$  is nonzero, for sufficiently large and small  $x$  this would give probabilities less than 0 or greater than 1. Furthermore, linear regression assumes the same variances for every observation, but for Bernoulli data  $\text{Var}[Y_i] = p_i(1 - p_i)$ .

So linear regression is not appropriate for these data. In this section, we discuss a type of regression suitable for 0–1 data, *logistic regression*. We begin by introducing odds.

**Definition 9.5** Let  $p$  denote the probability of some event. The *odds* of the event is defined by  $p/(1 - p)$ . ||

For instance, if  $p = 0.8$  is the probability of a soccer team winning its next game, then  $0.8/(1 - 0.8) = 4$  is its odds of winning the next game: The odds of the team winning (to not winning) the next game is 4 to 1. If  $p = 0.25$  is the probability of dying from a certain disease, then  $0.25/0.75 = 0.33$ : the odds of dying (to not dying) is 1 to 3.

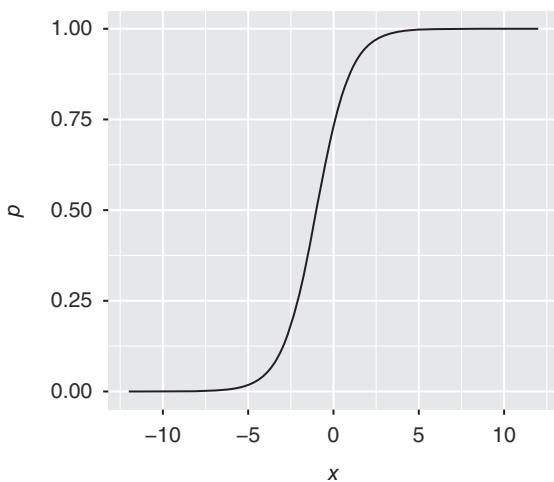
Let  $(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)$  be a set of ordered pairs, where  $x_1, x_2, \dots, x_n$  are fixed and  $Y_1, Y_2, \dots, Y_n$  are Bernoulli random variables with  $P(Y_i = 1) = p_i$ . In logistic regression, we model the logarithm of the odds as a linear function of  $x$ :

$$\ln\left(\frac{p_i}{1 - p_i}\right) = \alpha + \beta x_i \quad i = 1, 2, \dots, n. \quad (9.19)$$

Equivalently,

$$p_i = \frac{e^{\alpha + \beta x_i}}{1 + e^{\alpha + \beta x_i}}. \quad (9.20)$$

This gives an S-shaped relationship between  $x$  and  $E[Y]$  (see Figure 9.22). The predictions are bounded by  $0 < p_i < 1$ , which is appropriate for probabilities.



**Figure 9.22** Plot of a typical logistic curve, Equation (9.20).

The slope coefficient  $\beta$  describes how quickly the estimated probability increases; the maximum slope is  $\beta/4$  (where  $p = 0.5$ ). We can also interpret  $\beta$  by evaluating the odds at two different values, say  $x$  and  $x + \Delta x$ ,

$$\frac{p_1}{1-p_1} = e^{\alpha+\beta x},$$

$$\frac{p_2}{1-p_2} = e^{\alpha+\beta(x+\Delta x)}.$$

The *odds ratio* is

$$\frac{p_2/(1-p_2)}{p_1/(1-p_1)} = \frac{e^{\alpha+\beta(x+\Delta x)}}{e^{\alpha+\beta x}} = e^{\beta\Delta x}$$

and the *log odds ratio* is  $\beta\Delta x$ . So  $\beta$  measures how quickly the log odds ratio changes.

The parameters  $\alpha$  and  $\beta$  are estimated using maximum likelihood. The likelihood function is given by

$$L(\alpha, \beta) = \prod_{i=1}^n p_i^{Y_i} (1-p_i)^{1-Y_i},$$

then taking the logarithm, we find

$$\begin{aligned} \ln(L(\alpha, \beta)) &= \sum_{i=1}^n Y_i \ln(p_i) + (1 - Y_i) \ln(1 - p_i) \\ &= \sum_{i=1}^n Y_i \ln\left(\frac{p_i}{1-p_i}\right) + \ln(1-p_i). \end{aligned}$$

Setting the partial derivatives with respect to  $\alpha$  and  $\beta$  (using the chain rule, because  $p_i$  is a function of  $\alpha$  and  $\beta$ ) equal to 0 and simplifying yields equations:

$$\frac{\partial \ln(L)}{\partial \alpha} = \sum_{i=1}^n y_i - \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}} = \sum_{i=1}^n y_i - p_i = 0,$$

$$\frac{\partial \ln(L)}{\partial \beta} = \sum_{i=1}^n y_i x_i - \frac{e^{\alpha+\beta x_i} x_i}{1+e^{\alpha+\beta x_i}} = \sum_{i=1}^n (y_i - p_i) x_i = 0.$$

There is no closed-form solution to these two equations so a numerical algorithm must be used to find estimates  $\hat{\alpha}$  and  $\hat{\beta}$ . For instance, the `glm` function in R uses a procedure called iteratively reweighted least squares, a multivariate version of Newton's method for finding the zero of a function; for those who have taken linear algebra, it uses the gradient and Hessian of  $\ln L(\alpha, \beta)$ .

**Example 9.12** For the FARS fatalities data from Pennsylvania, the estimated logistic equation is

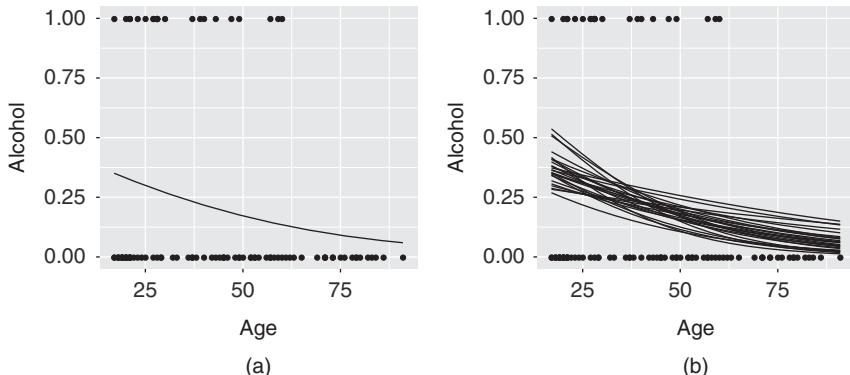
$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = -0.123 - 0.029x.$$

See Figure 9.23. For a 25-year-old driver, the estimated probability that alcohol was involved is

$$\hat{p} = \frac{e^{-0.123-0.029\times 25}}{1+e^{-0.123-0.029\times 25}} = 0.3$$

and the odds of alcohol being involved to not being involved is

$$\frac{\hat{p}}{1-\hat{p}} = e^{-0.123-0.029\times 25} = 0.428.$$



**Figure 9.23** (a) Plot of estimated probability of alcohol being involved against age of driver.  
(b) Estimates from 25 bootstrap samples (see Section 9.6.1).

Similarly, we find the odds of alcohol being involved for a 35-year-old driver,  $\hat{p}_2/(1 - \hat{p}_2) = 0.32$ . The odds ratio is  $0.428/0.32 = 1.34$ . The odds of alcohol being involved in the fatal accident of a 25-year-old driver is 1.34 times greater than the odds of alcohol being involved in the fatal accident of a 35-year-old driver. Equivalently, we can say that the odds of alcohol being involved in the fatal accident of a 25-year-old driver is 34% higher than the odds of alcohol being involved in the fatal accident of a 35-year-old driver.

### R Note

The FARS data are in a file called `Fatalities.csv`. In R, logistic regression is performed using the `glm` function. The syntax is similar to using `lm`, except that we must specify that the  $Y$  variable follows a binomial (Bernoulli) distribution:

```
> glm(Alcohol ~ Age, data = Fatalities, family = binomial)
...
Coefficients:
(Intercept)          Age
-0.12262        -0.02898

f <- function(x) exp(-0.123 - 0.029*x) / (1+exp(-0.123 - 0.029*x))

ggplot(Fatalities, aes(x = Age, y = Alcohol)) + geom_point() +
  stat_function(fun = f)
```

The `plogis` function computes the cumulative distribution function (cdf) for a logistic random variable and is a handy way to compute  $\exp(x)/(1 + \exp(x))$ . Thus, the function `f` above can be defined by

```
f <- function(x){plogis(-0.123 - 0.029*x)}
```

□

**Example 9.13** Suppose a hospital conducts a study to see if there is a link between patients getting an infection ( $y = 1$  if yes) and their length of stay in the hospital ( $x$ , in days). A logistic regression performed on their data gives

$$\ln\left(\frac{\hat{p}}{1 - \hat{p}}\right) = -1.942 + 0.023x.$$

How do the odds of getting an infection change for somebody who stays an additional week at this hospital?

### Solution

To compare the odds of infection for somebody who stays  $\Delta x = 7$  days longer than another patient, compute  $e^{0.023 \cdot 7} = 1.175$ . Thus, staying an additional week increases the odds of infection by about 17.5%. Alternatively, we can express

this result: the odds of infection are 1.175 times greater for every extra 7 days in the hospital.  $\square$

Logistic regression is a special case of a *generalized linear model*; another common version is *Poisson regression* in which  $Y$  has a Poisson distribution with mean given by  $\exp(\alpha + \beta x)$ ; see (Collett (2003); Dobson (2002); Kutner et al. (2005); McCullagh and Nelder (1989)) for details. One of us (Tim) consulted at Google on a project to predict web traffic for billions of search phrases based on time of day and day of week, using Poisson regression. Searches that receive more traffic than predicted may be flagged for *Google Trends*. For example, a trending search in 2021 in the United States was “How to help Afghan refugees.”

### 9.6.1 Inference for Logistic Regression

Standard errors in logistic regression are based on conditions similar to those in Section 9.4, that

- The  $x$  values are fixed, not random.
- The relationship between the  $x$  values and  $\log(p/(1 - p))$  is linear.
- The  $Y$  values have Bernoulli distributions, with parameters  $p_i$ .
- The  $Y$  values are independent.

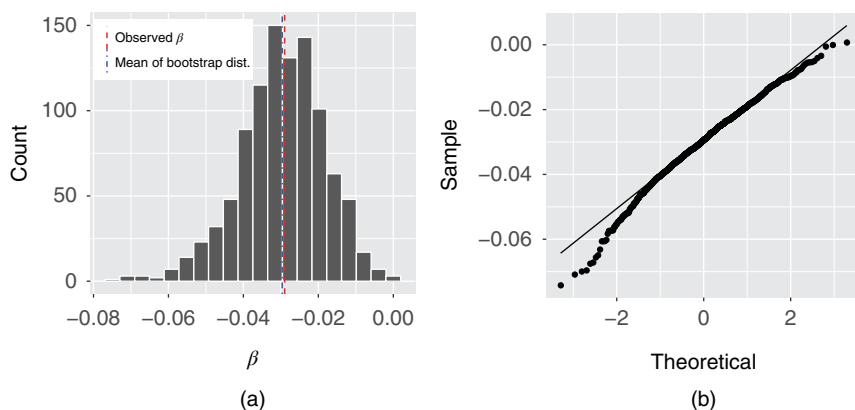
We rely on software to do the calculations, for standard errors for coefficients and predictions ( $\hat{p}_i$ ).

We can use the standard errors to produce confidence intervals,  $t$  statistics,  $P$ -values, and hypothesis tests, but be warned that sample sizes may need to be quite large before those are accurate. Some software calculates  $t$  statistics, but then admirably declines to print  $P$ -values based on those  $t$  statistics because the  $P$ -values cannot be trusted.

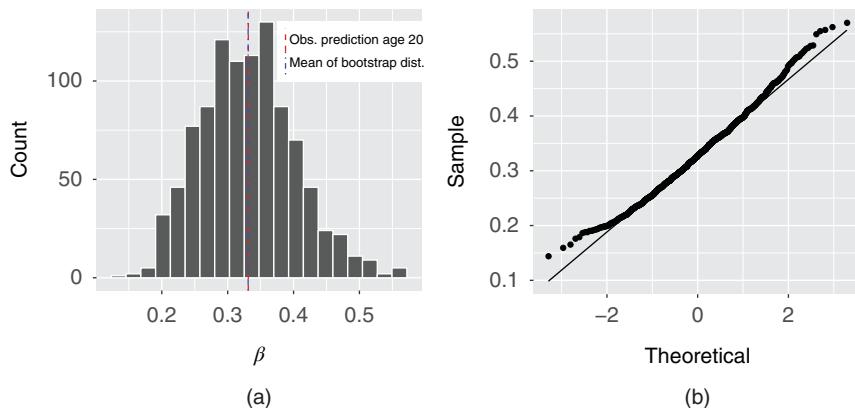
Alternately, we may bootstrap for standard errors and confidence intervals. We resample individuals, that is, resample paired values (age, alcohol). For each bootstrap data set, we estimate the logistic regression parameters, and calculate any desired predictions. Figure 9.23a shows the predictions from 25 bootstrap samples. This gives a rough idea of variability and suggests that some distributions are skewed – for example, predictions for the probability of alcohol involvement at age 80 are mostly near zero, but with a few larger values.

Figure 9.24 shows a histogram and normal quantile plot for  $b$  for 1000 bootstrap samples, and Figure 9.25 shows a histogram and normal quantile plot for the probabilities of alcohol involvement at age 20.

Bootstrap percentile confidence intervals are  $(-0.056, -0.008)$  for  $\beta$  and  $(0.20, 0.47)$  for the prediction of alcohol involvement at age 20. These are quite wide, a range of 20–47% for the probability, and a ratio of 7:1 for the slope. For comparison, we note that intervals based on the formula standard errors are similar:  $(-0.056, -0.002)$  for  $\beta$  and  $(0.187, 0.475)$  for the probability.



**Figure 9.24** (a) Histogram and (b) normal quantile plot for  $b$ , the logistic regression slope coefficient for alcohol involvement versus age of driver.



**Figure 9.25** (a) Histogram and (b) normal quantile plot for probability of involvement at age 20.

### R Note

The command to extract coefficients from a `glm` object is `coef`.

```
fit <- glm(Alcohol ~ Age, data = Fatalities,
            family = binomial)
data.class(fit) # is a "glm" object, so for help use:
help(glm)

fit           # prints the coefficients and other basic info
coef(fit)     # the coefficients as a vector
summary(fit) # gives standard errors for coefficients, etc.
```

```

# Full bootstrap - slope coeff. and prediction at age 20
N <- 10^3
n <- nrow(Fatalities) # number of observations
alpha.boot <- numeric(N)
beta.boot <- numeric(N)
pPred.boot <- numeric(N)

for (i in 1:N)
{
  index <- sample(n, replace = TRUE)
  Fatal.boot <- Fatalities[index, ] # resampled data

  fit.boot <- glm(Alcohol ~ Age, data = Fatal.boot,
                  family = binomial)
  alpha.boot[i] <- coef(fit.boot)[1] # new intercept
  beta.boot[i] <- coef(fit.boot)[2] # new slope
  pPred.boot[i] <- plogis(alpha.boot[i] + 20 * beta.boot[i])
}

quantile(beta.boot, c(.025, .975)) # 95% percentile CI
df <- data.frame(alpha.boot, beta.boot, pPred.boot)
ggplot(df, aes(x = beta.boot)) +
  geom_histogram(bins = 20, color = "white")
ggplot(df, aes(sample = beta.boot)) + geom_qq() + geom_qq_line()

```

Repeat for the intercept and prediction.

The predict function can also be used to give predicted values. The required arguments of predict are the model object (the output from the glm function) and a data frame containing the values of the explanatory variable at which you wish to predict. By default, predict returns the predicted  $a + bx$ . To obtain the predicted probabilities, provide the argument type = "response".

For example, to predict the probability of a fatal accident for somebody of age 20,

```

df <- data.frame(Age = 20)
predict(fit, newdata = df, type = "response")

#compare to
plogis(-.123 - .029*20)

```

## Exercises

- 9.1** Let  $X$  and  $Y$  be random variables with joint probability density function given by

$$f(x, y) = \begin{cases} \frac{6}{5}(x + y^2), & 0 \leq x \leq 1, 0 \leq y \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

Find the covariance  $\text{Cov}[X, Y]$ .

- 9.2** Let  $X$  be a random variable with mean  $\mu$  and variance  $\sigma^2$ . Let  $Y = C$ , a constant. Find the covariance between  $X$  and  $Y$ .
- 9.3** Let  $X$  and  $Y$  be random variables with  $\text{Var}[X] = 4$ ,  $\text{Var}[Y] = 9$  and  $\text{Cov}[X, Y] = 3$ . Find the variance of  $2X + 3Y$ .
- 9.4** Let  $X$  and  $Y$  be random variables with  $\text{Var}[X] = 5$ ,  $\text{Var}[Y] = 7$  and  $\text{Cov}[X, Y] = 2$ . Find the variance of  $2X - 5Y$ .
- 9.5** Let  $X, Y$ , and  $Z$  be random variables with  $\text{Var}[X] = 3$ ,  $\text{Var}[Y] = 4$ ,  $\text{Var}[Z] = 24$ , and  $\text{Cov}[X, Y] = -2$ ,  $\text{Cov}[X, Z] = -4$  and  $\text{Cov}[Y, Z] = 7$ . Find  $\text{Var}[5X - Y + 2Z]$ .
- 9.6** Import the data set `Olympics2012` (see Exercise 7.11).
  - Find the correlation between weight and height.
  - Create a scatter plot of height against weight. What do you observe?
  - Remove the outliers and recompute the correlation. Are the outliers influential?

- 9.7** In R, type the following commands:

```
x <- seq(-2, 2, length = 10) #10 points from -2 to 2
y <- x^2
plot(y ~ x)
```

- Are  $x$  and  $y$  related?
  - Find the correlation between  $x$  and  $y$ . Does a correlation of 0 indicate that there is no relationship between  $x$  and  $y$ ?
- 9.8** The data set `RangersTwins2016` has information on baseball players (excluding pitchers) who played at least 50 games in 2016 for the Texas Rangers or the Minnesota Twins. The Rangers had the best winning percentage in the American League (0.586), while the Twins had the worst (0.364).
  - Create a scatter plot of runs batted in (RBI) against batting average (BA), and then find the correlation between these two variables.
  - Find the correlation between RBI and BA for each team and compare the results to part (a).

Use the `group_by` and `summarize` functions in the `dplyr` package to get the correlations for each team.

```
RangersTwins2016 %>% group_by(Team) %>%
  summarize(corr = cor(BA, RBI))
```

- 9.9** The data set `NBA1617` has information on basketball players from four National Basketball Association (NBA) teams who played a minimum of 100 min during the 2016–2017 season.

- Find the correlation between field goal percentage (`PercFG`) and offensive rebounds (`OffReb`).
- Find the mean of `PercFG` for each team; find the mean of `OffReb` for each team.
- Create a scatter plot of the means of the field goal percentage against the means of the offensive rebounds and then compute the correlation. Compare to (a).

*Ecological correlation:* Correlations based on rates or groups are often higher than correlations based on individuals. This is a common problem in the social/behavior sciences where many data sets are based on summaries (e.g. census data for the 50 states: mean income levels, mean literacy rate, etc.).

- 9.10** Compare the round-off error of two ways of computing sample variances. Write functions that compute  $(\text{mean}(x^2) - \text{mean}(x)^2) * n / (n-1)$  and  $\text{mean}((x - \text{mean}(x))^2) * n / (n-1)$  and calculate the variance of  $x_1 = c, x_2 = c + 1, x_3 = c + 2$  for  $c = 0, c = 10^5, c = 10^6, \dots$  using both. What do you find? Compare to the R `var` function.

- 9.11** Verify that linear regression residuals satisfy  $\sum_{i=1}^n (y_i - \hat{y}_i) = 0$  – that is, the sum of the residuals is 0. *Hint:* Use the fact that  $a$  and  $b$  are solutions to  $\partial g / \partial a = 0$ , where  $g(a, b) = \sum_{i=1}^n (y_i - (a + bx_i))^2$ .

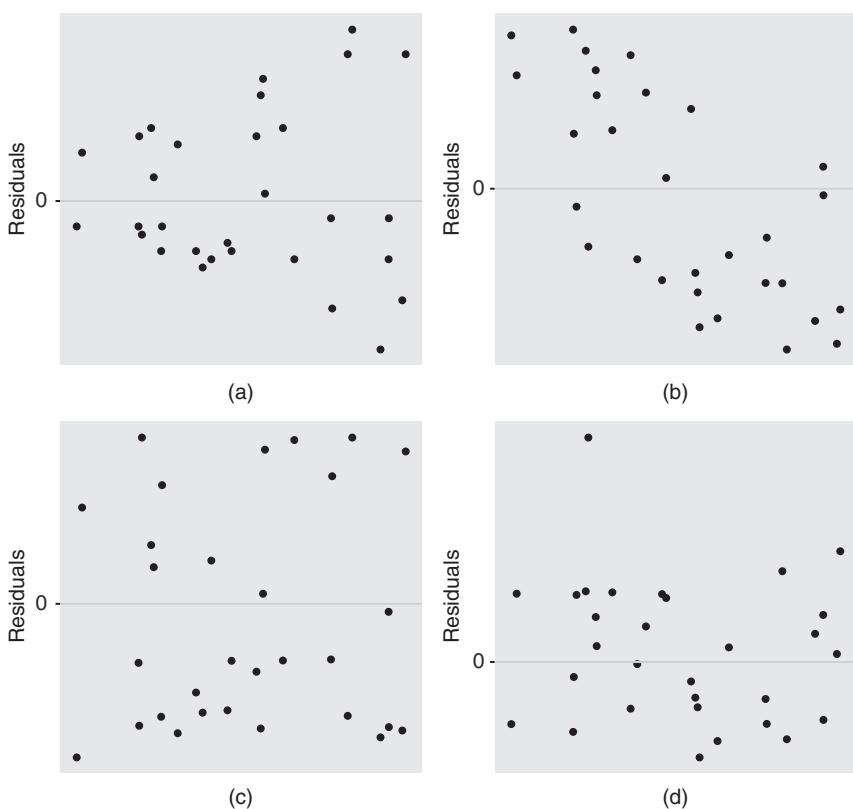
- 9.12** Let  $s_y$  denote the standard deviation of the predicted  $y$ 's; that is, the standard deviation of  $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n$ .
- Verify that  $s_y = |r|s_e$  (Equation (9.13)).
  - Let  $s_e$  denote the standard deviation of the residuals,  $e_i = y_i - \hat{y}_i, i = 1, 2, \dots, n$ . Show that

$$s_e = \sqrt{1 - r^2}s_y.$$

- 9.13** Suppose the height and weight of 30 girls in Sodor are measured. The mean and standard deviation of the heights is 46 and 7 in., respectively, and the mean and standard deviation of the weights are 94 and 15 lb, respectively. Suppose the correlation between height and weight is 0.75.
- Find the equation of the least-squares regression line of weight against height.
  - Find the predicted weight of a girl who is 5 ft tall.
  - Find  $R$ -squared and give a sentence interpreting this statistic.
- 9.14** Import the data set `Quetzal` (refer to Exercise 8.10). The biologists also studied the relationship between the heights of the nests and the snags.
- Create a scatter plot of the nest heights against the snag heights and describe the relationship.
  - Find the equation of the least-squares regression line of nest height against snag height. Give a sentence interpreting the slope.
  - What is the predicted height of a resplendent quetzal's nest that is in a snag of height 8 m?
  - Find  $R$ -squared and give a sentence interpreting this statistic.
- 9.15** (Exercise 9.14 continued)
- Compute a 95% confidence interval for the true slope  $\beta$ .
  - Use the bootstrap to find a 95% confidence interval for the slope.
- 9.16** Return to the beer and hot wings case study in Section 1.9.
- Create a scatter plot of beer consumed against hot wings eaten and find the correlation between these two variables.
  - Find the equation of the least-squares regression line (take `Hotwings` as the predictor variable) and give a sentence interpreting the slope.
  - Compute  $R$ -squared and state the interpretation of this statistic.
- 9.17** Figure 9.26 contains residual plot for several least-squares regression models. Describe the unusual features.
- 9.18** Is there a relationship between female literacy and birth rate? The data set `Illiteracy`<sup>4</sup> contains data on a sample of countries where female illiteracy is more than 5%. The variable `Illit` is the percentage of women over 15 years old who are illiterate (2003) and the variable `Births` is the number of births per woman in that country (2005).

---

<sup>4</sup> [www.unesco.org](http://www.unesco.org), [data.worldbank.org](http://data.worldbank.org).



**Figure 9.26** Residuals plot for regression models.

- (a) Create a scatter plot of birth rate against illiteracy and comment on the relationship.
  - (b) Find the equation of the least-squares line and interpret the slope and  $r^2$ .
  - (c) Create a residual plot and comment on the appropriateness of a straight line model.
  - (d) Can we say that improving literacy (reducing illiteracy) will cause the birth rate to go down? Explain.
- 9.19** The data set `Volleyball2009`<sup>5</sup> contains data on 30 Division I women volleyball teams from the 2009 season.<sup>6</sup>

<sup>5</sup> © 2009 American Volleyball Coaches Association, © 2009 National Collegiate Athletic Association.

<sup>6</sup> <http://www.ncaa.org/championships/statistics/womens-volleyball-statistics>.

- (a) Create a scatter plot of the number of kills per set (`Kills`) against assists per set (`Assists`) and describe the relationship.
- (b) Find the least-squares equation for the line and interpret the slope and  $r^2$ .
- (c) Create the residual plot and comment on the appropriateness of a straight line model.
- 9.20** Refer to the previous exercise, Exercise 9.19.
- Find a 95% confidence interval for the slope.
  - Suppose a team not listed records 12.4 assists per set. Find the predicted number of kills per set and a 95% confidence interval.
- 9.21** Return to the North Carolina births case study in Section 1.2.
- Create a scatter plot of weight against gestation period and compute the correlation between the two variables.
  - Find the least-squares regression line.
  - Give a sentence with the interpretation of the slope and the  $R$ -squared.
  - Create the residual plot and comment on any unusual features. Is a linear model appropriate for the relationship between gestation period and weight?
- 9.22** North Carolina Births (continued)
- What is the estimate of  $\sigma$  for the linear model of weight against gestation period?
  - Find a 95% confidence interval for the true slope  $\beta$ .
- 9.23** In the ice cream example (Example 9.6), the sum of squares for the sugar variable ( $ss_x$ ) is 324.446, and the residual standard error is 2.838. Find a 95% confidence interval for the true slope (use the model including all the data).
- 9.24** Suppose a company offers tutoring to students interested in improving their math SAT scores. From a random sample of 100 previous customers, they find that  $\text{Score} = 502.7 + 1.45 \cdot \text{Hours}$ , where `Score` is the SAT math score and `Hours` is the number of hours the student was tutored. The  $R$ -squared and residual standard error for this model are 0.855 and 16.54, respectively.
- For every additional 10 h of tutoring, what is the corresponding change in the test score?
  - If the standard deviation of the hours variable is 27.5, what is the standard deviation of the score variable?
  - Find a 95% confidence interval for the true slope.

- (d) Find a 95% confidence interval for the mean score for students who are tutored 50 h. Assume the mean number of hours tutored is 55 h.
- 9.25** The Mauna Loa Observatory located near the Hawaiian volcano Mauna Loa, specializes in research in the atmospheric sciences. The facility has been collecting data on carbon dioxide levels since the 1950s. The data file `Maunaloa` contains average CO<sub>2</sub> levels (ppm) for the month of May from 1990 to 2010.<sup>7</sup>
- (a) Create a scatter plot of CO<sub>2</sub> levels against year and describe the relationship.
  - (b) Find the equation for the least-squares regression line.
  - (c) Plot the residuals against year. Is a straight line model appropriate? Discuss.
- 9.26** The data set `Walleye` from the Minnesota Pollution Control Agency contains length (inches) and weight (pound) measurements for a sample of 60 walleye caught in Minnesota Lakes during the 1990s (B. Monson, private communication).
- (a) Create a scatter plot of weight against length. Does the relationship appear linear?
- In fact, biologists have determined that the relationship between length and weight of fish is given by  $W = aL^b$ , where  $a$  and  $b$  depend on the species (Ricker (1973, 1975)). We will consider  $\log(W) = \log(a) + b \log(L)$  (base 10).
- (b) Transform the weight and length variables by log base 10 and then create a scatter plot of  $\log(\text{Weight})$  against  $\log(\text{Length})$ . Describe the relationship.
  - (c) Use least-squares to find estimates of  $a$  and  $b$  based on this sample.
  - (d) What is the 95% confidence interval for  $b$ ?
- 9.27** Import the data set `Aleelager` that contains alcohol content (per volume) and calories for a sample of beers (12 oz). Find the correlation between alcohol and calories and then compute a 95% bootstrap percentile confidence interval for the true correlation.
- 9.28** Using the data set `Illiteracy`,
- (a) Find the correlation between illiteracy rates and births. Use the bootstrap to find a 95% bootstrap percentile interval.
  - (b) Use a permutation test to test whether illiteracy rates and births are independent.

---

<sup>7</sup> <https://www.esrl.noaa.gov/gmd/ccgg/trends>.

- 9.29** The data set `Eyes` contains measurements (mm) of the pupillary distances (PD) for each eye on a sample of volunteers: this is the distance from the pupil of the eye to the middle of the bridge of the nose. Is there a relationship between age and the PD?
- Create a scatter plot of (right) PD against age and describe the relationship.
  - Find the correlation between the PD measurement of the right eye and the age of the volunteer.
  - Compute a 95% bootstrap confidence interval for the correlation. What does this suggest about the relationship between PD measurements and age?
  - Conduct a permutation test to see if PD measurements and age are independent.
- 9.30** Exercise 5.15 describes a study of school-age children to examine gender differences in math anxiety. In addition to completing the Abbreviated Math Anxiety Scale (AMAS), the students took an arithmetic test. The data are in `MathAnxiety`.
- For the girls only, find the correlation between the scores on the arithmetic test (`Arith`) and the scores on the AMAS.
  - Use the bootstrap to find a 95% confidence interval for the correlation. Is it plausible that there is no correlation between the arithmetic score and the AMAS?
  - Repeat the above for the boys.
- 9.31** Exercise 5.15 describes a study of school-age children to examine gender differences in math anxiety. In addition to completing the AMAS questionnaire, the students also completed the Revised Children's Manifest Anxiety Scale: Second Edition (RCMAS) which measures general anxiety.
- Create a scatter plot of the AMAS score against the RCMAS score for the girls. Is the relationship linear?
  - Find the equation of the least-squares regression line.
  - Use the bootstrap to find a 95% confidence interval for the slope.
  - Repeat the above for the boys.
- 9.32** Prove Proposition 9.2
- 9.33** Prove the second part of Proposition 9.4.
- 9.34** Prove Proposition 9.5.
- 9.35** Prove the results about  $\bar{Y}$  in Theorem 9.6.

- 9.36** In Theorem 9.3, we stated without proof that  $\hat{\beta}$  and  $\bar{Y}$  are independent. Instead, prove that they are uncorrelated.
- 9.37** A campaign manager conducts a survey to gauge voter support for his candidate Senator Lopez. He gathers data on the age of a registered voter ( $x$ ) and whether this person supports Senator Lopez ( $Y = 1$ ) or somebody else ( $Y = 0$ ). An analysis yields the following logistic equation:
- $$\ln(\hat{p}/(1 - \hat{p})) = -0.324 + 0.012x,$$
- where  $p$  is the probability of a vote for Lopez.
- Find the estimated probability that a 21-year-old will vote for Senator Lopez.
  - Compare the odds of support for Senator Lopez between two people who are 10 years apart in age and give your answer in a complete sentence.
  - At what age is the expected response equal to 0.5?
- 9.38** On January 28, 1986, the space shuttle Challenger exploded during lift-off, killing all seven astronauts aboard.<sup>8</sup> In the follow-up investigation, attention was focused on the rubber O-rings that sealed the booster rockets. Engineers had concerns earlier that the ambient temperature at the time of lift-off could affect the integrity of the O-ring. The data set Challenger contains data on 23 Challenger flights before the January 21 flight. The binary variable Incident records 1 if one of the O-rings on one of the booster rockets was damaged on this flight. The variable Temperature records the temperature (°F) at the time of lift-off.
- Find the logistic regression equation modeling the log-odds of an O-ring incident against temperature. Plot the graph of the predicted probabilities against temperature and add the observed incidents also.
  - How does a 10° decrease in temperature affect the odds of an incident? State your answer in a complete sentence.
  - The day of the Challenger accident, the temperature was 33°. What is the predicted probability of an O-ring incident?
  - Some would argue that it is not appropriate to use this model to predict an O-ring incident at 33°. Why not?
- 9.39** The biologist in the black spruce case study in Section 1.10 was also interested in the relationship between seedling growth and water table depth. The data set Watertable contains data on a sample of the seedlings. The variable binary Alive indicates 1 if the seedling was alive

<sup>8</sup> <http://history.nasa.gov/sts51l.html>.

at the end of the second year of the study and 0 otherwise. The variable `Depth` gives the depth of the water table (cm).

- (a) Find the logistic equation modeling the log-odds of a seedling being alive against water table depth. Plot the graph of the predicted probabilities against depth and add the observed data points.
- (b) Interpret the slope of the regression equation (in terms of odds).
- (c) For a seedling growing in soil with a water table depth of 15 cm, find the predicted probability of being alive at the end of the second year.

**9.40** Import the data set `Phillies2009`, which contains data for the Philadelphia Phillies baseball team.

- (a) Find the logistic equation modeling the log-odds of the team winning (`Outcome2`) against the number of hits in a game (`Hits`). The variable `Outcome2` is 1 if the team won the game and 0 if the team lost.
- (b) Interpret the slope of the equation (in terms of odds).
- (c) Find a 95% bootstrap percentile interval for the slope.
- (d) Predict the probability of winning if the team has 17 hits and then find a 95% bootstrap percentile interval.
- (e) For inference, we need to assume that the observations are independent. Is that condition met here?

**9.41** The data set `Titanic` contains information about the male passengers on the *Titanic*. The variable `Survived` is 1 if the passenger survived the sinking and 0 if the passenger died.

- (a) Find the logistic equation modeling the log-odds of a male passenger surviving against age.
- (b) Compare the odds of survival for a 30-year-old to a 40-year-old.
- (c) Find a 95% bootstrap percentile interval for the slope.
- (d) Estimate the probability of a 69-year-old male surviving, and find a 95% bootstrap percentile interval for the probability.

**9.42** Verify that Equation (9.20) is equivalent to  $p_i = 1 / (1 + e^{-(\alpha + \beta x)})$ .

# 10

## Categorical Data

One question in the General Social Survey (GSS) case study in Section 1.7 is whether someone favors or opposes the death penalty for murder. A  $5 \times 2$  contingency table summarizing the responses by degree (education level) is given in Table 10.1.<sup>1</sup>

The percentage of people who favor the death penalty varies by degree, from 54% to 71%. Can these differences easily be explained by chance variation, or do these data suggest that support for the death penalty depends on education?

In Chapters 3 and 8, we learned ways to compare two proportions, but here there are five proportions. In this chapter, we will learn methods to handle contingency tables with more than two rows and columns and other categorical data.

### 10.1 Independence in Contingency Tables

To analyze two categorical variables with  $I$  and  $J$  groups respectively, we first summarize the data into an  $I \times J$  contingency table, as in Table 10.1, with row and column sums to see the distribution of each variable. We then look at row or column percentages, to get a better idea of how the variables are related. Table 10.1 includes one of the row percentages, the percent who favor; the other is redundant.

An important question for contingency tables is whether the two variables are independent. Here, if opinion and degree are independent, we would expect about the same percentage of each education level to favor the death penalty. Of the 232 people with graduate degrees, we would expect 63.2% to favor the death penalty and 36.8% to oppose, the same proportions as the whole sample. Similarly for the other cells.

<sup>1</sup> There are only 2193 respondents here compared to the total 2348 originally interviewed because many people did not respond to the death penalty question. They have been removed from this analysis.

**Table 10.1** Counts of death penalty opinions grouped by degree.

Degree	Death penalty for murder?			% Favor
	Favor	Oppose	Row sum	
Less than high school	142	99	241	58.9
High school	759	345	1104	68.8
Junior college	129	52	181	71.3
Bachelor	235	200	435	54.0
Graduate	120	112	232	51.7
Column sum	1385	808	$n = 2193$	63.2

To test for independence between two categorical variables, we need a test statistic and a reference distribution. If there were just two groups, we could proceed as in Chapters 3 and 8 and use the difference in proportions as a statistic. Here we need a statistic that takes into account all differences between groups.

We could define a test statistic in terms of differences in proportions (see Exercise 10.28), but it is simpler to work with counts. Let  $N_{ij}$  be the observed number of people in row  $i$  and column  $j$ , let  $R_i$  be the total for row  $i$ ,  $C_j$  the total for column  $j$ , and  $n$  the overall total. If rows and columns are independent, then the probability of falling in a cell is the probability of the row times the probability of the column; we estimate this by  $(R_i/n)(C_j/n)$ , and multiply by the sample size to get the *expected count*  $E_{ij}$  for the cell:

$$E_{ij} = n(R_i/n)(C_j/n) = R_i C_j / n.$$

For instance, the expected count for the  $(3, 2)$ -cell is  $181 \times 808 / 2193 = 66.689$ . The expected counts, and differences between observed and expected, are shown in Table 10.2. Even though the percentage who favor was greatest for the junior college group, the difference is greatest for the high school

**Table 10.2** Expected counts of death penalty opinions grouped by degree.

Degree	Death penalty?		Observed–Expected	
	Favor	Oppose	$O - E$	$O - E$
Less than high school	152.2	88.8	-10.2	10.2
High school	697.2	406.8	61.8	-61.8
Junior college	114.3	66.7	14.7	-14.7
Bachelors	274.7	160.3	-39.7	39.7
Graduate	146.5	85.5	-26.5	26.5

group; the junior college group is so small that the higher percentage does not correspond to many extra people.

It seems intuitive that the observed and expected counts should be similar if education and support for the death penalty are independent. Thus, it seems plausible to look for a test statistic that takes into account the differences *observed count – expected count* for all cells. But just adding the differences does not work; the differences always add to zero (check this!). And both positive and negative differences should contribute to the test statistic in the same way – both large positive and large negative differences suggest dependence. One idea is to let the test statistic be the sum of absolute values of the differences, or the sum of squared differences. These are reasonable test statistics, but are not ideal – they do not take size into account. A difference of say 20 in a cell with an expected count of 10 matters more than a difference of 20 in a cell with an expected count of 2000.

There is a standard test statistic in this setting that does take cell size into account. In the 1900's, Karl Pearson proposed the statistic

$$C = \sum_{\text{all cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \sum (O - E)^2 / E. \quad (10.1)$$

This combines all the differences in an appropriate way, and has other nice properties we will see later. This statistic is common enough to merit a name, the *chi-square test statistic*.

For the GSS example, the value of the statistic is

$$\begin{aligned} c &= \frac{(142 - 152.2)^2}{152.2} + \frac{(99 - 88.8)^2}{88.8} + \frac{(759 - 697.2)^2}{697.2} + \frac{(345 - 406.8)^2}{406.8} \\ &\quad + \frac{(129 - 114.3)^2}{114.3} + \frac{(52 - 66.7)^2}{66.7} + \frac{(235 - 274.7)^2}{274.7} + \frac{(200 - 160.3)^2}{160.3} \\ &\quad + \frac{(120 - 146.5)^2}{146.5} + \frac{(112 - 85.5)^2}{85.5} \\ &= 50.429. \end{aligned}$$

Next, we need a reference distribution to judge how extreme the observed test statistic is. We will consider three approaches – permutation testing, chi-square distributions, and exact calculations.

## 10.2 Permutation Test of Independence

To do permutation testing, we start with the GSS2018 data set with two columns, degree and opinion on the death penalty. If the null hypothesis that degree and opinion are independent is correct, then we could permute the values on one column while keeping the other column fixed, and any

permutation would be equally likely (this is the same procedure illustrated in Table 3.3). For each such permutation resample, we can cross-tabulate to obtain a contingency table and compute the chi-square test statistic for this table. Note that for every resample, the row and column totals of the contingency table are the same, as are the expected values; only the cells in the table change. By forming many such resamples and computing the corresponding chi-square test statistic, we obtain the permutation distribution of the chi-square test statistic. We follow this algorithm:

#### Permutation Test of Independence of Two Categorical Variables

Store the data with one row per observation, and one column per variable  
 Calculate a test statistic for the original data. Normally large values of the test statistic suggest dependence.

**repeat**

Randomly permute the rows in one of the columns  
 Create a contingency table for the resampled data  
 Calculate the test statistic for the new contingency table

**until** we have enough samples

Calculate the  $P$ -value as the fraction of times the random statistics exceed the original statistic.

Optionally, plot a histogram of the resampled statistic values.

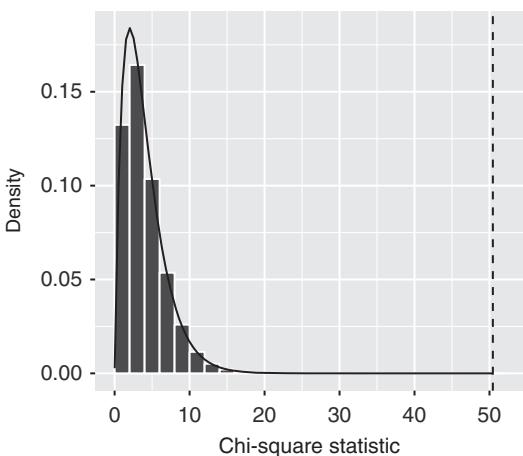
For instance, in the GSS2018 data set, one permutation of the values in the DeathPenalty column while leaving the Degree column fixed results in the contingency table Table 10.3.

The corresponding chi-square statistic (from Equation (10.1)) is  $c = 50.449$ . Repeating this permutation many times and computing the chi-square statistic each time gives the permutation distribution of this statistic, shown in Figure 10.1. The estimated  $P$ -value, based on  $10^5 - 1$  replications, is 0.00001,

**Table 10.3** Contingency table after permuting DeathPenalty column.

Degree	Death penalty?	
	Favor	Oppose
Less than high school	147	94
High school	687	417
Junior college	119	62
Bachelors	280	155
Graduate	152	80

**Figure 10.1** Null distribution for chi-square statistic for death penalty opinions; the overlaid density is a chi-square distribution with 4 degrees of freedom.



near zero. Thus, we conclude that there is an association between the education level of a person and his/her support for the death penalty.

The GSS draws a random sample of participants for its surveys, so we can draw an inference to a population. However, in this example, we did exclude all people who did not provide a response to the death penalty question (155 people). Do you think this affects the results?

The permutation test of independence of two categorical variables when the corresponding contingency table is larger than  $2 \times 2$  is implemented in R by the function `chisq.test`, by specifying `simulate.p.value=TRUE` and the number of resamples, say  $B=10^5-1$  (the default of 2000 is too low).

We have also provided a different implementation of the R code at our website <https://github.com/lchihara/MathStatsResamplingR>.

### R Note

The `chisq.test` function accepts raw data as well as contingency tables.

```
chisq.test(GSS2018$Degree, GSS2018$DeathPenalty,
           simulate.p.value = TRUE, B = 10^5-1)
mat <- table(GSS2018$Degree, GSS2018$DeathPenalty)
chisq.test(mat, simulate.p.value = TRUE, B = 10^5-1)
```

The `chisq.test` function also removes rows with missing values.

## 10.3 Chi-Square Test of Independence

The permutation test approach is easy now, with fast computers, but was not easy in Pearson's day. Work in the 1920s by Pearson (1900, 1922, 1923) and

Fisher (1922) led to a shortcut—if the expected counts are all reasonably large, then the null distribution of the chi-square test statistic is approximately equal to a chi-square distribution, with  $(I - 1)(J - 1) = 4$  degrees of freedom, where  $I$  and  $J$  are the number of rows and columns, respectively. This is apparent in Figure 10.1 where the overlaid chi-square distribution is a close match to the permutation distribution.

If the chi-square statistic has a chi-square distribution with 4 degrees of freedom, then the  $P$ -value, the probability of exceeding  $c = 50.449$ , is  $2.8 \times 10^{-10}$ . We conclude that education level and support for the death penalty are not independent.

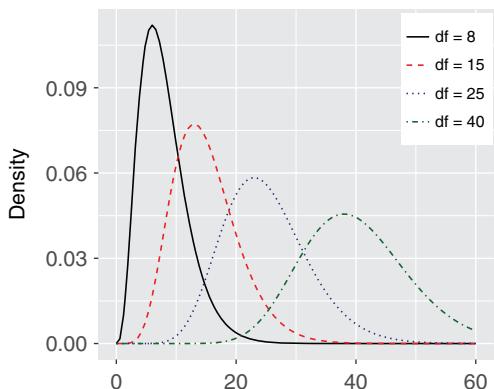
The number of zeros in these  $P$ -values does not matter; neither the formula nor the permutation test are particularly accurate this far in the tail. In any case, the  $P$ -values are small enough that dependence is discernible.

This reference distribution—the chi-square distribution—is described more fully in Section B.10. The pdf for a chi-square random variable is given by

$$f(x) = \frac{x^{m/2-1} e^{-x/2}}{2^{m/2} \Gamma(m/2)}, \quad 0 \leq x < \infty.$$

Chi-square distributions have a parameter called the *degrees of freedom*, written here as  $m$ . We write  $X \sim \chi_m^2$  to denote chi-square random variables. Some are shown in Figure 10.2. The mean and variance are proportional to  $m$ :  $E[X] = m$  and  $Var[X] = 2m$ , so a chi-square statistic must be larger to be statistically discernible when the degrees of freedom is larger.

The name “degrees of freedom” corresponds to how many extra parameters can vary freely under the alternative hypothesis, compared to the null hypothesis. Here, once the top left  $(I - 1)(J - 1)$  are determined the others are constrained by the row and column totals.



**Figure 10.2** Densities for the chi-square distribution.

**R Note**

The `pchisq` function computes cumulative probabilities for the chi-square distribution: `pchisq(x, m)` gives  $P(\chi_m^2 \leq x)$ .

For the death penalty example above,

```
> 1 - pchisq(50.449, 4)
[1] 2.909686e-10
```

### 10.3.1 Model for Chi-Square Test of Independence

Here, we describe the chi-square test of independence for a two-way table more formally.

Consider two categorical variables  $A$  and  $B$  with  $I$  and  $J$  levels, respectively. We use  $A_1, A_2, \dots, A_I, B_1, B_2, \dots, B_J$  to denote the levels of  $A$  and  $B$ , respectively. For example, in the GSS case study, the education variable has  $I = 5$  levels and the death penalty variable has  $J = 2$  levels.  $A_1$  would be those who have a Graduate degree while  $B_1$  would be those who favor the death penalty.

From a sample of  $n$  randomly selected individuals, let the random variable  $N_{ij}$  denote the number classified by the  $i$ th level of  $A$  and the  $j$ th level of  $B$ , respectively, with observed value  $n_{ij}$ . Also, let  $p_{ij}$  denote the probability that a randomly selected individual from some population is classified by the  $i$ th level of  $A$ , and the  $j$ th level of  $B$ . The  $N_{ij}$  are multinomial random variables with  $E[N_{ij}] = np_{ij}$ . (See Section B.2 for more information on the multinomial Distribution.)

The row and column sums of the contingency table (Table 10.4) are

$$R_i = \sum_{j=1}^J N_{ij} \quad \text{and} \quad C_j = \sum_{i=1}^I N_{ij}.$$

In addition, let

$$p_{i\cdot} = \sum_{j=1}^J p_{ij} \quad \text{and} \quad p_{\cdot j} = \sum_{i=1}^I p_{ij}$$

denote the marginal probabilities; for example,  $p_{1\cdot}$  is the population proportion in the first row. Clearly,

$$\sum_{i=1}^I p_{i\cdot} = \sum_{j=1}^J p_{\cdot j} = 1.$$

**Table 10.4** Observed counts.

A	B				Row sum
	$B_1$	$B_2$	...	$B_J$	
$A_1$	$N_{11}$	$N_{12}$	...	$N_{1J}$	$R_1$
$A_2$	$N_{21}$	$N_{22}$	...	$N_{2J}$	$R_2$
$\vdots$					
$A_I$	$N_{I1}$	$N_{I2}$	...	$N_{IJ}$	$R_I$
Column sum	$C_1$	$C_2$	...	$C_J$	$n$

The hypotheses to test whether or not the categorical variables are independent is given by

$$\begin{aligned} H_0: p_{ij} &= p_i p_j, \quad i = 1, 2, \dots, I, j = 1, 2, \dots, J, \\ H_A: p_{ij} &\neq p_i p_j, \quad \text{for some } i, j. \end{aligned} \quad (10.2)$$

Now,  $p_i$  and  $p_j$ ,  $i = 1, 2, \dots, I$ ;  $j = 1, 2, \dots, J$ , are unknown, but it seems reasonable to estimate them by the sample proportions

$$\hat{p}_i = \frac{R_i}{n} \quad \text{and} \quad \hat{p}_j = \frac{C_j}{n}.$$

Thus, an estimate of the expected value of  $N_{ij}$  is

$$\hat{E}[N_{ij}] = n\hat{p}_i\hat{p}_j = n \frac{R_i}{n} \frac{C_j}{n} = \frac{R_i C_j}{n}.$$

The random test statistic is

$$C = \sum_{i=1}^I \sum_{j=1}^J \frac{(N_{ij} - \hat{E}[N_{ij}])^2}{\hat{E}[N_{ij}]} \quad (10.3)$$

with observed value

$$c = \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{ij} - R_i C_j / n)^2}{R_i C_j / n}. \quad (10.4)$$

We often refer to Equation (10.4) more colloquially by Equation (10.1).

In 1924, R. A. Fisher proved that under the condition of independence,  $p_{ij} = p_i p_j$  for all  $i$  and  $j$ , the distribution of  $C$  approaches a chi-square distribution with  $(I-1)(J-1)$  degrees of freedom as the sample size goes to infinity. In practice, this means that if the expected counts are large, then the chi-square approximation is reasonably accurate, except in the tails of the distribution.

**Example 10.1 (GSS continued)**

For the GSS example,

$H_0$ : Opinion on the death penalty is independent of degree (there is no association between degree and opinion on the death penalty).

$H_A$ : Opinion on the death penalty is not independent of degree level (there is an association between degree and opinion on the death penalty).

**R Note**

The function `chisq.test` computes the chi-square test statistic and compares it to a  $\chi^2$  distribution to find the  $P$ -value, when given two factor variables.

```
> chisq.test(GSS2018$Degree, GSS2018$DeathPenalty)
  Pearson's Chi-squared test

  data: Degree and DeathPenalty
  X-squared = 50.449, df = 4, p-value = 2.91e-10
```

□

**Remark** We usually express hypotheses for an independence test in words, rather than multiplying probabilities as in Equation (10.2). ||

The chi-square test statistic has an *approximate* chi-square distribution if the null hypothesis is true. There are various rules of thumb that give conditions under which usage of this test is appropriate. One rule suggests that  $E \geq 5$  for all cells; another (Cochran, 1954) recommends that at least 80% of the cells have  $E \geq 5$ .

Frankly, those old rules are inaccurate. With modern computers, when the expected counts are small, it is better to use a permutation test (Section 10.2) or exact calculations (Section 10.3.3). However, these do not avoid the universal problem of testing with small samples – that the power is low, possibly too low to have reasonable chance of discerning dependence. This is exacerbated in discrete problems like this where not every  $P$ -value is possible; for example, in a  $2 \times 2$  contingency with 10 in each row and column, the possible  $P$ -values for an exact test are 0.00001, 0.0011, 0.023, 0.179, ..., so the critical value would correspond to a  $P$ -values of 0.023. The discreteness makes the critical value harder to reach.

**Example 10.2** Is there a relationship between gender and support for the death penalty? The GSS (case study in Section 1.7) asked participants whether or not they were in favor of or opposed to the death penalty.

Gender now	Death penalty?	
	Favor	Oppose
Man	411	201
Woman	420	276
Transgender	1	1
A gender now listed here	0	1

We note the extremely small counts in the cells for the transgender and non-binary respondents. In fact, of the subgroup of GSS participants who were asked for their current gender, three did not answer either man or woman, which is only 0.2% of the sample. For the two transgender cells in the above table, the expected counts are 1.27 and 0.73. A chi-square test should not be used. A permutation test could be used: the  $P$ -value would not be inflated but even here, the test may give undue weight to the small cells. In cases of extremely low cell counts, it would be better to omit those rows.

This example highlights the importance of exploring your data before blindly applying a statistical procedure.  $\square$

### 10.3.2 $2 \times 2$ Tables

An important special case of the chi-square test for contingency tables is  $2 \times 2$  tables. There are a number of things different about this case. Two are in terms of testing – there is an exact test, or if we use the chi-square approximation then we should apply a continuity correction. The other is that here we are not limited to just testing independence. This case is equivalent to comparing two proportions (either row proportions or column proportions). We can describe the difference, either by the difference of proportions as we did earlier, or using a new descriptive statistic, the odds ratio (see Exercise 10.29).

**Example 10.3** Two researchers studied whether or not being bullied in school was associated with being short. The table below summarizes their findings for 209 pupils, from “Bullying in school: are short pupils at risk? Questionnaire study in a cohort” (Voss and Mulligan, 2000).

Height	Bullied?	
	Yes	No
Short	42	50
Not short	30	87

The hypotheses to be tested are as follows:

$H_0$ : Being bullied is independent of height (there is no association between being bullied and height).

$H_A$ : Being bullied is not independent of height (there is an association between being bullied and height).

We first approach this using the chi-square reference distribution introduced in Section 10.3.

Here are the expected counts.

Height	Bullied?	
	Yes	No
Short	31.7	60.3
Not short	40.3	76.7

Thus, the test statistic is

$$c = \frac{(42 - 31.7)^2}{31.7} + \frac{(50 - 60.3)^2}{60.3} + \frac{(30 - 40.3)^2}{40.3} + \frac{(87 - 76.7)^2}{76.7} = 9.12.$$

We compare this value to a chi-square distribution with  $(2 - 1)(2 - 1) = 1$  degree of freedom. The  $P$ -value is 0.0025, so the data support the hypothesis that being bullied and height are not independent!  $\square$

### R Note

The `chisq.test` function also accepts a contingency table as an argument. Use the `rbind` function to bind the values in the contingency table by row.

```
> mat <- rbind(c(42, 50), c(30, 87))
> chisq.test(mat)
Pearson's Chi-squared test with Yates' continuity correction
data: mat
X-squared = 8.2683, df = 1, p-value = 0.004034
```

Notice that the  $P$ -value returned by `chisq.test` is different than we had computed. In the case of  $2 \times 2$  tables, the `chisq.test` function uses the *Yates continuity correction* by default, producing an adjusted chi-square test statistic:

$\sum (|O_i - E_i| - 0.5)^2 / E_i$ . This is to handle the accuracy issues associated with approximating discrete quantities with continuous distributions (recall Section 4.3.2). To obtain the non-corrected version of the test, add the argument `correct = FALSE`.

**Table 10.5** General  $2 \times 2$  contingency table.

Variable 1	Variable 2		Row sum
	C	D	
A	$a$	$b$	$a + b$
B	$c$	$d$	$c + d$
Column sum	$a + c$	$b + d$	$N = a + b + c + d$

**Remark** In the case of a  $2 \times 2$  table,  $C = Z^2$  where  $Z$  is the standard normal statistic for comparing two proportions (see Exercise 10.27). ||

### 10.3.3 Fisher's Exact Test

Another approach when working with  $2 \times 2$  tables is to perform an exact test – that is, compute an exact probability of obtaining an outcome as or more extreme than the one observed. Fisher's exact test conditions on the margin totals; in other words, it assumes that the row and column sums in the original  $2 \times 2$  table are fixed.

The probability of obtaining this particular distribution of cell counts in Table 10.5, conditional on the row and column totals, is given by the hypergeometric distribution (see Section B.5). If we let  $X$  denote the count in the (1,1) cell, then

$$P(X = a) = \frac{\binom{a+b}{a} \binom{c+d}{c}}{\binom{N}{a+c}}. \quad (10.5)$$

For Fisher's exact test, assuming that  $a + b, c + d, a + c$ , and  $b + d$  (in Table 10.5) are fixed, we look at all table configurations with the (1,1) cell entry as or more extreme than the observed (1,1) cell entry (the  $a$ ). The  $P$ -value is 2 times the smaller of  $P(X \geq a)$  or  $P(X \leq a)$ .

#### R Note

The `fisher.test` function performs Fisher's exact test for a  $2 \times 2$  table and outputs the result in terms of odd ratios. See Exercise 10.29 for the definition of odds ratio.

```
> mat <- rbind(c(42, 50), c(30, 87))
> fisher.test(mat)
Fisher's Exact Test for Count Data
```

```
data: mat
p-value = 0.003292
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
1.305198 4.557041
sample estimates:
odds ratio
2.425225
```

The  $P$ -value using Fisher's exact test for the bullying example is 0.0033 compared to 0.0025 using the chi-square reference distribution or 0.004 using a continuity corrected test statistic with the chi-square reference distribution.

### Remark

- In the case of  $2 \times 2$  tables, the hypothesis that the two variables are independent is equivalent to the odds ratio being equal to 1.
- There is more than one way to compute the  $P$ -value for the two-sided Fisher's exact test. In fact, R's implementation of this test uses a different one than what we defined above. Agresti (2012) gives four different approaches with some pros and cons of each.
- Researchers have developed exact calculations or nearly exact approximations for  $P$ -values for tables larger than  $2 \times 2$  with small expected counts, some of which are implemented in R's `fisher.test`. ||

#### 10.3.4 Conditioning

Fisher's exact test is based on the condition that both row and column totals are fixed. When comparing two proportions, this means that both sample sizes and the total number of successes are fixed – the only thing random is how many of those successes occur in the first group. Yet in practice it is fine to use this test when none of those are fixed – sample sizes, or total successes. This is called *conditioning*, to compute standard errors,  $P$ -values, or other quantities conditional on the actual value of sample sizes or other quantities that are not directly related to  $\theta$ , the quantity of interest. We have actually been doing this without even thinking about it – for example, we compute standard errors for a single mean or difference in means as if the sample size(s) are fixed, even though in practice they may have been random. We are computing standard errors given the observed sample sizes, not for other sample sizes that might have happened. For instance, in the North Carolina babies case study (Example 8.5), of the 1009 babies in the sample,  $n_1 = 898$  babies were born to mothers who did not smoke

while  $n_2 = 111$  babies were born to mothers who did smoke. These sizes, the 898 and 111 are random since another random sample of 1009 babies from the population of North Carolina babies born in 2004 would result in a different set of numbers of babies in each group. But in using the  $t$  test for comparing means, we treat the 898 and 111 as fixed.

For testing  $2 \times 2$  contingency tables, Fisher's exact test computes the  $P$ -value given the actual number of successes, not other numbers that might have occurred. Fisher's test is in fact the permutation test for testing independence between two variables, but with the  $P$ -value computed exactly rather than by simulation.

## 10.4 Chi-Square Test of Homogeneity

In the death penalty example in Section 10.1, we drew a sample from a single population, created a contingency table based on values of two factor variables, and tested for independence between the factors. In other situations, we may instead draw samples from two or more different populations, classify the observations by the levels of a single factor variable, and test whether that factor has the same distribution in each population.

For example, suppose a candy company wants to know whether boys or girls differ in their taste preferences for three new candy flavors that will be sold next year (Table 10.6). The company obtains a random sample of 100 boys and a random sample of 110 girls, gets each child to taste the three flavors, and asks them to name their favorite. We want to know if boys and girls have the same distribution of favorites.

This seems very similar to the death penalty example, except that here the number of boys and girls is fixed in advance; in the death penalty example, the number of people in each education level depended on the original random survey (in other words, a different random sample would have resulted in a different number of people in each education level). This distinction turns out not to matter – we will use exactly the same test statistic and procedures for computing  $P$ -values and null distributions.

There is a second difference in that we may express the parameters and hypotheses differently. In the earlier example, we worked with a matrix of parameters  $p_{ij}$  that added to 1 for the whole table. Here we work with parameters for each population.

Let  $\pi_{ij}$  denote the proportion of gender  $i$  that prefer candy  $j$ , where  $i = B, G; j = 1, 2, 3$ ; within each row, these probabilities add to 1. We test

$$H_0: \pi_{B1} = \pi_{G1}, \quad \pi_{B2} = \pi_{G2}, \quad \pi_{B3} = \pi_{G3};$$

$H_A$ : at least one of the equalities does not hold.

**Table 10.6** Counts of candy preferences.

Gender	Candy			Row sum
	Flavor 1	Flavor 2	Flavor 3	
Boys	42	20	38	100
Girls	33	27	50	110
Column sum	75	47	88	210

If the null hypothesis is true, there is a single set of preferences  $\pi_1 = \pi_{B1} = \pi_{G1}$ ,  $\pi_2 = \pi_{B2} = \pi_{G2}$ , and  $\pi_3 = \pi_{B3} = \pi_{G3}$ . We estimate these probabilities using the sample proportions

$$\hat{p}_1 = \frac{75}{210} = 0.3571, \quad \hat{p}_2 = \frac{47}{210} = 0.2238, \quad \hat{p}_3 = \frac{88}{210} = 0.4190.$$

The estimated expected counts of the six cells are obtained by multiplying those proportions by either 100 (for the boys row) or 110 (for the girls) row.

In spite of the different parameterization, the expected counts are the same as using  $R_i C_j / n$ . For example, the expected number of boys who like flavor 1 is  $\hat{p}_1 \times 100 = (75/210) \times 100 = R_1 C_1 / n = 35.71$ . We use the same test statistic:

$$\begin{aligned} C &= \sum_{\text{all cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \quad (10.6) \\ &= \frac{(42 - 35.71)^2}{35.71} + \frac{(20 - 22.38)^2}{22.38} + \frac{(38 - 41.90)^2}{41.90} \\ &\quad + \frac{(33 - 39.281)^2}{39.281} + \frac{(27 - 24.618)^2}{24.618} + \frac{(50 - 46.09)^2}{46.09} \\ &= 3.2902. \end{aligned}$$

As before, we may compare this statistic to a permutation distribution, or a  $\chi^2$  distribution with  $(2 - 1)(3 - 1) = 2$  degrees of freedom. The  $P$ -value using the  $\chi^2$  approximation is 0.193; so if the null hypothesis is true, then about 19.3% of samples from the same two populations would result in a test statistic as large or larger than this one. We conclude that the two populations could well be the same in their preferences for the three flavors. If there are differences, they are not large enough to distinguish from the random noise of sampling, with samples this small.

More generally, consider samples of size  $R_i$ ,  $i = 1, 2, \dots, I$  from  $I$  independent populations. Suppose each individual can be classified as one of  $J$  different types and let  $N_{ij}$  denote the number of individuals from population  $i$  of type  $j$ . The data can be summarized as in Table 10.4.

Let  $\pi_{ij}$  be the probability that an observation from population  $i$  falls in column  $j$  (like the conditional probability of column  $j$  given row  $i$ ). We test

$$H_0: \pi_{1j} = \pi_{2j} = \cdots = \pi_{Ij}, \quad j = 1, 2, \dots, J,$$

$$H_A: \text{Equality does not hold for some } i, j$$

using the test statistic

$$c = \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{ij} - R_i C_j / n)^2}{R_i C_j / n}.$$

If the null hypothesis is true and sample sizes are large, then the test statistic has an approximate  $\chi^2$  distribution with  $(I - 1)(J - 1)$  degrees of freedom.

The R `chisq.test` function can be used for tests of homogeneity also.

## 10.5 Goodness-of-Fit Tests

The chi-square statistic is useful in other situations where we wish to compare differences between observed and expected counts, including *goodness-of-fit tests*, to check whether the data fit a probability model.

### 10.5.1 All Parameters Known

**Example 10.4** Barnsley et al. (1992) investigated the relationship between month of birth and achievement in sport. Birth dates were collected for players in teams competing in the 1990 World Cup soccer games.

	Birth month			
	Aug-Oct	Nov-Jan	Feb-Apr	May-Jul
Observed	150	138	140	100

We wish to test whether these data are consistent with the hypothesis that birthdays of soccer players are uniformly distributed across the four quarters of the year.

Let  $p_i$  denote the probability of a birth occurring in the  $i$ th quarter; the hypotheses are as follows:

$$H_0: p_1 = 1/4, \quad p_2 = 1/4, \quad p_3 = 1/4, \quad p_4 = 1/4,$$

$$H_A: p_i \neq 1/4 \text{ for at least one } i.$$

There were a total of  $n = 528$  players considered for this study, so the expected count for each quarter is  $528/4 = 132$ . Thus, the test statistic is

$$c = \frac{(150 - 132)^2}{132} + \frac{(138 - 132)^2}{132} + \frac{(140 - 132)^2}{132} + \frac{(100 - 132)^2}{132} = 10.97.$$

We cannot use permutation resampling to obtain a null distribution (there is nothing to permute, we are not testing the independence of two variables), so we will use a chi-square approximation, with  $4 - 1 = 3$  degrees of freedom (we will discuss the degrees of freedom below.) The resulting  $P$ -value  $P(C \geq 10.97) = 0.012$  supports the hypothesis that birthdays are not uniformly distributed across the four quarters. One explanation is that players born shortly after the yearly cutoff for school enrollment are relatively old for their grade and competing against younger classmates. They enjoy more success early, and ultimately do better in the sport.  $\square$

The degrees of freedom for a goodness-of-fit test with  $k$  cells and no parameters estimated from the data is  $k-1$ .

Here is an explanation of the degrees of freedom. If no parameters are estimated from the data, the degrees of freedom would be  $k - 1$ : pick any  $k - 1$  cells and numbers may be placed in these cells freely, but the number in the final cell must make the sum equal to the sample size, so there is no freedom in what number goes in that cell.

**Example 10.5** Suppose you draw 100 numbers at random from an unknown distribution. Thirty values fall in the interval  $(0, 0.25]$ , 30 fall in  $(0.25, 0.75]$ , 22 fall in  $(0.75, 1.25]$  and the rest fall in  $(1.25, \infty)$ . Your friend claims that the distribution is exponential with parameter  $\lambda = 1$ . Do you believe her?

### Solution

The hypotheses we wish to test are

$$\begin{aligned} H_0: & \text{The data are from an exponential distribution, } \lambda = 1, \\ H_A: & \text{The data are not from an exponential distribution, } \lambda = 1. \end{aligned}$$

Let  $X \sim \text{Exp}(1)$ . The probabilities for each interval are as follows:

$$p_1 = P(0 \leq X \leq 0.25) = \int_0^{0.25} e^{-x} dx = 0.22,$$

$$p_2 = P(0.25 < X \leq 0.75) = \int_{0.25}^{0.75} e^{-x} dx = 0.306$$

$$p_3 = P(0.75 < X \leq 1.25) = \int_{0.75}^{1.25} e^{-x} dx = 0.186$$

$$p_4 = P(1.25 < X < \infty) = \int_{1.25}^{\infty} e^{-x} dx = 0.287.$$

Then for a sample of  $n = 100$  numbers, the expected counts are

Interval	(0, 0.25]	(0.25, 0.75]	(0.75, 1.25]	(1.25, $\infty$ )
Observed count	30	30	22	18
Expected count $np_i$	22	30.6	18.6	28.7

Thus, the chi-square test statistic is

$$c = \frac{(30 - 22)^2}{22} + \frac{(30 - 30.6)^2}{30.6} + \frac{(22 - 18.6)^2}{18.6} + \frac{(18 - 28.7)^2}{28.7} = 7.53.$$

Under the null hypothesis, the test statistic comes from a chi-square distribution with  $4 - 1 = 3$  degrees of freedom, so the  $P$ -value is  $P(c \geq 7.53) = 0.057$ , so it is plausible that the data do come from  $\text{Exp}(1)$ .  $\square$

**Example 10.6** Is it possible that the following 50 numbers are a random sample from a chi-square distribution with 10 degrees of freedom?

1.85	2.68	2.84	3.76	3.86	4.96	5.42	6.50	6.65	6.81
6.95	7.42	7.48	7.99	8.50	8.54	8.65	8.71	8.80	9.47
9.82	9.91	10.09	10.30	10.62	10.63	10.70	10.79	10.94	11.92
12.14	12.22	12.96	13.29	13.43	14.14	14.29	14.36	14.65	14.68
14.87	15.00	15.91	16.15	16.18	17.21	19.06	20.81	22.82	23.55

### Solution

We will use R to compute the 0.2, 0.4, 0.6, and 0.8 quantiles of the chi-square distribution with 10 degrees of freedom – that is, those points that mark off probabilities (areas) equal to 0.2.

#### R Note

The `qchisq` function computes quantiles of the chi-square distribution.

```
> qchisq(c(.2, .4, .6, .8), 10)
[1] 6.179079 8.295472 10.473236 13.441958
```

Thus, we expect to see 20% of the 50 values fall into each subinterval determined by the above quantiles. We will compare the expected count of 10 for each interval with the observed number of values that fall into each of the subintervals.

Interval	(0, 6.179]	(6.179, 8.295]	(8.295, 10.473]	(10.473, 13.442]	(13.442, $\infty$ )
Observed	7	7	10	11	15
Expected	10	10	10	10	10

The chi-square test statistic is

$$c = \frac{(7 - 10)^2}{10} + \frac{(7 - 10)^2}{10} + \frac{(10 - 10)^2}{10} + \frac{(11 - 10)^2}{10} + \frac{(15 - 10)^2}{10} = 4.4.$$

If the data are from a chi-square distribution with 10 degrees of freedom, then the test statistic  $c = 4.4$  comes from a chi-square distribution with  $5 - 1 = 4$  degrees of freedom, so the  $P$ -value is  $P(c \geq 4.4) = 0.355$ . We cannot rule out the possibility that the data are a random sample from a chi-square distribution with 10 degrees of freedom.  $\square$

### 10.5.2 Some Parameters Estimated

In other situations, some parameters must be estimated from the data. The testing procedure is similar, but with adjusted degrees of freedom.

**Example 10.7** A home run in baseball is an exciting event, a majestic flight that can dramatically alter the course of a game, driving in as many as four runs with one swing of a bat. Table 10.7 displays a summary of the home run data for the Philadelphia Phillies in their 2009 season. For instance, the Phillies hit 2 home runs in 40 of the 162 games played, but 5 home runs in only 1 game.

Since we have counts data and home runs are relatively rare, the Poisson distribution is a natural candidate for modeling these data. If the random variable  $X$  denotes the number of home runs in a game, then the probability mass function is given by  $f(x) = P(X = x) = \lambda^x e^{-\lambda} / x!$ ,  $x = 0, 1, 2, \dots$ , and parameter  $\lambda > 0$ . Since  $\lambda$  is unknown, we estimate it by the empirical average number of home runs per game,  $224/162 = 1.3827$ , the maximum likelihood estimate

**Table 10.7** Counts of home runs (in 162 games).

Number of home runs $x$	Number of games	Proportion of games
0	43	0.2653
1	52	0.3209
2	40	0.2469
3	17	0.1049
4	9	0.0555
5	1	0.0062

(Proposition 6.2). Thus, we will model the number of home runs per game with the probability density function  $P(X = x) = (1.3827^x e^{-1.3827})/x!$ ,  $x = 0, 1, 2, \dots$ .

To assess this model, we consider the following hypotheses:

$H_0$ : The home runs are modeled by a Poisson distribution,

$H_A$ : The home runs are not modeled by a Poisson distribution.

We assume that the distribution of home runs is the same for each game and independent between games. Neither of these is exactly true in practice – some pitchers are better than others, and a home run fest in one game might make the opponents pitch more conservatively in the next game – but for the current analysis, we ignore these effects.

Thus, under the null hypothesis, we compute the expected number of games in which there are  $x$  home runs,  $x = 0, 1, 2, \dots$  by  $P(X = x) \times 162 = (1.3827^x e^{-1.3827})/x! \times 162$  to obtain Table 10.8.

To test whether the actual counts are discernibly different than the expected counts, we will use a chi-square statistic and calculate the  $P$ -value using a chi-square approximation. But this approximation is reasonable only if the expected counts are reasonably large, and an expected count of 1.7 does not qualify. The expected counts for 6, 7, and more home runs are even smaller, and we need to include them somewhere. To work around this, we combine cells, with the final cell being the count for four or more home runs per game. The observed and expected counts, and their contributions to the chi-square statistic are shown in Table 10.9.

**Table 10.8** Probabilities for Poisson distribution with  $\lambda = 1.3827$ , and expected counts for 162 games.

$x =$	0	1	2	3	4	5
$P(X = x)$	0.251	0.347	0.240	0.111	0.038	0.011
Expected	40.6	56.2	38.9	17.9	6.2	1.7

**Table 10.9** Chi-square test for Poisson goodness-of-fit to home run data.

$x =$	0	1	2	3	4+
Observed count	43	52	40	17	10
Expected count	40.6	56.2	38.9	17.9	8.4
$(O - E)^2/E$	0.14	0.31	0.03	0.05	0.31

The chi-square statistic is 0.84; the  $P$ -value is  $P(\chi_3^2 > 0.84) = 0.84$ . We conclude that the number of home runs per game is consistent with a Poisson distribution.  $\square$

The degrees of freedom for a goodness-of-fit test with  $k$  cells and  $\ell$  parameters estimated from the data is  $k - \ell - 1$ .

Estimating parameters reduces the degrees of freedom, because estimated parameters tend to “overfit,” making the model fit the data better than the true (unknown) parameters would.

### R Note

```
> Homeruns <- subset(Phillies, select = Homeruns, drop = T)
> lambda <- mean(Homeruns) # average number home runs/game
> dpois(0:4, lambda)       # theoretical model
[1] 0.25089618 0.34691818 0.23984466 0.11054569 0.03821332
> table(Homeruns)
Homeruns
  0   1   2   3   4   5
43  52  40  17   9   1
> table(Homeruns)/162      # empirical probabilities
HomeRuns
  0           1           2           3           4           5
0.26543210 0.32098765 0.24691358 0.10493827 0.05555556 0.00617284
```

**Example 10.8** In Section 6.1.3, we modeled the wind speeds from a turbine with the Weibull distribution. The plots (Figure 6.4) seem to suggest that this is a good fit, but here we will check this more formally. We find the deciles (the points marking the 10%, 20%, ... probabilities) for the Weibull distribution with the estimated parameters  $\hat{k} = 3.169$ ,  $\hat{\lambda} = 7.661$  and count the number of data points that fall into each of the intervals determined by these deciles. Since there are 168 data points, we expect 16.8 points to fall into each of these intervals (Table 10.10).

The chi-square test statistic is  $c = 3.071$  and if the data do come from the proposed Weibull distribution, then  $c$  comes from a chi-square distribution with  $10 - 2 - 1 = 7$  degrees of freedom. The corresponding  $P$ -value is 0.878, so we conclude that the distribution of the wind speed data is consistent with a Weibull distribution with parameters  $\hat{k} = 3.169$  and  $\hat{\lambda} = 7.661$ .  $\square$

**Table 10.10** Distribution of wind speed data into intervals.

Interval	[0.0, 3.77]	(3.77, 4.77]	(4.77, 5.53]	(5.53, 6.20]	(6.20, 6.80]
Observed	17	18	19	13	20
Expected	16.8	16.8	16.8	16.8	16.8

Interval	(6.80, 7.45]	(7.45, 8.12]	(8.12, 8.90]	(8.90, 9.97]	(9.97, $\infty$ )
Observed	14	17	17	14	19
Expected	16.8	16.8	16.8	16.8	16.8

## 10.6 Chi-Square and the Likelihood Ratio\*

In Section 8.6.2, we discussed the likelihood ratio test. For contingency tables, the null hypothesis is that rows and columns of an  $r \times c$  table are independent,  $p_{ij} = p_i p_j$ , and the alternative hypothesis is that they are not. Both are composite hypotheses. Under the null hypothesis, the maximum likelihood estimates for the parameters  $p_{ij}$  are  $(R_i/n)(C_j/n) = E_{ij}/n$ , and under the alternative hypothesis they are the observed proportions  $N_{ij}/n$ . The likelihood ratio is

$$T(x) = \frac{\max_{\Omega_0} L(\hat{\theta} \mid x)}{\max_{\Omega} L(\hat{\theta} \mid x)} = \frac{\prod_{ij} (E_{ij}/n)^{N_{ij}}}{\prod_{ij} (N_{ij}/n)^{N_{ij}}}.$$

Then

$$\ln(T(x)) = \sum_{ij} N_{ij} \ln(E_{ij}/n) - \sum_{ij} N_{ij} \ln(N_{ij}/n) = - \sum_{ij} N_{ij} \ln(N_{ij}/E_{ij}).$$

Let

$$G(x) = -2 \ln(T(x)) = 2 \sum_{ij} N_{ij} \ln(N_{ij}/E_{ij}) = 2 \sum O \ln(O/E).$$

This is known as the  $G$  statistic. The reference distribution can be computed in the same ways as the chi-square statistic – exact calculations, permutation tests, or a chi-square distribution with  $(I - 1)(J - 1)$  degrees of freedom.

The  $G$  statistic for the death penalty example is 23.58, compared to the chi-square statistic 23.45.

Similarly, the  $G$  statistic for goodness of fit is  $2 \sum O \ln(O/E)$ , and  $P$ -values can be approximated by chi-square distributions with the same degrees of freedom.

The chi-square statistic  $C$  (Equation (10.1)) was originally developed by Pearson as an approximation to the  $G$  statistic because logs were too difficult to compute. That is no longer an issue and the  $G$  statistic produces rejection regions with a better shape. So why do we not ditch the chi-square statistic in favor of the  $G$  statistic? Because the chi-square statistic is easier to remember and to interpret. The individual terms  $(O - E)^2/E$  are large in the cells that

demonstrate the biggest deviations from independence. In contrast, with the  $G$  statistic, the terms  $O \ln(O/E)$  are positive and negative and are not easy to interpret.

## Exercises

For problems that call for the permutation test, use `chisq.test` or `fisher.test` with the `simulate.p. value = TRUE, B = 10^5-1` option. To check your understanding, consider writing your own simulation test code!

- 10.1** In the Iowa Recidivism data set (Section 1.4), we have data on the gender of the prisoners who were released.
- Create a table summarizing the relationship between gender and whether or not the prisoner recidivated.
  - To determine whether there is a relationship between these two variables, will you conduct a test of independence or a test of homogeneity? State the appropriate hypothesis.
  - Calculate the expected counts for each cell.
  - Compute the test statistic (Equation (10.1)) and then compare to the appropriate chi-square distribution. State your conclusion in a complete sentence.
- 10.2** For frontline healthcare workers in the battle against COVID-19, proper usage of personal protective equipment (PPE) is critical. Researchers administered a questionnaire to a sample of healthcare workers in Nepal to determine factors associated with satisfactory knowledge of putting on and taking off PPE (Pandey et al., 2021).

Profession	Satisfactory knowledge?	
	Yes	No
Doctors	67	25
Nurses	96	19

- Write down the appropriate hypothesis to test to see if there is a relationship between profession and knowledge about PPE.
- Calculate the expected counts for each cell.
- Compute the test statistic (Equation (10.1)) and then compare to the appropriate chi-square distribution. State your conclusion in a complete sentence.

- 10.3** Researchers in Brazil were interested in the prevalence of alcohol or drug use among 376 patients who were admitted to the emergency room of a hospital in Sao Paulo for traumatic injuries due to traffic accidents, falls, or violence (Bombana et al., 2021). Of the 259 patients who were admitted to the emergency room (ER) during the day, 68 tested positive for alcohol or drug use compared to 51 of the 177 patients admitted during the night.

Drugs or alcohol?		
Time	Yes	No
Day	68	191
Night	51	66

- (a) Write down the appropriate hypothesis to test to see if there is a relationship between time of day and use of drugs or alcohol.  
 (b) Calculate the expected counts for each cell.  
 (c) Compute the test statistic (Equation (10.1)) and then compare to the appropriate chi-square distribution. State your conclusion in a complete sentence.  
 (d) Can the results be generalized to a population? Explain.
- 10.4** The Pew Research Center conducted a survey in 2016 on issues of biomedical technologies to “enhance” human abilities. One question asked about the use of a synthetic blood substitute to give healthy people much greater speed, strength and stamina.<sup>2</sup>

Race/ethnicity	Response		
	Morally acceptable	Morally unacceptable	Not sure
White	829	1262	1442
Black	57	126	215
Hispanic	79	136	159

(Whites and Blacks include only non-Hispanics; Hispanics are of any race.)

- (a) Write down the appropriate hypothesis to test whether there is a relationship between race/ethnicity and belief about morality of synthetic blood use for enhancement of human abilities.  
 (b) Compute the expected counts for each cell.

<sup>2</sup> <http://www.pewinternet.org/2016/07/26/u-s-public-wary-of-biomedical-technologies-to-enhance-human-abilities>.

- (c) Compute the test statistic (Equation (10.1)) and then compare it to the appropriate chi-square distribution. State your conclusion in a complete sentence.
- 10.5** Carlos conducts a small study in his dormitory to see if there is a relationship between a student's class year and preference for Instagram or Snapchat. After deciding to use a 5% alpha level for his study, he randomly polls 15 students and finds
- | Class  | Which app? |          |
|--------|------------|----------|
|        | Instagram  | Snapchat |
| Junior | 2          | 5        |
| Senior | 7          | 1        |
- (a) Is this a test of independence or a test of homogeneity? Write down the appropriate hypotheses.  
 (b) Conduct the chi-square test by using `chisq.test`. What conclusion would Carlos draw?  
 (c) Conduct the test by using Fisher's exact test and by using a permutation test. What conclusion would Carlos draw? and  $P\text{-value} = 0.0389$   
 (d) What conclusion do you draw?
- 10.6** Two students went to a local supermarket and collected data on cereals; they classified cereals by their target consumer (children versus adults) and the placement of the cereal on the shelf (bottom, middle, top). The data are given in `Cereals`.
- (a) Create a table to summarize the relationship between age of target consumer and shelf location.  
 (b) Conduct a chi-square test using R's `chisq.test` function.  
 (c) R returns a warning message. Compute the expected counts for each cell to see why.  
 (d) Conduct a permutation test of independence and state your conclusion.
- 10.7** The California Department of Game and Fish published a report on a study of jack mackerel fish from three different regions off the waters of California: near Guadalupe Island and Cedros Island off Baja California and near San Clemente Island in southern California (Gregory and Tasto, 1976). One characteristic of the fish that they were interested in was the number of rays on the second dorsal fin; in particular, is the number of rays different for fish from the different regions?

Habitat	Fin ray count					
	$\geq 36$	35	34	33	32	$\leq 31$
Guadalupe Island	14	30	42	78	33	14
Cedro Island	11	28	53	66	27	9
San Clemente Island	10	17	61	53	22	10

- (a) Does this setting call for a test of independence or a test of homogeneity?  
(b) Write down the appropriate hypothesis and carry out the test.
- 10.8** A researcher in Hong Kong conducted a study on children's perception of advertising and brands on television (Chan, 2008). For one part of the study, she analyzed surveys sent to 1481 children from rural areas of Mainland China and received responses from 726 boys and 755 girls. The responses to the question about their feeling of commercials on TV are summarized below:
- | Feeling toward TV ads |                |      |         |         |                   |
|-----------------------|----------------|------|---------|---------|-------------------|
| Gender                | Like very much | Like | Neither | Dislike | Dislike very much |
| Boys                  | 180            | 260  | 137     | 96      | 52                |
| Girls                 | 210            | 266  | 145     | 85      | 49                |
- (a) Will this be a test of independence or a test of homogeneity?  
(b) Conduct the appropriate test to determine the relationship between sex and feelings toward TV commercials.
- 10.9** For the flight delays case study in Section 1.1 what proportion of flights American flights were delayed for more than 30 min (`Delayed30`)? For United? Is this difference statistically discernible?  
(a) Does this call for a test of independence or a test of homogeneity?  
(b) Carry out the test.
- 10.10** From the GSS 2018 case study in Section 1.7, we will consider participants by their income level. First, remove from the data set those income entries listed as "Not applicable" – these refer to those who are unemployed or retired.  
(a) Create a table to summarize the relationship between a person's income and their view on how much the government spends on alternative energy.

- (b) Use the `chisq.test` function to see if there is a relationship between a person's income and their view on government spending on alternative energy. R returns a warning, why?  
 (c) Conduct a permutation test and state your conclusion.

**10.11** Researchers investigated the effects of testosterone on the behavior of male European starlings, including whether or not males assist their mates in incubating eggs (De Ridder et al., 2000). Twenty-one male starlings had a silastic tube inserted in the neck region. Ten of the tubes were filled with testosterone while 11 were empty. The assignment to the two treatments were random. All 11 males in the control group assisted their female mate during the incubation period while 6 of the 10 males receiving the testosterone assisted their mates.

- (a) Conduct a chi-square test using the `chisq.test` function to determine if this difference in behavior is statistically discernible. Confirm the possible issue that arises in the procedure.  
 (b) Repeat the test but use simulation instead. Do you reach the same conclusion?  
 (c) Repeat the test using Fisher's exact test.  
 (d) What conclusion will you report?

**10.12** Researchers conduct a pilot study to test the effectiveness of a drug in preventing a certain disease. Of 20 patients in the study, 10 are randomly assigned to receive the drug and 10 to receive a placebo. After 1 year, suppose five patients in the control group contract the disease, while two patients who took the drug contract the disease.

Outcome	Response	
	Disease	No disease
Drug	2	8
Placebo	5	5

- (a) For a test of homogeneity, what are the expected cell counts?  
 (b) If the drug is not effective, then every patient is equally likely to contract the disease. In that case, if 7 patients out of 20 contract the disease, what is the probability that 2 of them are in the treatment group?  
 (c) In that case, what is the probability that two or fewer of them are in the treatment group?

**10.13** At a university, 15 juniors and 20 seniors volunteer to serve as a special committee that requires 8 members. A lottery is used to select the

committee from among the volunteers. Suppose the chosen students consists of six juniors and two seniors.

- (a) For a test of homogeneity, what are the expected counts?
- (b) If the selection were random, what is the probability of the committee having exactly two seniors?
- (c) If the selection were random, what is the probability that the committee would have two or fewer seniors?
- (d) Is there evidence that the selection was not random?

**10.14** For a given  $r \times c$  contingency table, we have the test statistic  $C$  given by Equation (10.1).

- (a) What happens to the value of  $C$  if every entry in the contingency table is multiplied by the same integer  $k > 1$ ? Do the marginal probabilities change? Does the degrees of freedom change?
- (b) What is the implication of this fact? That is, if the probabilities stay the same, but the actual counts in each cell increase (multiplicatively) by the same amount, what happens to our conclusion from the test?

**10.15** Of a sample of 70 random numbers, thirty fall in the interval  $[1, 1.5)$ , 18 fall in the interval  $[1.5, 2)$ , 9 fall in the interval  $[2, 3)$ , 10 fall in the interval  $[3, 5)$  and the rest are greater than 5. Is it plausible that these numbers were drawn from a distribution with  $\text{pdf } f(x) = 2/x^3$  for  $x \geq 1$ ?

**10.16** Suppose you randomly draw 75 values from a distribution that your friend claims has  $\text{pdf } f(y) = (1/9)y^2$ ,  $0 < y \leq 3$ . If 2 of the values fall in the interval  $(0, 1.25]$ , 6 fall in the interval  $(1.25, 1.75]$ , 10 fall in the interval  $(1.75, 2.25]$ , 32 fall in the interval  $(2.25, 2.75]$ , and the rest fall in the interval  $(2.75, 3]$ , perform a goodness-of-fit test to see if your data supports his claim.

**10.17** Suppose you randomly draw 50 values from an unknown distribution.

1.28	4.53	5.50	7.91	8.23	9.67	9.82	10.28	10.45	11.91
12.57	13.75	13.80	14.00	14.05	16.02	16.18	16.25	16.58	16.68
16.87	17.61	17.63	17.71	18.13	18.42	18.43	18.44	19.62	20.401
20.73	20.74	21.29	21.51	21.66	21.87	22.67	23.11	24.40	24.55
24.66	25.30	25.46	25.91	26.12	26.61	26.72	29.28	31.93	36.94

Could these data have come from the normal distribution  $N(22, 7^2)$ ?

- (a) Use the `qnorm` function in R to find the 0.2, 0.4, 0.6, and 0.8 quantiles of the normal distribution – that is, those points that mark off equal probabilities (equal areas) of 0.2.

- (b) Use these quantiles to determine your intervals and count the number of values that fall in each interval.  
 (c) Finish the goodness-of-fit test.

**10.18** Suppose you randomly draw 60 values from an unknown distribution.

16.21	16.96	17.07	17.81	19.66	21.16	21.95	22.76	23.81	23.94
24.12	24.26	25.10	25.15	25.22	25.47	25.62	25.91	27.34	27.51
28.05	28.67	28.76	28.89	28.93	29.45	29.54	29.64	30.38	30.60
31.49	31.52	32.25	32.26	32.40	32.52	32.54	32.66	33.01	33.02
33.91	34.32	34.83	34.88	34.93	35.05	35.33	35.84	36.18	36.33
37.27	37.84	38.24	38.33	38.42	38.74	38.83	40.87	41.77	43.91

- (a) Conduct a test to see if these data are consistent with a normal distribution with  $\mu = 25$ ,  $\sigma = 10$ .  
 (b) Suppose you suspect the data are from a normal distribution but do not know  $\mu$  or  $\sigma$ . Using the sample mean of 30.324 and standard deviation 6.54 as estimates of  $\mu$  and  $\sigma$ , conduct a goodness-of-fit test.

**10.19** For the Philadelphia Phillies data (*Phillies2009*), consider the number of doubles hit per game. Model this using a Poisson distribution, and perform a goodness-of-fit test to compare the theoretical model with the empirical data.

**10.20** The geometric distribution has been used to model the number of consecutive dry days (no rain) over a period of time or capture-recapture rates for wildlife (Hershfield, 1971; Romesburg and Marshall, 1979). Let  $X \sim \text{Geom}(p)$  where  $P(X = k) = p(1 - p)^k$ ,  $k = 0, 1, 2, \dots$ . Suppose the follow data are observed (so, for example,  $X = 0$  was observed 80 times):

$X =$	Distribution of values				
	0	1	2	3	4
Frequency	80	40	22	15	10

- (a) Use maximum likelihood to estimate  $p$ .  
 (b) Is it plausible that these 167 values came from a geometric distribution?

The R command `dgeom(k, p)` can be used to compute geometric probabilities.

- 10.21** In a 2012 paper, Adrian et al. (2012) studied the prevalence of a flatworm (*Dactylogyrus*) that infects the gills of a species of minnow, the Telescope Shiner. Part of their data included counting the number of fish that were infected by this parasite over a 4 month period (the non-breeding season of October through January). For the 102 minnows inspected, they found 57 of the fish had no parasites while 20 of the fish had 1 parasite in its gills. The rest of the data are below.

	Number of parasites per fish						
	0	1	2	3	4	5	$\geq 6$
Number of fish	57	20	11	4	6	2	2

Model this using a Poisson distribution and perform a goodness-of-fit test to compare the theoretical model with the empirical data.

- 10.22** California, like many states, sponsors lotteries to raise revenue. In one popular game, Fantasy 5, a player tries to match 5 numbers chosen from 1 through 39. For instance, on August 15, 2010, the 5 winning numbers were 29, 19, 37, 34, and 07. California uses a random mechanism to draw the numbers each day; how good is it? The file `Lottery` contains the winning numbers for the daily games from May 5, 2010 through August 15, 2010.<sup>3</sup> Determine whether or not the winning numbers are drawn randomly.
- 10.23** In Exercise 6.18, we modeled the distribution of service times at a snack bar using the gamma distribution. Use a goodness-of-fit test to see whether the gamma distribution is an adequate model for these data.
- 10.24** In Section 6.1.3 and Exercise 6.19, we modeled the distribution of times between earthquakes with the Weibull distribution. Conduct a goodness-of-fit test to see whether the Weibull distribution is an adequate model for these data.
- 10.25** Consider a  $2 \times 2$  contingency table. Using the notation of Table 10.4,
- Show that  $(N_{ij} - \hat{E}[N_{ij}])^2$  has the same value for all  $i, j$ .
  - Using (a), show that  $C = n(N_{11}N_{22} - N_{12}N_{21})^2 / (R_1 R_2 C_1 C_2)$ .
  - Verify that (b) yields the same value of  $C$  as Equation (10.3) for the following  $2 \times 2$  table

<sup>3</sup> <http://www.calottery.com/play/draw-games/fantasy-5>.

	$B_1$	$B_2$
$A_1$	6	8
$A_2$	10	12

- 10.26** Consider a test of independence for two variables that have the following  $2 \times 2$  table:

	$B_1$	$B_2$
$A_1$	$m$	10
$A_2$	10	$m$

What value(s) of  $m$  would lead to a conclusion that the two variables are not independent at the  $\alpha = 0.05$  alpha level?

- 10.27** In the case of a  $2 \times 2$  table, show that the chi-square test statistic  $C$  in (Equation (10.1)) satisfies  $C = Z^2$  where  $Z$  is the two-sample  $Z$  statistic for proportions given in Section 8.3.2.
- 10.28** In this exercise, you will show that the chi-square test statistic could be expressed in terms of differences of proportions.
- Show that the chi-square test statistic for  $C$  (Equation (10.1)) can be expressed in terms of differences between the individual column fractions in each row  $f_{ij} = N_{ij}/R_i$  and the overall column fractions  $f_j = C_j/n$ . The answer should be of the form  $C = \sum_{ij} a_{ij}(f_{ij} - f_j)^2$  for some  $a_{ij}$ .
  - Express  $f_j$  in terms of the  $f_{ij}$ . Substituting the answer for (b) into the answer for (a) yields the desired (but messy) result.

- 10.29** In comparing proportions,  $\pi_1, \pi_2$ , we have been looking at their difference,  $\pi_1 - \pi_2$ . Another option is to use the *odds ratio*. If  $\pi_i$  denotes the probability of an event for two groups  $i = 1, 2$ , then  $\mathcal{O}_i = \pi_i/(1 - \pi_i)$  is the odds of the event for each group, and the odds ratio is  $\mathcal{O}_1/\mathcal{O}_2$ .

Suppose  $\pi_i$ ,  $i = 1, 2$  denotes the probability (risk) of a certain disease for two groups A and B, each with a population of 1000 people. We look at the situation where  $\pi_1 - \pi_2 = 0.05$ .

- If  $\pi_1 = 0.35$  and  $\pi_2 = 0.30$ , how many people in each group will get the disease, on average?
- Verify that the odds ratio is 1.256. This is interpreted as “the odds that a person in group A will get the disease is 1.256 times greater than the odds that a person in group B will get the disease.”

(c) If  $\pi_1 = 0.06$  and  $\pi_2 = 0.01$ , how many people in each group will get the disease, on average?

(d) Compute the odds ratio and express this number in a sentence.

*Remark:* In both cases above, on average 50 more people in group A will get the disease than in group B, but the increase is more dramatic in the second case where both proportions are close to 0.

- 10.30** Refer to Exercise 10.29 and Example 10.3. Compute the odds of being bullied for the short and the not short pupils and then compute the odds ratio. Express this number in a sentence.
- 10.31** Refer to Exercise 10.29 and Table 10.5. Show that the odds ratio  $\mathcal{O}_A/\mathcal{O}_B = ad/bc$ .
- 10.32** Compute the  $G$  statistic for the tables in Exercise 10.1 and compare to the chi-square statistic.

# 11

## Bayesian Methods

You may have opened your e-mail to find an inbox filled with offers for cheap Viagra, eye-popping financial opportunities overseas, alluring (?) new companions, and warnings that your bank account was hacked (“click here to go to your account!”) The problem used to be far worse. Many e-mail providers have made a huge dent using *Bayesian spam filtering*.

Since certain words are more likely to occur in spam mail than legitimate mail, spam filters assign messages containing these words a higher probability of being spam. Certain senders get lower or higher probabilities. The Bayesian filtering combines information from all the words and other characteristics of the e-mail to assign a probability that each message is spam. If that probability exceeds a certain threshold, the message may be sent directly to the trash.

More generally, Bayes offers a mechanism for combining information from multiple sources. We do this all the time in our real lives (though we may not do it well). What is the probability that our team will win the next game? We combine what we know about the strengths of the two teams, any injuries, where the game is played, how hot key players have been, and so on, to come up with an estimate.

In statistics, Bayes offers a way to combine *prior information* with information provided by the data. The Bayesian approach contrasts to the *frequentist* approach. The confidence intervals and hypothesis tests we have done up to this point are frequentist techniques, based on what would happen under repeated sampling from the population.

Bayesian answers are often easier to understand. For example, a 95% confidence interval means that if the sampling procedure and interval calculation procedure were repeated many times, 95% of the intervals would include the true parameter. In contrast, using Bayes we produce an interval that has a 95% probability of including the parameter. Similarly, in hypothesis testing, a *P*-value is the probability that repeated sampling would give a result as extreme as the actual result, if the null hypothesis is true. In contrast, using Bayes we compute the probability that each hypothesis is true.

## 11.1 Bayes Theorem

The starting point for Bayesian methods is Bayes' theorem,

$$P(\theta | X) = \frac{P(\theta)P(X | \theta)}{P(X)}, \quad (11.1)$$

where  $\theta$  is the parameter and  $X$  the data. On the face of it, this is trivial, following from two applications of the definition of conditional probability,  $P(A | B) = P(AB)/P(B)$ , or equivalently (also called the multiplication rule)  $P(AB) = P(B)P(A | B)$ :

$$P(\theta | X) = \frac{P(\theta X)}{P(X)} = \frac{P(X\theta)}{P(X)} = \frac{P(\theta)P(X | \theta)}{P(X)}.$$

Yet this result has some far-reaching implications.

First, Bayesian methods treat  $\theta$  as random, rather than a fixed (but unknown) parameter, implying that  $\theta$  has a probability distribution. This is both a major strength and a major weakness of the approach. It allows a scientist to assign his or her own probabilities, a *prior distribution*, to  $\theta$ , to take advantage of past experience. For instance, in drug discovery companies screening thousands of compounds for efficacy against a particular cancer may use any prior information they have about which compounds are more or less likely to be effective in deciding which compounds to study further. But conversely, allowing a scientist to assign his or her own probabilities means that different people may get different results when analyzing the same data – just as different people have different estimates of the home team winning. Hence, Bayesian methods are viewed with suspicion in some settings. For instance, in phase III clinical trials, the final phase in testing new drugs, the results should stand on their own, based just on the data rather than on prior beliefs of the sponsor. One counter to that disadvantage is the use of “noninformative” prior information, a type of artificial prior information. In some cases these methods yield the same answers as frequentist methods. Bayesian methods have some other advantages. They are not subject to the same multiple testing issues as frequentist approaches, and they offer a logical self-consistent framework.

## 11.2 Binomial Data: Discrete Prior Distributions

We begin with cases where both  $\theta$  and  $X$  are discrete, in particular where the data are binomial and  $\theta$  is the probability of success. In Bayes' theorem, Equation (11.1), we refer to  $P(\theta)$  as the *prior*,  $P(X | \theta)$  as the *likelihood*, and  $P(\theta | X)$  as the *posterior*. We may restate the equation as

$$\text{Posterior} = \frac{\text{Prior} \times \text{Likelihood}}{\text{Data}}. \quad (11.2)$$

Suppose that  $\theta$  has  $k$  possible values,  $\theta_1, \theta_2, \dots, \theta_k$ , and let  $A_j$  be the event that  $\theta = \theta_j$ ; these are mutually exclusive events whose union is the whole sample space. By the law of total probability,

$$\begin{aligned} P(X) &= P(XA_1) + P(XA_2) + \cdots + P(XA_k) \\ &= P(A_1)P(X | A_1) + P(A_2)P(X | A_2) + \cdots + P(A_k)P(X | A_k) \\ &= \sum_{j=1}^k P(A_j)P(X | A_j). \end{aligned} \quad (11.3)$$

Thus, Equation (11.1) can be expressed as follows:

$$\begin{aligned} P(\theta = \theta_j | X) &= \frac{P(\theta = \theta_j)P(X | \theta = \theta_j)}{\sum_{i=1}^k P(A_i)P(X | A_i)} \\ &= \frac{P(\theta = \theta_j)P(X | \theta = \theta_j)}{\sum_{i=1}^k P(\theta = \theta_i)P(X | \theta = \theta_i)} \end{aligned} \quad (11.4)$$

$$= \frac{\text{Prior} \times \text{Likelihood}}{\sum \text{Prior} \times \text{Likelihood}}. \quad (11.5)$$

This means that the posterior is proportional to the product of the prior and the likelihood, and that the denominator is obtained by adding the numerator across all possible values for  $\theta$ . We call the denominator a *normalizing constant*. In general, a normalizing constant makes the distribution add or integrate to 1.

**Example 11.1** Suppose you are playing with someone new in tennis and you do not know who is stronger, but you suspect you are. Let  $\theta$  be the probability that you win a single game. For now we will treat  $\theta$  as discrete, with possible values  $0, 0.1, \dots, 1.0$ , and suppose you guess that the corresponding probabilities are

$\theta:$	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Prior $P(\theta):$	0.00	0.02	0.03	0.05	0.10	0.15	0.20	0.25	0.15	0.05	0.00

In other words, before you play, you believe that there is a 2% chance that your probability of winning any given game is 10%, a 3% chance that your winning probability is 20%, and so on. Suppose you win the first time (and you're feeling pretty smug), but then you lose twice (ouch). How does this information change your belief about the probabilities for  $\theta$ ?

For example, what is the posterior probability that  $\theta = 0.2$ ? The likelihood of one win and two losses (WLL) given  $\theta = 0.2$  is  $(0.2)^1(0.8)^2 = 0.128$ . So the posterior probability is

$$P(\theta = 0.2 | \text{WLL}) = \frac{P(\theta)P(\text{WLL} | \theta)}{P(\text{WLL})} = \frac{(0.03)(0.128)}{P(\text{WLL})}.$$

To complete this calculation, we need the normalizing constant, the marginal probability  $P(\text{WLL})$ . The following table shows the calculations:

$\theta$	Prior	Likelihood $\theta(1 - \theta)^2$	Prior $\times$ Likelihood	Posterior Previous / 0.0862
0.0	0.00	0.000	0.0000	0.0000
0.1	0.02	0.081	0.0016	0.0188
0.2	0.03	0.128	0.0038	0.0446
0.3	0.05	0.147	0.0073	0.0853
0.4	0.10	0.144	0.0144	0.1671
0.5	0.15	0.125	0.0188	0.2176
0.6	0.20	0.096	0.0192	0.2228
0.7	0.25	0.063	0.0158	0.1827
0.8	0.15	0.032	0.0048	0.0557
0.9	0.05	0.009	0.0004	0.0052
1.0	0.00	0.000	0.0000	0.0000
Sum	1		<b>0.0862</b>	1

By the law of total probability,

$$\begin{aligned} P(\text{WLL}) &= P(\theta = 0.1)P(\text{WLL} | \theta = 0.1) + P(\theta = 0.2)P(\text{WLL} | \theta = 0.2) \\ &\quad + \dots + P(\theta = 0.9)P(\text{WLL} | \theta = 0.9) \\ &= 0.0862, \end{aligned}$$

which is the sum of the entries in the fourth column (Prior  $\times$  Likelihood). So the posterior probability for  $\theta = 0.2$  is

$$P(\theta = 0.2 | \text{WLL}) = \frac{(0.03)(0.128)}{0.0862} = 0.0445.$$

In other words, prior to seeing any data, you thought there was a 3% chance that your long-term winning proportion is 0.2. After winning just one of the three games, you now believe that there is a 4.46% chance that the proportion is 0.2.

Similarly, you thought that  $P(\theta = 0.7) = 0.25$ , that is, there was a 25% chance that in the long run you would win an average of 70% of the games.

$$\begin{aligned} P(\theta = 0.7 \mid \text{WLL}) &= \frac{\text{Prior} \times \text{Likelihood}}{P(\text{WLL})} \\ &= \frac{0.25 \times 0.063}{0.0862} \\ &= 0.1827, \end{aligned}$$

so now you think there is an 18.28% chance of winning 70% of games.

Before the games, you thought your overall chances of winning were

$$E[\theta] = \sum_{i=0}^{10} \theta_i \times \text{Prior} = \sum_{i=0}^{10} \theta_i \times P(\theta_i) = 0.598,$$

almost 60%. Now, you think they are

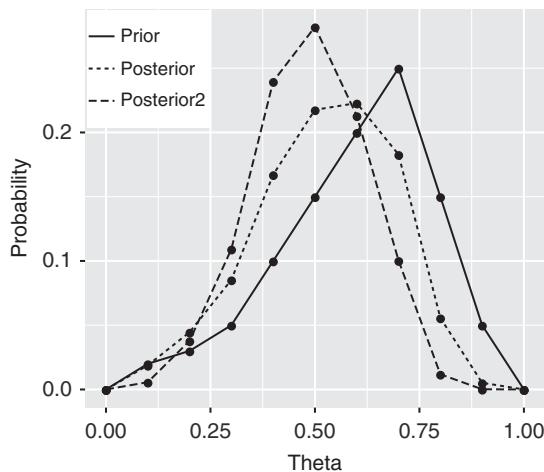
$$E[\theta \mid \text{WLL}] = \sum_{i=0}^{10} \theta_i \times \text{Posterior} = \sum_{i=0}^{10} \theta_i \times P(\theta_i \mid \text{WLL}) = 0.523;$$

you still think you are better!

### R Note

```
> theta <- seq(0, 1, by = .1)
> prior <- c(0, .02, .03, .05, .1, .15, .2, .25, .15, .05, 0)
> likelihood <- theta * (1 - theta)^2
> constant <- sum(prior * likelihood)
> posterior <- prior * likelihood / constant
> posterior
[1] 0.000000000 0.018802228 0.044568245 0.085306407
[5] 0.167130919 0.217618384 0.222841226 0.182799443
[9] 0.055710306 0.005222841 0.000000000
> sum(theta * prior) # prior mean
[1] 0.598
> sum(theta * posterior) # posterior mean
[1] 0.5229805
```

But what if you play five more times and lose three more? There are two different ways to calculate this – either start from scratch and use all the data, or use the current posterior as a new prior and work with just the new data. They are equivalent. Results are shown in Figure 11.1, using the following R code. Your posterior mean is now 0.49 – you are not so cocky anymore!



**Figure 11.1** Prior distribution and posterior distributions after three games (one win) and eight games (three wins), for tennis example.

### R Note

```
> likelihood2 <- theta^3 * (1 - theta)^5 # 3 success, 5 fail
> constant2 <- sum(prior * likelihood2)
> posterior2 <- prior * likelihood2 / constant2
> posterior2
[1] 0.0000000000 0.0056870025 0.0378705884 0.1092609179
[5] 0.2396498173 0.2821578712 0.2130220598 0.1003416593
[9] 0.0118345589 0.0001755248 0.0000000000
```

Now, using the previous posterior as a prior, we add two wins and three losses:

```
> likelihood3 <- theta^2 * (1 - theta)^3
> constant3 <- sum(posterior2 * likelihood3)
> posterior3 <- posterior2 * likelihood3 / constant3
> posterior3 # not shown, same as posterior2
> sum(theta*posterior2) # posterior mean
[1] 0.485538
```

For a plot:

```
ggplot(df, aes(x = theta, y = prior)) +
  geom_point() + geom_line(lty = 1) +
  geom_point(aes(y = posterior)) +
  geom_line(aes(y = posterior), lty = 2) +
  geom_point(aes(y = posterior2)) +
  geom_line(aes(y = posterior2), lty = 3)
```



**Example 11.2** Suppose you are conducting an exit poll during a state governors election. You survey every fifth voter leaving the polling station and ask him or her whether they voted for candidates Cobb or Moore. Before the election there appears to be a slight edge for Cobb, so you decide to assume that there is a 50% chance that Cobb will get 50% of the votes, but a 35% chance that he will get 40% and a 15% chance of getting 60% of the votes. Let  $\theta$  be Cobb's percentage of the votes.

If the first three voters that you survey tell you they voted for Cobb, Cobb, and Moore, respectively, what is the posterior probability of  $\theta = 0.4$ ?

### Solution

Given  $\theta = 0.4$ , the likelihood of CCM is  $0.4^2(1 - 0.4)^1 = 0.096$ . Thus, the posterior probability is

$$P(\theta = 0.4 \mid \text{CCM}) = \frac{P(\theta = 0.4)P(\text{CCM} \mid \theta = 0.4))}{P(\text{CCM})} = \frac{(0.30)(0.096)}{P(\text{CCM})}.$$

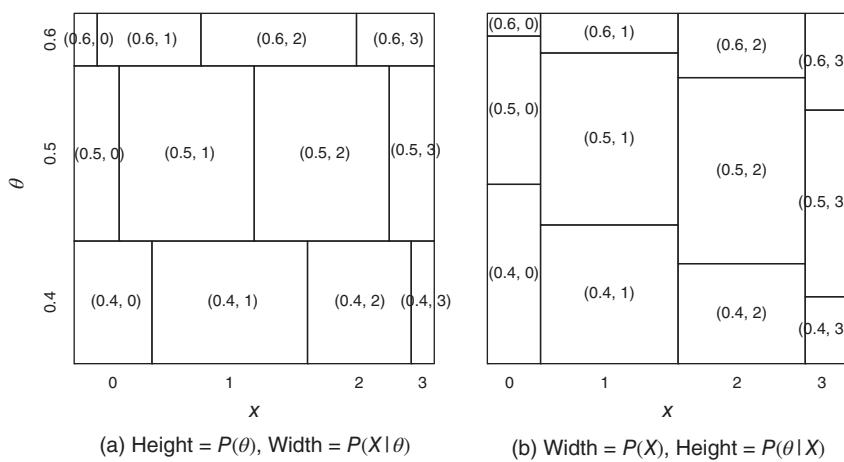
Again,  $P(\text{CCM})$  is found by summing the entries in the column labeled Prior  $\times$  Likelihood (see Table 11.1), so the posterior probability is  $P(\theta = 0.4 \mid \text{CCM}) = 0.2855$ . Thus, after surveying three voters, your belief that Cobb will get only 40% of the vote drops from 35% to 28.6%.

Figure 11.2 shows two Venn diagrams for the governor's exit poll (Example 11.2). In both sides, the area of a box equals the joint probability  $P(\theta, X)$  for the values of  $\theta$  and  $X$  (the number of people who voted for Cobb) shown in the box. Figure 11.2a has box heights equal to the prior probabilities, and widths equal to conditional probabilities for  $X$  given  $\theta$ . To find  $P(X = 2)$ , we add the areas for boxes with  $X = 2$ . Now look at Figure 11.2b at the column of boxes above the "2" at the bottom. These are versions of those  $X = 2$  boxes from Figure 11.2a – same area, but different shapes. Their width is  $P(X = 2)$ , and areas are  $P(\theta = \theta_j, X = 2)$ , so their heights must be the posterior probabilities  $P(\theta = \theta_j \mid X = 2)$ .

Compare the heights in Figure 11.2a (the prior) to the heights above  $X = 2$  in Figure 11.2b (the posterior given two votes for Cobb). We see that two Cobb votes makes the small values of  $\theta$  less likely and large values of  $\theta$  more likely.  $\square$

**Table 11.1** Calculations for Example 11.2.

$\theta$	Prior	Likelihood		Posterior
		$\theta^2(1 - \theta)$	Prior $\times$ Likelihood	Previous/0.1177
0.4	0.35	0.096	0.0336	0.2855
0.5	0.50	0.125	0.0625	0.5310
0.6	0.15	0.144	0.0216	0.1835
Sum	1		0.1177	1



**Figure 11.2** Venn diagrams with box areas showing joint probabilities for  $\theta$  and  $X$ . Each box has area equal to the joint probability  $P(\theta = \theta_j, X = x_i)$ . (a) Box heights are  $P(\theta)$  and widths are  $P(X | \theta)$ . (b) Box widths are  $P(X)$  and heights are  $P(\theta | X)$ , the result of Bayes' theorem.

### Remark

- The posterior must always add to 1 (or integrate to 1, for densities).
- Multiplying all likelihoods by the same constant does not change the posterior.

For instance, the calculations in the previous example were done assuming that order matters – you recorded voter preferences in the order they came out of the polling station. Suppose instead that we just know that out of the first three voters surveyed, two voted for Cobb. Then, if we assume that  $\theta = 0.40$ , then the likelihood is  $\binom{3}{2} 0.4^2(1 - 0.4)^1$ . So all the likelihoods in Table 11.1 would have an extra factor of  $c = \binom{3}{2}$ . However, this extra factor would carry over to the next column and to the calculation of the marginal probability  $P(CCM)$  and thus be canceled out in the end.

$\theta$	Prior	Likelihood	Prior $\times$ Likelihood	Posterior
		$\theta^2(1 - \theta)$		Previous / 0.1177c
0.4	0.35	0.096c	0.0336c	0.2855
0.5	0.50	0.125c	0.0625c	0.5310
0.6	0.15	0.144c	0.0216c	0.1835
Sum	1		0.1177c	1

- Multiplying all the priors by the same constant also does not change the posterior probability. Hence, we can be careless about specifying priors; they need not add to 1. We may use an *improper prior*. This is particularly useful in some calculations with continuous distributions below, where we let the prior be a function that integrates to  $\infty$ . So the posterior is unaffected by constants in either the prior or likelihood; what is important is that it is proportional to their product:

**Posterior  $\propto$  Prior  $\times$  Likelihood**

The posterior distribution is proportional to the prior times the likelihood,  $P(\theta | X) \propto P(\theta)P(X | \theta)$ .

- We can be careless about prior distributions in another way – they can include some impossible values, for example negative values of the rate parameter  $\lambda$  for an exponential distribution. The likelihood is zero for negative  $\lambda$ , so it will not affect the posterior. Still, including such negative values adds nothing but confusion and may be a sign of an error.

||

**Example 11.2 (continued)**

Suppose after having observed CCM, you poll four more voters and their preferences are CMCC. What is the new posterior distribution?

**Solution**

After having observed CCM, you have updated your beliefs about the probabilities for  $\theta$  – they are the values in the posterior column of Table 11.1. These values become your new *priors*, and the likelihoods are computed via  $\theta^3(1 - \theta)$ .

$\theta$	Prior	Likelihood $\theta^3(1 - \theta)$	Prior $\times$ Likelihood	Posterior Previous/0.0600
0.4	0.2855	0.0384	0.0109	0.1827
0.5	0.5310	0.0625	0.0332	0.5531
0.6	0.1835	0.0864	0.0159	0.2642
Sum	1		0.0600	1

Thus, we now think there is a 18.3% chance that Cobb will receive only 40% of the votes. Alternately, we could start over, work with the original prior, and

analyze all the data at once, using the sequence CCMCMCC. Then, we can fill out the table using the likelihood  $\theta^5(1 - \theta)^2$ :

$\theta$	Prior	Likelihood $\theta^5(1 - \theta)^2$	Prior $\times$ Likelihood	Posterior Previous / 0.0071
0.4	0.35	0.0037	0.0013	0.1827
0.5	0.50	0.0078	0.0039	0.5531
0.6	0.15	0.0124	0.0019	0.2642
Sum	1		0.0071	1

Note that we obtain the exact same posterior probabilities whether we analyze the data sequentially or altogether. We come back to this point in Section 11.5.  $\square$

### 11.3 Binomial Data: Continuous Prior Distributions

In the above examples, we used discrete prior distributions. This may be fine as a rough approximation but is bad in the longer term. We essentially claimed that it is impossible for the probability  $\theta$  to be 0.45 or any other value between 0.4 and 0.5, for example. That is just wrong.

To avoid this, we need to use a continuous prior distribution. Instead of writing  $P(\theta)$ , we will write  $\pi(\theta)$ , where  $\pi$  is some density (or an improper prior), and write  $p(\theta | X)$  for the posterior density.

For the data, we will work with the binomial model  $X \sim \text{Binom}(n, \theta)$ , so the likelihood is

$$\begin{aligned} P(X = x | \theta) &= \binom{n}{x} \theta^x (1 - \theta)^{n-x} \\ &\propto \theta^x (1 - \theta)^{n-x}. \end{aligned}$$

Then the density for the posterior distribution is

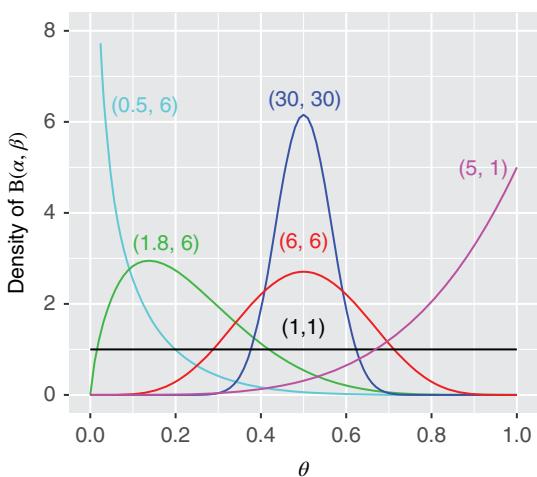
$$p(\theta | X = x) \propto \pi(\theta) \theta^x (1 - \theta)^{n-x}. \quad (11.6)$$

To get rid of the  $\propto$  and have an exact formula, we need to divide by the integral of the product of the prior and likelihood

$$p(\theta | X = x) = \frac{\pi(\theta) \theta^x (1 - \theta)^{n-x}}{\int_0^1 \pi(\theta) \theta^x (1 - \theta)^{n-x} d\theta},$$

a continuous analog of Equation (11.5). Depending on what we choose for  $\pi(\theta)$ , this integral can get nasty. However, it is relatively simple if  $\pi(\theta) \propto \theta^c (1 - \theta)^d$  for some constants  $c$  and  $d$ . In this case, we can combine terms to obtain

**Figure 11.3** Beta( $\alpha, \beta$ ) with different  $\alpha$ 's and  $\beta$ 's.



$p(\theta | x) \propto \theta^{x+c} (1 - \theta)^{n-x+d}$ . There is one family of distributions with that form, the beta distributions (see Section B.12). Some examples of the densities corresponding to different values of the parameters are shown in Figure 11.3. This is a flexible family of distributions, able to accommodate not only expected values – the prior  $E[\theta]$  estimates of your overall chance of winning – but also levels of uncertainty.

For example, compare the curves for (30, 30) and (6, 6) – they both have a mean of 50%, but the narrower curve expresses a much stronger prior belief that the long-term probability must be close to 50%, while the wider curve expresses much more uncertainty. The curves may be pessimistic (Beta(0.5, 6) has a vertical asymptote at  $\theta = 0$ ) or optimistic (e.g. Beta(5, 1)). This flexibility is useful for specifying prior distributions.

Now, using the beta family for prior distributions, we have  $\theta \sim \text{Beta}(\alpha, \beta)$ , so the prior density is

$$\pi(\theta) \propto \theta^{\alpha-1} (1 - \theta)^{\beta-1},$$

the likelihood is

$$P(X = x | \theta) \propto \theta^x (1 - \theta)^{n-x}$$

and hence the posterior density is

$$\begin{aligned} p(\theta | X = x) &\propto \pi(\theta) \times P(X = x | \theta) \\ &\propto \theta^{\alpha+x-1} (1 - \theta)^{\beta+n-x-1}, \end{aligned}$$

which is the density for Beta( $\alpha + x, \beta + n - x$ ).

The posterior distribution looks similar to the prior distribution, but with the number of successes  $x$  added to  $\alpha$ , and the number of failures  $n - x$  added to  $\beta$ . We can interpret  $\alpha$  and  $\beta$  as giving prior successes and failures, with the

posterior parameters giving total numbers of successes and failures. From Theorem B.20, we have

### Binomial Data, Beta Prior

For the binomial model  $X \sim \text{Binom}(n, \theta)$ , suppose we place the continuous prior distribution  $\text{Beta}(\alpha, \beta)$  on  $\theta$ . Then the prior mean and variance for  $\theta$  is

$$\text{E}[\theta] = \frac{\alpha}{\alpha + \beta}, \quad (11.7)$$

$$\text{Var}[\theta] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \quad (11.8)$$

and the posterior distribution of  $\theta$  given  $X = x$  is

$$\theta | x \sim \text{Beta}(\alpha + x, \beta + n - x)$$

with posterior mean and variance

$$\text{E}[\theta | x] = \frac{\alpha + x}{\alpha + \beta + n}. \quad (11.9)$$

$$\text{Var}[\theta | x] = \frac{(\alpha + x)(\beta + n - x)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)}. \quad (11.10)$$

### Remark

- The prior  $\text{Beta}(0, 0)$  is known as a *noninformative* prior with a posterior mean  $x/n$  determined solely by the data. This is an example of an improper prior that integrates to infinity.
- The prior  $\text{Beta}(1, 1)$  is the standard uniform distribution and is referred to as a *flat prior*.
- We say that the beta prior is a *conjugate family* to the binomial likelihood. A conjugate family is one for which the posterior distribution belongs to the same family as the prior distribution.

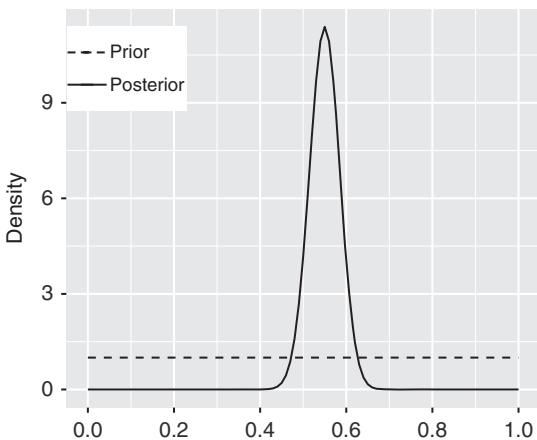
||

**Example 11.3** You conduct a poll of students and ask “Do you have a Twitter account?” Of the  $n = 200$  randomly chosen students you survey,  $x = 110$  say “yes.” Let  $\theta$  be the true proportion of students at your institution who would say yes (and assume that 200 is a negligible fraction of the population).

Let us compare the frequentist approach and two Bayesian approaches to this problem.

*Frequentist:* The estimated proportion of students who watch is  $\hat{\theta} = x/n = 110/200 = 0.55$ . The frequentist Matt computes a 95% confidence interval using the Agresti–Coull interval in Section 7.4.1, obtaining  $\tilde{\theta} \pm 1.96\sqrt{\tilde{\theta}(1 - \tilde{\theta})/\tilde{n}}$  where  $\tilde{n} = n + 4 = 204$  and  $\tilde{\theta} = (x + 2)/\tilde{n} = 112/204$ . Thus, the 95% confidence interval is  $(0.4807, 0.6174)$ .

**Figure 11.4** Pdf for Beta(111, 91), resulting from a uniform prior and 110 successes in 200 observations. The posterior distribution is concentrated near 0.55.



*Bayesian A:* One Bayesian analyst, Yuta, has no information about  $\theta$  and uses a flat prior. The prior density  $\pi(\theta)$  is from Beta(1, 1) with  $\alpha = 1$ ,  $\beta = 1$ . Hence, the posterior density  $p(\theta | X = 110)$  is from Beta( $1 + 110, 1 + 200 - 110$ ) = Beta(111, 91). This density is shown in Figure 11.4. This distribution is much narrower than the prior because the data provide a lot of information about  $\theta$ .

The posterior mean for  $\theta$  is

$$E[\theta | x = 110] = \frac{1 + 110}{1 + 1 + 200} = 111/202 = 0.5495.$$

This corresponds to a sample proportion with one artificial success and one failure added to the data. Yuta computes the middle 95% of the posterior distribution and finds (0.4807, 0.6174), nearly identical to the frequentist interval. There is a 95% probability that  $\theta$  is within this interval. This is known as a *credible interval*, rather than a confidence interval.

### Credible Interval

The Bayesian analog of a confidence interval is a credible interval. The range of the middle 95% of a posterior distribution, between the 0.025 and 0.975 quantiles, is a 95% credible interval.

Since we have the pdf for the posterior distribution, technically, we can actually find any probability we wish, evaluating the integral using statistical software or otherwise.

For instance,

$$P(\theta \geq 0.5 | X = 110) = \int_{0.5}^1 \frac{\Gamma(202)}{\Gamma(111)\Gamma(91)} \theta^{110}(1-\theta)^{90} d\theta = 0.9209,$$

where software is used to evaluate the integral. There is a 92% probability that at least 50% of the students at this school have a Twitter account.

### R Note

```
> qbeta(.025, 111, 91)
[1] 0.4806705
> qbeta(.975, 111, 91)
[1] 0.6174106
> 1-pbeta(.5, 111, 91)
[1] 0.9209173
```

To create the densities for the prior and posterior distributions in Figure 11.4,

```
ggplot(data.frame(x = c(0,1)), aes(x = x)) +
  stat_function(fun = dbeta, aes(lty = "2"),
                args = list(shape1 = 1, shape2 = 1)) +
  stat_function(fun = dbeta, aes(lty = "1"),
                args = list(shape1 = 111, shape2 = 91)) +
  scale_linetype_manual(values = c("2" = 2, "1" = 1),
                        labels = c("Posterior", "Prior"),
                        guide = guide_legend(reverse = TRUE)) +
  scale_x_continuous(breaks = seq(0, 1, by = .2)) +
  labs(x = "", y = "Density") +
  theme(legend.title = element_blank(),
        legend.position = c(.1, .85),
        legend.key = element_blank())
```

*Bayesian B:* Based on previous polls at this or other institutions, Saahithi expects that 58% of students have a Twitter account with an uncertainty quantified by a standard deviation of 0.03.

She will use a beta prior. As is common in practice, the prior is given in a form that makes sense for the person stating the prior, rather than in a form convenient for the analysis. Here, she needs to find parameters for the beta prior that match the given mean and standard deviation. Using Equations (11.7) and (11.8), she solves the following:

$$E[\theta] = \frac{\alpha}{(\alpha + \beta)} = 0.58.$$

$$\text{Var}[\theta] = \frac{\alpha \cdot \beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} = 0.03^2.$$

This yields  $\alpha = 156.4$  and  $\beta = 113.26$ .

Saahithi has quite a bit of prior information corresponding to about  $156 + 113 = 269$  prior observations.

### Solving Two Equations in Two Unknowns

Computer algebra systems such as Mathematica™ or Maple™ can solve the system of two equations and two unknowns above. The website <http://www.wolframalpha.com> provides another means for students to solve two equations with two unknowns:

```
Solve[{a/(a+b) == .58, (a*b)/((a+b)^2*(a+b+1)) == .03^2}, {a,b}]
```

The resulting posterior probability distribution (Equation (11.3)) is

$$\text{Beta}(110 + 156.4, 200 - 110 + 113.26) = \text{Beta}(266.4, 203.26).$$

The moments of the posterior, using Equations (11.9) and (11.10), are:

$$E[\theta | X = 110] = \frac{266.4}{266.4 + 203.26} = 0.567,$$

$$\text{Var}[\theta | X = 110] = \frac{(266.4)(203.26)}{(266.4 + 2.3.26)^2(266.4 + 203.26 + 1)} = 0.00005.$$

The posterior mean for  $\theta$  is 0.567, a bit lower than the prior mean, with a standard deviation of 0.0228, 24% smaller than the prior standard deviation of 0.03. The prior and posterior are shown in Figure 11.5.

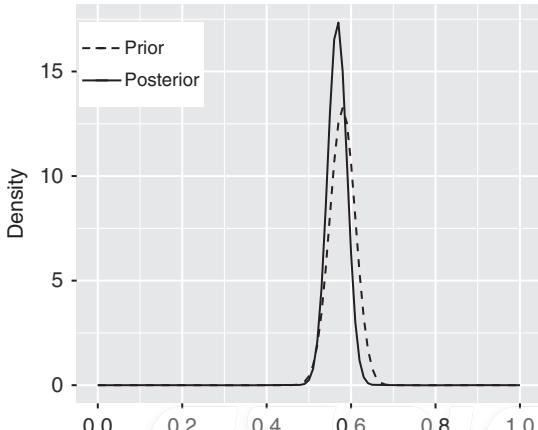
The 95% credible interval is the range from the 0.025 and 0.975 quantiles of Beta(266.4, 203.26), or (0.5222, 0.6117). The interval is narrower than either of the previous intervals, reflecting the amount of Saahithi's prior information. It is centered to the right of the previous intervals, reflecting her prior belief about the true probability.

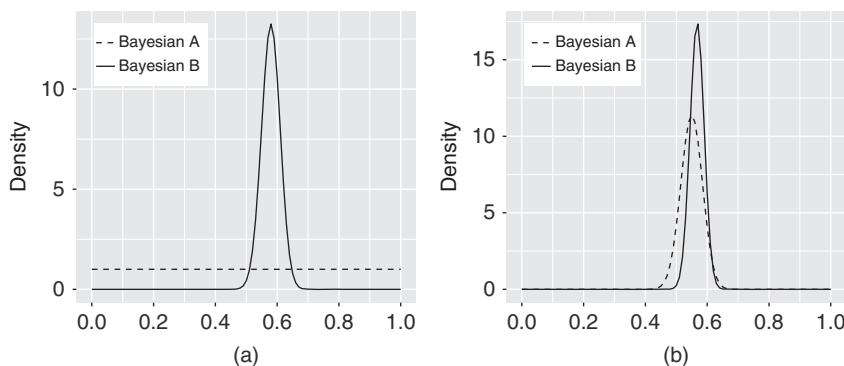
Figure 11.6 compares the prior and posterior densities for the two Bayesian approaches.

Even though the prior distributions are quite different, after observing the data,  $x = 110$ , the posterior distributions are similar. In practice, the data usually swamps the prior – with a reasonable amount of data, the posterior usually depends more on the data than on the prior, as long as the prior distribution is not unreasonable.  $\square$

**Remark** In the previous example, we saw that the posterior distribution depends more on the data than on the prior. This is usually the case in practice – the prior provides a framework to get an answer, but most of the information comes from the data.

**Figure 11.5** Pdfs for the prior Beta(156, 113.26) and posterior Beta(266.4, 203.26) for Bayesian B.





**Figure 11.6** Comparison of pdfs for the priors of Bayesians A and B, as well as the posteriors of Bayesians A and B.

One exception is when the prior is badly chosen – for instance, the prior at some values of  $\theta$  is zero when they are actually possible. Then the posterior is always zero at these values, regardless of the data. One example of this is the use of a discrete prior distribution for a binomial success probability  $\theta$ . For example, our exit poll in Example 11.2 allowed only three values,  $\theta = 0.4, 0.5$ , or  $0.6$ , so the posterior says that any other  $\theta$  is impossible.

In the Twitter account example, if a prior distribution had been unreasonable, say indicating zero chance of  $\theta > 0.1$ , then no amount of data would change that conclusion. (Perhaps you know someone like that?) ||

## 11.4 Continuous Data

Most of what we learned previously about Bayesian methods for discrete data also applies to continuous data, with one notational change – instead of working with probabilities for the data, we work with densities for the data. In this section, we also focus on continuous  $\theta$ , the more common case.

Let  $\pi(\theta)$  denote the pdf for the prior distribution and  $f$  the density for the data, given  $\theta$ . Then Bayes' theorem becomes

$$\begin{aligned}
 p(\theta | x) &= \frac{\pi(\theta)f(x | \theta)}{f(x)} \\
 &= \frac{\pi(\theta)f(x | \theta)}{\int \pi(\theta)f(x | \theta)d\theta} \\
 &\propto \pi(\theta)f(x | \theta).
 \end{aligned} \tag{11.11}$$

We begin with the case of normal distributions. Suppose the data  $x_1, x_2, \dots, x_n \sim N(\mu, \sigma^2)$ , where  $\mu$  is unknown but  $\sigma^2$  is known. The likelihood of  $\mu$  is

$$L(\mu) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x_i - \mu)^2}{2\sigma^2}\right].$$

As in the discrete case, the analysis is much simpler if we use conjugate prior distributions. It turns out that for normal data with known  $\sigma^2$ , the normal distributions are a conjugate family. We will assume a normal prior  $\mu \sim N(\mu_0, \sigma_0^2)$  so that  $\pi(\theta) = 1/(\sqrt{2\pi}\sigma_0)e^{-(\mu-\mu_0)^2/(2\sigma_0^2)}$ . Thus,

$$\begin{aligned} p(\mu | x_1, x_2, \dots, x_n) &\propto \exp\left[-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right] \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x_i - \mu)^2}{2\sigma^2}\right] \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left[-\frac{(\mu - \mu_0)^2}{2\sigma_0^2} - \sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2}\right]. \end{aligned}$$

Now, a tedious algebraic calculation involving multiplying out the products in the exponents, collecting like terms, and completing the square results in the posterior density simplifying to

$$p(\mu | x_1, x_2, \dots, x_n) \propto \exp\left[-\frac{(\mu - \mu_1)^2}{2\sigma_1^2}\right], \quad (11.12)$$

where  $\mu_1$  and  $\sigma_1^2$  are given below (Equations (11.13) and (11.14)). Thus, the posterior distribution of  $\mu$  is normal,  $N(\mu_1, \sigma_1^2)$ .

### Normal Data, Normal Prior

Let  $x_1, x_2, \dots, x_n \sim N(\mu, \sigma^2)$ , where  $\mu$  is unknown but  $\sigma^2$  is known. Let  $\bar{x}$  denote the sample mean.

If the prior distribution of  $\mu$  is  $N(\mu_0, \sigma_0^2)$ , then the posterior distribution of  $\mu$  is

$$\mu | x_1, x_2, \dots, x_n \sim N(\mu_1, \sigma_1^2),$$

where

$$\mu_1 = \frac{\frac{1}{\sigma_0^2}}{\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)} \mu_0 + \frac{\frac{n}{\sigma^2}}{\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)} \bar{x} \quad (11.13)$$

and

$$\frac{1}{\sigma_1^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}. \quad (11.14)$$

The updated mean  $\mu_1$  is a weighted average of the prior mean and the observed data.

**Definition 11.1** The reciprocal of the variance,  $1/\sigma^2$ , is called the *precision* for a normal distribution. ||

You can think of precision as information. The information of the posterior is the sum of the information provided by the prior and the data. Furthermore, the updated mean  $\mu_1$  is a weighted average, with weights proportional to the information provided by the prior and the data.

**Example 11.4** A biologist is investigating a species of trout in a certain area of California. She assumes the lengths of these fish are normally distributed with mean  $\mu$  (cm) and variance  $8^2$ . She obtains a random sample of 15 fish and records their lengths,  $x_1, x_2, \dots, x_{15} \sim N(\mu, 8^2)$ . Based on her knowledge of this species at other locations, she assumes that the prior distribution is  $\mu \sim N(50, 6^2)$ . Suppose the mean of this random sample is  $\bar{x} = 45$  cm.

We have  $n = 15$ ,  $\sigma^2 = 8^2$ ,  $\mu_0 = 50$ ,  $\sigma_0^2 = 6^2$ , and  $\mu$  is unknown. Thus,

$$\begin{aligned}\mu_1 &= \frac{1/36}{1/36 + 15/64} \times (50) + \frac{15/64}{1/36 + 15/64} \times (45) \\ &= 0.106(50) + 0.894(45) \\ &= 45.53\end{aligned}$$

$$\begin{aligned}\frac{1}{\sigma_1^2} &= \frac{1}{36} + \frac{15}{64} \\ &= 0.2622,\end{aligned}$$

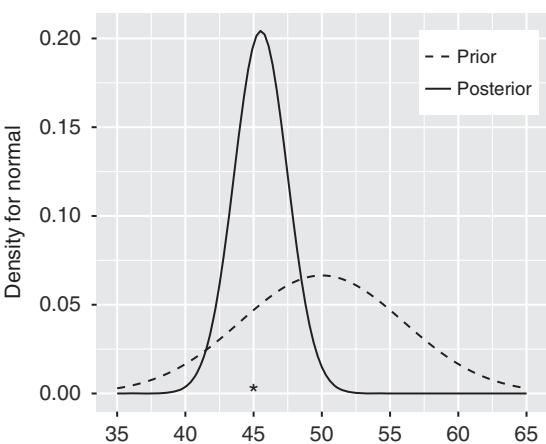
which results in a variance of  $\sigma_1^2 = 3.8146 = 1.953^2$ . Her updated belief about  $\mu$  is  $\mu \mid x_1, x_2, \dots, x_{15} \sim N(45.53, 1.953^2)$ . In this example, the observed mean receives more weight in determining  $\mu_1$ . The prior and posterior are shown in Figure 11.7.

A 95% credible interval for  $\mu$  can be obtained by finding the 0.025 and 0.975 quantiles of  $N(45.53, 1.953^2)$ : the probability that the true  $\mu$  lies in the interval (41.70, 49.36) is 0.95. □

### Remark

- The noninformative prior for the normal distribution is  $\pi(\theta) = 1$  for all  $\theta$  (an improper prior).
- The case of normal distributions with unknown  $\sigma^2$  is more complicated; the conjugate family has inverse Gamma distributions for  $\sigma^2$  (i.e. where the precision  $1/\sigma^2$  has a Gamma distribution). This is beyond the scope of this book; we encourage interested students to see Carlin and Louis (2009) or Gelman et al. (2010) for more details. ||

**Figure 11.7** Prior  $N(50, 6^2)$  and posterior  $N(45.53, 1.953^2)$  distributions of  $\mu$ . The observed mean  $\bar{x} = 45$  has more weight in determining the mean of the posterior distribution than does the prior mean.



## 11.5 Sequential Data

We commented at the end of Section 11.2 that when data arrives over time, we may either analyze all the data at once, using the original prior, or may use the prior from an earlier step and the new data.

This is easiest to show using the notation from continuous distributions (though everything that follows also applies to discrete data, if we interpret  $f$  as being either a density or probability mass function, for continuous and discrete distributions, respectively).

The posterior distribution after observing data  $x_1$  is

$$p(\theta \mid x_1) = \frac{\pi(\theta)f(x_1 \mid \theta)}{\int \pi(\theta)f(x_1 \mid \theta)d\theta}.$$

The posterior distribution based on data  $x_2$ , using the first posterior as a prior, is

$$\begin{aligned} p_2(\theta \mid x_1, x_2) &= \frac{p(\theta \mid x_1)f(x_2 \mid \theta)}{\int p(\theta \mid x_1)f(x_2 \mid \theta)d\theta} \\ &= \frac{\left( \pi(\theta)f(x_1 \mid \theta) / \int \pi(\theta)f(x_1 \mid \theta)d\theta \right) f(x_2 \mid \theta)}{\left( \int \pi(\theta)f(x_1 \mid \theta) / \int \pi(\theta)f(x_1 \mid \theta)d\theta \right) f(x_2 \mid \theta)d\theta} \\ &= \frac{\pi(\theta)f(x_1 \mid \theta)f(x_2 \mid \theta)}{\int \pi(\theta)f(x_1 \mid \theta)f(x_2 \mid \theta)d\theta} \\ &= \frac{\pi(\theta)f(x_1, x_2 \mid \theta)}{\int \pi(\theta)f(x_1, x_2 \mid \theta)d\theta}, \end{aligned}$$

where the last step assumes that  $x_1$  and  $x_2$  are independent, given  $\theta$ . This is Bayes' formula for the combined data. In other words, we get the same posterior whether we analyze the data in one step or two. We can iterate this and find that the posterior distribution is the same whether we analyze data one observation at a time or all at once. This self-consistency of the answers is a valued property of Bayesian estimates.

**Example 11.5** Suppose from a random sample of 38 students, 10 approve of the new graduation requirements. To estimate  $\theta$ , the true proportion of students who approve, we will use a beta distribution and assume a flat prior. Thus, the posterior distribution, with  $x = 10$ ,  $n = 38$ , and  $\alpha = 1$ ,  $\beta = 1$  is  $P(\theta | x = 10) = \text{Beta}(10 + 1, 1 + 38 - 10) = \text{Beta}(11, 29)$ . Thus,  $E[\theta | x = 10] = 11/40 = 0.275$ .

Suppose we poll an addition 15 students and find 5 approve of the new graduation requirements. We update the posterior using  $x = 5$ ,  $n = 15$  and  $\alpha = 11$ ,  $\beta = 29$ , to find  $P(\theta | x = 5) = \text{Beta}(5 + 1, 29 + 15 - 5) = \text{Beta}(16, 39)$ . Thus,  $E[\theta | x = 10] = 15/53 = 0.283$ . On the other hand, suppose we wait until we have all the data. In this case, we have  $x = 10 + 5 = 15$  out of  $n = 38 + 15 = 53$  who approve of the graduation requirements. Then, starting with the flat prior, the posterior distribution is  $P(\theta | x = 15) = \text{Beta}(15 + 1, 1 + 53 - 15) = \text{Beta}(16, 39)$ , which matches the two-step answer.  $\square$

Sequential data are important in a wide range of applications, including the following:

- Clinical trials, where data from new patients arrives at different times. Rather than waiting for the end of a multiyear trial to analyze all data, the Food and Drug Administration (FDA) mandates that the data be analyzed periodically. If a new drug or treatment appears to be harming patients, the trial may be stopped early. A trial may be stopped early because the new drug is convincingly better than anything else available, though this is rare because the FDA also wants enough data to look for adverse side effects.
- Google continuously updates its algorithms for producing search results.
- Google Analytics Content Experiments, described in Section 8.5, reports results to website owners on demand using the latest data.

Bayesian analysis offers substantial advantages in these situations. Frequentist testing suffers from multiple testing – if you do a hypothesis test many times, each at the 5% level, and keep collecting new data, then eventually a test will come out rejecting the null, even when the null hypothesis is true. To compensate for that, frequentist clinical trials must be designed and analyzed with special software that calculates how to adjust critical values and  $P$ -values

due to sequential testing. (One of the authors worked on such software, *S+SeqTrial*; it is complicated and expensive.)

In contrast, Bayesian analysis may be performed as often as desired. Say that we are interested in the difference between  $\theta$  values for the treatment and control groups,  $\theta_t - \theta_c$ . Say that a difference of 30 would be important in practice. We might stop a trial whenever there is a high probability of an improvement that large, say  $P(\theta_t - \theta_c > 30) \geq 80\%$ . Or we might stop when the probability of improvement is at least 90% and the potential expected loss  $E(\max(0, \theta_c - \theta_t)) \leq 15$ . This prevents stopping too early, when the posterior for  $\theta_t - \theta_c$  is wide.

Bayes does not completely avoid the inflated false positive issue – using a non-informative prior and stopping whenever  $P(\theta > 0) \geq 0.95$  is exactly the same as stopping when the  $P$ -value  $\leq 0.05$ , with the same inflated false positive rate if  $\theta = 0$ . But from a Bayesian viewpoint that is unimportant because there is zero probability that  $\theta$  exactly equals zero. Good Bayesian stopping rules care not just about  $P(\theta > 0)$ , but by how much.

At Instacart, we are considering switching to Bayesian rules for deciding when to stop experiments. We are a small company: we do not have as much data as Google, and we want to stop experiments quickly to incorporate successful experiments into our product and to stop unsuccessful experiments quickly. I (Tim) just started working here; we have not decided yet – stay tuned.

Bayesian analysis also offers advantages in *multiple-arm trials*,<sup>1</sup> when three or more things are being compared.

- In clinical trials, instead of just one treatment arm and a control arm, there may be multiple treatment arms at different levels of a drug, or perhaps using combinations of drugs.
- In Google Analytics Content Experiments, website owners can test multiple arms (different versions of their page) simultaneously, using different combinations of graphics and text content.

For example, a website owner may choose to stop the experiment if there is a 60% probability that a particular version of the page is better than all other versions being tested. If the current posterior distribution for each  $\theta_j$  (for arm  $j$ ) has a beta distribution with parameters  $(\alpha_j + x_j, \beta_j + n_j - x_j)$ , and the results from different arms are independent, then a simple simulation using values randomly generated from the posterior distributions can estimate that probability.

For another example, suppose a content publisher (a website owner) is testing six versions of her home page (arms), and the numbers of impressions (visitors)

<sup>1</sup> The terminology comes from casinos with *multi-armed bandits*, slot machines with more than one arm – you get to choose which arm to pull to lose your money.

on each arm and corresponding successes (e.g. purchases or donations) to date are the following:

<i>n</i>	1874	1867	1871	1868	1875	1875
<i>X</i>	52	41	55	49	39	39

(This is artificial data, motivated by Content Experiments.) To estimate the probability that each arm is best, using prior consisting of independent noninformative beta priors, we use the following code:

### R Note

```
n <- c(1874, 1867, 1871, 1868, 1875, 1875)
X <- c(52, 41, 55, 49, 39, 39)
alpha <- X      # vector of posterior parameters
beta <- n - X  # vector of posterior parameters
N <- 10^5          # replications
theta <- matrix(0.0, nrow = N, ncol = 6)
for (j in 1:6)
{
  theta[, j] <- rbeta(N, alpha[j], beta[j])
}
probBest <- numeric(6)      # vector for results
best <- apply(theta, 1, max) # maximum of each row
for (j in 1:6)
{
  probBest[j] <- mean(theta[, j] == best)
}
```

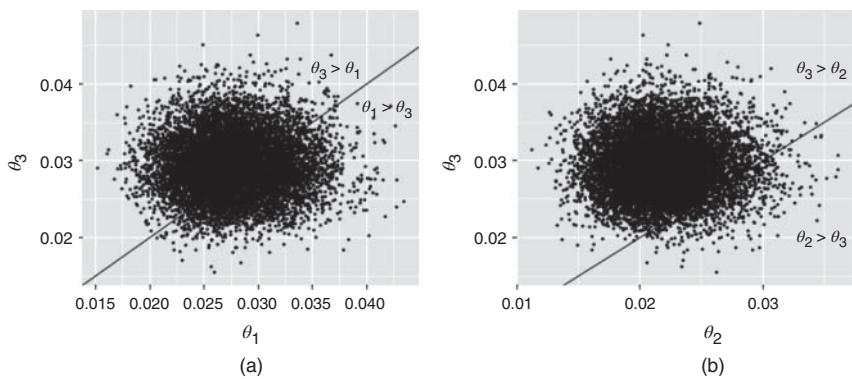
The vector `probBest` contains the probabilities of each of the six arms being best.

```
> probBest
[1] 0.29254 0.02027 0.49754 0.17027 0.00975 0.00963
```

These commands reproduce Figure 11.8a below.

```
df <- data.frame(theta[1:10^4,])
names(df)
ggplot(df, aes(x = X1, y = X3)) + geom_point(size = .5) +
  geom_abline(slope = 1, intercept = 0) +
  annotate("text", x = 0.037, y = 0.042, parse = TRUE,
           label = "theta[3] > theta[1]" ) +
  annotate("text", x = 0.042, y = 0.037, parse = TRUE,
           label = "theta[1] > theta[3]" ) +
  labs(x = expression(theta[1]), y=expression(theta[3]))
```

Figure 11.8a shows the joint posterior distribution for two of the parameters  $\theta_1$  and  $\theta_3$ . While  $\theta_3$  has a better chance of being the better of the two, there



**Figure 11.8** Posterior distribution for artificial Google Analytics Content Experiments example. (a) Joint posterior for  $\theta_1$  and  $\theta_3$ , with regions, where each is better than the other. (b)  $\theta_2$  and  $\theta_3$ .

is a good chance that  $\theta_1$  is better, and possibly substantially so. In contrast, Figure 11.8b shows the joint posterior distribution for  $\theta_2$  and  $\theta_3$ ;  $\theta_2$  has only a small chance of being the better of the two, and even smaller chance of being substantially better.

We estimate that arm 3 has almost a 50% chance of being the best, arm 1 almost 30%, and the other arms smaller chances. Google Analytics Website Experiments estimates like this to direct more traffic to the better-performing arms (50% to arm 3, etc.) As more data are collected, a higher proportion of traffic is sent to the best-performing arm, while still collecting some data from all arms, in case the real best arm was just unlucky at the start.

## Exercises

- 11.1** We return to your favorite game. A friend thinks the possible values for  $\theta$ , your long-term probability of winning, are 0.4, 0.5, 0.6, with probabilities  $1/5, 3/5, 1/5$ , respectively. Suppose you win four games out of five.

$\theta$	Prior	Likelihood	Prior $\times$ Likelihood	Posterior
0.4	$1/5$			
0.5	$3/5$			
0.6	$1/5$			
Sum	1			

- (a) Fill out the rest of this table.  
 (b) Give a sentence interpreting the posterior probability for  $\theta = 0.6$ .  
 (c) Find the expected value of  $\theta$  and the posterior expected value of  $\theta$ .

- 11.2** Another friend is positive that the long-term probabilities are one of 0.4, 0.5, 0.6, with probabilities  $1/10, 3/10, 1/10$ . (He's not too good at math, these don't add to 1.) The corresponding table, with  $c = 1/2$ , is

$\theta$	Prior	Likelihood of WLL	Prior $\times$ Likelihood	Posterior
0.4	$c/5$	0.144	0.0288c	
0.5	$3c/5$	0.125	0.0750c	
0.6	$c/5$	0.096	0.0192c	
Sum				

Another friend insists on using binomial probabilities with this table (with  $d = 3$ ):

$\theta$	Prior	Likelihood of WLL	Prior $\times$ Likelihood	Posterior
0.4	$c/5$	0.144d	0.0288cd	
0.5	$3c/5$	0.125d	0.0750cd	
0.6	$c/5$	0.096d	0.0192cd	
Sum				

Verify that the posterior probabilities are the same.

- 11.3** Suppose the possible values of  $\theta$  for your winning probabilities in tennis are  $\theta$  are  $0, 0.1, \dots, 1$ , with equal probabilities.

$\theta$	Prior	Likelihood	Prior $\times$ Likelihood	Posterior
0.0	$1/11$			
0.1	$1/11$			
0.2	$1/11$			
0.3	$1/11$			
0.4	$1/11$			
0.5	$1/11$			
0.6	$1/11$			
0.7	$1/11$			
0.8	$1/11$			
0.9	$1/11$			
1	$1/11$			
Sum	1			

- (a) Fill out the rest of this table, if you win four game out of seven. Hint:  
Adapt the R code in Example 11.1.
- (b) Give a sentence interpreting the posterior probability for  $\theta = 0.2$ .
- (c) Find the expected value of  $\theta$  and the posterior expected value of  $\theta$ .
- 11.4** (Exercise 11.3 continued) Suppose you play another six games and you win five of them.
- (a) Compute the new posterior probabilities using the posteriors from Exercise 11.3 as your new priors.
- (b) Compute the posterior probabilities by considering the data altogether, that is, by considering 13 games of which you won 9.
- 11.5** According to a Pew Research Center survey conducted in 2021, 281 out of 1124 participants in Singapore cited their occupation and career when describing what gives them meaning in life.<sup>2</sup>
- (a) Find a 90% confidence interval for the true proportion of adults in Singapore (in 2021) for whom their career and occupation is what gives them meaning in life.
- (b) A Bayesian analyst thinks (prior to seeing the data) that the true proportion has a beta distribution with a mean of 0.3 and standard deviation 0.05. Find a 90% credible interval for the true proportion.
- (c) For this Bayesian, find the probability that  $\theta > 0.30$ , given the data.
- (d) Suppose the analyst discovers another 500 survey results that they forgot to include. In this sample, 145 cite their career and occupation when describing what gives them meaning in life. Update the 90% credible interval.
- 11.6** *Toxoplasma gondii* is a parasite that can infect humans who eat raw or undercooked meat. Dogs can carry the parasite so researchers conducted a study to determine prevalence of *T. gondii* in pet dogs in Shenyang, China (Yang et al., (2013)). In a sample of 328 pet dogs, 33 tested positive for *T. gondii* antibodies in their blood.
- (a) Assuming this is a random sample of pet dogs in Shenyang, find a 90% confidence interval for the true proportion of pet dogs in Shenyang who are positive for *T. gondii*.
- (b) Based on a previous study done in Shanghai, a Bayesian analyst thinks (prior to seeing the data) that the true proportion has a beta distribution with a mean of 0.039 and standard deviation 0.02. Find a 90% credible interval for the true proportion as well as the posterior mean.

---

<sup>2</sup> <https://www.pewresearch.org/global/2021/11/18/what-makes-life-meaningful-views-from-17-advanced-economies>.

- (c) Suppose the researchers collect blood samples from an additional 258 pet dogs and find that 28 of these test positive for the *T. gondii* antibodies. Compute the 90% credible interval and the posterior mean.
- 11.7** The Pew Research Center conducted a survey in the fall of 2005 on attitudes toward pets. For those who owned dogs, they asked “Do you think of your dog as a member of your family?” Of the 178 people who were between 18 and 29 years old, 160 responded yes.<sup>3</sup>
- Find a 95% confidence interval for the true proportion of 18–29 years old who responded yes.
  - A Bayesian statistician at this polling company suspects that the prior for  $\theta$ , the true proportion, is a beta distribution with a mean  $\theta$  of 0.85 and a variance of 0.0025. A second statistician uses a flat prior. A third statistician thinks the prior is Beta(6, 4). Find the estimates for  $\theta$  (the posterior means) and 95% credible intervals.
  - For each statistician, plot their prior and posterior distributions on one graph.
  - For each statistician, find the probability, given the data, that  $\theta > 0.90$ .
- 11.8** Analyze the Twitter survey question (Example 11.3) using the noninformative prior.
- 11.9** Suppose you want to find the mean  $\mu$  of the math SAT scores for the class of 2019 high school seniors in your city. You know the distribution of scores is normal with a standard deviation of 117, which is the standard deviation for the national distribution of scores. On the basis of information on previous classes, you put a normal prior on  $\mu$ , say  $\mu \sim N(600, 25^2)$ . If a sample of size 60 yields a mean score of 538,
- Find the posterior distribution of  $\mu$ .
  - Find a 95% credible interval for the true  $\mu$ .
  - Find the probability that the posterior mean math SAT score is greater than 599.
- 11.10** Bone fractures are a common medical problem as people age, so there is much interest in determining risk factors. One assessment tool is bone mineral density (BMD), a measure of the amount of certain minerals such as calcium in the bone (in  $\text{g}/\text{cm}^2$ ). The lower the BMD, the higher the risk for bone fractures. Suppose a researcher wants to determine the mean BMD  $\mu$  in female vegetarians between

---

<sup>3</sup> <http://pewresearch.org/pubs/303/gauging-family-intimacy>.

30 and 39 years old. On the basis of other research, he believes  $\mu \sim N(0.72, 0.08^2)$ . He measures the BMD in the lumbar spine for 18 female vegetarians aged 30–39 years old and finds a mean BMD of  $\bar{x} = 0.85 \text{ g/cm}^2$ . He will assume that BMD measurements for this population come from a normal distribution with unknown mean  $\mu$  but known standard deviation  $\sigma = 0.15$ .

- (a) Find the posterior distribution of  $\mu$ .
- (b) Find a 99% credible interval for  $\mu$ .
- (c) Find the posterior probability that  $\mu$  is less than or equal to  $0.8 \text{ g/cm}^2$ .

- 11.11** A biologist is trying to determine the true mean weight  $\mu$  of a certain species of fish in Lyman Lake. She is certain that the distribution of weights is normal with known standard deviation  $\sigma = 100 \text{ g}$ . Based on prior work, her belief about  $\mu$  is  $\mu \sim N(900, 160^2)$ . A random sample of  $n = 10$  fish yields a sample mean of  $\bar{x} = 970 \text{ g}$ .
- (a) Find the posterior distribution of  $\mu$ .
  - (b) Find a 95% credible interval for  $\mu$ .
  - (c) Suppose she goes out and gets another sample of  $n = 15$  fish and computes  $\bar{x} = 940 \text{ g}$ . Find the new posterior distribution and a 95% credible interval for  $\mu$ .

- 11.12** Suppose  $x_1, x_2, \dots, x_n \sim N(\mu, 6^2)$  with mean  $\bar{x} = 19$ .
- (a) If  $n = 15$  and the prior is  $\mu \sim N(25, 5^2)$ , find the posterior distribution (mean, standard deviation) of  $\mu$  and the posterior precision.
  - (b) If  $n = 50$  and the prior is  $\mu \sim N(25, 5^2)$ , find the posterior distribution of  $\mu$  and the posterior precision.
  - (c) If  $n = 15$  and the prior is  $\mu \sim N(25, 10^2)$ , find the posterior distribution of  $\mu$  and the posterior precision.
  - (d) Compare the above three outcomes, and discuss the impact of sample size and the standard deviation of the prior distribution on the posterior mean, standard deviation, and precision.

- 11.13** Suppose  $x_1, x_2, \dots, x_{15}$  are data from a normal distribution and  $\bar{x} = 30$  and the prior distribution is normal with mean  $\mu_0 = 40$  and  $\sigma_0^2 = 5^2$ , compute the posterior distribution and posterior precision if
- (a) The data distribution is  $N(\mu, 3^2)$ .
  - (b) The data distribution is  $N(\mu, 10^2)$ .
  - (c) Compare the two outcomes and discuss the impact of the standard deviation of the data distribution.

- 11.14** Name two disadvantages for using a normal distribution as the prior distribution for binomial data.

- 11.15** Show that the posterior distribution resulting from the noninformative prior  $\pi(\theta) = 1$  for a normal mean (with known  $\sigma^2$ ) is equal to the limit as  $\sigma_0^2 \rightarrow \infty$  of the posterior resulting from the informative prior  $\mu \sim N(\mu_0, \sigma_0^2)$ .
- 11.16** Let  $X_1, X_2, \dots, X_N \stackrel{\text{i.i.d.}}{\sim} \text{Unif}[0, \theta]$ .
- Write down the likelihood function  $f(\theta)$ .
  - Suppose the prior distribution for  $\theta$  is from a Pareto distribution with pdf  $\pi(\theta | \alpha, \beta) = \alpha\beta^\alpha / \theta^{\alpha+1}$  for  $\theta \geq \beta > 0, \alpha > 0$ . Write down the posterior density and conclude that the Pareto family of distributions is a conjugate family to the uniform distribution.
  - Suppose the random sample 3, 6, 8, 10 is drawn from  $\text{Unif}[0, \theta]$ , and you use the Pareto distribution with  $\alpha = 0.3, \beta = 5$  as your prior distribution. Find the probability that  $\theta \geq 15$ .
- 11.17** Let  $X_1, X_2, \dots, X_N$  be a random sample from the Poisson distribution with pdf  $f(x) = \theta^x e^{-\theta} / x!, x = 0, 1, 2, \dots$
- Write down the likelihood function  $f(\theta)$ .
  - Suppose the prior for  $\theta$  is the gamma distribution with parameters  $r, \lambda$ . Let  $\pi(\theta)$  denote the pdf for this prior. Find the posterior density  $f(\theta)\pi(\theta)$ .
  - Recognize this posterior density as the pdf for what known distribution with what parameters?
  - Suppose you observe the values 6, 7, 9, 9, 16 and you believe the prior is gamma with parameters  $r = 15, \lambda = 3$ . Find the posterior density.
  - Find a 95% credible interval for the  $\theta$ .
- 11.18** Let  $X_1, X_2, \dots, X_N$  be a random sample from the exponential distribution with pdf  $f(x) = \theta e^{-\theta x}$ .
- Write down the likelihood function  $L(\theta)$ .
  - Suppose the prior for  $\theta$  is the gamma distribution with parameters  $r, \lambda$ . Let  $\pi(\theta)$  denote the pdf for this prior. Find the posterior density  $L(\theta)\pi(\theta)$ .
  - Recognize this posterior density as the pdf for what distribution with what parameters?
  - Suppose you observe the values 1, 2, 4, 6 and you believe the prior is gamma with parameters  $r = 10, \lambda = 4$ . Find the posterior density.
  - Find a 95% credible interval for the  $\theta$ .
- 11.19** In Section 11.5, we noted that a big advantage of Bayesian statistics is the ability to analyze data all at once, or in steps as data are collected over time. Why is this not possible with a frequentist approach?

- (a) Suppose we draw a sample of size  $n_1$  from a population. If we do not reject the null, what happens if we go out and get another sample, say of size  $n_2$ , and combine the data sets?

In particular, suppose we have a sample from  $N(\mu, 1)$  and we wish to test

$$H_0: \mu = 0 \quad \text{versus} \quad H_A: \mu \neq 0.$$

Let us see what happens if in fact, the null hypothesis *is* true, that is, the population really is  $N(0, 1)$ . We will compute the actual Type I error rate if we try sequential testing in this frequentist case.

```

counter1 <- 0      #counter for reject null, sample 1
counter2 <- 0      #counter for reject null, sample 2

N <- 10^5          #number of initial draws

n1 <- 25           #size of sample 1
n2 <- 25           #size of sample 2

```

Now we run the algorithm: We set the Type I error rate to be  $\alpha = 0.05$ .

```

for (i in 1:N)
{
  x <- rnorm(n1) #initial sample from N(0, 1)
  pvalue <- t.test(x, mu = 0)$p. value
  if (pvalue < 0.05)
    { # reject null--false positive
      counter1 <- counter1 +1 #keep track of false positives
    }
  else {      #do not reject, draw another sample
    y <- rnorm(n2)   #draw another sample from N(0, 1)
    w <- c(x,y)      #combine samples
    pvalue2 <- t.test(w, mu = 0)$p. value
    if (pvalue2 < 0.05) #reject null-false positive
      counter2 <- counter2 +1 #keep track of false positives
    } #end else
} #end for loop
(counter1 + counter2)/N      #total prop. of false positives

```

What is the actual Type I error rate?

- (b) Find theoretically, the probability of a Type I error using this approach: that is, suppose  $X_1, X_2, \dots, X_n \sim N(0, 1)$ , and you conduct the hypothesis test  $H_A H_0: \mu = 0$  versus  $H_A: \mu > 0$ . If you do not reject the null, draw another sample of the same size  $n$  (to keep it simple):  $Y_1, Y_2, \dots, Y_n \sim N(0, 1)$  and combine with the first sample and conduct the hypothesis test again.

- 11.20** In the Google Analytics Content Experiments example at the end of Section 11.5, the number of visitors assigned to each arm was similar: 1874, 1867, 1871, 1868, 1875, 1875. These seem almost too good to be true – closer together than we would expect from random chance. Perform a two-sided goodness-of-fit test for the hypothesis that the customers were assigned randomly with equal probabilities. What is your conclusion?

If the test fails on the lower end, this indicates that the numbers are closer together than would reasonably occur by random chance, and we conclude that the assignment was not purely random. (In practice, the assignment is semirandom; assignment is performed by a number of different data centers operating independently, and each data center does a systematic round-robin.)

# 12

## One-Way ANOVA

In Chapters 3, 7, and 8, we encountered methods for comparing two populations. In particular, we learned procedures for comparing the means of two independent populations. In this chapter, we will compare the means of three or more populations. The classical approach is called the analysis of variance, or ANOVA for short. Contrary to the name given to this method, we will not be comparing the variances of the populations, but rather the variability in their means. We will start off with an approach that relies on a theoretical model for the sampling distribution of a test statistic, and end with a permutation test approach.

### 12.1 Comparing Three or More Populations

We have seen that the weight of a newborn baby is affected by whether or not the mother smokes cigarettes (see Example 7.9). Does the age of the mother also affect the birth weight of a baby? Figure 12.1 displays the distribution of birth weights of a random sample of boys born in Illinois in 2004 to mothers in the age ranges of 15–19, 20–24, and 25–29 years.<sup>1</sup> There appears to be a trend of weight increasing as age increases, but is this difference statistically discernible?

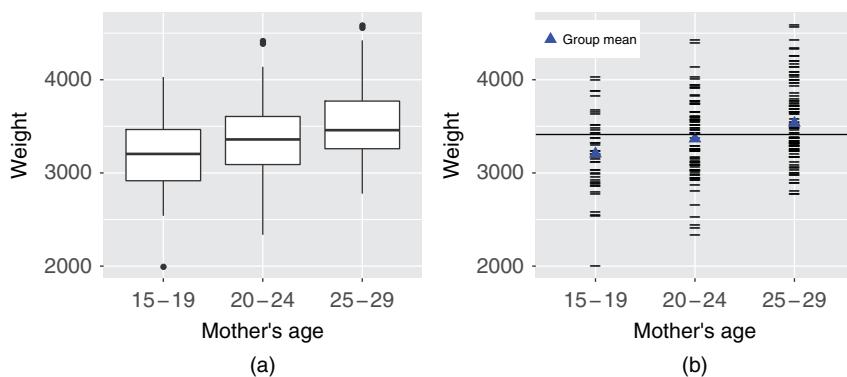
ANOVA takes into account how the birth weights in each age group vary from the groups mean, and how the weights vary between the age groups, using the information in Table 12.1.

#### 12.1.1 The ANOVA F Test

The ANOVA  $F$  test is for comparing means of populations.

Assume we have independent random samples drawn from  $G$  groups (populations). In the birth weight example, the  $G = 3$  populations are baby boys born to mothers in each of the three age groups. In a medical experiment

<sup>1</sup> The births are also restricted to single births only and gestation lengths of at least 37 weeks.



**Figure 12.1** (a) Distribution of the birth weights of boys born in Illinois in 2014.  
(b) Distribution of birth weights with horizontal line at overall mean and triangle at the means within each age group.

**Table 12.1** Summary statistics for birth weights.

Age	15–19	20–24	25–29
Mean	3207.205	3367.000	3535.363
$d$	422.003	410.682	413.589
$n$	44	90	107

investigating weight loss under different treatments, the populations might represent obese individuals on the Atkins diet, a vegan diet, a high carbohydrate diet, and a control group ( $G = 4$ ).

Let  $\mu_g$  be the true mean in group  $g$ ,  $g = 1, 2, \dots, G$ . The hypothesis of interest is

$$H_0: \mu_1 = \mu_2 = \dots = \mu_G$$

versus

$$H_A: \mu_i \neq \mu_j \text{ for some } i \neq j.$$

Suppose there are  $n_g$  observations in the sample from the  $g$ th group and  $n = n_1 + n_2 + \dots + n_G$ . Let  $Y_{gk}$  denote the response of the  $k$ th observation in the  $g$ th group (Table 12.2).

We can describe the model using the *cell means model*,

$$Y_{gk} = \mu_g + \epsilon_{gk},$$

with all  $Y$ 's independent,  $E[Y_{gk}] = \mu_g$  and  $\text{Var}[Y_{gk}] = \sigma^2$ ,  $k = 1, 2, \dots, n_g$ ,  $g = 1, 2, \dots, G$ . The  $\epsilon_{gk}$  represents random error; it follows that  $E[\epsilon_{gk}] = 0$  and  $\text{Var}[\epsilon] = \sigma^2$ .

**Table 12.2** Observations drawn from the  $G$  populations.

Group	Observations			Group mean
1	$Y_{11}$	$Y_{12}$	$\dots$	$Y_{1n_1}$
2	$Y_{21}$	$Y_{22}$	$\dots$	$Y_{2n_2}$
$\vdots$				$\vdots$
$G$	$Y_{G1}$	$Y_{G2}$	$\dots$	$Y_{Gn_G}$
				$\bar{Y}_G$

Let  $\bar{Y}_{g\cdot}$  denote the mean for the sample from the  $g$ th group,

$$\bar{Y}_{g\cdot} = \frac{1}{n_g} \sum_{k=1}^{n_g} Y_{gk}$$

and  $\bar{Y}_{..}$  denote the overall sample mean (often called the grand mean),

$$\bar{Y}_{..} = \frac{1}{n} \sum_{g=1}^G \sum_{k=1}^{n_g} Y_{gk} = \frac{1}{n} \sum_{g=1}^G n_g \bar{Y}_{g\cdot}$$

The idea is to compare the variability between each group to the variability within each group. Thus, for variability between the groups, we look at  $(\bar{Y}_{g\cdot} - \bar{Y}_{..})$ , the amount that the group means differ from the overall mean. For within group variability, we look at  $(Y_{gk} - \bar{Y}_{g\cdot})$ , that is, the amount that the  $n_g$  individuals in the  $g$ th group differ from that group's mean. Thus, if the means  $\mu_g$ ,  $g = 1, 2, \dots, G$ , are all the same, then the variability between the groups should be small compared to the variability within the groups.

We now define the *treatment sum of squares*,

$$SSTR = \sum_{g=1}^G \sum_{k=1}^{n_g} (\bar{Y}_{g\cdot} - \bar{Y}_{..})^2 = \sum_{g=1}^G n_g (\bar{Y}_{g\cdot} - \bar{Y}_{..})^2,$$

the *error sum of squares*,

$$SSE = \sum_{g=1}^G \sum_{k=1}^{n_g} (Y_{gk} - \bar{Y}_{g\cdot})^2$$

and the *total sum of squares*,

$$SST = \sum_{g=1}^G \sum_{k=1}^{n_g} (Y_{gk} - \bar{Y}_{..})^2,$$

the last sum representing the variability of each observation from the overall mean. We can partition this overall variability (SST) into the between group variability plus the within group variability.

**Theorem 12.1**  $SST = SSTR + SSE$ .

*Proof.*

$$\begin{aligned}
 SST &= \sum_{g=1}^G \sum_{k=1}^{n_g} (Y_{gk} - \bar{Y}_{..})^2 \\
 &= \sum_{g=1}^G \sum_{k=1}^{n_g} [(\bar{Y}_{g.} - \bar{Y}_{..}) + (Y_{gk} - \bar{Y}_{g.})]^2 \\
 &= \sum_{g=1}^G \sum_{k=1}^{n_g} (\bar{Y}_{g.} - \bar{Y}_{..})^2 + 2 \sum_{g=1}^G \sum_{k=1}^{n_g} (\bar{Y}_{g.} - \bar{Y}_{..})(Y_{gk} - \bar{Y}_{g.}) \\
 &\quad + \sum_{g=1}^G \sum_{k=1}^{n_g} (Y_{gk} - \bar{Y}_{g.})^2 \\
 &= \sum_{g=1}^G \sum_{k=1}^{n_g} (\bar{Y}_{g.} - \bar{Y}_{..})^2 + 0 + \sum_{g=1}^G \sum_{k=1}^{n_g} (Y_{gk} - \bar{Y}_{g.})^2.
 \end{aligned}$$

The proof that

$$\sum_{g=1}^G \sum_{k=1}^{n_g} (\bar{Y}_{g.} - \bar{Y}_{..})(Y_{gk} - \bar{Y}_{g.}) = 0$$

is left as an exercise. □

**Remark** We have seen the idea of partitioning the variability of a quantity before in the context of least-squares regression (Section 9.3.2). ||

The sample variance of the observations in group  $g$  is

$$S_g^2 = \frac{1}{n_g - 1} \sum_{k=1}^{n_g} (Y_{gk} - \bar{Y}_{g.})^2,$$

so the error sum of squares can be expressed as

$$SSE = \sum_{g=1}^G (n_g - 1) S_g^2.$$

We can pool these estimates of the sample variances across all groups

$$\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 + \cdots + (n_G - 1)S_G^2}{(n_1 - 1) + (n_2 - 1) + \cdots + (n_G - 1)} = \frac{\sum_{g=1}^G (n_g - 1) S_g^2}{n - G} = \frac{SSE}{n - G} \quad (12.1)$$

to obtain an unbiased estimate of the common variance  $\sigma^2$ .

**Theorem 12.2** Let  $Y_{gk}$ ,  $k = 1, 2, \dots, n_g$ ,  $g = 1, 2, \dots, G$ , denote independent random variables with  $E[Y_{gk}] = \mu_g$ ,  $\text{Var}[Y_{gk}] = \sigma^2$ . Let  $\mu = \sum_{g=1}^G n_g \mu_g / n$ .

Then

$$E\left[\frac{\text{SSTR}}{G-1}\right] = \sigma^2 + \frac{1}{G-1} \sum_{g=1}^G n_g (\mu_g - \mu)^2 \quad (12.2)$$

$$E\left[\frac{\text{SSE}}{n-G}\right] = \sigma^2. \quad (12.3)$$

*Proof.* We prove Equation (12.2); the proof of Equation (12.3) follows similar reasoning.

First, it is easy to check that  $E[\bar{Y}_{..}] = \mu$  and  $\text{Var}[\bar{Y}_{..}] = \sigma^2/n$ . Thus, by Theorem A.7,  $E[(\bar{Y}_{..} - \mu)^2] = \text{Var}[\bar{Y}_{..}] = \sigma^2/n$ .

In addition,  $\text{Var}[\bar{Y}_{g.} - \mu] = E[(\bar{Y}_{g.} - \mu)^2] - (E[(\bar{Y}_{g.} - \mu)])^2$ , by Proposition A.2. Thus, since  $\text{Var}[\bar{Y}_{g.} - \mu] = \text{Var}[\bar{Y}_{g.}] = \sigma^2/n_g$ , we have  $\sigma^2/n_g = E[(\bar{Y}_{g.} - \mu)^2] - (E[(\bar{Y}_{g.} - \mu)])^2$ , or rewriting,

$$E[(\bar{Y}_{g.} - \mu)^2] = \sigma^2/n_g + (\mu_g - \mu)^2.$$

Thus,

$$\begin{aligned} E[\text{SSTR}/(G-1)] &= \frac{1}{G-1} E\left[\sum_{g=1}^G n_g (\bar{Y}_{g.} - \bar{Y}_{..})^2\right] \\ &= \frac{1}{G-1} E\left[\sum_{g=1}^G n_g \left[(\bar{Y}_{g.} - \mu) - (\bar{Y}_{..} - \mu)\right]^2\right] \\ &= \frac{1}{G-1} E\left[\sum_{g=1}^G n_g \left[(\bar{Y}_{..} - \mu)^2 - 2(\bar{Y}_{g.} - \mu) \right.\right. \\ &\quad \times (\bar{Y}_{..} - \mu) + (\bar{Y}_{..} - \mu)^2\left.\right]\Bigg] \\ &= \frac{1}{G-1} E\left[\sum_{g=1}^G n_g (\bar{Y}_{g.} - \mu)^2 - 2(\bar{Y}_{..} - \mu) \right. \\ &\quad \times \sum_{g=1}^G n_g (\bar{Y}_{g.} - \mu) + \sum_{g=1}^G n_g (\bar{Y}_{..} - \mu)^2\Bigg] \\ &= \frac{1}{G-1} E\left[\sum_{g=1}^G n_g (\bar{Y}_{g.} - \mu)^2 - 2(\bar{Y}_{..} - \mu) \right. \\ &\quad \times n(\bar{Y}_{..} - \mu) + n(\bar{Y}_{..} - \mu)^2\Bigg] \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{G-1} E \left[ \sum_{g=1}^G n_g (\bar{Y}_{g\cdot} - \mu)^2 \right] - n \frac{1}{G-1} E[(\bar{Y}_{..} - \mu)^2] \\
&= \frac{1}{G-1} \sum_{g=1}^G n_g (\sigma^2/n_g + (\mu_g - \mu)^2) - \frac{1}{G-1} n \sigma^2/n \\
&= \sigma^2 + \frac{1}{G-1} \sum_{g=1}^G n_g (\mu_g - \mu)^2.
\end{aligned}$$

□

Thus, if the population means are all the same,  $\mu_1 = \mu_2 = \dots = \mu_G = \mu$ , then the sum in Equation (12.2) is zero, so on average, the ratio  $(SSTR/(G-1))/(SSE/(n-G))$  should be equal to 1. Otherwise, if  $(\mu_g - \mu)^2 > 0$  for at least one  $g$ , then on average the ratio is greater than 1.

If the populations are normally distributed we can say more.

**Theorem 12.3** Let  $Y_{gk}$ ,  $k = 1, 2, \dots, n_g$ ,  $g = 1, 2, \dots, G$ , denote independent random variables,  $Y_{gk} \sim N(\mu_g, \sigma^2)$ .

Then

1. SSE and SSTR are independent.
2.  $SSE/\sigma^2$  has a chi-square distribution with  $n - G$  degrees of freedom.
3. If  $\mu_1 = \mu_2 = \dots = \mu_G$ , then  $SSTR/\sigma^2$  has a chi-square distribution with  $G - 1$  degrees of freedom.

*Proof.* The proofs of parts (1) and (2) are left as exercises, and the proof of part (3) is omitted. □

**Definition 12.1** The treatment sum of squares divided by its degrees of freedom is called the *mean square for treatments*,

$$\text{MSTR} = \frac{\text{SSTR}}{G-1}.$$

The error sum of squares divided by its degrees of freedom is called the *mean squared error* (or *mean squared residual*),

$$\text{MSE} = \frac{\text{SSE}}{n-G}. \tag{12.4}$$

||

**Theorem 12.4** Let  $Y_{gk}$ ,  $k = 1, 2, \dots, n_g$ ,  $g = 1, 2, \dots, G$ , denote independent random variables,  $Y_{gk} \sim N(\mu_g, \sigma^2)$ . If  $\mu_1 = \mu_2 = \dots = \mu_G$ , then the  $F$  statistic

$$F = \frac{\text{MSTR}}{\text{MSE}}$$

has an  $F$  distribution with  $G - 1$  and  $n - G$  degrees of freedom.

*Proof.*

$$F = \frac{\text{MSTR}}{\text{MSE}} = \frac{\text{SSTR}/(G-1)}{\text{SSE}/(n-G)} = \frac{\frac{(\text{SSTR}/\sigma^2)}{(G-1)}}{\frac{(\text{SSE}/\sigma^2)}{n-G}}. \quad (12.5)$$

By Theorem 12.3, SSTR and SSE are independent, and  $\text{SSTR}/\sigma^2$  and  $\text{SSE}/\sigma^2$  have chi-square distributions with degrees of freedom  $G-1$  and  $n-G$ , respectively. Thus, the result follows by Definition B.9.  $\square$

**Example 12.1** For the Illinois boys, the overall mean birth weight is  $\bar{Y} = 3412.564$  g. The treatment sum of squares is the sum (over observations) of the squared difference between the group mean for the observation and the overall mean (refer to Table 12.1). This simplifies to a sum (over groups) of the group size times the squared difference between group mean and overall mean:

$$\begin{aligned} \text{SSTR} &= \sum_{g=1}^3 \sum_{k=1}^{n_g} (\bar{Y}_{g\cdot} - 3412.564)^2 \\ &= \sum_{k=1}^{44} (3207.205 - 3412.564)^2 + \sum_{k=1}^{90} (3367.000 - 3412.564)^2 \\ &\quad + \sum_{k=1}^{107} (3535.336 - 3412.6)^2 \\ &= 44(3207.205 - 3412.564)^2 + 90(3367.000 - 3412.564)^2 \\ &\quad + 107(3535.336 - 3412.6)^2 \\ &= 3\,655\,236. \end{aligned}$$

The degrees of freedom for SSTR is  $G-1 = 3-1 = 2$ . Thus, the mean square for treatment is  $3\,655\,236/2 = 1\,827\,628$ .

The error sum of squares is the sum of squared difference between each observation and its group mean:

$$\begin{aligned} \text{SSE} &= \sum_{g=1}^3 \sum_{k=1}^{n_g} (Y_{gk} - \bar{Y}_{g\cdot})^2 \\ &= \sum_{k=1}^{44} (Y_{1k} - 3207.205)^2 + \sum_{k=1}^{90} (Y_{2k} - 3367.00)^2 + \sum_{k=1}^{107} (Y_{3k} - 3535.336)^2 \\ &= 40\,800\,419. \end{aligned}$$

The degrees of freedom is  $n-G = 241-3 = 238$ , and the mean square for error is  $40\,800\,419/238 = 171\,430.3$ .

**Table 12.3** ANOVA table.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Mother's age	2	3 655 256	1 827 628	10.661	0.00004
Residuals	238	40 800 419	171 430		
Total	240	44 455 675			

Thus, the  $F$  statistic is  $1\ 827\ 628/171\ 430.3 = 10.661$  and is compared to an  $F$  distribution with 2 and 238 degrees of freedom. The corresponding  $P$ -value is approximately 0.00004 so we conclude that the true means between the age groups are indeed different.

The calculations can be summarized in an ANOVA table (Table 12.3). □

### R Note

The `lm` function that we used for linear regression in Chapter 9 used in conjunction with the `anova` function can be used to perform the ANOVA  $F$  test; or we can combine `summary` and `aov`.

The data for this example are in the file `ILBoys`.

```
> anova(lm(Weight ~ MothersAge, data = ILBoys))
      Df  Sum Sq  Mean Sq F value    Pr(>F)
MothersAge   2 3655256 1827628 10.661 3.679e-05 ***
Residuals  238 40800419 171430
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> anova(lm(Weight ~ MothersAge, data = ILBoys))$F[1]
[1] 10.66105                      #Extract F statistic

> summary(aov(Weight ~ MothersAge, data = ILBoys)) #same
...
```

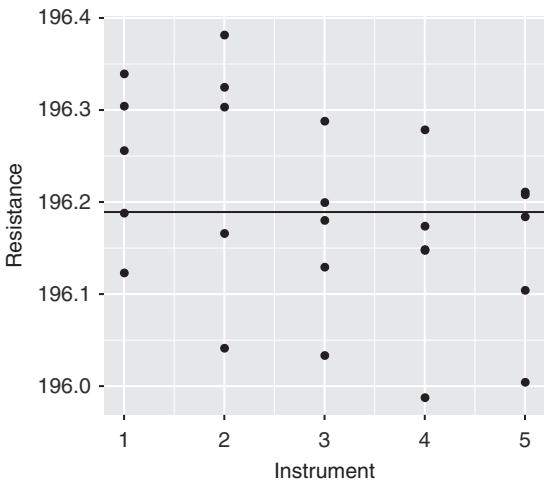
**Example 12.2** Measurements of the resistivity (ohms-centimeter) of silicon wafers were made at the National Institute of Standards and Technology (NIST) with five instruments on each of 5 days. Each of the 25 observations is the average of 6 measurements (Figure 12.2).<sup>2</sup>

If we let  $\mu_i$ ,  $i = 1, 2, \dots, 5$  denote the mean of the resistivity measurements for each of the instruments, then we wish to test

$$H_0: \mu_1 = \mu_2 = \dots = \mu_5 \text{ versus } H_A: \mu_i = \mu_j \text{ for some } i \neq j.$$

<sup>2</sup> [http://www.itl.nist.gov/div898/strd/anova/SiRstv\\_info.html](http://www.itl.nist.gov/div898/strd/anova/SiRstv_info.html).

**Figure 12.2** Distribution of the resistivity measurements across the five instruments. The horizontal line marks the overall mean.



The overall mean resistance is  $196.1892 \Omega\text{-cm}$  with standard deviation  $0.1056 \Omega\text{-cm}$ .

The treatment sum of squares is  $SSTR = 0.051146$  on 4 degrees of freedom, so the mean square for treatments is  $MSTR = 0.051146/4 = 0.012787$ .

The residual sum of squares is  $SSE = 0.216637$  on  $25 - 5 = 20$  degrees of freedom, so the mean square for error is  $MSE = 0.216637/20 = 0.010832$ .

Thus, the  $F$  statistic is  $0.012787/0.010832 = 1.1805$  and we compare this to an  $F$  distribution with 4 and 20 degrees of freedom. The resulting  $P$ -value is 0.35 so we have no reason to reject the null hypothesis. We conclude that the mean resistance levels are the same across instruments.  $\square$

#### 12.1.1.1 Conditions

The conditions needed to use the  $F$  tests parallel those described in Section 9.4 for linear regression, except for the condition of linearity.

The conditions are:

- The residuals are all independent.
- The residuals have constant variance.
- The residuals are normally distributed.

Of those conditions, the independence condition is critical, and the others usually less important. The reasoning is similar to that in Section 9.4.3, but two conditions merit extra comments here: the conditions of constant variance and of normal distributions.

We motivate both comments by considering the case  $G = 2$ . Using ANOVA with only two groups is equivalent to doing a two-sample pooled-variance

two-sided  $t$ -test. The  $t$ -test is preferred; it is easier to understand, allows for one-sided tests, and does not require pooling the variances.

The big problem with non-normality in  $t$ -tests is the effect of skewness on one-sided tests. But ANOVA tests are inherently two-sided (we are testing for *any* differences between means, not differences in one direction) so non-normal distributions in general and skewness in particular have little effect as long as the sample sizes are reasonably large.

With ANOVA we are forced to assume that variances are equal. If the sample sizes  $n_g$  are roughly equal, then unequal variances do not hurt much, but if the population variances differ, then the actual sampling distribution of the  $F$  statistic could be very different from an  $F$  distribution. In particular, if there is a small sample from a population with large variance, then the  $F$  statistic can explode. To see this, consider an extreme case, where one sample is size 1, so the observation from that sample has no effect on SSE or MSE. If the variance for that population is a billion times the other population variances, then the SSTR will tend to be much larger than we would expect based on the MSE. With less extreme situations, the effects will be more subtle, but the actual Type I error rate could be substantially different than the nominal value.

### 12.1.2 A Permutation Test Approach

The ANOVA test (Theorem 12.4) conditions include constant variance and normal distributions. Alternately, we can use permutation test techniques from Chapter 3 to form a permutation distribution of an appropriate test statistic, under the null hypothesis that every group has the same population. We randomly assign the values of the numeric variable to the  $k$  groups and compute the corresponding  $F$  statistic. We then note how extreme the observed  $F$  statistic is relative to the permutation distribution of the  $F$  statistic.

#### R Note

```
#Checking the normality condition
ggplot(ILBoys, aes(sample = Weight)) + geom_qq() +
  geom_qq_line() + facet_wrap(. ~ MothersAge)

#Permutation test
observed <- anova(lm(Weight ~ MothersAge, data = ILBoys))$F[1]
n <- length(ILBoys$Weight)
N <- 10^4 - 1
results <- numeric(N)
for (i in 1:N)
{
  index <- sample(n)
  Wt.perm <- ILBoys$Weight[index]
  results[i] <- anova(lm(Wt.perm ~ MothersAge, data = ILBoys))$F[1]
}
(sum(results >= observed) + 1) / (N + 1) # P value
```

## Exercises

- 12.1** In the early 1900's, Latter (1902) investigated the behavior of female cuckoos that lay their eggs on the ground and then move them to the nests of other birds. In particular, Latter gathered data on the lengths of the cuckoo eggs found in these foster-nests. Data based on this work is used in the book *The Methods of Statistics* by L. H. C. Tippett (Tippett, 1952) and is located in the file *Cuckoos*. The data contains the lengths, in millimeters, of cuckoo eggs and the species of the nests where the eggs were placed.
- (a) Conduct some exploratory data analysis (EDA): Compute the mean and standard deviation of the lengths across the different species. How many eggs of each species are represented in the data? Create side-by-side boxplots to compare the distributions of lengths across the different species.
  - (b) Conduct an ANOVA test to see if the mean lengths of the cuckoo eggs are the same across the species.
- 12.2** The data set *ChiMarathonMen* contains data on a sample of the men, ages between 20 and 39 years, who completed the Chicago Marathon in 2015. The runners were classified into age groups, 20–24, 25–29, 30–34, and 34–39. Their finishing times are given in minutes.
- (a) Conduct some EDA: How many runners in each age division? What are the average times and standard deviations of the finishing times in each division? Create side-by-side boxplots to visualize the distributions of times across divisions. Create quantile normal plots as well. Is the normality condition reasonably met?
  - (b) Conduct an ANOVA test to see if the mean finishing times are the same across age groups.
- 12.3** Starcraft is a popular strategy video game with a science fiction military theme. Players choose to be one of three races – the Terrans, Zergs or Protoss' – and compete for dominance in a distance part of the Milky Way galaxy. The file *Starcraft* contains information on a sample of the top Korean players from the database <http://www.teamliquid.net/tlpd/players> (J. Evans, private communication). In addition to the player's chosen race, the file contains age as well as the number of games won (out of his most recent 40 games).
- (a) Conduct some EDA: What are the mean and standard deviation of the ages of the players across races? How many players are there in each of the races? Create a boxplot to compare the distributions of the ages across races. Create quantile normal plots to check if distributions are approximately normal.

- (b) Conduct an ANOVA test to determine whether or not the mean age of the players is the same across races.
- (c) Repeat (a) and (b) for the mean number of wins across the three races.
- 12.4** Recall the flight delays case study in Section 1.1, and the distribution of flight delay times for United Airlines across days (see Figure 2.6).
- Conduct some EDA: what are the means and standard deviations of flight delay times across the days? Are the distributions of times normal across days? Create normal quantile plots to check.
  - Use a permutation test to determine whether or not the mean delay times across days of the week are the same.
- 12.5** Recall the General Social Survey (GSS) 2018 case study in Section 1.7. We will investigate the relationship between a person's age (`Age`) and their view on how harshly courts treat criminals (`Courts`).
- Conduct some EDA: What are the mean and standard deviation of the ages of the participants across views of the courts? Are the distribution of ages normal across views of the courts? (Note: Use the `drop_na` function in the `tidyverse` package to remove the missing values in the `age` variable.)
  - Use a permutation test to determine whether or not the mean ages across the views of the courts are the same.
- 12.6** We will run simulations to explore the conditions for ANOVA: in particular, how does unbalancedness (sample sizes not the same) and unequal population variances affect the outcome? We consider the hypotheses  $H_0: \mu_A = \mu_B = \mu_C$  versus  $H_A$ : at least one pair of means is not equal.
- Run the code below to draw three random samples from populations (called A, B, C) with the same mean and standard deviation and then perform an ANOVA test. What proportion of times do you reject the null hypothesis (false positive)? Now change the sample size for sample a to 10 (`nA <- 10`) and repeat.
  - Repeat (a) by increasing the standard deviation of population A to 9 and trying samples of size  $nA = 50$  and  $nA = 10$  (and keeping the other sample sizes at 50). What proportion of times do you reject the null hypothesis?
  - Explore other scenarios: What if the population means are all different, but the population variance are the same. How do sample sizes affect the outcome? Try with all sample sizes the same and then unequal. Now try different variances and again, with balanced and unbalanced samples.

```

# Set sample sizes and create groups
nA <- 50
nB <- 50
nC <- 50
Group <- rep(c("A", "B", "C"), c(nA, nB, nC))

N <- 10^4
counter <- 0

for (i in 1:N)
{
  a <- rnorm(nA, 20, 3)      # Draw samples
  b <- rnorm(nB, 20, 3)
  c <- rnorm(nC, 20, 3)
  X <- c(a, b, c)           # Combine into one vector

  Pvalue <- anova(lm(X ~ Group))$P[1]
  if (Pvalue < 0.05)          # Reject H0?
    counter <- counter + 1   # If yes, increase counter
}
counter/N                      # proportion of times H0 rejected

```

**12.7** In Theorem 12.1, prove  $\sum_{g=1}^G \sum_{k=1}^{n_g} (\bar{Y}_{g\cdot} - \bar{Y}_{..})(Y_{gk} - \bar{Y}_{g\cdot}) = 0$ .

**12.8** Prove parts (1) and (2) of Theorem 12.3.



# 13

## Additional Topics

This chapter includes a variety of topics. We will begin by considering the fundamental bootstrap principle – plugging in an estimate for the population in place of the population and looking at two new possibilities for what to plug in – in Sections 13.1 (smoothed bootstrap) and 13.2 (parametric bootstrap). We next consider a pair of topics that are important in the design and analysis of experiments – stratified sampling and control variates. We then consider some computational methods; one motivation is for Bayesian analysis in nonconjugate situations, but the methods are useful in general. These lead into Monte Carlo integration and importance sampling; the latter is useful in experimental design and analysis as well as Bayesian analysis. We then conclude with the EM algorithm, useful for handling an ugly fact of life, missing data.

Many of the design and analysis applications are based on one author's experience at Google, Pacific Gas & Electric Company, and (recently joined) Instacart.

Instacart (IC) offers grocery delivery and pickup in the United States and Canada. It is a relatively young, small, and rapidly growing company. IC works with:

- *Stores*: IC partners with grocery stores to make their inventory available for purchase and delivery.
- *Customers*: Individuals (or sometimes restaurants) order food from IC via the web or a mobile app, specifying what groceries from which store.
- *Shoppers*: Shoppers are the freelance contractors who do the shopping and deliver the groceries to the customer. IC offers them work via a web app: shoppers can choose to fill a single order or batches of orders, and the app provides details about the order(s) (which store, number of items, distance to customer), pay and anticipated tip.
- *Advertisers*: Instacart works with advertisers to show relevant ads to customers.

IC operations are extraordinarily complex for a company its size. IC is running many different experiments to improve operations and balance different factors such as incentivizing shoppers to pick and deliver orders quickly and paying them a reasonable price; determining when to offer more pay for an order or batch because no shopper is picking it; grouping multiple orders with similar destinations together into a batch (without making the batch too large that it hampers speedy delivery); keeping delivery fees low by displaying ads but in a way that does not impede customers from quickly finding what they want. It is an exciting place to work, with a lot of interesting problems that we are just figuring out.

## 13.1 Smoothed Bootstrap

Recall that a sampling distribution is the distribution of a statistic when drawing random samples from a population. In practice drawing thousands or millions of repeated sample from the population is impossible, so the fundamental bootstrap idea is to draw samples from an estimate of the population.

In earlier chapters, we sampled from the observed data, with empirical cumulative distribution function (ecdf)  $\hat{F}_n(x) = (1/n)\#\{x_i \leq x\}$ , where  $\#\{x_i \leq x\}$  is the number of  $x_i$  values that are below  $x$ .

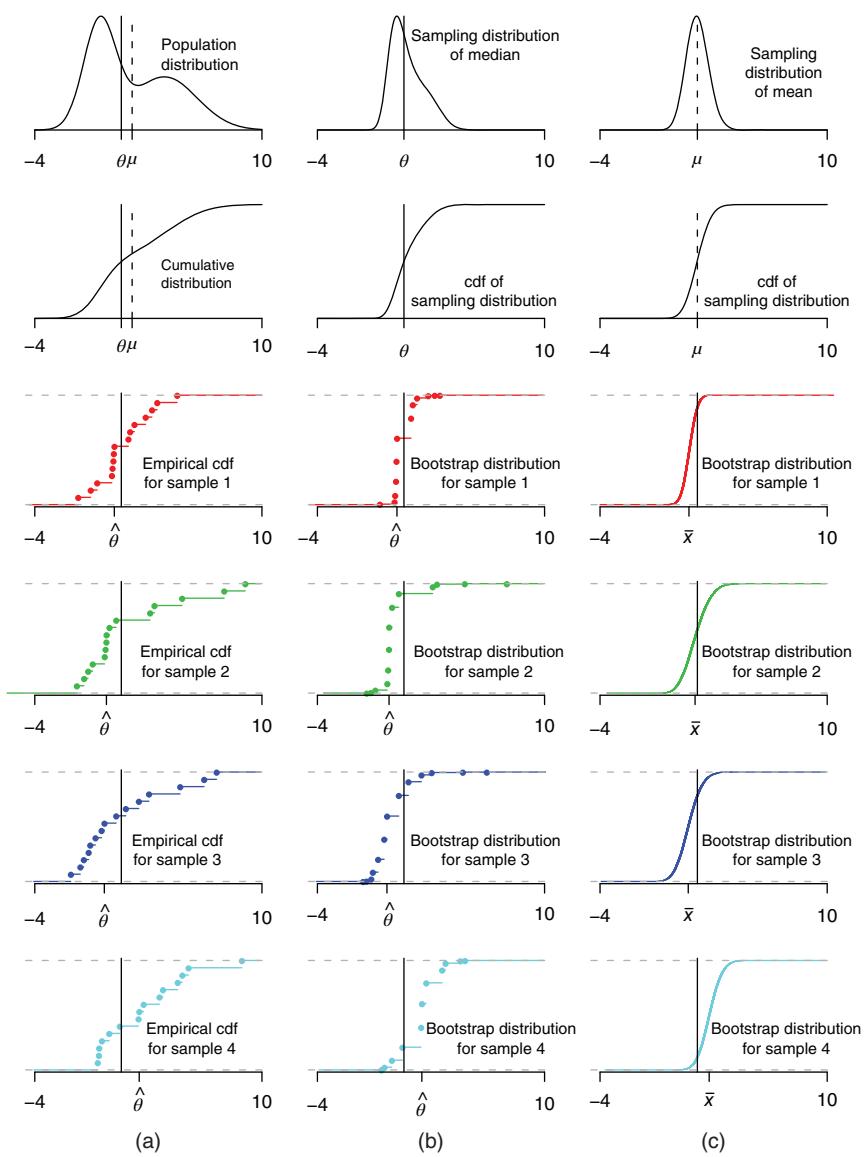
But that is not always a good choice. Recall Section 5.8, where we found that the ordinary bootstrap does not work well for the median; the bootstrap distributions in Figure 5.19 look nothing like the sampling distribution.

Figure 13.1 shows another view of why that happened. The population and sampling distribution are continuous, but the ecdfs are discrete, so the bootstrap distributions for the median are also discrete.

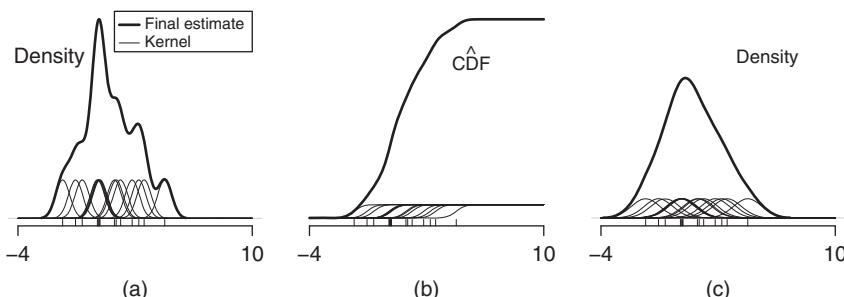
For comparison, the mean is less sensitive to whether the distribution is discrete or continuous. The right column of Figure 13.1 displays the bootstrap distributions for the mean from the same ecdfs. Even though the ecdfs are discrete, the bootstrap distributions are nearly continuous. In fact, the theoretical (exhaustive) bootstrap distribution has  $\binom{2n-1}{n} = \binom{29}{15} = 77\,558\,760$  jumps, corresponding to the number of distinct unordered bootstrap samples.

### 13.1.1 Kernel Density Estimate

The median is sensitive to discreteness, so we will try drawing samples from a continuous distribution, in particular from a smoothed version of the empirical distribution, using what is known as a *kernel density estimate*. Figure 13.2a illustrates the idea. Think of this as a *cow-pie estimate* – drop a cow-pie at every data point, and see how high they pile up. Where there are data points close together, such as four observations beginning with the fourth smallest, the pile



**Figure 13.1** Ordinary bootstrap distributions for the median and mean,  $n = 15$ .  
 (a) The population density and cdf, as well as the cdfs for four samples. (b) The mean and  
 (c) the median, respectively, in each case showing the density and cdf for the sampling  
 distribution, as well as bootstrap distributions for the four samples.



**Figure 13.2** Kernel density estimates. (a) A kernel density estimate using the first sample from Figure 13.1, with a normal kernel with standard deviation  $s/\sqrt{n}$ . (b) The corresponding cdf estimate. (c) A kernel twice as wide as the first one.

gets high. More formally, at each observation we center  $1/n$  times a normal density with a small standard deviation  $\sigma_K$  and then add those up to obtain the density estimate,

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n g(x - x_i), \quad (13.1)$$

where  $g$  is the normal density with mean 0 and standard deviation  $\sigma_K$  (the “kernel”). The cumulative distribution function (cdf) estimate (Figure 13.2b) is

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^n G(x - x_i) = \frac{1}{n} \sum_{i=1}^n \Phi((x - x_i)/\sigma_K), \quad (13.2)$$

where  $G$  is the corresponding normal cdf.

Figure 13.2a is jarring, with a number of wiggles. We might prefer to use a wider kernel, as shown in Figure 13.2c. There is a tradeoff – a wider kernel reduces the variability in the density estimate (i.e. fewer wiggles) – but at the cost of increased bias – the resulting density estimate gets flatter and wider, with increasing standard deviation, eventually beyond what could be justified by the data. As the sample size increases, we typically want  $\sigma_K$  to decrease, but not too fast. The choice  $\sigma_K = s/\sqrt{n}$  goes to zero at a reasonable rate, is easy to remember, and has another nice property for bootstrapping that we will discuss shortly.

**Remark** The kernel density estimate is useful in its own right as a way to look at data. It is an alternative to a histogram. Histograms can change dramatically as the bar widths change. Kernel density estimates are more stable as the kernel standard deviation varies. ||

To generate random observations from the distribution, we draw an ordinary bootstrap sample and then add noise to each observation, independently, from

a normal distribution with mean 0 and variance  $\sigma_K^2$ ; that is,  $Y_i = X_i^* + V_i$ , where  $X_i^*$  denotes a bootstrap observation and  $V_i$  denotes the “noise.”

### Smoothed Bootstrap for a Single Population

Given a sample of size  $n$  from a population,

1. Draw a resample of size  $n$  with replacement from the sample.
2. Add independent random normal values with mean zero and variance  $\sigma_K^2$  to each observation in the resample.
3. Proceed as for ordinary bootstrapping – calculate the statistic, repeat many times, and construct and use the bootstrap distribution.

The variance of the kernel smooth distribution is the variance of the empirical distribution plus the variance of the noise,

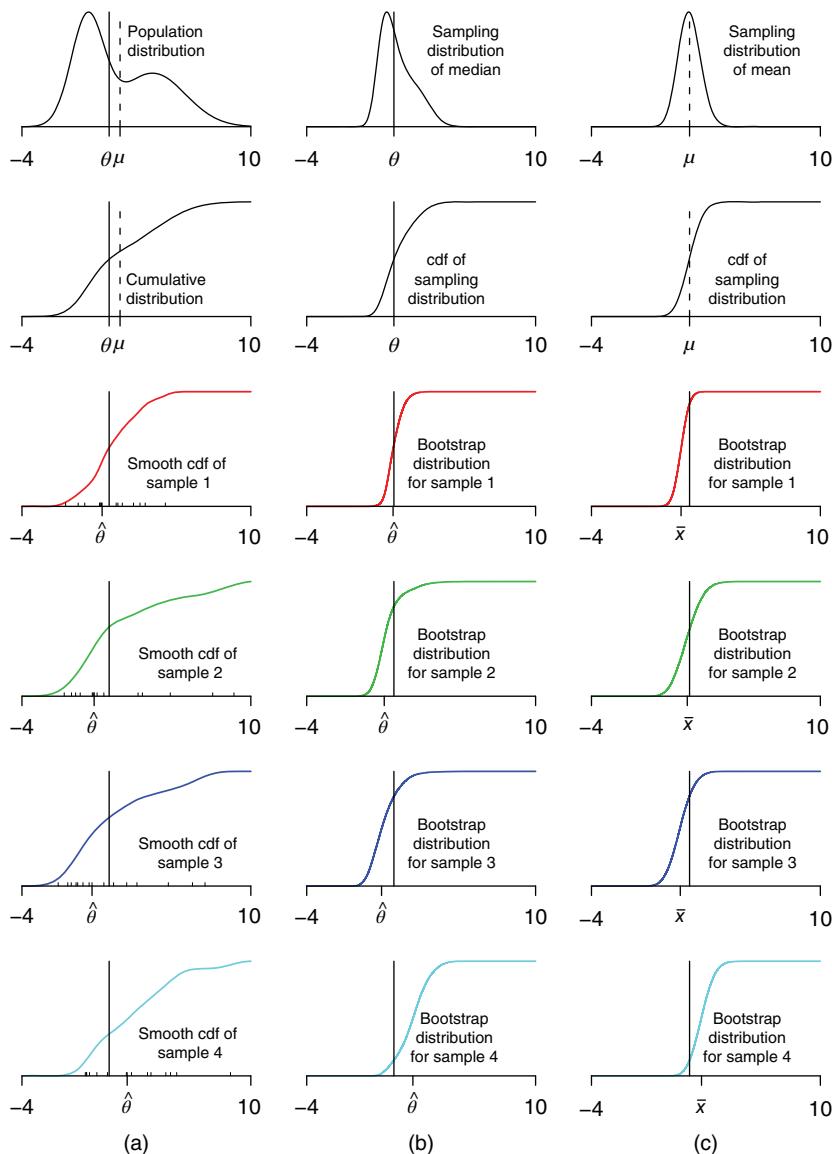
$$\begin{aligned}\text{Var}[Y] &= \text{Var}[X^* + V] \\ &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 + \sigma_K^2 \\ &= \frac{n-1}{n} s^2 + \sigma_K^2.\end{aligned}$$

The choice

$$\sigma_K = s/\sqrt{n} \tag{13.3}$$

makes  $\text{Var}[Y] = s^2$ . This is particularly useful for bootstrapping. One problem with bootstrapping using the empirical distribution is that the bootstrap distributions tend to be too narrow. In the case that the statistic is the sample mean, the exhaustive (non-Monte Carlo) bootstrap standard error is  $\sqrt{(n-1)/n} (s/\sqrt{n})$ . It is narrower than the common formula standard error by  $\sqrt{(n-1)/n}$ , because empirical distributions tend to be narrower than the population. Adding the right amount of noise by using  $\sigma_K = s/\sqrt{n}$  corrects for that.

Figure 13.3 shows the smoothed bootstrap estimates for the mean and the median. The bootstrap distributions are much improved for the median; they are now continuous and the spreads are closer to the spread of the sampling distribution. As always, the centers are centered at the sample median  $\hat{\theta}$  rather than the population median  $\theta$ . The spreads vary substantially because with samples this small, the corresponding samples vary substantially. In contrast, smoothing does not make as much difference for the mean; the biggest difference is that the bootstrap distributions are slightly wider, with standard error larger by a factor  $\sqrt{15}/14$ .



**Figure 13.3** Smoothed bootstrap distributions for the median and the mean,  $n = 15$ .  
 (a) The population density and cdf, as well as cdfs for four samples. (b) The mean and  
 (c) the median, respectively, in each case showing the density and cdf for the sampling  
 distribution, as well as bootstrap distributions for the four samples.

**Remark** There is a problem with applying kernel density estimates to non-negative data like arsenic levels or Verizon repair times; some of the density falls to the left of 0, and doing smoothed bootstrap sampling by adding noise may make some values negative.

A remedy is to transform the data, say  $y = \log(x)$ , draw bootstrap samples from the  $y$  values, add noise to the bootstrap  $y$  values using kernel standard deviation  $s_y/\sqrt{n}$ , and then transform back to the original scale. Other transformations may be used, for example,  $y = \sqrt{x}$ . If  $x$  has some zero values, then a transformation like  $\log(x + 0.1)$  prevents taking the log of zero. An example of R code for this is available at <https://github.com/lchihara/MathStatsResamplingR>. ||

## 13.2 Parametric Bootstrap

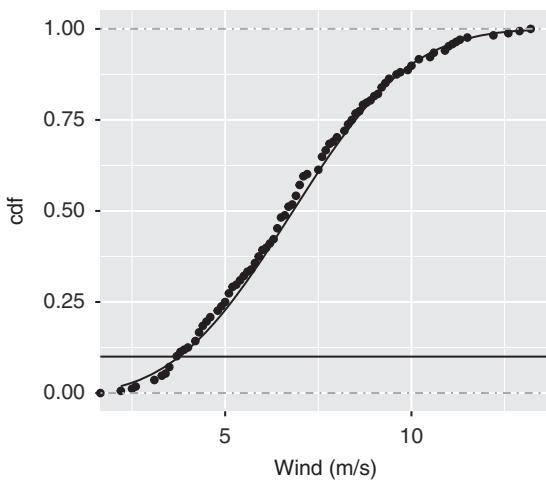
In the ordinary bootstrap, we assume essentially nothing about the underlying population and sample from the empirical distribution. In the smoothed bootstrap, we assume that the population has a density, and we sample from an estimate of this density. Suppose we are willing to make an even stronger assumption – that the underlying population has some parametric distribution. Then, we can use a *parametric bootstrap*: we estimate parameters based on the data, for instance, by maximum likelihood and then draw bootstrap samples from the corresponding parametric distribution.

**Example 13.1** Recall the wind energy case study in Section 6.1.3, where we modeled wind speed using a Weibull distribution. We previously estimated the parameters to be  $\hat{k} = 3.169$  and  $\hat{\lambda} = 7.661$ . Figure 13.4 shows the cdf of the data and the cdf of a Weibull distribution with these parameters.

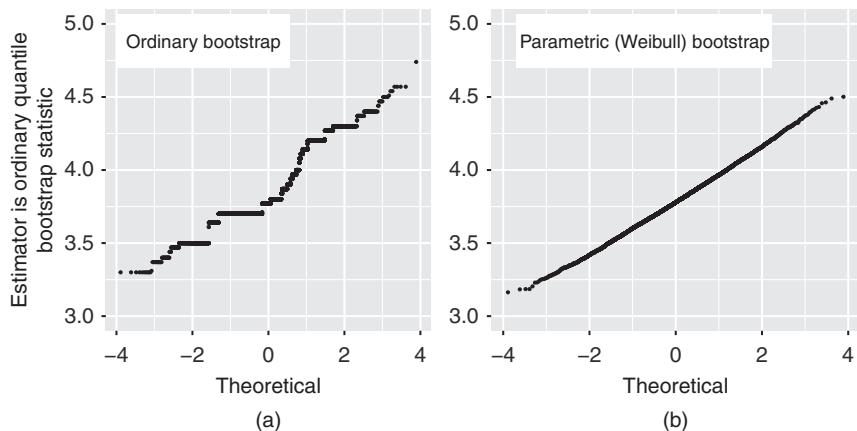
We will compare two bootstraps, the ordinary bootstrap and a parametric bootstrap based on a Weibull distribution.

The statistic we will use is the 10% sample quantile,  $\hat{\eta}_1$  in Figure 13.4. We saw in Section 5.8 that the ordinary bootstrap does not work well for the median in small samples. We may expect similar problems with the 10% quantile. We are looking at the 10% quantile rather than the median, because we are interested in a measure of reliability, the minimum amount of energy that the turbine will generate 90% of the time. While the sample size is 168, there are only 17 observations at or below the quantile, a small effective sample size, so the ordinary bootstrap may perform poorly. The two bootstrap distributions are shown in Figure 13.5. The ordinary bootstrap does badly. The parametric distribution seems more reasonable.

For comparison, we will also consider an alternative estimator for the 10% quantile of the distribution – given a set of data (original data, or bootstrap

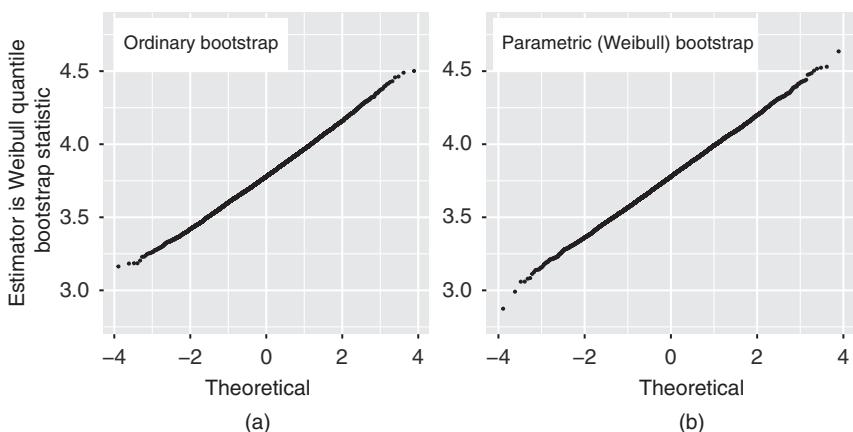


**Figure 13.4** Empirical cumulative distribution function of wind speeds (m/s), with cdf for Weibull superimposed.  $\hat{\eta}_1$  is the quantile where the ecdf crosses  $y = 0.1$  and  $\hat{\eta}_2$  is the quantile where the Weibull cdf crosses  $y = 0.1$ .



**Figure 13.5** Bootstrap distributions for 10% quantile of wind speed. (a) The ordinary bootstrap. (b) The Weibull parametric bootstrap. The estimator is the 10% empirical quantile (i.e. of each bootstrap sample). Each panel contains a normal quantile plot.

data), fit a Weibull distribution to the data, then calculate the 10% quantile of that Weibull distribution. This is indicated as  $\hat{\eta}_2$  in Figure 13.4. Again, we will do both the ordinary and Weibull bootstraps. The results are shown in Figure 13.6. For this estimator, the ordinary bootstrap works just fine, there is little difference between the ordinary and Weibull bootstraps. So the problem with the ordinary bootstrap was limited to the sample quantile as an estimator.



**Figure 13.6** Bootstrap distributions for fitted Weibull distribution estimates of the 10% quantile of wind speed. (a) The ordinary bootstrap. (b) The Weibull parametric bootstrap. The estimator is the 10% quantile of a Weibull distribution fitted to the data (i.e. of each bootstrap sample).

If the estimator is the Weibull quantile rather than the ordinary sample quantile, then the ordinary bootstrap works fine. Still, if we are confident enough that the Weibull distribution fits the data well enough to use the Weibull estimator, then we might as well also do the parametric Weibull bootstrap.

It is also worth comparing the performance of the two estimators, the raw data quantile and the fitted Weibull quantile. Compare the vertical spread of the bootstrap statistics in Figures 13.5a and 13.6b; the standard deviation of the  $y$  values is the bootstrap standard error. The standard error is much smaller for the Weibull quantile. This agrees with our findings in Section 6.3.2, where we compared two estimators for the fraction of time that wind speed exceed 5 m/s, the ordinary sample fraction and a probability based on estimated Weibull parameters. In both cases, the Weibull estimator is more efficient.  $\square$

When doing a parametric bootstrap, computing standard errors is straightforward: we use the sample standard deviation of the bootstrap values as we did for the ordinary bootstrap. Computing bias has a twist – we use the mean of the bootstrap values minus the parameter corresponding to the plug-in parametric distribution. Recall that the definition of bias is

$$\text{bias} = E[\hat{\theta}] - \theta = E_F[\hat{\theta}] - \theta_F,$$

the expected value of a statistic when sampling from a population  $F$  minus the corresponding parameter for that population. We estimate this using the bootstrap by plugging in an estimate for  $F$ ,

$$E_{\hat{F}}[\hat{\theta}] - \theta_{\hat{F}}$$

and we use the same  $\hat{F}$  in both places, whether that  $\hat{F}$  is the empirical data or a parametric distribution.

### 13.3 Stratified Sampling

We first described stratified sampling in the context of sample surveys, Section 1.8. Suppose a health clinic is running a survey from a population with 14% smokers and 86% nonsmokers; they can stratify and draw a sample with exactly 14% smokers and 86% non-smokers or draw observations randomly from the population and risk getting an unbalanced sample.

Stratified sampling is useful in computer simulations. For example, suppose we want to estimate  $E[e^X]$ , where  $X$  has a standard normal distribution. We can divide the population (the normal distribution) into two strata, say  $X < 0$  and  $X > 0$ , and draw exactly half the observations from each.

In all cases, sampling from a finite population or infinite population, real data or computer simulation, stratified sampling can reduce the variance of the result. In this section, we will quantify this.

Consider the general case where a population has  $J$  strata (subpopulations), and that stratum  $j$  represents a fraction  $\pi_j$  of the population, with  $\sum_j \pi_j = 1$ . For now assume that  $n\pi_j$  is a whole number for each  $j$ , so we could draw exactly the “right” number of observations from each stratum.

Let  $\mu_j$  and  $\sigma_j^2$  be the mean and variance for stratum  $j$  and let  $\mu$  and  $\sigma^2$  be the mean and variance for an observation chosen from the whole population; then  $\mu = \sum_{j=1}^J \pi_j \mu_j$  and

$$\sigma^2 = \sum_{j=1}^J \pi_j (\sigma_j^2 + (\mu_j - \mu)^2). \quad (13.4)$$

If observations are drawn from the whole population without regard to strata, then it is as if strata did not exist and

$$\text{Var}[\bar{Y}] = \sigma^2/n. \quad (13.5)$$

In contrast, when stratifying, if exactly  $n_j = n\pi_j$  observations are chosen for stratum  $j$ ,  $Y_{ji}$  for  $i = 1, \dots, n_j$ , then

$$\bar{Y} = (1/n) \sum_{j=1}^J \sum_{i=1}^{n_j} Y_{ji}$$

and the variance is

$$\text{Var}[\bar{Y}] = (1/n^2) \sum_{j=1}^J n_j \sigma_j^2 = (1/n) \sum_{j=1}^J \pi_j \sigma_j^2. \quad (13.6)$$

Comparing Equations (13.5) and (13.6), with a view to Equation (13.4), we see that the variance of the stratified estimate is smaller than that of the unstratified estimate by  $(1/n)\sum_j \pi_j(\mu_j - \mu)^2$ . If the strata means differ, the stratified estimate has smaller variance. This is terrific – it cannot make the variance of the estimates worse, it can only make them better.

The standard error for the stratified estimate is obtained by plugging in sample variances  $s_j^2$  for  $\sigma_j^2$  in Equation (13.6).

Stratification can also be used in delta method situations with similar gains. For example, for a ratio estimate  $\hat{\eta} = \bar{Y}/\bar{X}$  for bivariate data, the asymptotic variance in the unstratified case is  $\text{Var}[(Y - \eta X)/(n\mu_X^2)]$ , and in the stratified case

$$\text{Var}[\bar{Y}/\bar{X}] \approx \frac{1}{n\mu_X^2} \sum_{j=1}^J \pi_j \text{Var}_j[Y - \eta X].$$

Standard errors are obtained by substituting sample estimates where needed. In particular, for  $\text{Var}_j[Y - \eta X]$ , we substitute the sample variances of  $Y_{ji} - \hat{\eta}X_{ji}$ .

Another option is to bootstrap to obtain standard errors from a stratified sample, sampling separately from each stratum.

### 13.3.1 Post-stratification

There are three flavors of stratified sampling. The first involves how we pick the sample, the second how we produce estimates, and the third has elements of both.

The first, “proportional stratified sampling,” is described above: the number of observations take from each stratum is proportional to the population size.

The second is “post-stratification:” we draw the sample without determining how many units we should draw from each stratum; instead, we correct for differences between the population and sample proportions afterwards. For example, if a sample had too many smokers, say 25%, when there are only 14% in the population, we would downweight those smokers.

Suppose that the true stratum proportions are  $\pi_j$  and that the sample contains  $n_j$  observations from stratum  $j$  ( $j = 1, 2, \dots, J$ ). To compute an overall sample mean, we calculate the mean for each stratum and combine them:

$$\hat{\mu} = \sum_{j=1}^J \pi_j \bar{x}_j. \quad (13.7)$$

The variance of this estimate is  $\sum_j \pi_j^2 \sigma_j^2 / n_j$ , where  $\sigma_j^2$  is the variance for stratum  $j$ , and we plug in sample variances to obtain the standard error

$$\text{SE}(\hat{\mu}) = \sqrt{\sum_{j=1}^J \pi_j^2 s_j^2 / n_j}.$$

Similarly, estimates for population proportions are combinations of the stratum proportion estimates  $\hat{p} = \sum_{j=1}^J \pi_j \hat{p}_j$ . For other quantities we may compute weights  $w_j = \pi_j/n_j$ , use those weights for everyone in the stratum, then compute a weighted estimates. Note that if  $n_j$  happens to be relatively large, then the weights for subjects in that stratum will be smaller.

When subjects are chosen randomly from the overall population so that differences between population and stratum proportions are small and random, then the variance reduction offered by post-stratification is similar to proportional stratified sampling. Post-stratification is commonly used at companies such as Google and Instacart to get improved estimates from experiments. We randomly assign people to treatment and control groups, typically without considering other factors such as region, number of searches in the last week, or number of orders in the past week. We may end up with random imbalances between the treatment and control groups in these factors. Using post-stratification to control for those random differences reduces the overall variability and gives a better signal-to-noise ratio.

### 13.3.2 Optimal Stratified Sampling

The third flavor, (non-proportional) stratified sampling, combines sampling and post-sampling correction; we explicitly choose the sample sizes  $n_j$  to give lower variance. In particular, we over-sample strata that we think have higher variance. For example, to estimate the overall income in a state, we may over-sample from high-income precincts, then downweight observations from those strata to avoid bias.

If we optimize the stratum proportions, this is called “optimal stratified sampling.” Suppose we are allowed an overall sample size of  $n$ , and want to allocate that among strata to minimize the overall variance. We start by deriving the optimal allocation based on the unknown stratum variances  $\sigma_j^2$ . We minimize  $\text{Var}[\hat{\mu}]$  subject to the sampling budget:

$$\min \sum_{j=1}^J \frac{\pi_j^2 \sigma_j^2}{n_j} \quad \text{subject to} \quad \sum_{j=1}^J n_j = n.$$

We begin by treating the  $n_j$  as continuous values. Using Lagrange multipliers we set partial derivatives equal to zero:

$$\frac{\partial}{\partial n_k} \sum_{j=1}^J \frac{\pi_j^2 \sigma_j^2}{n_j} + \lambda \sum_{j=1}^J n_j = -\frac{\pi_k^2 \sigma_k^2}{n_k^2} + \lambda = 0$$

and solve for  $n_k$ , yielding  $n_k = (1/\sqrt{\lambda})\pi_k \sigma_k$  for some arbitrary  $\lambda$ . In other words, the optimal  $n_k$  is proportional to  $\pi_k \sigma_k$  and the solution given the sampling budget is

$$n_k^* = n \frac{\pi_k n_k}{\sum_{j=1}^n \pi_j n_j}.$$

In practice we need to round these to integers. In addition, given the uncertainty in our estimates of the  $\sigma_j^2$ , we may be conservative and choose an allocation that is somewhat closer to proportional stratified sampling.

There are other cases where we pick stratum sizes not to optimize a single overall variance, but to ensure that the sample contains reasonable coverage of all strata. For example, in tech companies it is common to use “human evaluation” to obtain training data for regression models. For example, at IC if a customer searches for “coffee,” we want to provide the most relevant results. “Folger’s coffee” would be relevant, “Dove Beauty Bar” not so much. We send a variety of query/product pairs to human raters, who rate the relevance of each product to the corresponding query. But we cannot pay humans to evaluate every possible pair; instead we use those human ratings as data to train (estimate coefficients for) a regression model (also known as a machine learning model). For this purpose, we want to provide a wide variety of combinations, so we may use stratified sampling to determine how many pairs to evaluate in different categories. See Exercise 13.15. In addition, over time we can refine the models with data that have what customers actually select.

## 13.4 Control Variates and Casual Modeling

Much of statistics is the study of relationships. And even when we are primarily interested in one quantity, we can use those relationships to our advantage via the technique of *control variates*. This is useful in both experiments and observational settings, sometimes yielding terrific variance reductions, or in letting us obtain reasonable estimates in what would otherwise be hopelessly biased situations. (Do not get confused between *control variate* and *control group*.)

In the simplest control variate situation, we wish to estimate  $E[Y]$  when there is another variable  $X$ , a *covariate*, that is correlated with  $Y$ ; in addition,  $E[X]$  is known. Using the simple algebraic relation  $Y = Y - bX + bX$ , where  $b$  is a constant, we can then write  $E[Y] = E[Y - bX] + bE[X]$ , which we estimate using

$$\hat{\mu}_{Y:b} = \bar{Y} - b(\bar{X} - \mu_X). \quad (13.8)$$

The variance of this estimate is

$$\begin{aligned} \text{Var}[\bar{Y} - b\bar{X} + bE[X]] &= \text{Var}[\bar{Y} - b\bar{X}] \\ &= \text{Var}[\bar{Y}] + b^2\text{Var}[\bar{X}] - 2b\text{Cov}[\bar{Y}, \bar{X}] \\ &= (\text{Var}[Y] + b^2\text{Var}[X] - 2b\text{Cov}[Y, X])/n. \end{aligned}$$

This variance is minimized when  $b = \beta = \text{Cov}[X, Y]/\text{Var}[X]$  – the slope of the regression line  $Y$  against  $X$ . In practice, we estimate this from the data with the least-squares estimate  $\hat{\beta}$  of the slope given in Equation (9.4). This gives

$$\hat{\mu}_{Y:\hat{\beta}} = \bar{Y} - \hat{\beta}(\bar{X} - \mu_X). \quad (13.9)$$

We can interpret this as correcting the simple estimate  $\bar{Y}$  by the difference between sample average and known mean of  $X$ .

The variance of this estimate depends on the residual variance,  $\text{Var}[Y - \beta X]$  (or equivalently  $\text{Var}[Y - (\alpha + \beta X)]$ ) where  $(\alpha, \beta)$  are the slope and intercept of the true regression line. The variance improves by a factor  $r^2$ , where  $r$  is the correlation between  $Y$  and  $X$ :  $\text{Var}[\hat{\mu}_{Y:\beta}] \approx (1 - r^2)\text{Var}[\bar{Y}]$ . If  $r$  is close to 1, that is,  $X$  and  $Y$  are highly correlated, this gives a good variance reduction.

**Example 13.2** Suppose we wish to estimate  $E[\sin(U)]$  for  $U \sim \text{Unif}[0, 0.5]$ . We will use  $U$  as a covariate.

Using a fixed  $b = 1$ , the estimate of  $E[\sin(U)]$  is  $\bar{Y} - \bar{X} + E[U]$ . The variance of this estimate is approximately  $0.000035/n$ , whereas the variance of the simple average  $\bar{\sin(U)}$  is about  $0.20/n$ ; the control variate variance is about 0.0018 as large.

Or we can set  $b = \hat{\beta} = 0.96$ , in which case the estimated variance of the estimate is  $0.0000055/n$ , smaller by an additional factor of 0.16.

```
> n <- 10^4
> U <- runif(n, 0, .5)
> Y <- sin(U)
> mean(Y)
[1] 0.244911
> mean(Y - U) + 0.25
[1] 0.244785
> var(Y)
[1] 0.01962409
> var(Y - U)
[1] 3.503318e-05
> betahat <- lm(Y ~ U)$coef[2]
> betahat
0.9626484
> var(Y - betahat * U)
[1] 5.497232e-06
```

□

We can also use multiple control variates,  $\hat{\mu}_Y = \bar{Y} - \sum_j b_j(\bar{X}_j - \mu_j)$ , and estimate the optimal coefficients using multivariate regression. In the special case that the covariates are dummy variables corresponding to strata, the resulting estimate is equivalent to post-stratification, Equation (13.7) (see Exercise 13.6).

Sometimes  $E[X]$  is not known, but we have another estimate for it that is better than  $\bar{X}$  from the sample, such as the mean from a larger sample.

For example, we might have collected mean blood PCB levels from a large sample of people last year, and a smaller subset of that sample this year; a rough estimate of average level now would be the mean from last year, plus the change in the smaller set. In such cases control variates provide a way to combine information from the two samples.

### 13.4.1 Control Variates in Experiments

Control variates are useful in reducing variability in experiments. We can randomly assign people to different arms of an experiment; the groups are similar on average, but there may still be random fluctuations. By controlling for the differences we can reduce the random variation and improve the signal-to-noise ratio. Researchers at Bing (Deng et al., 2013) use control variates in controlled experiments to reduce variance by about 50%, and we are incorporating a similar approach into the experimental infrastructure at IC.

Suppose we are trying an experiment with the goal of increasing the number of orders IC customers make. Let  $Y$  be the number of orders for a customer during a 14-day experiment, and we randomly assign customers to treatment or control (T or C) groups.

One way to reduce the variability is to use a *difference in differences estimate*. Let  $Y_{\text{pre}}$  denote the same metric as  $Y$  for the pre-period (period before the experiment), number of orders during 14 days before the experiment starts. Then instead of using the simple difference  $\bar{Y}_T - \bar{Y}_C$  to estimate the treatment effect, we could use

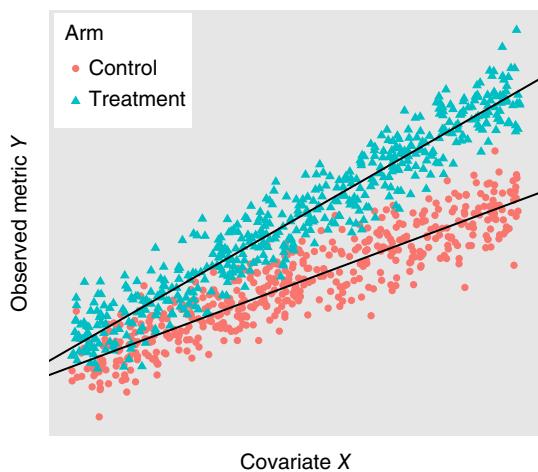
$$\bar{Y}_T - \bar{Y}_C - c(\bar{Y}_{T,\text{pre}} - \bar{Y}_{C,\text{pre}}),$$

for some constant  $c$ , say  $c = 1$ .

However, a different value of  $c$  which might give better results – perhaps the pre and post periods are not perfectly correlated in which case we might use a longer pre-period to reduce noise, or we might use some other metric altogether than  $Y_{\text{pre}}$ , in particular for an experiment with new customers. We need a data-driven way to choose the coefficient.

Figure 13.7 shows a typical situation where there is one linear relationship between the metric being measured,  $Y$ , and the covariate,  $X$ , for the treatment group,  $Y = a_T + b_T X$  and another for the control group  $Y = a_C + b_C X$ . The slopes of the lines may differ because the average magnitude of the treatment effect might depend on  $X$ . For example, if  $X$  is the number of orders in the pre-period, the experiment may have a greater effect when  $X$  is larger.

We cannot apply either Equation (13.8) or (13.9) directly because we do not know the population mean for the covariate. However, we do know that it should be the same for T and C, so we can estimate it using the



**Figure 13.7** A/B Trial, linear relationships between outcome metric and covariate.

average across both arms. Let  $\bar{x} = (n_C \bar{x}_c + n_T \bar{x}_T)/n$  be the common mean. Using Equation (13.9), our estimates for both groups and the treatment effect are:

$$\begin{aligned}\hat{E}[\bar{Y}_T | E[X] = \bar{x}] &= \bar{Y}_T - \hat{\beta}_T(\bar{x}_T - \bar{x}) \\ \hat{E}[\bar{Y}_C | E[X] = \bar{x}] &= \bar{Y}_C - \hat{\beta}_C(\bar{x}_C - \bar{x}) \\ \hat{\Delta} &= \hat{E}[\bar{Y}_T | E[X] = \bar{x}] - \hat{E}[\bar{Y}_C | E[X] = \bar{x}] \\ &= (\bar{Y}_T - \hat{\beta}_T(\bar{x}_T - \bar{x})) - (\bar{Y}_C - \hat{\beta}_C(\bar{x}_C - \bar{x})).\end{aligned}\quad (13.10)$$

Here is another way to interpret this estimate that is easier to remember and also extends to nonlinear models. In Figure 13.7, the vertical distance between the two lines at any value of  $x$  gives an estimate of the treatment effect at that  $x$ . Averaging those treatment effects across the combined sample gives an estimate of the *average treatment effect* (ATE) for the population.

$$\hat{\Delta} = \frac{1}{n} \sum_{i=1}^n \hat{Y}_T(x_i) - \hat{Y}_C(x_i), \quad (13.11)$$

where  $\hat{Y}$  are model predictions. Some algebra shows that this is equivalent to Equation (13.10) in the case that both models use linear regression.

The basic approach – fit two prediction models, one for the treatment group and one for the control group, let the individual treatment effect be the difference in predictions, and compute the ATE by averaging across the combined data – is useful in a wide variety of experimental analysis settings. We can have more than one covariate and use multiple regression. We can handle missing data using dummy variables to indicate missingness. We can handle categorical

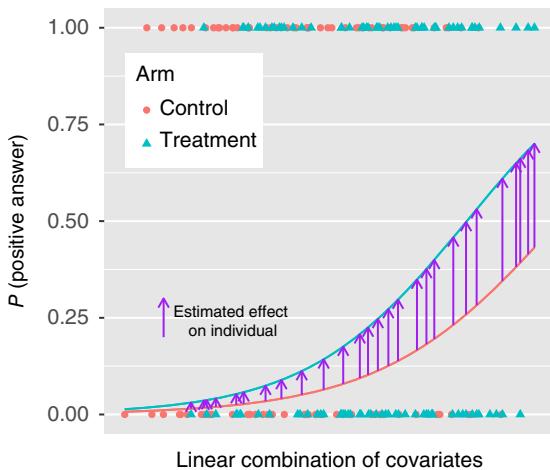
covariates by converting them to dummy variables. And we can use nonlinear models such as logistic or Poisson regression.

**Example 13.3** Google Survey Lift estimates the effectiveness of display advertising using surveys (Fan et al., 2018). Display ads pay for most of the free content on the web. Website owners reserve part of the space on their page for ads. When you visit a page, the website owner contacts a middleman like Google to place ads. Google picks a relevant ad, based on how relevant the ad is to the website and what it knows about your interests. Measuring how effective the ads are is difficult, because they are primarily there to build brand awareness, not to generate concrete actions like clicks. For example, an ad for a movie may try to make you aware of the movie and get you excited about going to see it.

To analyze a campaign, Google randomly assigns people to the treatment group (people who are allowed to see campaign ads) or a control group (campaign ads are held out or they may see a different ad). Later some people visit sites where Google can give a survey to ask questions such as whether people have heard of the movie or are thinking of going to see it. Different average responses between the treatment and control groups indicate the effectiveness of the ad.

Google uses logistic regression models to estimate the probability of a positive response ( $Y = 1$  if yes, 0 if no) using covariates such as age and gender. Here it is not possible to use the pre-period value of  $Y$  as a covariate because the survey is only given after displaying a campaign ad. Figure 13.8 shows an example of two logistic regression prediction models, and the estimated individual treatment effects (as arrows). The ATE is the average of all estimated individual treatment effects.

**Figure 13.8** Logistic relationships between outcome metric and covariate. The  $X$  axis corresponds to a linear combination of covariates  $L = \sum_{j=1}^p b_j X_j$ , the individual  $Y$  values to 0 or 1, and each the prediction is of the form  $\hat{Y} = \exp(L)/(1 + \exp(L))$ .



In practice the two models may have different coefficients, and inspecting the coefficients can yield valuable insights. For example, if the coefficient for the 25–34 age group in the C model is higher than for other age groups, that indicates that despite not seeing an ad, that group is relatively aware of the movie.

If the difference between T and C model coefficients is higher for the 18–24 age group than for other groups, that indicates that the ad is most effective for that age group.

In practice, advertisers are interested in the actual effect of their campaign – on all people who saw campaign ads – but the initial estimates were done on a small subset of that, people who answered surveys. We need to extrapolate to the larger population. For example, perhaps the campaign was especially effective among younger people, but a relatively small share of young people answered the survey – then the campaign was probably more effective overall than among those who answered.

Let  $S_0$  be the whole population,  $S_1$  be the subset who saw a campaign ad (or controls who would have),  $S_2$  the subset of  $S_1$  who later visit a site where a survey can be given,  $S_3$  be the random subset who are actually given a survey, and  $S_4$  be those who answer the survey. The population of interest is  $S_1$ , but the initial estimates are based on  $S_4$ .  $S_2$  is not a random subset of  $S_1$ , and  $S_4$  is not a random subset of  $S_3$ , so the answers obtained from  $S_4$  may be biased when extrapolated to  $S_1$ . Google fits T and C models to the  $S_4$  data, uses those models to predict T and C for  $S_1$ , and averages the T – C difference across  $S_1$ . This approach reduces bias when extrapolating from  $S_4$  to  $S_1$ , controlling for differences between  $S_1$  and  $S_4$  in age, gender, and other covariates.

Standard errors are difficult to compute analytically; Google uses bootstrapping.

The estimates are not perfect. Tracking is based on cookies, small files on your device. If you visit a website and see an ad, Google records that that cookie saw an ad. If you later visit a website where a survey can be shown, and have not cleared cookies, Google may show a survey, and record the survey answer. Two people sharing a device appear to be the same person. If you have multiple devices they are treated as separate people. The ad may be further down the page, and the browser may or may not indicate whether the person scrolled down that far. There are interesting challenges in estimating and adjusting for such effects.  $\square$

### 13.4.2 Potential Outcomes Framework

We wrote Equation (13.11) as the average of the difference of two predictions. Alternately, we could use actual values instead of predictions where we have them, e.g. for the T observations use  $Y_T$  instead of  $\hat{Y}_T(x)$ , and similarly for

controls. Plugging that into Equation (13.11) gives

$$\begin{aligned}\hat{\Delta} &= \frac{1}{n} \left( \sum_{i=1}^{n_T} (Y_{Ti} - \hat{Y}_C(x_{Ti})) + \sum_{i=1}^{n_C} (\hat{Y}_T(x_{Ci}) - Y_{Ci}) \right) \\ &= \frac{1}{n} \left( n_T \bar{Y}_T - \sum_{i=1}^{n_T} \hat{Y}_C(x_{Ti}) + \sum_{i=1}^{n_C} \hat{Y}_T(x_{Ci}) - n_C \bar{Y}_C \right).\end{aligned}\quad (13.12)$$

Averaging these differences give an identical ATE estimate because both linear and logistic regression have the nice property that the average of the predicted values is equal to the average  $Y$  value (Proposition 9.5)! So it did not matter that we used the actual values instead of the predictions.

But logically, we should use actual values when we have them. That brings us to the *potential outcomes* framework, and *Neyman–Rubin causal models*, originally described in a 1923 thesis in French by Jerzy Neyman, and extended by Rubin (1974).

The idea is that every experimental unit has two potential outcomes, what their  $Y$  value would be if they received the T treatment or the C treatment. However when the experiment is run, we observe only one of them. For instance, for those in the T group, we only observe the T outcome  $Y_T$ . The other potential outcome is the *counterfactual*, what would have occurred had that experimental unit been assigned to the control group.

We estimate these counterfactuals using data from the other group, the predictions from the model for the other group.

Group	Observed	Counterfactual	Estimate of counterfactual
Treatment	$Y_T$	$Y_C$	$\hat{Y}_C(x)$
Control	$Y_C$	$Y_T$	$\hat{Y}_T(x)$

### 13.4.3 Observational Data – Causal Modeling

In the examples above, the experimental units are assigned to the treatment or control randomly, and the regression modeling makes estimates of the  $Y$  less variable but is not required to obtain reasonable answers. In contrast, with observational data, assignment is not random. There may be substantial differences between the treatment and control groups, and modeling can help to reduce or ideally eliminate bias.

For example, an earlier Google project to estimate the effectiveness of display ads was not a randomized trial (Chan et al., 2010). Instead, for analyzing any given ad campaign, say for a new movie, the treatment group were defined as people who saw an ad from the campaign. There was no holdout group to use as controls. Instead, controls were people who visited the same websites where

the ads were (sometimes) shown and saw at least one ad (they were not using an ad blocker) but did not see the campaign ad.<sup>1</sup> One of the ads they saw was chosen as a reference ad.

The two groups were then compared based on whether they searched for a term related to the ad after seeing the ad, e.g. “Marvel Universe” (The data were from a group of people who had given Google permission to track their web activity; the data no longer exist.) (No surveys were used in this project.)

A simple comparison would suggest that the ads were wildly successful – the treatment group did far more relevant searches than the control group. But the comparison was badly biased – the reason that some people saw the ad while others did not is that some people just spent far more time on the Internet, performing more searches, visiting a wide variety of websites. They would have done more relevant searches, even without seeing a campaign ad.

We reduced the bias using a potential outcomes regression framework, controlling for every covariate we thought might be relevant and for which we had data, including whether and how many relevant searches the people in the survey did before the ad, how many total searches and page visits they made (whether relevant or not), what kinds of things they were interested in (based on the types of searches they did, pages they visited), their age, gender, and others covariates.

The model would be unbiased if it would properly include every variable that might affect both whether someone saw a campaign ad and the outcome. However, we cannot be sure of that. We tested for bias using dummy outcomes – e.g. seeing a movie ad should have no effect on whether someone searches for wool socks – and the model performed well there. But there could still be bias for more relevant outcomes.

It is impossible with a causal model to be sure there is not something missing from the model, some bias not removed. This is the attraction of randomized trials – the random assignment protects against bias.

There is one additional twist to this study – that the goal was to estimate the campaign effect, on people who actually saw the ads. Hence in place of Equation (13.11) or (13.12) we average only over the treatment group:

$$\hat{\Delta} = \bar{Y}_T - (1/n_T) \sum_{i=1}^{n_T} \hat{Y}_C(x_{Ti}).$$

## 13.5 Computational Issues in Bayesian Analysis

The examples in Chapter 11, involving beta and normal distributions, use conjugate priors. In these cases the posterior distributions of the parameters are from known distributions and computations of expected values and credible intervals are straightforward, using R or other software packages.

---

<sup>1</sup> Web pages are dynamic – in particular, different people may see different ads. Google records which ads it showed you (to the device with your Google cookie).

In practice, many applications are too complicated for conjugate priors, or the analyst may have prior knowledge that suggests using another prior. In most nonconjugate applications and in some conjugate applications, closed-form solution are not available, so approximations must be found.

Bayesian calculations typically involve integrals. Recall from Equation (11.11) that the posterior density may be written

$$p(\theta | x) = \frac{\pi(\theta)f(x | \theta)}{\int_{\Theta}\pi(\theta)f(x | \theta)d\theta},$$

where  $\Theta$  is the set of possible values for  $\theta$ . The denominator is the normalizing constant. To compute an expected value, say  $E_{p(\theta | x)}[\theta]$ , requires

$$\int_{\Theta}\theta p(\theta | x)d\theta = \int_{\Theta}\theta\pi(\theta)f(x | \theta)d\theta$$

and the normalizing constant. Similarly, to find a probability, say  $P(\theta \in A | x)$ , requires

$$\int_A\pi(\theta)f(x | \theta)d\theta$$

and the normalizing constant. Finding a credible interval requires finding values  $\theta_1$  and  $\theta_2$  that satisfy  $\int_{-\infty}^{\theta_1}\theta\pi(\theta)f(x | \theta)d\theta = \int_{\theta_2}^{\infty}\theta\pi(\theta)f(x | \theta)d\theta = C\alpha/2$ , where  $C$  is the normalizing constant. In the content experiments example at the end of Chapter 11,  $\theta$  is six dimensional, and to find the probability that arm 3 is the best, given the current data, requires

$$\int_{\Theta_1}\int_{\Theta_2}\int_{\Theta_3}\int_{\Theta_4}\int_{\Theta_5}\int_{\Theta_6}^{\infty} \pi(\theta)f(x | \theta)d\theta_3 d\theta_6 d\theta_5 d\theta_4 d\theta_2 d\theta_1$$

and the normalizing constant. In all cases, calculations require evaluating integrals with respect to  $\theta$ .

For univariate  $\theta$ , the integrals may be evaluated with techniques you may remember from calculus, such as Simpson's rule or the trapezoid rule. In R, the `integrate` function uses an *adaptive* version of such methods: for a finite region, it starts by evaluating points on a grid and then splits the grid into finer grids until it estimates that the accuracy satisfies a level specified by the user. For an infinite region, it first does a transformation (a " $u$ -substitution," in calculus terms) to express the problem as an integral over a finite region, before proceeding with the methods for a finite region.

Though these methods can be extended to functions of more than one variable, the number of evaluations required to achieve a desired level of accuracy increases exponentially in the number of dimensions: this is the "curse of dimensionality," a term coined by applied mathematician Richard Bellman (Bellman, 1966). For Simpson's method, for example, we can improve the accuracy by cutting the grid spacing in half, but in  $d$  dimensions this requires approximately  $2^d$  times as many points; in practice,  $d$  is often large enough to make this infeasible.

Alternatively, we may use simulation. We will describe some simulation methods next. One approach, known as *Monte Carlo integration*, uses simulation to estimate definite integrals. A variation of this approach, *importance sampling*, can improve efficiency and solve some otherwise intractable problems. We describe these methods next.

Another approach, *Markov Chain Monte Carlo* (known as MCMC or MC<sup>2</sup>), involves drawing sequences of dependent observations; for more information, see Gilks et al. (1996), Robert and Casella (2004), Robert and Casella (2010), or Suess and Trumbo (2010). MCMC has exploded in popularity; Google Scholar reports about 290 000 results for MCMC. MCMC is beyond the scope of this text. The R `nimble` package (de Valpine et al., 2021) and related packages offer a good balance of flexibility and ease of use, see <https://r-nimble.org>.

## 13.6 Monte Carlo Integration

Consider the problem of estimating a univariate integral over a finite region,  $\int_a^b h(x)dx$ , with  $b > a$ . Let  $X_i \stackrel{\text{i.i.d.}}{\sim} U(a, b)$  for  $i = 1, 2 \dots$ , and let  $Y_i = (b - a)h(X_i)$ . By Theorem A.1,

$$E[Y_i] = E[h(X_i)] = \mu = \int_a^b (b - a)h(x) \frac{1}{b - a} dx = \int_a^b h(x)dx.$$

By the strong law of large numbers,

$$\lim_{N \rightarrow \infty} \bar{Y}_N = \mu = \int_a^b h(x)dx$$

with probability 1, where  $\bar{Y}_N = (1/N)(Y_1 + Y_2 + \dots + Y_N)$  is the average after  $N$  replications. Thus, for large  $N$ ,

$$\bar{Y}_N \approx \int_a^b h(x)dx.$$

In other words, to approximate  $\int_a^b h(x)dx$ , we randomly draw  $N$  points from the interval  $[a, b]$ , evaluate  $h$  at each of these points, and compute the average times the length of the interval (since  $Y_i = (b - a)h(X_i)$ ).

If  $\text{Var}[Y] < \infty$  (the usual case), we can compute the standard error for  $\bar{Y}$  as  $s/\sqrt{n}$  and calculate a confidence interval for  $\mu$ .

**Example 13.4** Estimate  $\int_1^3 e^{-x^2} dx$ .

### Solution

The following R code draws random numbers from the uniform distribution, evaluates  $h(x) = (3 - 1)e^{-x^2} = 2e^{-x^2}$  at each random point, and averages.

**R Note**

```
> N <- 10^5
> x <- runif(N, 1, 3)    # draw from Unif[1, 3]
> out <- 2*exp(-x^2)     # evaluate h at each random x
> mean(out)
[1] 0.1401142
> sd(out) / sqrt(N)      # standard error
[1] 0.0006160375
```

The `integrate` function computes definite integrals. The first argument must be an R function.

```
> integrate(function(x) exp(-x^2), 1, 3)  # adaptive
0.1393832 absolute error < 1.5e-15
```

One run of this simulation gives an estimate of 0.1401 with standard error 0.00062. Thus, a 95% confidence interval for the true value of the integral is  $0.1401 \pm 1.96 \times 0.00062 = (0.1395, 0.1407)$ . For comparison, the value obtained by adaptive numerical integration is 0.1394.  $\square$

**Example 13.5** According to a poll conducted by the Pew Research Center for the People and the Press (February 2019), 31% of 920 teenagers, between 13 and 17 years of age, surveyed said that they get help or advice from parents with homework or school projects on an almost daily basis.<sup>2</sup>

We will consider the binomial model for the data and assume  $X \sim \text{Binom}(n, \theta)$  with  $X = 285, n = 920, \hat{\theta} = 0.31$ .

Consider the following prior density for  $\theta$ :

$$\pi(\theta) = ce^{-25|\theta-0.5|}, \quad 0 \leq \theta \leq 1,$$

where  $c$  is a normalizing constant. Then by Equation (11.3), the posterior density is

$$\begin{aligned} p(\theta | X = 285) &= \frac{c e^{-25|\theta-0.5|} \theta^{285} (1-\theta)^{635}}{\int_0^1 c e^{-25|u-0.5|} u^{285} (1-u)^{635} du} \\ &= e^{-25|\theta-0.5|} \theta^{285} (1-\theta)^{635} / K, \end{aligned} \tag{13.13}$$

where  $K$  is another normalizing constant.

---

<sup>2</sup> <https://www.pewsocialtrends.org/2019/02/20/most-u-s-teens-see-anxiety-and-depression-as-a-major-problem-among-their-peers/>

To compute the expected value of  $\theta$ , given the data,

$$\begin{aligned} E[\theta | X = 285] &= \int_0^1 \theta p(\theta | X = 285) d\theta \\ &= \int_0^1 \theta e^{-25|\theta - 0.5|} \theta^{285} (1 - \theta)^{635} d\theta / K. \end{aligned}$$

Thus, we have two integrals to compute: one for the numerator and another for the normalizing constant  $K$ .

### R Note

We define three functions in R, draw random numbers from the uniform distribution on [0,1], then evaluate each of these functions at these random numbers.

```
#first function computes prior, except for the constant
f0 <- function(u) exp(-25 * abs(u-0.5))
f1 <- function(u) f0(u) * u^285 * (1-u)^635 # prior*like
f2 <- function(u) u * f1(u) # for expected value
x <- runif(10^6)
a <- mean(f0(x)) # constant for prior
K <- mean(f1(x)) # constant in denom. of E[theta|x=215]
b <- mean(f2(x)) # numerator of E[theta|x=215]
c(a, K, b, b/K) # output all

# Plot the prior and posterior
df <- data.frame(x = x, y = f0(x)/a, w = f1(x)/K)
ggplot(df) +
  geom_line(aes(x = x, y = y, color = "Prior")) +
  geom_line(aes(x = x, y = w, color = "Post"), lty = 2) +
  scale_color_manual(name = NULL,
  values = c("Prior" = "black", "Post" = "red"))
```

In one run of this simulation, we find that the estimated mean of the posterior distribution is 0.316, with density shown in Figure 13.9.  $\square$

We may draw random variables from any distribution, not just the uniform distribution.

### General Monte Carlo Integration

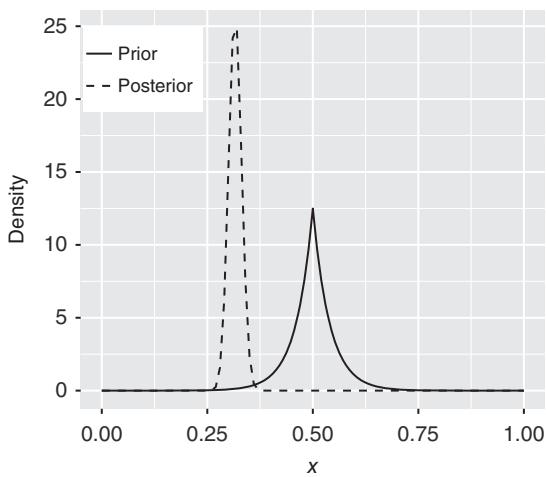
Suppose that an integral can be written in the form

$$\int_C h(x)f(x)dx = E[h(X)], \quad (13.14)$$

where  $f$  is a density with  $f(x) > 0$  over  $C$  and zero elsewhere, and that the integral exists and is finite. Then, we can estimate  $E[h(X)]$  by generating  $X_1, X_2, \dots, X_N$  i.i.d. from a distribution with pdf  $f$  and using the sample mean

$$\frac{1}{N} \sum_{i=1}^N h(X_i). \quad (13.15)$$

**Figure 13.9** Prior and posterior densities for Pew survey example.



We leave the verification of this as an exercise (Exercise 13.9).

We may need to do a bit of manipulation to express a problem in the form required by Equation (13.14). For example, to integrate  $\int_C r(x)dx$ , choose a density  $f$  with the proper domain, then let  $h(x) = r(x)/f(x)$ . For example, to integer  $\int_0^\infty \sin(x^2) \exp(-x^2)dx$ , we may choose an exponential distribution  $f(x) = \exp(-x)$ , then  $h(x) = \sin(x^2) \exp(-x^2 + x)$ .

Or, suppose the integral appears to be in the proper form, but with the wrong domain, for example,  $\int_0^5 x^2 \exp(-x)$  looks like  $h(x) = x^2$  and  $f(x) = \exp(-x) I(x > 0)$  (a standard exponential distribution) would work, except that the domain is wrong. One option is to let  $f$  be a uniform distribution over  $(0, 5)$  and define  $h$  accordingly. Another option is to let  $f$  be the standard exponential and incorporate the domain restriction into  $h$ ,  $h(x) = x^2 I(0 < x < 5)$ .

In Bayesian applications, two natural choices for  $f$  are the prior and posterior distributions, if these are tractable.

**Example 13.6** Refer to Example 11.4. Based on the data, the biologist's posterior distribution for the mean lengths (cm) of trout is  $\mu \sim N(45.53, 1.953^2)$ . Biologists have determined that the relationship between the weight ( $W$ ) and

length ( $X$ ) of a fish is approximately  $W = aX^b$  for constant  $a$  and  $b$  that are determined empirically for any give species (Ricker, 1973, 1975). Suppose for this species,  $W = 0.088 \times X^{3.069}$  g. If  $X \sim N(45.53, 1.953^2)$ , find the expected value of the weight of the trout.

### Solution

We want  $E[W] = E[0.088X^{3.069}] = \int_{-\infty}^{\infty} 0.088x^{3.069} \cdot f(x)dx$ . where  $f(x)$  is the pdf for the normal distribution with mean 45.53 and standard deviation 1.953.

#### R Note

```
x <- rnorm(10^6, 45.53, 1.953)
out <- 0.088*x^(3.069)
mean(out)
```

One run of the simulation gives an estimated mean of 10 872.55 g (23.9 lb).  $\square$

For functions of a single variable, estimating integrals using Monte Carlo integration is generally less effective than deterministic methods such as Simpson's rule. But such deterministic methods suffer from the curse of dimensionality, whereas Monte Carlo does not; doubling the accuracy requires four times as many points, regardless of the dimension.

Monte Carlo has another advantage – we can use it to obtain not only the expected value but also the whole distribution. In the previous example, where  $f$  is the posterior distribution for fish lengths, the values  $h(X_i)$  are a sample from the posterior distribution of fish weights  $h(X)$ .

But in many applications it is not possible to draw samples from the posterior. We will use a technique called *importance sampling* to draw observations from some other distribution, but still obtain answers based on the posterior. We will start with basic importance sampling for estimating integrals or expected values and then turn to estimating distributions.

## 13.7 Importance Sampling

The Monte Carlo procedure may be inefficient if  $h(X)$  is skewed, for example, if  $h(X)$  is small most of the time and only rarely takes on large values. Then, the estimate of an integral may be largely determined by a relatively small number of points with large  $h(X_i)$ .

We are not limited to sampling from the prior, posterior, or uniform distributions. The original motivation behind *importance sampling* is to choose the distribution to emphasize important regions, to oversample these regions, and to obtain more observations there, increasing the effective sample size.

Importance sampling turns out to be useful in a variety of applications, Bayesian and otherwise, including many cases where it is the best way to get answers, regardless of efficiency.

**Example 13.7** My (Tim) first career job was for Pacific Gas and Electric Co., estimating how much fuel oil PG&E should carry in inventory at the start of a winter. PG&E had a diverse system, with electricity generated by hydroelectric power in California and the northwest, natural gas fired generators, nuclear, geothermal, wind, and burning oil. Power demand and availability is random, depending on factors such as temperature, rainfall, and breakdowns. There were about 900 input random variables.

Burning oil was the fuel of last resort, more expensive and dirtier than other options, and PG&E would rarely burn oil. Oil inventories were very expensive (about US\$200 million in inventory), so we did not want to hold more than necessary. However, a combination of high demand and low generation from other sources, over a sustained period, could require burning enough oil to exhaust the oil stocks and result in a potentially major, long-term shortage, where the company could not generate sufficient electricity. Naturally, this would probably occur at the worst possible time, in the middle of a cold winter.

Oil has to be in inventory at the start of winter. PG&E used low-sulfur crude that required special delivery from Indonesia; this takes 2 months, and delivery in mid-winter may be impossible due to winter storms.

The ratio between the marginal cost of running one barrel short and the marginal cost of holding an extra barrel was about 300 to 1, so an optimal inventory would result in around 99.7% reliability – holding enough oil to run even one barrel short only about once every 300 years.

The model was large and complex – I would often exceed quota and be kicked off the company mainframe, a multimillion dollar machine. Early versions of the model generated random data in a way that mimicked real life, generating synthetic years of temperature, rainfall, breakdowns, and so on. But only 1 of every 300 such replicates yielded useful information.

We switched to importance sampling. We over-sampled cold dry weather with more breakdowns, so that more replicates yielded useful information. We then used importance sampling formulas to adjust for the sampling bias. The work is described in greater detail in Hesterberg (1988, 1995). □

We will start with the classical importance sampling approach, see what is wrong with that, and discuss an alternative. Let  $f$  be a pdf, and  $h$  a function. We write

$$\mathbb{E}[h(X)] = \mu = \int h(x)f(x)dx = \int h(x)\frac{f(x)}{g(x)}g(x)dx = \int h(x)w(x)g(x)dx, \quad (13.16)$$

where  $g$  is another density with  $g(x) > 0$  when  $f(x)h(x) \neq 0$ , and the ratio  $w(x) = f(x)/g(x)$  is a weighting function.

Instead of drawing observations from  $f$ , we may instead draw observations  $X_i \stackrel{\text{i.i.d.}}{\sim} g$  and, referring to General Monte Carlo Integration, compute the classical importance sampling estimate

$$\hat{\mu}_{\text{IS}} = \frac{1}{N} \sum_{i=1}^N h(X_i)w(X_i). \quad (13.17)$$

$f$  is called the *target distribution* and  $g$  the *design distribution* or *importance function*. The target is what we want answers for; we design our simulation as a means to get there. In the case of PG&E, we choose  $g$  with  $g(x) > f(x)$  for cold, dry years with more breakdowns. But that sampling is biased, and if we then just took a sample average of the resulting costs, it would be biased. The weighting factor  $w$  counteracts the sampling bias by downweighting the over-sampled cases and upweighting the under-sampled cases.

Write  $Y_i = h(X_i)w(X_i)$ ; then  $\hat{\mu}_{\text{IS}} = \bar{Y}$  is just a sample average, with standard error  $s_Y/\sqrt{n}$ . Thus, we may use the standard error or confidence intervals to gauge the accuracy of this estimate.

**Example 13.8** In a *European Option*, an investor purchases the right to buy a stock on a certain future date at a given price  $K$ , the “strike price.” Suppose the stock price on that date is  $X$ ; the value of the option is then  $h(X) = \max(X - K, 0)$ . Suppose the strike price is  $K = 700$  and that  $X \sim N(500, 120^2)$ , based on historic volatility and current market conditions. What is the expected value of  $h(X)$ ? What is the distribution of the payout?

### Solution

Here, the target distribution  $f$  is the normal density with mean 500 and standard deviation 120. That implies that  $X$  could be negative; in practice it could not, but we ignore the discrepancy because the probability is small. The probability that the option has any value is  $P(X > 700) = 1 - \Phi((700 - 500)/120) = 0.048$ , slightly under 5%.

The expected value is

$$\mathbb{E}[h(X)] = \mu = \int_{-\infty}^{\infty} h(x)f(x)dx = \int_{700}^{\infty} (x - 700)f(x)dx = 2.38. \quad (13.18)$$

The average payout is 2.38, even though 95% of the time the payout is zero; hence, the average payout given that there is a payout is  $2.38/0.048 = 49.8$ .

Now consider solving this problem using simulation, without importance sampling. We generate  $X_i \stackrel{\text{i.i.d.}}{\sim} N(500, 120^2)$ , compute the payout  $h(X_i)$  for each, and calculate the sample average, as well as standard errors and confidence intervals. Using  $10^5$  replications, we estimate a mean payout of 2.33 with standard error 0.04.

As a bonus, we get an estimate of the whole distribution of the payout – the values of  $h(X_i)$  are a sample from that distribution. This makes it easy to compute such quantities as the mean payout, given that the payout is nonzero (49.5) or the probability that the payout exceeds 100 (0.0056) – the person selling the option may be interested in that or similar quantities.

As a side note, the payout distribution is neither discrete nor continuous, but has elements of both.

Now consider using importance sampling. Instead of sampling from  $N(500, 120^2)$ , we will sample from  $N(700, 120^2)$ : that is, we set  $g$  to be the density for  $N(700, 120^2)$ . With this design, half the observations have nonzero payout. The weighting factor  $w(x) = f(x)/g(x)$  is then much less than one for those observations with positive payout. Thus, drawing  $X_i \stackrel{\text{i.i.d.}}{\sim} N(700, 120^2)$ , our estimate is of the form

$$\hat{\mu}_{\text{IS}} = \frac{1}{N} \sum_{i=1}^N \max(X_i - 700, 0) w(X_i).$$

With  $10^5$  replications, we estimate a mean of 2.36, with standard error 0.0085. The standard error is about five times smaller than for simple Monte Carlo; in other words, we achieve comparable accuracy with about  $5^2$  times fewer replications (actually 27.6). This is the *relative efficiency*, or variance under simple Monte Carlo divided by variance with this importance sampling distribution.

### R Note

```
K <- 700      # strike price
mu <- 500     # mean of the stock price at option date
sigma <- 120 # sd

# P(option has value)
pnorm(K, mu, sigma, lower.tail = FALSE)

# define function g(x) = h(x)*f(x)
g <- function(x) (x - K) * dnorm(x, mu, sigma)
# expected payout
integrate(g, K, Inf)
# expected payout given there is a payout
integrate(g, K, Inf)$value /
  pnorm(K, mu, sigma, lower.tail = FALSE)
```

The option `lower.tail = FALSE` to `pnorm` gives  $1 - \text{pnorm}(K, \mu, \sigma)$ .

We next estimate the integral using Monte Carlo integration without importance sampling. The `pmax` function computes the coordinate-wise maximum of two vectors.

```
# Simulation, no ImpSamp
N <- 10^5
X <- rnorm(N, mu, sigma)
payout <- pmax(X-K, 0)
mean(payout)
sd(payout) / sqrt(N)
mean(payout[payout > 0])
mean(payout > 100)
```

`payout > 0` returns a vector of TRUE or FALSE's (depending on whether or not a particular value is greater than 0. Thus, `payout (payout > 0)` returns those payout values corresponding to the TRUE's.

`payout > 100` return a vector of TRUE's or FALSE's depending on whether or not a particular value is greater than 100. Thus, `mean (payout > 100)` returns the proportion of values greater than 100.

Now we integrate using importance sampling:

```
# ImpSamp, normal with mean = K
X2 <- rnorm(N, K, sigma) # drawing from g
w2 <- dnorm(X2, mu, sigma) / dnorm(X2, K, sigma) # w2(x) = f(x)/g(x)
Y2 <- pmax(0, X2-K) * w2 # h(x) * w(x)
mean(Y2)
sd(Y2) / sqrt(N)
sd(Y2) / sd(payout)
var(payout) / var(Y2) # relative efficiency
```

□

**Definition 13.1** Let  $\hat{\theta}_1$  and  $\hat{\theta}_2$  be estimates of  $\theta$  under two Monte Carlo methods. The *relative efficiency* of these two methods is the ratio of the variances under these methods:  $\text{Var}[\hat{\theta}_1]/\text{Var}[\hat{\theta}_2]$ . ||

Choosing to sample from  $N(700, 120^2)$  in the previous example was a bit ad hoc. Let us look a bit deeper and figure out what makes a good importance function  $g$ .

We may try to choose  $g$  to minimize the variance of the estimate,  $\sigma_Y^2/N$ .

**Theorem 13.1** Let  $E_g[Y]$  and  $\text{Var}_g[Y]$  denote the expected value and variance of  $Y$  when sampling from  $g$ . Then,  $\text{Var}_g[Y]$  is minimized when

$$g_{\text{IS}}^*(x) \propto |h(x)f(x)|. \quad (13.19)$$

*Proof.* We give here a non-rigorous proof based on calculus of variations; readers not familiar with this method may skip the proof.

$$\begin{aligned} \text{Var}_g[Y] &= E_g[Y^2] - \mu^2 \\ &= \int h(x)^2 w(x)^2 g(x) dx - \mu^2 \\ &= \int h(x)^2 f(x)^2 / g(x) dx - \mu^2. \end{aligned}$$

Write  $H(x) = h(x)f(x)$ ; the objective is to minimize  $\int H(x)^2 / g(x) dx$  subject to  $\int g(x) dx = 1$  and  $g(x) \geq 0$  for all  $x$ . Consider minimizing

$$\int \frac{H(x)^2}{g(x)} + \lambda g(x) dx$$

and setting the derivative with respect to  $g$  in the integral equal to zero,

$$\frac{-H(x)^2}{g(x)^2} + \lambda = 0,$$

the solution has  $g(x) = \lambda|H(x)|$  for some  $\lambda$ . □

In a case like this, where  $h(x)$  is non-negative for all  $x$ , Equation (13.19) reduces to

$$g^*(x) = \frac{h(x)f(x)}{\int_{-\infty}^{\infty} h(u)f(u)du}, \quad (13.20)$$

which makes

$$Y = h(X)w(X) = h(x) \frac{f(x)}{g(x)} = \int_{-\infty}^{\infty} h(u)f(u)du.$$

In other words, the optimal  $g$  makes  $Y$  a constant, so the estimate has zero variance. Unfortunately, this distribution requires the normalizing constant  $\int_{-\infty}^{\infty} h(u)f(u)du$ , which happens to be the answer we are looking for (Equation (13.18)). In other words, we have to know the answer in order to get the answer. This is a problem!

Even though it is impossible to generate values from the optimal  $g^*$ , maybe we can come close. We can use Equation (13.19) as a guide – we would like  $g$  to

be roughly proportional to  $|h(x)f(x)|$ . So  $g$  should be bigger than  $f$  when  $h$  is large – the sampling should be biased to produce more of the important cases where  $h$  is large. Conversely,  $g$  can be smaller than  $f$  where  $h$  is small.

### Example 13.8 (continued)

In the European option example,  $f$  can even be zero where  $h$  is zero, for  $x < 700$ . Let's try a *translated exponential distribution*, an exponential distribution translated right 700 units:

$$g(x) = \begin{cases} \lambda \exp(-\lambda(x - 700)), & x \geq 700, \\ 0, & x < 700. \end{cases} \quad (13.21)$$

This “wastes” no observations for values of  $x$  where  $h(x) = 0$ . To use this, we must choose  $\lambda$ . For a start, we will try using an exponential distribution with the same standard deviation as the normal distribution  $f$ . The relative efficiency is about 102, about four times better than using a normal distribution for importance sampling.

#### R Note (European Option Continued)

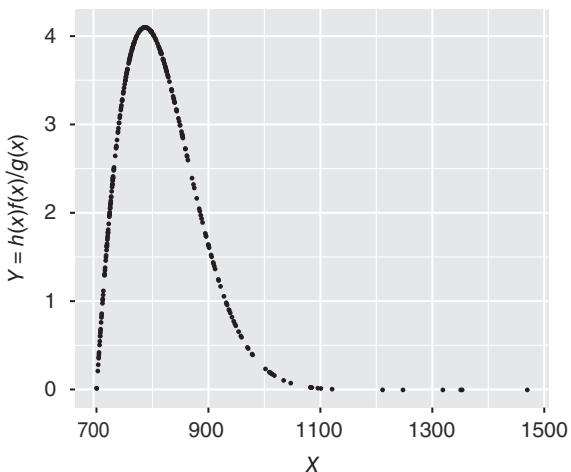
```
lambda <- 1 / sigma      # same st deviation as the normal
X3 <- K + rexp(N, lambda)  # g ~ shifted exponential
w3 <- dnorm(X3, mu, sigma) / dexp(X3 - K, lambda)
Y3 <- pmax(0, X3-K) *w3
mean(Y3)
sd(Y3) / sqrt(N)
sd(Y3) / sd(payout)
var(payout) / var(Y3)

df <- data.frame(x = X3[1:300], y = Y3[1:300])
ggplot(df, aes(x = x, y = y)) + geom_point(size = .5) +
  labs(x = "X", y = "Y = h(x)f(x)/g(x)")
```

A scatter plot of the resulting  $Y$  versus  $X$  is shown in Figure 13.10.  $Y$  is initially small, then large, then small, indicating that the  $g$  was too large at both ends, relative to  $g^*$ . Changing  $\lambda$  could help at either end, at the cost of making the other end worse.

We could put in more effort to find a better  $g$ . To reduce the number of replications, we would choose  $g$  to make  $Y(x) = h(x)f(x)/g(x)$  as flat as possible. The next step might be to try a shifted gamma distribution with shape parameter  $r = 2$  chosen to mimic the shape of  $h \times f$  near 700, namely,  $g(x) \approx c(x - 700)$  for some  $c$ , and scale parameter chosen to mimic  $h \times f$  over a broader range. We would not do this here; in practice, there is always a trade-off between human and computer time, and for this application more effort is not needed.  $\square$

**Figure 13.10** Importance sampling for the European option example, where  $X \stackrel{\text{i.i.d.}}{\sim} g$  is exponential and  $Y = h(X)f(X)/g(X)$ .



### 13.7.1 Ratio Estimate for Importance Sampling

We hinted above that there is something wrong with the classical importance sampling approach. Actually, there are a couple of problems. Let us see what they are and consider remedies.

In practice, we usually want to estimate more than one quantity. In the fuel inventory example, we want to estimate quantities such as the probability of running short, expected cost, probability of cost exceeding US\$300 million, probability of running short in each of several months, and the expected oil burn in each month, and the expected amount of oil left in inventory. For the European option, we want to estimate the expected value, probabilities of the value exceeding different levels, and the probability of a gain.

One problem with the classical approach is that a design distribution that works well for one quantity may be terrible for another; we will consider this below. The second problem is that its estimates for various quantities are inconsistent, for example,

$$\hat{P}(\text{run out of oil}) + \hat{P}(\text{no run out of oil}) \neq 1$$

and

$$\hat{E}[\text{amount of oil burnt}] + \hat{E}[\text{amount of oil left over}] \neq \text{total amount of oil.}$$

The reason is that the estimate is a weighted average with weights that do not sum to 1, so probabilities do not add to 1. The importance sampling estimate of a probability is obtained by letting  $h$  be a zero–one variable in Equation (13.17). For example, the estimated probability that an option has value is

$$\hat{P}_{\text{IS}}(X > 700) = \frac{1}{N} \sum_{i=1}^N I(X_i > 700) w(X_i),$$

where  $I(A)$  is the *indicator function*, with value 1 when  $A$  occurs and 0 otherwise. The probability that an option has no value is similar, and the probability of one or the other is

$$\begin{aligned}\hat{P}_{\text{IS}}(X > 700) + \hat{P}_{\text{IS}}(X \leq 700) &= \frac{1}{N} \sum_{i=1}^N (I(X_i > 700) + I(X_i \leq 700))w(X_i) \\ &= \frac{1}{N} \sum_{i=1}^N w(X_i).\end{aligned}$$

In general, this is not equal to 1. In our simulation, the two probabilities add to 0.988 when using the normal design distribution and 0.048 when using the exponential design. In fact, with the latter, the estimated probability of the option having no value is zero!

The problem is that the classical estimate is a weighted average, with weights  $w(X_i)/N$  that do not sum to 1. A remedy is to normalize the weights to sum to 1, using  $w_i / \sum_{j=1}^N w_j$ ; this simplifies to the *importance sampling ratio estimate*:

$$\hat{\mu}_{\text{ratio}} = \sum Y_i / \sum w(X_i) = \bar{Y} / \bar{W}. \quad (13.22)$$

As long as  $g(x) > 0$  when  $f(x) > 0$ , then  $E[\bar{X}] = 1$  (the proof is an exercise), and this estimate is usually similar to the earlier estimate Equation (13.17). The ratio estimate does require  $g(x) > 0$  when  $f(x) > 0$  to be consistent; the translated exponential distribution does not qualify.

One extra benefit of using the ratio estimate is that we get the whole distribution for free, as a weighted empirical distribution. For any output quantity  $h$ , the estimate of the distribution of  $h(X)$  has cdf

$$\hat{F}(v) = \frac{\sum_{i=1}^N w_i I(h(x_i) \leq v)}{\sum_{i=1}^N w_i}. \quad (13.23)$$

A second extra benefit of the ratio estimate is that  $f$  need not be completely known, it may have an unknown normalizing constant. This is useful in Bayesian applications.

**Remark** Now a personal note. When I (Tim) first applied importance sampling in the fuel inventory problem, I was not familiar with the importance sampling literature and did what made sense to me. Then, at one point my boss asked me, “You mean you normalize the weights to add to 1?” My answer was roughly, “Of course, doesn’t everyone?” Well, actually no, they did not. It was a bit embarrassing. However, this realization and the follow-up work turned into my PhD dissertation (Hesterberg, 1988).

One reason that people used the classical estimate instead of the ratio estimate is that the classical estimate is unbiased, while the ratio estimate is in general biased. However, the bias is small when  $N$  is large (bias  $\leq c/N$  for some  $c$ ), and we can make  $N$  large running computer simulations. This is a case where people were too concerned about unbiasedness. ||

As long as  $g(x) > 0$  when  $f(x) > 0$ ,  $E[W] = 1$ , and the ratio estimate is approximately normal with mean  $r = \mu_Y$  and variance  $\text{Var}[Y - \mu_Y W]/N$ , from Equation (7.7). The standard error is  $s/\sqrt{N}$  where  $s$  is the sample standard deviation of the residuals  $Y - \hat{\mu}_{\text{ratio}} W$ .

The asymptotically optimal design for the ratio estimate minimizes

$$\begin{aligned}\text{Var}_g[Y - \mu_Y W] &= E_g[(Y - \mu_Y W)^2] \\ &= \int (h(x) - \mu_Y)^2 w(x)^2 g(x) dx \\ &= \int (h(x) - \mu_Y)^2 f(x)^2 / g(x) dx\end{aligned}$$

and is

$$g_{\text{ratio}}^*(x) \propto |h(x) - \mu| f(x). \quad (13.24)$$

This is often more reasonable than  $g_{\text{IS}}^*$  from Equation (13.19).  $g_{\text{ratio}}^*$  is large when  $h(X)$  is far from its mean, whereas  $g_{\text{IS}}^*$  is large when  $h(X)$  is far from zero. Consider two examples:

- To estimate  $P(\text{option has value})$ ,  $h(x) = I(x > 700)$ ,  $g_{\text{IS}}^*$  is zero for  $x \leq 700$ ; while to estimate  $P(\text{option has no value})$ ,  $h(x) = I(x \leq 700)$ ,  $g_{\text{IS}}^*$  is zero for  $x > 700$ . Hence, to estimate equivalent quantities, the design would be the exact opposite.  $g_{\text{ratio}}^*$  is the same for estimating both probabilities and draws half of its observations from each region.
- Consider estimating the average amount of oil left in inventory at the end of the winter. This is usually identical to the initial amount.  $g_{\text{IS}}^*$  tries not to draw observations from the rare difficult cases (cold, dry years with power plant breakdowns) where a lot of oil is burned. In contrast,  $g_{\text{ratio}}^*$  tries to draw more observations from these cases, exactly as if we were estimating the average amount of oil burned.

A strategy that usually works well for both the classical and ratio estimates and is often nearly optimal for the ratio estimate is to use a *defensive mixture distribution* (Hesterberg, 1988, 1995). Pick a distribution that emphasizes the

rare cases and call it  $g_1$ . But rather than rely completely on  $g_1$ , we use a mixture of 50% of samples from  $g_1$  and 50% from  $f$ ,

$$g(x) = 0.5g_1(x) + 0.5f(x). \quad (13.25)$$

Taking half the replications from  $f$  defends against everything that can go wrong if  $g_1$  happens to be bad, either for the main output or for a secondary output. It bounds the weight ratio  $w(x)/g(x) \leq 2$ , so the estimate cannot be dominated by one or a few observations with huge weights.

For estimating probabilities,  $g_{\text{ratio}}^*$  puts 50% of its probability on the cases where the event occurs and 50% where it does not (the proof is an exercise). The 50% defensive mixture tends to approximate this.

To generate replications from Equation (13.25), it is best to stratify (Section 13.3) and draw exactly  $N/2$  observations from each of the two mixture components.

### 13.7.2 Importance Sampling in Bayesian Applications

Now turn back to Bayesian applications. For importance sampling in a Bayesian context, we consider the posterior density for some parameter  $\theta$ ,

$$p(\theta | x) = \frac{\pi(\theta)L(\theta)}{\int \pi(\theta)L(\theta)d\theta},$$

where  $L(\theta)$  is the likelihood. The target distribution is  $p(\theta | x)$ . To compute the expected value of any function  $h$  of  $\theta$ , sampling from  $g$ , we estimate

$$\mathbb{E}[h(\theta) | x] = \frac{\int h(\theta)\pi(\theta)L(\theta)d\theta}{\int \pi(\theta)L(\theta)d\theta} \quad (13.26)$$

$$= \frac{\int h(\theta)w(\theta)g(\theta)d\theta}{\int w(\theta)g(\theta)d\theta}, \quad (13.27)$$

where

$$w(\theta) = \frac{\pi(\theta)L(\theta)}{g(\theta)}.$$

This  $w$  is not equal to the ratio of the target divided by the design distribution,  $p(\theta | x)/g(\theta)$ , because it is missing a normalizing constant; that constant cancels out.

We can view Equation (13.26) as a ratio of two classical importance sampling estimates, or as a single ratio importance sampling estimate. In either case we draw  $\theta_1, \theta_2, \dots, \theta_N$  from  $g$ , and obtain the weighted empirical distribution with weights  $w_i / \sum_{j=1}^N w_j$ . The corresponding estimate of the mean of  $h(\theta)$  is

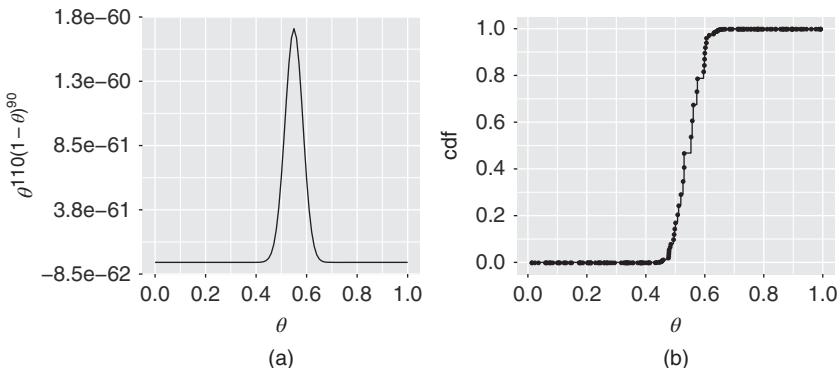
$$\hat{\mathbb{E}}[h(\theta) | x] = \bar{Y}/\bar{w},$$

where  $Y = hw$ .

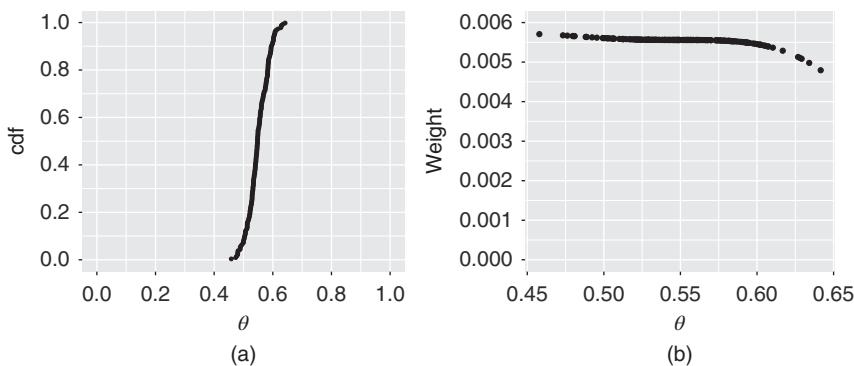
**Example 13.9** Return to Example 11.3 (Twitter account survey), to the “Bayesian A” analysis with a flat (uniform) prior,  $n = 200$  and  $x = 110$ . Suppose that we did not know about conjugate families and we use importance sampling with observations drawn from the uniform distribution. In this case, the prior  $\pi(\theta)$  and design distribution  $g$  are both uniform and the posterior and  $w$  are both proportional to the likelihood  $L(\theta) = \theta^{110}(1 - \theta)^{90}$ . The likelihood is quite close to 0 outside a narrow range around  $110/200$  (Figure 13.11a). Figure 13.11b shows the corresponding weighted ecdf. Most of the observations are almost wasted, with weights almost zero. We use a small  $N = 180$  in this example, so the jumps in the ecdf are visible, but in practice  $N$  would be much larger, say  $N = 10^6$ .

We should be able to find a better design distribution than the uniform, one that is shaped more like the posterior, resulting in more even weights. The shape of  $L$  in Figure 13.11a appears approximately normal, and the posterior has the same shape as  $L$  here. Furthermore, in practice, posterior distributions are often approximately normal when sample sizes are large, so we will consider a normal distribution. We will find the location and value of the maximum of  $L$  and its second derivative at the maximum. We will pick the normal mean to match the maximum and standard deviation so that  $g''/g = L''/L$  at the maximum. For the latter, it is easier to work with logs; we set  $\log''(g(\theta)) = \log''(L(\theta))$ . (In general, we would work with the product  $\pi(\theta)L(\theta)$  rather than  $L$ ; here they are the same.) This procedure of matching derivatives to choose the normal variance is the Laplace method (Carlin and Louis, 2009). The second derivative of the log of a normal density is  $-1/\sigma^2$ .

In this simple example, we can find the derivatives of  $\log(L)$  using calculus. The maximum occurs at  $\theta = 110/200$ , and the second derivative of the



**Figure 13.11** Importance sampling for the Twitter account example,  $n = 200$  and  $x = 110$ , where the prior and design distributions are uniform. (a) The likelihood  $L$ . (b) The importance sampling weighted ecdf.



**Figure 13.12** Weighted ecdf from importance sampling for the Twitter account example, using a normal design distribution. (a) The weighted cdf. (b) The corresponding weights.

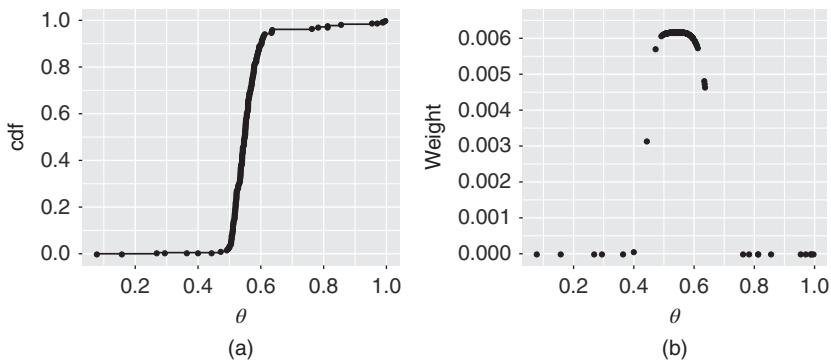
log-likelihood at this value is  $-200^2(1/110 + 1/90)$ . In practice, we usually need to find the maximum numerically, for example, using the `optimize` function in R, and then evaluate the second derivative numerically.

The resulting weighted ecdf is shown in Figure 13.12a. Compare this with Figure 13.11a – in the middle there are many more steps, and they are smaller. This succeeded in sampling more in the important region in the middle. Figure 13.12b shows a plot of the weights; it is always a good idea to look at the weights, especially to make sure that the weights are not blowing up in the tails of the distribution. This appears not to be the case, the weights are nearly constant.

The normal design distribution does sometimes produce observations outside the range of  $(0, 1)$ . This does not invalidate the analysis, but it is a waste of effort; any observation that occurs there get a weight of zero. In this example, there are so few of them that we would not worry about it. In general, it is alright if  $g(x) > 0$  where  $f(x) = 0$ , but not the reverse.

The normal design appears to work well here, but is often risky; the normal distribution drops to zero very quickly outside the  $\pm 3$  standard deviations; if the target distribution does not drop as quickly, there could be huge weights in the tails. In practice, people often prevent this using a translated scaled  $t$  distribution or using a mixture distribution. Here, we will demonstrate using a mixture of 90% from the same normal distribution as before and 10% from a uniform distribution.

The resulting weighted ecdf is shown in Figure 13.13a. Compared to the pure normal, this has more observations in the tails, each receiving small weight. Figure 13.13b shows the weights; here they are smaller in the tails, which tends to be safe.



**Figure 13.13** Weighted ecdf from importance sampling for the Twitter account example using a mixture distribution. (a) The weighted cdf. (b) The corresponding weights.

### R Note

These commands reproduce Figure 13.11.

```
likelihood <- function(theta) theta^110 * (1-theta)^90
logLikelihood <- function(theta) 110 * log(theta) +
    90 * log(1-theta)

N <- 180
theta <- sort(runif(N)) # Sampling from uniform dist.

ggplot(data.frame(x = c(0,1)), aes(x = x)) +
  stat_function(fun = likelihood)

w <- likelihood(theta)
weight <- w / sum(w)
cumsumwt <- cumsum(weight)
df <- data.frame(x = theta, y = cumsumwt, w = weight)

ggplot(df, aes(x = x, y = y)) + geom_step() + geom_point()
```

Using calculus, we find that the maximum of the log-likelihood occurs at  $\theta = 110/200$ . The second derivative here is  $-(110/\theta^2 + 90/(1-\theta)^2) = -200^2 \times (1/90 + 1/110) = -1/\sigma^2$ .

```
mu <- 110/200
sigma <- 1/(200 * sqrt(1/90 + 1/110))
```

To find the maximum and second derivative numerically in R:

```
thetaMax <- optimize(logLikelihood, interval = 0:1,
                      maximum = TRUE)$maximum
```

```

epsilon <- .001
(logLikelihood(thetaMax + epsilon) +
logLikelihood(thetaMax - epsilon)

- 2 * logLikelihood(thetaMax)) / epsilon^2 # 2nd derivative
-200^2*(1/110 + 1/90) # comparison, is very close

```

These commands reproduce Figure 13.12.

```

# Sampling from a normal distribution (no mixture)
theta2 <- sort(rnorm(N, mu, sigma))
w2 <- likelihood(theta2) / dnorm(theta2, mu, sigma)
weight2 <- w2 / sum(w2)
cumsumwt2 <- cumsum(weight2)
df2 <- data.frame(x = theta2, y = cumsumwt2, w = weight2)
ggplot(df2, aes(x = x, y = y)) + geom_step() + geom_point()
ggplot(df2, aes(x = x, y = w)) + geom_point()
# make sure weights don't blow up anywhere, esp. in tails

```

Here is code to estimate a probability, and standard error for the estimate:

```

y2 <- (theta2 > .6) * w2 # h(theta) = (theta > .6)
r2 <- mean(y2) / mean(w2) # ratio estimate for P(theta > .6)
r2
sd(y2 - r2 * w2) / (mean(w2) * sqrt(N)) # standard error

```

This reproduces Figure 13.13:

```

# Mixture of 90% normal and 10% uniform
theta3 <- c(rnorm(.9*N, mu, sigma), runif(.1*N))
out <- order(theta3)
theta3b <- theta3[out] # sorted
w3 <- likelihood(theta3) /
(0.9 * dnorm(theta3, mu, sigma) + 0.1)

weight3 <- (w3 / sum(w3))[out]
cumsumwt3 <- cumsum(weight3[out])

df3 <- data.frame(x = theta3b, y = cumsumwt3, w = weight3)
ggplot(df3, aes(x = x, y = y)) + geom_step() + geom_point()
# check if weights blow up
ggplot(df4, aes(x = x, y = w)) + geom_point()

```

To estimate a probability and standard error for the estimate based on stratified sampling:

```

y3 <- (theta3 >.6) * w3           # h(theta) = (theta >.6)
r3 <- mean(y3) / mean(w3)         # ratio estimate for P(theta >.6)
r3
sqrt(.9 * var((y3 - r3 * w3)[1:(.9*N)]) +
    .1 * var((y3 - r3 * w3)[(.9*N+1):N])) / (mean(w3) * sqrt(N))

```

□

## 13.8 The EM Algorithm

Maximum likelihood parameter estimates often do not have closed form solutions. The EM algorithm (*Expectation-Maximization*) is an iterative approach to estimating a parameter obtained by maximum likelihood. This approach is especially useful when there is missing data. We present an introduction to this algorithm by starting with a simple example from Dempster et al. (1977).

A model in genetics considers the problem of the distribution of  $n$  animals of five different types and with certain population probabilities. If we let  $X_i, i = 1, 2, \dots, 5$  denote the numbers of each type, we consider the multinomial distribution (see Appendix B.2) of  $n$  objects with five types:

$$(X_1, X_2, X_3, X_4, X_5) \sim \text{Multinom}\left(n, \frac{1}{2}, \frac{\pi}{4}, \frac{1-\pi}{4}, \frac{1-\pi}{4}, \frac{\pi}{4}\right),$$

where  $0 < \pi < 1$ ; the pdf is

$$f(x_1, x_2, x_3, x_4, x_5; \pi) = \binom{n}{x_1, x_2, x_3, x_4, x_5} \left(\frac{1}{2}\right)^{x_1} \left(\frac{\pi}{4}\right)^{x_2+x_3} \left(\frac{1-\pi}{4}\right)^{x_3+x_4} \cdot \quad (13.28)$$

We would like to estimate  $\pi$  using maximum likelihood, but suppose the data we observe are  $(y, x_3, x_4, x_5) = (125, 18, 20, 34)$  where  $y = x_1 + x_2$ ; we're missing the breakdown between  $x_1$  and  $x_2$ .

We could ignore the missing  $(x_1, x_2)$  and instead consider the data as coming from a multinomial distribution of  $n$  objects into four types, by combining the first two types together:

$$(Y, X_3, X_4, X_5) \sim \text{Multinom}\left(n, \frac{1}{2} + \frac{\pi}{4}, \frac{1-\pi}{4}, \frac{1-\pi}{4}, \frac{\pi}{4}\right).$$

The pdf for this distribution is

$$g(y, x_3, x_4, x_5; \pi) = \binom{n}{y, x_3, x_4, x_5} \left(\frac{1}{2} + \frac{\pi}{4}\right)^y \left(\frac{1-\pi}{4}\right)^{x_3+x_4} \left(\frac{\pi}{4}\right)^{x_5}. \quad (13.29)$$

To find the MLE of  $\pi$ , we take logarithms of both sides, differentiate with respect to  $\pi$  and set the result equal to 0. For example, with the given observation  $(y, x_3, x_4, x_5) = (125, 18, 20, 34)$  (we will use these numbers throughout this section), then  $\hat{\pi} = 0.628$  (see Exercise 13.16).

In this example, we are able to solve explicitly for the MLE, but that is often not possible – the equation(s) obtained by differentiating the log likelihood for the *incomplete data problem* may be intractable.

The *EM algorithm* is often useful in cases like this, where we if we had complete data we could calculate the *complete data* log likelihood, and find the value of  $\pi$  that maximizes that. Here, if we take the logarithm of Equation (13.28), differentiate with respect to  $\pi$  and set the result to 0, we get

$$\hat{\pi} = \frac{x_2 + x_5}{x_2 + x_3 + x_4 + x_5} = \frac{(y - x_1) + x_5}{(y - x_1) + x_3 + x_4 + x_5}, \quad (13.30)$$

in particular, for our observed data

$$\hat{\pi} = \frac{(125 - x_1) + 34}{(125 - x_1) + 72} = \frac{159 - x_1}{197 - x_1}.$$

Unfortunately, we do not know  $x_1$ . But given that  $X_1 + X_2 = 125 = y$ ,  $X_1$  is binomial (see Exercise 13.17)

$$X_1 \sim \text{Binom}\left(y, \frac{1/2}{1/2 + \pi/4}\right)$$

with expected value

$$\mathbb{E}[X_1] = y \frac{1/2}{1/2 + \pi/4} = \frac{2y}{2 + \pi}. \quad (13.31)$$

The idea behind the EM algorithm is to replace  $x_1$  in (13.30) with its *expected value* to obtain an estimate of  $\pi$  that *maximizes* the log likelihood. But wait, to compute the expected value in Equation (13.31), we need to know  $\pi$ ! This is a chicken-and-egg problem. We will just start with something, maybe a dinosaur, and eventually we get a chicken.

In EM, we start with an initial guess, then alternate between **E** (expected value) and **M** (maximum likelihood parameter estimation) steps. Say we start with  $\pi^{(0)}$ , then

$$\mathbf{E} : \quad x_1^{(k)} = \mathbb{E}[X_1] = \frac{2y}{2 + \pi^{(k-1)}}. \quad (13.32)$$

$$\begin{aligned} \mathbf{M} : \quad \hat{\pi}^{(k)} &= \frac{y - x_1^{(k)} + x_5}{(y - x_1^{(k)}) + x_3 + x_4 + x_5} \\ &= \frac{159 - x_1^{(k)}}{197 - x_1^{(k)}}. \end{aligned} \quad (13.33)$$

For instance, if we use an initial guess of  $\pi^{(0)} = 0.5$ , then  $x_1^{(1)} = 100$ . Then,  $\hat{\pi}^{(1)} = \frac{159-100}{197-100} = 0.6082$ . After five iterations,  $\hat{\pi}^{(k)}$  settles down to approximately 0.6268.

### 13.8.1 EM in General

Suppose we have data  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  from a distribution with pdf  $f(\mathbf{x}; \theta)$ . Suppose one or more of the observations  $x_i$  are missing. Let  $\mathbf{y} = (y_1, y_2, \dots, y_m)$  denote the observed values, which are typically a subset of the  $x$  values, or sometimes elements of  $y$  are combinations of the  $x$ 's. Let  $g(\mathbf{y}; \theta)$  denote the corresponding pdf. We call  $\mathbf{x}$  the complete data and  $\mathbf{y}$  the incomplete data. To estimate  $\theta$ , we may form the log likelihood (called the incomplete data log likelihood)

$$\ln(L(\theta | \mathbf{y})) = \ln(g(\mathbf{y}; \theta))$$

and try to maximize this with respect to the unknown parameter  $\theta$ . However, in many cases, it is often simpler to work with the complete data log likelihood

$$\ln(L(\theta | \mathbf{x})) = \ln(f(\mathbf{x}; \theta)).$$

The EM algorithm uses an iterative procedure on the complete data log likelihood to estimate  $\theta$ . Let  $\theta^{(k)}$  denote the estimate that maximizes the log likelihood  $f(\mathbf{x} | \theta)$  at the  $k$ th step. We let  $Q(\theta | \theta^{(k)})$  be the expected value of the log likelihood of the complete data conditioned on the observed data  $\mathbf{y}$ .

$$\begin{aligned} Q(\theta | \theta^{(k)}) &= E[\ln(L(\theta | \mathbf{x})) | \mathbf{y}, \theta^{(k)}] \\ &= E[\ln(f(\mathbf{x}; \theta)) | \mathbf{y}, \theta^{(k)}]. \end{aligned}$$

The EM algorithm alternates between two steps,

*E step:* Calculate  $Q(\theta | \theta^{(k)})$

*M step:* Maximize  $Q(\theta | \theta^{(k)})$  with respect to  $\theta$ . Let  $\theta^{(k+1)}$  denote the value that maximizes  $Q$ .

Return to the E step until you reach your stopping criterion, say

$$|\theta^{(k+1)} - \theta^{(k)}| < \epsilon \text{ for some pre-determined value of } \epsilon.$$

The key to the fact that the estimates  $\theta^{(k)}$  will converge to an MLE for the incomplete data MLE is given below.

**Theorem 13.2** Let  $\{\theta^{(k)}\}$  be a sequence of values obtained by the EM algorithm. Then

$$L(\theta^{(k+1)} | \mathbf{y}) \geq L(\theta^{(k)} | \mathbf{y}).$$

Under certain regularity conditions, Wu (1983) proved that the estimates converge to a stationary point of the likelihood. (See Dempster et al. (1977) or McLachlan and Krishnan (1997) for more information.) Since a stationary point might be a saddle point instead of a maximum, check your estimate by using different initial values.

**Example 13.10** We return to our previous example to clarify the notation. The incomplete data are  $(y, x_3, x_4, x_5)$  and the complete data are  $(x_1, x_2, x_3, x_4, x_5)$ , where  $x_1 + x_2 = y$ . The complete data log likelihood is (see Equation (13.28))

$$\ln(L(\pi \mid \mathbf{x})) = C + (y - x_1 + x_5) \ln(\pi) + (x_3 + x_4) \log(1 - \pi),$$

where  $C$  denotes a constant independent of  $\pi$ .

Thus, if  $\pi^{(k)}$  is the estimate of  $\pi$  at the  $k$ th stage of the algorithm, then we compute  $Q(\pi \mid \pi^{(k)})$ , the expected value of the above log likelihood conditioned on the observed data,  $\mathbf{y} = (125, 18, 20, 34)$ ,

$$\begin{aligned} Q(\pi \mid \pi^{(k)}) &= E[C + (y - X_1 + x_5) \ln(\pi) + (x_3 + x_4) \log(1 - \pi) \mid \mathbf{y}, \pi^{(k)}] \\ &= C + (y - E[X_1 \mid \mathbf{y}, \theta^{(k)}] + x_5) \ln(\pi) + (x_3 + x_4) \ln(1 - \pi) \\ &= C + \left(y - \frac{2y}{2 + \pi^{(k)}} + x_5\right) \ln(\pi) + (x_3 + x_4) \ln(1 - \pi) \\ &= C + \left(159 - \frac{250}{2 + \pi^{(k)}}\right) \ln(\pi) + 38 \ln(1 - \pi). \end{aligned}$$

We use  $X_1$  to denote the missing value since it is a random variable. Maximizing  $Q$  by taking the derivative with respect to  $\pi$  and setting this equal to 0 leads to the Equations (13.32) and (13.33).

**Remark** While in Equation (13.32) we calculated the expected value of a number, in general the E step calculates the expected value of a log-likelihood.  $\square$

**Example 13.11** Let  $X_1 \sim \text{Pois}(\lambda_1), X_2 \sim \text{Pois}(\lambda_2), X_3 \sim \text{Pois}(\beta\lambda_1), X_4 \sim \text{Pois}(\beta\lambda_2)$  be independent random variables and  $\lambda_1, \lambda_2, \beta > 0$  unknown parameters. The joint pdf is given by

$$f(\mathbf{X}; \lambda_1, \lambda_2, \beta) = \frac{\lambda_1^{X_1} \lambda_2^{X_2} \lambda_1^{X_3} \lambda_2^{X_4} \beta^{X_3 + X_4} e^{-(\lambda_1 + \lambda_2)(\beta + 1)}}{X_1! X_2! X_3! X_4!}.$$

Suppose we observe  $(X_2, X_3, X_4) = (3, 5, 7)$ ;  $x_1$  is missing. Let  $\mathbf{y} = (3, 5, 7)$  denote the incomplete data.

The log likelihood of the complete data is

$$\begin{aligned} \ln(L(\lambda_1, \lambda_2, \beta \mid \mathbf{x})) &= (x_1 + x_3) \ln(\lambda_1) + (x_2 + x_4) \ln(\lambda_2) \\ &\quad + (x_3 + x_4) \ln(\beta) - (\lambda_1 + \lambda_2)(\beta + 1) + C, \end{aligned}$$

where the constant  $C$  is independent of  $\lambda_1, \lambda_2, \beta$ .

We first compute the expected value of the log-likelihood given  $\mathbf{y} = (3, 5, 7)$  and the estimate  $(\lambda_1^{(k)}, \lambda_2^{(k)}, \beta^{(k)})$ :

$$\begin{aligned} Q\left(\lambda_1, \lambda_2, \beta \mid \lambda_1^{(k)}, \lambda_2^{(k)}, \beta^{(k)}\right) \\ = E\left[\ln(L(\lambda_1, \lambda_2, \beta \mid \mathbf{x})) \mid \mathbf{y}, \lambda_1^{(k)}, \lambda_2^{(k)}, \beta^{(k)}\right] \end{aligned}$$

$$\begin{aligned}
&= \left( E \left[ X_1 | \mathbf{y}, \lambda_1^{(k)}, \lambda_2^{(k)}, \beta^{(k)} \right] + 5 \right) \ln(\lambda_1) + 10 \ln(\lambda_2) \\
&\quad + 12 \ln(\beta) - (\lambda_1 + \lambda_2)(\beta + 1) \\
&= \left( \lambda_1^{(k)} + 5 \right) \ln(\lambda_1) + 10 \ln(\lambda_2) + 12 \ln(\beta) - (\lambda_1 + \lambda_2)(\beta + 1),
\end{aligned}$$

where  $E[X_1 | \mathbf{y}, \lambda_1^{(k)}, \lambda_2^{(k)}, \beta^{(k)}] = E[X_1 | \lambda_1^{(k)}] = \lambda_1^{(k)}$ , since the  $X_i$ 's are assumed to be independent. Take derivatives of  $Q$  with respect to  $\beta$ ,  $\lambda_1$  and  $\lambda_2$ , then set the results equal to 0 to obtain

$$\lambda_1^{(k+1)} = \frac{\lambda_1^{(k)} + 5}{\beta^{(k)} + 1}.$$

$$\lambda_2^{(k+1)} = \frac{10}{\beta^{(k)} + 1}.$$

$$\beta^{(k+1)} = \frac{12}{\lambda_1^{(k)} + 3}.$$

Alternately, we may do the estimates sequentially, using new estimates in later steps, say starting with  $\beta$ :

$$\beta^{(k+1)} = \frac{12}{\lambda_1^{(k)} + 3}$$

$$\lambda_1^{(k+1)} = \frac{\lambda_1^{(k)} + 5}{\beta^{(k+1)} + 1}$$

$$\lambda_2^{(k+1)} = \frac{10}{\beta^{(k+1)} + 1}.$$

In this case we only need one starting value, say  $\lambda_1^{(0)} = 1$ , and iterating until convergence gives  $\beta = 2.333$ ,  $\lambda_1 = 2.143$ ,  $\lambda_2 = 3.000$ .

### R Note

```

lambda1 <- 1      #initial value for lambda1
for (i in 1:20)
{
  beta <- 12/(lambda1 + 3)
  lambda1 <- (lambda1 + 5)/(beta + 1)
  lambda2 <- 10/(beta + 1)
}
beta
lambda1
lambda2

```

□

**Remark** Often, the difficult part of the EM algorithm is determining the conditional distribution of the complete data given the observed data. ||

## Exercises

- 13.1 In Section 3.3, we performed a permutation test to determine if men and women consumed, on average, different amounts of hot wings.
  - (a) Do smoothed bootstraps for the mean consumption for men and the mean consumption for women. How do the distributions differ?
  - (b) Do a smoothed bootstrap for the difference in means.
  - (c) Do an ordinary bootstrap for the difference in means.
  - (d) How do the ordinary and smoothed bootstrap distributions differ?
- 13.2 Perform a smoothed bootstrap for the mean arsenic level from the Bangladesh data set (Example 5.3). Use a transformation to prevent the smooth from creating negative values.
- 13.3 Refer to Exercise 6.17 where, given the data, we used the method of moments to estimate  $\lambda$  and  $r$  for the Gamma distribution. It turns out that  $\hat{r} = 22.88$  and  $\hat{\lambda} = 3.138$ . Now perform a parametric bootstrap for the mean.
- 13.4 Refer to Exercise 6.19 where we modeled the time between successive earthquakes using the Weibull distribution. From the given data, we estimate  $k = 0.917$  and  $\lambda = 17.344$ . Do a parametric bootstrap for the mean time between earthquakes, and for the average number of earthquakes per year.
- 13.5 Show that when treatment and control models are fit by linear regression that average treatment effect in Equation (13.11) is given by Equation (13.10). *Hint:* Use Equation (9.5) for  $a$ .
- 13.6 Show that using a factor variable for control variates using multiple linear regression is equivalent to post-stratification, Equation (13.7). *Hint:* For the multiple regression, instead of using an intercept and  $J - 1$  dummy variables, where  $J$  is the number of levels of the factor, use  $J$  dummy variables, one for each factor level. Then the regression coefficients are  $\bar{Y}_k$ ,  $k = 1 \dots J$ . The observed and theoretical means for the dummy variables are proportions.

- 13.7** Estimate  $\int_0^1 \cos(x^2)dx$  by the following:
- using Monte Carlo integration.
  - using the `integrate` function in R.
  - expanding  $\cos(x^2)$  as a power series out to degree 12. What is the error?
- Hint:* This is an alternating series!
- using a computer algebra system.
- 13.8** Consider  $4 \int_0^1 1/(1+x^2)dx$ .
- Use calculus to evaluate the integral.
  - Use Monte Carlo integration to find an estimate for  $\pi$ .
- 13.9** Let  $X_1, X_2, \dots, X_N$  be i.i.d. from a distribution with density  $f$  and let  $h$  denote a real-valued function. Verify that Equation (13.15) holds. (Pick  $f$  and  $h$ , and verify using simulation that Equation (13.15) approximately equals Equation (13.14).)
- 13.10** Estimate  $\int_{-\infty}^{\infty} e^{-x^4}/2 dx$  using the design distribution  $g(x) = (1/\sqrt{2\pi})e^{-x^2/2}$ , the pdf for the standard normal distribution.
- 13.11** A computer algebra system gives the estimate  $\int_0^\infty \ln(1+x)e^{-x} dx = 0.596347$ .
- Estimate the integral and a confidence interval for the estimate using the pdf of the chi-square distribution with 3 degrees of freedom as the design distribution. Use the `rchisq` and `dchisq` functions to obtain random values and density values, respectively, for the chi-square distribution.
  - Compute an estimate of the integral and a confidence interval for the estimate, using the pdf of the exponential distribution with  $\lambda = 1/2$ . Use the `rexp` and `dexp` functions to obtain random values and density values, respectively, for the exponential distribution.
  - Write the integral in the form  $\int_0^\infty h(x)f(x)dx$ ; graph both  $h$  functions on one set of axes over the range (0, 10).
  - Which of the two  $f$  gives more accurate estimates. Explain why based on the graph.
- 13.12** In importance sampling,  $(1/N) \sum_{i=1}^N w(X_i)$  is in general not equal to 1. However, it may be equal to 1 on average.

- (a) Show that if  $g(x) > 0$  whenever  $f(x) > 0$ , then  $E\left[\frac{1}{N} \sum_{i=1}^N w(X_i)\right] = 1$ .  
 (b) In one example in this chapter this condition was not met; discuss what happened to  $(1/N) \sum_{i=1}^N w(X_i)$  in that example.

- 13.13** For estimating a probability  $P(A) = \int_A f(x) dx = \int h(x)f(x)dx$  with  $h(x) = 1$  if  $x \in A$  and 0 otherwise, show that if  $g(x) \propto |h(x) - P(A)| f(x)$ , then  $g$  has 50% of its probability on each of  $A$  and its complement.

This  $g$  is  $g^*_{\text{ratio}}$ , the optimal design for the ratio estimate; this implies that the optimal design samples equally from successes and failures.

- 13.14** Instacart (IC) personnel evaluate the display ads manually. IC plans to check the quality of these evaluations by randomly sampling some past evaluations and double-checking them. Let  $Q_i$  measures the quality of ad  $i$  that is viewed  $X_i$  times (impressions), and  $N$  is the number of ads to date. There are two metrics of interest: the equally-weighted average  $\mu_Q = N^{-1} \sum_{i=1}^n Q_i$ , and the impression-weighted average  $\mu_{QX} = \sum_{i=1}^n X_i Q_i / \sum_{i=1}^n X_i$ . Perform a simulation to estimate the accuracy of three importance sampling designs:

- (a) sample with equal probabilities,  
 (b) sample with probabilities proportional to impressions, and  
 (c) sample 50% with equal probabilities and 50% proportional to impressions.

Try each design for estimating both metrics. What would you recommend, and why?

In the simulation, let the population size be 3000, the sample size 200, let  $Q$  be uniformly distributed over a five-point scale (1, ..., 5), and let  $X$  have a Poisson distribution with mean 10. Generate  $Q$  and  $X$  once and use them throughout the simulation.

(In practice IC will estimate the difference in old and new ratings  $Q^{(\text{new})} - Q^{(\text{old})}$ , but that distinction does not matter much for the sampling design.)

- 13.15** When a customer shops at IC for “coffee,” IC offers choices such as brands of coffee or related products such as “creamer.” IC uses an external vendor to perform human evaluation to evaluate the quality of the query-product pairs. The vendor hires gig workers to do the ratings. IC is considering switching ratings frameworks, from a five-point scale based on relevance (e.g. “strongly relevant,” ..., “not relevant”) to choices “exact” (strongly relevant), “substitute” (can be used as a replacement), “complement” (can be used with a product, e.g. coffee and creamer), and “irrelevant.” (IC) will select a sample of past ratings to have the vendor re-evaluate using the new framework,

to check how old and new terms match up. They want a sample of size 500 that is representative of past ratings, including relevance level and type of query (for a specific brand, kind of food, a category, etc. [six possibilities]).

- (a) How would you perform this sample?
- (b) What are the advantages and disadvantages of that approach? What could go wrong?
- (c) In practice you would have the data to work with; what possible problems would you check for, and how would you deal with them?

This code generates artificial data for you to work with, with some characteristics of the real data (which is confidential):

```
set.seed(42)
N <- 30000
relevance <- sample(1:5, N, replace = TRUE,
                     prob = c(2, 1, 2, 3, 5)/13)
Z <- rnorm(N) + (relevance-3)/3
category <- cut(Z, c(-Inf, -2, -1, 0, 1, 2, Inf),
                  labels = paste0("cat", 1:6))
```

- 13.16** Verify that the MLE of  $\pi$  for the data with pdf given by Equation (13.29) is 0.628.
- 13.17** Let  $X_1, X_2, X_3, X_4, X_5 \stackrel{\text{i.i.d.}}{\sim} \text{Multinom}(n, p_1, p_2, p_3, p_4, p_5)$  where  $\sum_i p_i = 1$ . Show that  $X_1$  given that  $X_1 + X_2 = s$  is binomial,  $\text{Binom}(s, p_1/(p_1 + p_2))$ .
- 13.18** Let  $X_1, X_2, X_3, X_4, X_5 \sim \text{Exp}(\lambda)$  and  $X_6, X_7 \sim \text{Exp}(\beta\lambda)$  be independent random variables with  $\beta, \lambda > 0$  unknown parameters. Suppose we observe  $\mathbf{x} = (x_1, 5, 2, 2.5, 4, 0.25, 0.75)$ , where  $x_1$  is missing. Use the EM algorithm to find estimates for  $\beta$  and  $\lambda$ . Check your answer by finding the MLE estimates for the incomplete data problem.



## Appendix A

### Review of Probability

This section contains a brief review of some definitions and results from probability that are used in this textbook. Please consult a probability textbook for more in-depth information, for example, Ghahramani (2004), Pitman (1993), Ross (2009), or Scheaffer and Young (2010).

#### A.1 Basic Probability

Recall that the set of all possible outcomes of a random experiment is called a *sample space*,  $S$ . An event  $E$  is a subset of  $S$ .

**Proposition A.1 (Law of Total Probability)** Let  $A$  denote an event in a sample space  $S$  and let  $B_1, B_2, \dots, B_n$  be a disjoint partition of  $A$ . Then

$$P(A) = P(B_1)P(A | B_1) + P(B_2)P(A | B_2) + \cdots + P(B_n)P(A | B_n).$$

**Definition A.1** A *discrete random variable*  $X$  is a function from  $S$  into the real numbers  $\mathbf{R}$  with a range that is finite or countably infinite. That is,  $X : S \rightarrow \{x_1, x_2, \dots, x_m\}$ , or  $X : S \rightarrow \{x_1, x_2, \dots\}$ . ||

For instance, if we consider the experiment of rolling two dice, we can let  $X$  denote the sum of the two numbers that appear. Then  $X$  is a discrete random variable,  $X : S \rightarrow \{2, 3, \dots, 12\}$ .

The *probability mass function* (pmf) is a function  $p : \mathbf{R} \rightarrow [0, 1]$  such that  $p(x) = P(X = x)$ , for all  $x$  in the range of  $X$ . Note then that  $\sum_x p(x) = 1$ , where the sum is over the range of  $X$ . (We may sometimes also call this a density, or use density to refer to both discrete and continuous versions.)

**Definition A.2** A function  $X$  from  $S$  into the real numbers  $\mathbf{R}$  is a *continuous random variable* if there exists a non-negative function  $f$  such that for every

subset  $C$  of  $\mathbf{R}$ ,  $P(X \in C) = \int_C f(x)dx$ . In particular, for  $a \leq b$ ,  $P(a < X \leq b) = \int_a^b f(x)dx$ . ||

The function  $f$  is called the *probability density function* (pdf) of  $X$ . Note that  $\int_{-\infty}^{\infty} f(x)dx = 1$ . The (*cumulative*) *distribution function*  $F$  of a random variable  $X$  is the function  $F: \mathbf{R} \rightarrow [0, 1]$  that satisfies

$$F(x) = P(X \leq x), -\infty < x < \infty.$$

$F$  is a nondecreasing, right-continuous function with  $\lim_{x \rightarrow -\infty} F(x) = 0$  and  $\lim_{x \rightarrow \infty} F(x) = 1$ . In the case that  $X$  is a continuous random variable, then

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t)dt$$

and thus at every point  $x$  at which  $f(x)$  is continuous,

$$F'(x) = f(x),$$

by the fundamental theorem of calculus.

**Example A.1** Recall that  $X$  is an exponential random variable with  $\lambda > 0$  if its pdf is  $f(x) = \lambda e^{-\lambda x}$ ,  $x \geq 0$ . Then the distribution function is

$$F(x) = P(X \leq x) = \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x}.$$

We also have  $\lim_{x \rightarrow -\infty} F(x) = 0$  and  $\lim_{x \rightarrow \infty} F(x) = 1$  □

## A.2 Mean and Variance

**Definition A.3** Let  $X: S \rightarrow \mathbf{R}$  denote a random variable and  $f$  denote its density function. The *mean* of  $X$ , also known as the *expected value* of  $X$ , is

$$\mathbb{E}[X] = \mu = \int_{-\infty}^{\infty} xf(x)dx$$

and the *variance* is

$$\text{Var}[X] = \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx. ||$$

The mean and variance may not exist if the integrals are not finite.

**Definition A.4** The *standard deviation* of  $X$  is  $\text{SD}[X] = \sigma = \sqrt{\text{Var}[X]}$ . ||

If  $X: S \rightarrow \{x_1, x_2, \dots\}$  is a discrete random variable with  $P(X = x_i) = p_i$ , then  $E[X] = \mu = \sum_i x_i \cdot p_i$  and  $\text{Var}[X] = \sum_i (x_i - \mu)^2 \cdot p_i$ .

**Example A.2** Suppose  $X$  is a finite discrete random variable,  $X: S \rightarrow \{a_1, a_2, \dots, a_n\}$ , with  $p(a_i) = 1/n$ . That is, each outcome is equally likely. Then

$$E[X] = \mu = \frac{1}{n} \sum_{i=1}^n a_i, \quad (\text{A.1})$$

the “usual” average, and

$$\text{Var}[X] = E[(X - \mu)^2] = \sum_{i=1}^n (a_i - \mu)^2 \frac{1}{n} = \frac{1}{n} \sum_{i=1}^n (a_i - \mu)^2. \quad (\text{A.2}) \quad \square$$

**Proposition A.2**  $\text{Var}[X] = E[(X - \mu)^2] = E[X^2] - \mu^2$ .

**Theorem A.1** If  $X$  is a random variable with density  $f$  and  $g$  is any real-valued function, then

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

We state a special case of Theorem A.1 as its own theorem:

**Theorem A.2** If  $X$  is a random variable and  $a$  and  $b$  are constants, then  $E[a + bX] = a + bE[X]$  and  $\text{Var}[a + bX] = b^2\text{Var}[X]$ .

**Definition A.5** The random variables  $X$  and  $Y$  have a *joint density* if there exists a non-negative function  $f: \mathbf{R} \times \mathbf{R} \rightarrow \mathbf{R}$  such that for every subset  $C$  of the plane,

$$P((X, Y) \in C) = \iint_C f(x, y) dx dy.$$

||

**Definition A.6** Let  $X$  and  $Y$  be random variables. Then  $X$  and  $Y$  are *independent* if for any sets  $A$  and  $B$ ,

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B).$$

||

**Theorem A.3** If  $X$  and  $Y$  are random variables, then  $E[X + Y] = E[X] + E[Y]$ .

**Theorem A.4** If  $X$  and  $Y$  are independent, then  $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]$ .

Note that Theorem A.3 is true regardless of whether or not  $X$  and  $Y$  are independent, but Theorem A.4 is not. Failure to properly allow for the latter

point is a common error, which was a prime contributor to the global financial meltdown of 2008. Financial institutions calculated the risk of their portfolios based on variability calculations without allowing for the dependence between different assets. Variables that appear on the surface to be independent may in fact be dependent because of *lurking variables*, variables that affect both. For example, if  $X$  and  $Y$  are the amounts that two people default on their mortgages, then  $X$  and  $Y$  are dependent even if the people live in different states, because of economic conditions that affect both.

The combination of Theorems A.2 and A.3 or Theorem A.4 is very useful for working with two or more random variables. For example, for two variables,  $E[aX + bY] = aE[X] + bE[Y]$ , and for two independent variables,  $\text{Var}[aX + bY] = a^2\text{Var}[X] + b^2\text{Var}[Y]$ . For more variables, of central importance is the case of a mean of a set of variables.

### A.3 Marginal and Conditional Distributions

**Definition A.7** Let  $X$  and  $Y$  have joint density function  $f$ . The *marginal densities* of  $X$  and of  $Y$  are given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy \quad f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx,$$

respectively.

In the case of discrete random variables, replace the integral with a sum. ||

**Example A.3** Let  $X$  and  $Y$  have the joint density function

$$f(x, y) = e^{-y}, \quad 0 < x < y < \infty.$$

The marginal density of  $X$  is

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy = \int_x^{\infty} e^{-y} dy = e^{-x},$$

for  $x > 0$ . In particular, we see that  $X$  has an exponential distribution with  $\lambda = 1$ .

The marginal density of  $Y$  is

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx = \int_0^y e^{-y} dy = ye^{-y},$$

for  $y > 0$ . □

**Definition A.8** Let  $X$  and  $Y$  be jointly continuous random variables with joint density  $f$ . The *conditional density* of  $X$  given  $Y = y$  is

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)}, \tag{A.3}$$

for  $f_Y(y) > 0$ .

Similarly, the *conditional density* of  $Y$  given  $X = x$  is

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}, \quad (\text{A.4})$$

for  $f_X(x) > 0$ . ||

**Example A.4** Refer to the joint distribution in Example A.3. The conditional density of  $X$  given  $Y = y$  is

$$f_{X|Y}(x|y) = \frac{e^{-y}}{ye^{-y}} = \frac{1}{y},$$

for  $0 < x < y$ , the uniform distribution on  $(0, y)$ . □

## A.4 The Normal Distribution

The random variable  $X$  has a normal distribution with parameters  $\mu$  and  $\sigma$  if its pdf is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}.$$

We denote this by  $X \sim N(\mu, \sigma^2)$ .

**Theorem A.5** Let  $X \sim N(\mu, \sigma^2)$ . Then  $E[X] = \mu$  and  $\text{Var}[X] = \sigma^2$ .

**Theorem A.6** Let  $X \sim N(\mu, \sigma^2)$  and define  $Z = (X - \mu)/\sigma$ . Then  $Z \sim N(0, 1)$ .

Subtracting the mean and dividing by the standard deviation is called *standardization* (for any random variable, not just normal). For normal random variables, it is a key step in performing probability calculations.

$$\begin{aligned} P(X \leq x) &= \int_{-\infty}^x f(t)dt \\ &= P(Z \leq (x - \mu)/\sigma) \\ &= \Phi((x - \mu)/\sigma), \end{aligned} \quad (\text{A.5})$$

where

$$\Phi(z) = P(Z \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-t^2/2} dt \quad (\text{A.6})$$

is the cumulative distribution function (cdf) for the standard normal distribution.

**Remark** Recall the 68–95–99.7 rule: For a standard normal random variable  $Z \sim N(0, 1)$ ,

$$P(-1 < Z < 1) \approx 0.68.$$

$$P(-2 < Z < 2) \approx 0.95.$$

$$P(-3 < Z < 3) \approx 0.997.$$

||

## A.5 The Mean of a Sample of Random Variables

**Definition A.9** Let  $X_1, X_2, \dots, X_n$  be random variables. Then

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad (\text{A.7})$$

is called the (*sample*) *mean* of the  $X_1, X_2, \dots, X_n$ .

||

**Theorem A.7** Let  $X_1, X_2, \dots, X_n$  be identically distributed random variables with mean  $\mu$  and variance  $\sigma^2$ .

Then  $E[\bar{X}] = \mu$ .

If, in addition,  $X_1, X_2, \dots, X_n$  are independent, then  $\text{Var}[\bar{X}] = \sigma^2/n$

*Proof.*

$$\begin{aligned} E[\bar{X}] &= E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] \\ &= \frac{1}{n} \sum_{i=1}^n E[X_i] \\ &= \frac{1}{n} \sum_{i=1}^n \mu \\ &= \mu \end{aligned}$$

and

$$\begin{aligned} \text{Var}[\bar{X}] &= \text{Var}\left[\frac{1}{n} \sum_{i=1}^n X_i\right] \\ &= \frac{1}{n^2} \sum_{i=1}^n \text{Var}[X_i] \\ &= \frac{1}{n^2} n \sigma^2 \\ &= \frac{\sigma^2}{n}. \end{aligned}$$

□

**Example A.5** Let  $X_1, X_2, \dots, X_{20}$  be independent random variables, each normally distributed,  $N(3, 4^2)$ . Then  $E[\bar{X}] = 3$  and  $\text{Var}[\bar{X}] = 4^2/20$ .

□

Consider  $X_1, X_2, \dots, X_n$  independent Bernoulli random variables, each satisfying  $X_i = 1$  with probability  $p$ . That is,  $E[X_i] = p$  and  $\text{Var}[X_i] = p(1 - p)$  for  $i = 1, 2, \dots, n$ . Then  $\bar{X} = (1/n) \sum_{i=1}^n X_i = \# \text{ of } 1\text{'s}/n = \text{proportion of } 1\text{'s}$  (sample proportion). Then  $E[\bar{X}] = p$  and  $\text{Var}[\bar{X}] = p(1 - p)/n$ . We will use the notation  $\hat{p}$  to denote the sample proportion (i.e.  $\hat{p} = \bar{X}$  in the case of Bernoulli random variables.)

We summarize this:

**Corollary A.1** Let  $X_1, X_2, \dots, X_n$  be independent Bernoulli random variables with  $E[X_i] = p$  and  $\text{Var}[X_i] = p(1 - p)$  for  $i = 1, 2, \dots, n$ . Then  $E[\hat{p}] = p$  and  $\text{Var}[\hat{p}] = p(1 - p)/n$ .

In this book, we will work with independent random variables drawn from a common distribution.

**Definition A.10** Let  $X_1, X_2, \dots, X_n$  be independent random variables drawn from a common distribution. We say that the random variables are *independent and identically distributed* or i.i.d.. Let  $F$  and  $f$  denote the cdf and pdf, respectively, of the common distribution. Then we may also denote this by

$$X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} F \quad \text{or} \quad X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} f.$$

||

## A.6 Sums of Normal Random Variables

**Theorem A.8** Let  $X$  be a normal random variable with mean  $\mu_1$  and variance  $\sigma_1^2$ , and let  $Y$  be a normal random variable with mean  $\mu_2$  and variance  $\sigma_2^2$ . Assume  $X$  and  $Y$  are independent. Then  $X \pm Y$  is a normal random variable with mean  $\mu_1 \pm \mu_2$  and variance  $\sigma_1^2 + \sigma_2^2$ .

### Example A.6

In the town of Sodor, the weights of boys are normally distributed,  $N(100, 5^2)$ , while the weights of girls are normally distributed,  $N(90, 6^2)$  (in pounds). If a boy and a girl are selected independently and at random from the population, what is the probability that the boy will weigh at least 6 lb more than the girl?

### Solution

Let  $X$  and  $Y$  denote the weights of the selected boy and girl, respectively. We want  $P(X \geq Y + 6) = P(X - Y \geq 6)$ . By the theorem,  $X - Y$  is normally distributed with mean  $100 - 90 = 10$  and variance  $5^2 + 6^2 = 61$ . So,

$$\begin{aligned} P(X - Y \geq 6) &= P\left(\frac{X - Y - 10}{\sqrt{61}} \geq \frac{6 - 10}{\sqrt{61}}\right) \\ &= P(Z \geq -0.5121) \approx 0.6957. \end{aligned}$$

□

More generally,

**Theorem A.9** Let  $X_1, X_2, \dots, X_n$  be independent normal random variables with mean  $\mu_i$  and variance  $\sigma_i^2$ ,  $i = 1, 2, \dots, n$ , respectively. Let  $a_1, a_2, \dots, a_n$  be arbitrary constants. Then  $a_1X_1 + a_2X_2 + \dots + a_nX_n$  is a normal random variable with mean  $a_1\mu_1 + a_2\mu_2 + \dots + a_n\mu_n$  and variance  $a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \dots + a_n^2\sigma_n^2$ .

**Corollary A.2** Let  $X_1, X_2, \dots, X_n$  be independent normal random variables with common mean  $\mu$  and common variance  $\sigma^2$ . Let  $\bar{X}$  denote the sample mean. Then  $\bar{X}$  is normally distributed with mean  $\mu$  and variance  $(\sigma^2/n)$ .

### Example A.7

The amount of coffee a machine dispenses is normally distributed with mean 8 oz and variance 0.47 oz. If 10 cups of coffee dispensed from this machine are chosen at random, what is the probability that the average amount of coffee is more than 8.5 oz?

### Solution

The amount of coffee in one cup is a random variable with mean  $\mu = 8$  and standard deviation  $\sigma = \sqrt{0.47}$ . If  $\bar{X}$  denotes the mean of 10 cups of coffee, then  $\bar{X}$  is distributed normally with mean 8 oz and standard deviation  $\sqrt{0.47}/\sqrt{10} \approx 0.2168$ . Thus,

$$\begin{aligned} P(\bar{X} > 8.5) &= P\left(\frac{\bar{X} - 8}{0.2168} > \frac{8.5 - 8}{0.2168}\right) \\ &\approx P(Z > 2.3062) \\ &= 1 - P(Z \leq 2.3062) = 0.0105. \end{aligned}$$

□

## A.7 The Law of Averages

You may be familiar with the “law of averages,” the idea that in the long run, the sample average (or sample proportion) gets closer and closer to the true mean (or proportion). The more rigorous version of this is the *strong law of large numbers*:

**Theorem A.10 Strong Law of Large Numbers (SLLN)** Let  $X_1, X_2, \dots$  be independent and identically distributed random variables with  $E[X_i] = \mu < \infty$ . Then

$$P\left(\lim_{N \rightarrow \infty} \frac{X_1 + X_2 + \dots + X_N}{N} = \mu\right) = 1.$$

The SLLN is usually proved in a typical probability course (Casella and Berger, 2001; Ghahramani, 2004; Ross, 2009).

The SLLN works by *swamping*, not *compensation*. If you are flipping a fair coin and get 80 heads in the first 100 flips, then over the next 100 flips, the expected number of heads is 50 – not 20. As you continue to flip more coins – say  $10^8$  more – the extra 30 heads become insignificant. The remarkable thing about the SLLN is that it works even if the variance of the distribution is infinite. On the other hand, the SLLN does not say anything about how close the sample mean is to the true mean. If the variance is finite, then we can estimate how close using Chebyshev's inequality, or a combination of the central limit theorem (Section 4.3) and normal distributions.

**Theorem A.11 Chebyshev's Inequality** Let  $X$  be a random variable with finite mean  $\mu$  and variance  $\sigma^2$ . Then for any  $k > 0$ ,

$$P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}.$$

When  $X_1, X_2, \dots$ , are independent and identically distributed random variables with  $E[X_i] = \mu$  and  $\text{Var}[X_i] = \sigma^2 < \infty$ , then  $E[\bar{X}] = \mu$  and  $\text{Var}[\bar{X}] = \sigma^2/n$ , and by Chebyshev's inequality

$$P(|\bar{X} - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}.$$

This is typically quite conservative; a normal approximation is typically much more accurate.

## A.8 Higher Moments and the Moment Generating Function

The material in this section is used in Sections 2.7, B.9, and B.10. The mean and standard deviation of a normal distribution describe its *center* and *spread*. Every normal distribution has the same basic shape, so the mean and standard deviation completely determine a normal distribution. To describe non-normal variables, it is also useful to describe how asymmetrical the variable is, and how peaked. All four characteristics, center, spread, asymmetry, and peakedness, are based on *moments* of a variable.

**Definition A.11** Let  $X$  be a random variable with mean  $\mu$ . For a positive integer  $k$ , the  $k$ th moment of  $X$  is

$$\mu'_k = E[X^k] \tag{A.8}$$

and the  $k$ th central moment of  $X$  is

$$\mu_k = E[(X - \mu)^k], \tag{A.9}$$

provided  $\int_{-\infty}^{\infty} |x|^k f(x) dx < \infty$ , where  $f$  denotes the pdf of  $X$ .

||

Thus, the mean of  $X$  is its first moment, and the variance of  $X$  is its second central moment.

In many situations it is useful to characterize distributions using *moment generating functions*.

**Definition A.12** Let  $X$  be a random variable with cdf  $F$ . The *moment generating function (mfg)* of  $X$  is defined by

$$M(t) = E[e^{tX}],$$

provided the expected value exists in a neighborhood of 0. ||

**Example A.8** Let  $X \sim \text{Binom}(n, p)$ . Then

$$\begin{aligned} M(t) &= E[e^{tX}] \\ &= \sum_{x=0}^n e^{tx} \binom{n}{x} p^x (1-p)^{n-x} \\ &= \sum_{x=0}^n \binom{n}{x} (e^t p)^x (1-p)^{n-x} \\ &= (e^t p + 1 - p)^n, \end{aligned}$$

which holds for all  $t$ . □

**Example A.9** Let  $X \sim \text{Exp}(\lambda)$ , with  $\lambda > 0$ . Then

$$\begin{aligned} M(t) &= E[e^{tX}] \\ &= \int_0^\infty e^{tx} \lambda e^{-\lambda x} dx \\ &= \int_0^\infty \lambda e^{(t-\lambda)x} dx \\ &= \int_0^\infty \lambda e^{-u} \frac{1}{\lambda-t} du && \text{where } u = (\lambda-t)x \\ &= \frac{\lambda}{\lambda-t}, \end{aligned}$$

where  $t < \lambda$ . □

We now state the result that explains the name given to  $M(t)$ .

**Theorem A.12** Let  $X$  be a random variable with mgf  $M(t)$ . Then

$$\frac{d^n}{dt^n} M(t)|_{t=0} = E[X^n].$$

That is, the  $n$ th moment of  $X$  is the  $n$ th derivative of  $M(t)$  evaluated at  $t = 0$ .

**Example A.10** For  $X \sim \text{Binom}(n, p)$ , we have  $M(t) = (e^t p + 1 - p)^n$ . Thus,  $M'(t) = n(e^t p + 1 - p)^{n-1} e^t p$ , so  $M'(0) = np = E[X]$ .  $\square$

The following theorem allows us to use moment generating functions to characterize distributions.

**Theorem A.13** Let  $X$  and  $Y$  be random variables with corresponding cdf's  $F_X$  and  $F_Y$ , respectively. Suppose all the moments of  $X$  and  $Y$  exist. If the moment generating functions for  $X$  and  $Y$  exist and  $M_X(t) = M_Y(t)$  for all  $t$  in some neighborhood about 0, then  $F_X(s) = F_Y(s)$  for all  $s$ .

Moment generating functions for the sum of independent random variables can be found easily.

**Theorem A.14** Let  $X_1, X_2, \dots, X_n$  be independent random variables with moment generating functions  $M_{X_1}(t), M_{X_2}(t), \dots, M_{X_n}(t)$ , respectively. Then

$$M_{X_1+X_2+\dots+X_n}(t) = M_{X_1}(t)M_{X_2}(t)\dots M_{X_n}(t).$$



## Appendix B

### Probability Distributions

In this section, we gather results about some special probability distributions. Many of these are typically covered in a first probability course, so the presentations for them are brief. A document with information on computing probabilities and quantiles in R is provided at <https://github.com/lchihara/MathStatsResamplingR>.

#### B.1 The Bernoulli and Binomial Distributions

Consider a random experiment that has only two outcomes, such as tossing a coin and recording “heads” or “tails.” This experiment is called a *Bernoulli trial*. We define a random variable  $X$  on the set of these two outcomes, letting  $X$  have the value 0 with probability  $p$  for one outcome, and value 1 with probability  $1 - p$  for the other outcome. Then  $X$  is called a *Bernoulli* random variable. For instance, if we roll a die and consider seeing a 1 a “success” and anything else a “failure,” we can let  $X = 1$  for a success with probability  $p = 1/6$  and  $X = 0$  for a failure with probability  $1 - 1/6 = 5/6$ .

We denote a random variable  $X$  with probability  $p$  of success by  $X \sim \text{Bern}(p)$ .

$$\text{E}[X] = p, \quad \text{Var}[X] = p(1 - p). \quad (\text{B.1})$$

Now, let  $X_1, X_2, \dots, X_n$  be  $n$  independent Bernoulli random variables each with probability  $p$  of success, and let  $Y$  denote the number of successes. Then  $Y$  is a *binomial* random variable denoted  $Y \sim \text{Binom}(n, p)$  with probability mass function

$$f(y) = P(Y = y) = \binom{n}{y} p^y (1 - p)^{n-y}, \quad y = 0, 1, \dots, n. \quad (\text{B.2})$$

For instance, if we roll a die 10 times and consider seeing a 1 as a success, then  $Y \sim \text{Binom}(10, 1/6)$ . So the probability of seeing four 1’s is  $f(4) = \binom{10}{4} (1/6)^4 (5/6)^6 \approx 0.0543$ .

**Theorem B.1** Let  $Y \sim \text{Binom}(n, p)$ . Then

$$\mathbb{E}[Y] = np, \quad \text{Var}[Y] = np(1 - p). \quad (\text{B.3})$$

## B.2 The Multinomial Distribution

We saw that the binomial distribution can be used to model the number of heads in  $n$  tosses of a coin. Each coin toss has only two possible outcomes, heads or tails. What if an action has more than two outcomes, such as the tossing of a die? How might we model the distribution of the different possible outcomes in  $n$  tosses of the die?

First, we recall the multinomial coefficient. Suppose we have  $n$  objects of which  $x_1$  are of type  $O_1$ ,  $x_2$  are of type  $O_2$ , ...,  $x_r$  are of type  $O_r$ . Then the number of ways to arrange these  $n = x_1 + x_2 + \dots + x_r$  objects is

$$\binom{n}{x_1, x_2, \dots, x_r} = \frac{n!}{x_1! x_2! \dots x_r!}. \quad (\text{B.4})$$

To see why this is so, consider placing the objects in a row in positions 1, 2, ...,  $n$ . Then there are  $\binom{n}{x_1}$  ways to choose the  $x_1$  positions where the objects of type  $O_1$  can be placed. Now, there are  $n - x_1$  positions remaining, so there are  $\binom{n-x_1}{x_2}$  ways to choose the positions where the objects of type  $O_2$  can be placed. Continuing in this fashion, we have

$$\begin{aligned} & \binom{n}{x_1} \binom{n-x_1}{x_2} \binom{n-x_1-x_2}{x_3} \dots \binom{n-x_1-x_2-\dots-x_{r-1}}{x_r} \\ &= \frac{n!}{x_1!(n-x_1)!} \cdot \frac{(n-x_1)!}{x_2!(n-x_1-x_2)!} \\ &\quad \cdot \frac{(n-x_1-x_2)!}{x_3!(n-x_1-x_2-x_3)!} \dots \frac{(n-x_1-x_2-\dots-x_{r-1})!}{x_r!0!} \\ &= \frac{n!}{x_1! x_2! \dots x_r!}. \end{aligned}$$

Suppose a random experiment consists of  $n$  independent, identical trials, each of which has  $r$  possible outcomes, say  $O_1, O_2, \dots, O_r$ . Let  $X_i$  denote the number of times outcome  $O_i$  occurs and  $p_i = P(X_i = 1)$ . We say that  $X_1, X_2, \dots, X_r$  has a *multinomial distribution with parameters  $n$  and  $p_1, p_2, \dots, p_r$* . The joint probability mass function of  $X_1, X_2, \dots, X_r$  is

$$\begin{aligned} f(x_1, x_2, \dots, x_r) &= P(X_1 = x_1, X_2 = x_2, \dots, X_r = x_r) \\ &= \binom{n}{x_1, x_2, \dots, x_r} p_1^{x_1} p_2^{x_2} \dots p_r^{x_r}, \end{aligned} \quad (\text{B.5})$$

where  $n = \sum_{i=1}^r x_i$  and  $\sum_{i=1}^r p_i = 1$ .

We will denote this by

$$(X_1, X_2, \dots, X_r) \sim \text{Multinom}(n, p_1, p_2, \dots, p_r). \quad (\text{B.6})$$

### Example B.1

Toss a die 15 times. Find the probability of seeing three 1's, two 2's, five 3's, one 4, two 5's, and two 6's.

#### Solution

If  $p_i$ ,  $i = 1, 2, \dots, 6$ , denotes the probability of seeing an  $i$  on the die, then  $p_i = 1/6$ , for  $i = 1, 2, \dots, 6$ . Thus, the desired probability is

$$\binom{15}{3, 2, 5, 1, 2, 2} \left(\frac{1}{6}\right)^{3+2+5+1+2+2} = \frac{15!}{3!2!5!1!2!2!} \left(\frac{1}{6}\right)^{15} = 0.0005. \quad \square$$

### Example B.2

According to Mars Candy Company, the distribution of colors in the Milk Chocolate M&M™'s is blue 24%, orange 20%, green 16%, yellow 14%, red 13%, and brown 13%. Suppose in a well-mixed extremely large vat of M&M™'s, you draw out 30 pieces of candy. What is the probability you get 8 blue, 5 orange, 6 green, 5 yellow, 4 red, and 2 brown candies?

#### Solution

$$\binom{30}{8, 5, 6, 5, 4, 2} (0.24)^8 (0.20)^5 (0.16)^6 (0.14)^5 (0.13)^4 (0.13)^2 = 0.0002. \quad \square$$

### Example B.3

Suppose five numbers are drawn at random from a distribution with pdf  $f(x) = 3x^2$ ,  $0 \leq x \leq 1$ . What is the probability that one of the numbers lies in the interval  $I_1 = [0, 1/3)$ , two lie in the interval  $I_2 = [1/3, 2/3)$ , and two lie in the interval  $I_3 = [2/3, 1]$ ?

#### Solution

If  $X$  is a number drawn from this distribution, then

$$p_1 = P(0 \leq X < 1/3) = \int_0^{1/3} 3x^2 dx = 1/27,$$

$$p_2 = P(1/3 \leq X < 2/3) = \int_{1/3}^{2/3} 3x^2 dx = 7/27,$$

$$p_3 = P(2/3 \leq X \leq 1) = \int_{2/3}^1 3x^2 dx = 19/27.$$

Thus if  $X_i$  denotes the number of values that lie in interval  $I_i$ ,  $i = 1, 2, 3$ , then

$$P(X_1 = 1, X_2 = 2, X_3 = 2) = \binom{5}{1, 2, 2} \left(\frac{1}{27}\right) \left(\frac{7}{27}\right)^2 \left(\frac{19}{27}\right)^2 = 0.0369. \quad \square$$

Note that when there are only two types of outcomes ( $r = 2$ ), then the multinomial distribution reduces to the binomial distribution since we can consider outcome  $O_1$  as “heads” (or successes) and outcome  $O_2$  as “tails” (failures). We can generalize this when we have more than two outcomes. If  $X_1, X_2, \dots, X_r$  has a multinomial distribution with parameters  $n$  and  $p_1, p_2, \dots, p_r$ , consider  $X_i$  to be the number of occurrences of type  $O_i$ , each occurrence a “success” with probability  $p_i$ . Consider any occurrence of the other types to be a “failure.” This then occurs with probability  $1 - p_i$ . Thus  $X_i$  is a binomial random variable with parameters  $n$  and  $p_i$ , and

$$E[X_i] = np_i, \quad \text{Var}[X_i] = np_i(1 - p_i), \quad i = 1, 2, \dots, r. \quad (\text{B.7})$$

### Example B.4

Referring to Example B.2, suppose you draw out 38 candies at random from the vat of M&M™’s. What is the expected number of reds?

### Solution

Let  $X$  denote the number of red candies. Then  $X \sim \text{Binom}(38, 0.13)$  and  $E[X] = 38 \times 0.13 = 4.94$ , so about five red candies.  $\square$

## B.3 The Geometric Distribution

Toss a fair six-sided die. What is the probability that the first occurrence of “2” will occur on the 15th toss? More generally, consider a sequence of independent Bernoulli trials  $X_i$ , each with probability  $p$  of success. If  $X$  denotes the number of trials up to and including the first success, then  $X$  has a *geometric* distribution. In order for the first success to occur on the  $k$ th trial, we must have  $(k - 1)$  failures, each occurring with probability  $1 - p$ . Hence, the probabilities are given by

$$p(k) = P(X = k) = (1 - p)^{k-1}p, \quad k = 1, 2, 3, \dots. \quad (\text{B.8})$$

**Example B.5** The probability that the first occurrence of a “2” will occur on the 15th toss of a six-sided die is

$$P(X = 15) = (1 - 1/6)^{14}(1/6) = 0.01298. \quad \square$$

We denote a geometric random variable  $X$  with probability of success  $p$  by  $X \sim \text{Geom}(p)$ .

**Theorem B.2** Let  $X \sim \text{Geom}(p)$ . Then

$$\mathbb{E}[X] = \frac{1}{p}, \quad \text{Var}[X] = \frac{(1-p)}{p^2}. \quad (\text{B.9})$$

## B.4 The Negative Binomial Distribution

Toss a fair six-sided die until the 8th appearance of “2.” What is the probability that this will occur on the 30th toss? More generally, consider a sequence of independent Bernoulli trials with each trial having success probability  $p$ . What is the probability that the  $r$ th success will occur on the  $x$ th trial? Let  $X$  denote the trial at which the  $r$ th success occurs. Then  $X$  has a negative binomial distribution with parameters  $r$  and  $p$ , and its probability mass function is

$$P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \quad x = r, r+1, r+2, \dots \quad (\text{B.10})$$

If the  $r$ th success occurs on the  $x$  trial, then there must be  $(r-1)$  successes in the first  $(x-1)$  trials. There are  $\binom{x-1}{r-1}$  ways to choose the trials where the success occurs; in addition, the probability of  $r$  successes and  $(x-r)$  failures is  $p^r(1-p)^{x-r}$ .

**Example B.6** The probability that in tossing a fair six-sided die, the 8th appearance of “2” occurs on the 30th toss is  $P(X = 30) = \binom{30-1}{8-1} (1/6)^8 (5/6)^{22} = 0.0168$ .  $\square$

We check that Equation (B.10) is indeed a probability mass function:

$$\begin{aligned} \sum_{x=r}^{\infty} \binom{x-1}{r-1} p^r (1-p)^{x-r} &= \sum_{y=0}^{\infty} \binom{y+r-1}{r-1} p^r (1-p)^y \quad \text{set } y = x - r \\ &= \frac{p^r}{(1 - (1-p))^r} \\ &= 1, \end{aligned}$$

where we use the identity

$$\frac{1}{(1-w)^m} = \sum_{k=0}^{\infty} \binom{k+m-1}{m-1} w^k.$$

**Theorem B.3** Let  $X$  be a negative binomial random variable with parameters  $r$  and  $p$ . Then

$$\mathbb{E}[X] = \frac{r}{p}, \quad \text{Var}[X] = \frac{r(1-p)}{p^2}. \quad (\text{B.11})$$

*Proof.* Let  $X_1$  denote the geometric random variable giving the number of trials until the first success,  $X_2$  the geometric random variable giving the number of additional trials after the first success until the second success,  $\dots$ ,  $X_r$  the number of additional trials after the  $(r - 1)$ st success until the  $r$ th success. Then  $X = \sum_{i=1}^r X_i$  and since  $X_i$ ,  $i = 1, 2, \dots, r$ , are independent, we have

$$\begin{aligned}\mathbb{E}[X] &= \sum_{i=1}^r \mathbb{E}[X_i] = \sum_{i=1}^r \frac{1}{p} = \frac{r}{p}. \\ \text{Var}[X] &= \sum_{i=1}^r \text{Var}[X_i] = \sum_{i=1}^r \frac{1-p}{p^2} = \frac{r(1-p)}{p^2}.\end{aligned}$$

□

## B.5 The Hypergeometric Distribution

Suppose there are  $M$  women and  $N$  men at a college. You form a committee of size  $n$ , with  $n \leq M + N$ . There are  $\binom{M+N}{n}$  ways to choose a committee, of which  $\binom{M}{x} \binom{N}{n-x}$  have  $x$  women and  $(n - x)$  men. Suppose the committee is chosen randomly, and let  $X$  denote the number of women on the committee. Then  $X$  is called a *hypergeometric random variable*, with pmf

$$f(x) = P(X = x) = \frac{\binom{M}{x} \binom{N}{n-x}}{\binom{M+N}{n}}, \quad x = 0, 1, 2, \dots, n; x \leq \min\{M, n\}; x \geq n - M. \quad (\text{B.12})$$

It is helpful to visualize this using Table B.1. None of the four cells in the table  $(n, n - x, M - x, N - n + x)$  may be negative, leading to four constraints on the range of  $x$ . For example, if the committee is size  $n = 300$  and there are only  $N = 250$  men at the college, there must be at least  $n - N = 50$  women on the committee.

Since each committee of size  $n$  can have either  $0, 1, 2, \dots$ , or  $n$  women serving on it,  $\sum_{x=0}^n \binom{M}{x} \binom{N}{n-x} = \binom{M+N}{n}$ , the total number of committees of size  $n$  from  $M + N$  people. Thus  $\sum_{x=0}^n f(x) = 1$ .

**Table B.1** Table for hypergeometric distribution.

	Women	Men	Total
On committee	$x$	$n - x$	$n$
Off committee	$M - x$	$N - n + x$	$M + N - n$
Total	$M$	$N$	$M + N$

**Theorem B.4** For a hypergeometric variable  $X$  with parameters  $M, N$ , and  $n$ ,

$$\text{E}[X] = \frac{nM}{M+N} \quad \text{Var}[X] = \frac{nMN(M+N-n)}{(M+N)^2(M+N-1)}.$$

*Proof.* In a college of  $M$  women and  $N$  men, select a committee of size  $n$ ,  $n \leq \min\{M, N\}$ . Let  $X_i = 1$  if the  $i$ th person chosen is a woman,  $X_i = 0$  otherwise,  $i = 0, 1, \dots, n$ . Then  $P(X_i = 1) = M/M+N$ , so  $\text{E}[X_i] = 1 \times P(X_i = 1) + 0 \times P(X_i = 0) = M/M+N$ . Thus  $X = \sum_{i=1}^n X_i$  gives the number of women on the committee, and

$$\text{E}[X] = \sum_{i=1}^n \text{E}[X_i] = \sum_{i=1}^n \frac{M}{M+N} = \frac{nM}{M+N}.$$

Now, from Corollary 9.2 and since, by symmetry,  $\text{Cov}[X_i, X_j] = \text{Cov}[X_1, X_2]$ , we have

$$\text{Var}[X] = \sum_{i=1}^n \text{Var}[X_i] + 2 \sum_i \sum_j \text{Cov}[X_i, X_j]. \quad (\text{B.13})$$

$$= \frac{nNM}{(M+N)^2} + n(n-1)\text{Cov}[X_1, X_2]. \quad (\text{B.14})$$

Now, consider the case that  $n = M+N$ . Then  $X = M$  since *all* the women at the college must be selected. In this case,  $\text{Var}[X] = 0$ . Thus, from Equation (B.14), we have

$$\text{Cov}[X_1, X_2] = -\frac{MN}{(M+N)^2(M+N-1)}.$$

Thus, in the general case ( $n$  not necessarily equal to  $M+N$ ), Equation (B.14) gives

$$\text{Var}[X] = \frac{nMN}{(M+N)^2} \cdot \frac{M+N-n}{M+N-1},$$

which establishes the result.  $\square$

## B.6 The Poisson Distribution

**Definition B.1** A discrete random variable  $X$  with range  $0, 1, 2, \dots$  has a Poisson distribution with parameter  $\lambda > 0$  if its pmf is

$$f(x) = P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad \text{for } x = 0, 1, 2, \dots. \quad (\text{B.15})$$

||

We see that Equation (B.15) is a probability mass function since

$$\sum_{x=0}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} e^{\lambda} = 1.$$

The Poisson distribution is often used to model counts.

**Example B.7**

Suppose the number of typographical errors on a web page for a certain company follows a Poisson random variable with  $\lambda = 1.5$ . Find the probability that there are at most two typographical errors on a given page.

**Solution**

Let  $X$  denote the number of typographical errors on a web page. Then the probabilities are given by  $f(x) = 1.5^x e^{-1.5} / x!$  for  $x = 0, 1, 2, \dots$ , so  $F(2) = e^{-1.5} (1 + 1.5 + 1.5^2 / 2!) \approx 0.8088$ .  $\square$

We leave as an exercise the proof of the following.

**Theorem B.5** Let  $X$  be a Poisson random variable with  $\lambda > 0$ . Then,

$$\mathbb{E}[X] = \lambda \quad \text{Var}[X] = \lambda. \quad (\text{B.16})$$

The following theorem allows us to work with sums of Poisson random variables.

**Theorem B.6** Let  $X_1, X_2, \dots, X_n$  be independent Poisson random variables with parameters  $\lambda_1, \lambda_2, \dots, \lambda_n$ , respectively. Then  $X = X_1 + X_2 + \dots + X_n$  is Poisson with parameter  $\lambda_1 + \lambda_2 + \dots + \lambda_n$ .

*Proof.* We will prove the result for  $n = 2$ .

$$\begin{aligned} P(X_1 + X_2 = m) &= \sum_{j=0}^m P(X_1 = j, X_2 = m-j) \\ &= \sum_{j=0}^m P(X_1 = j)P(X_2 = m-j) \\ &= \sum_{j=0}^m \frac{\lambda_1^j e^{-\lambda_1}}{j!} \frac{\lambda_2^{m-j} e^{-\lambda_2}}{(m-j)!} \\ &= e^{-(\lambda_1+\lambda_2)} \sum_{j=0}^m \frac{\lambda_1^j \lambda_2^{m-j}}{j!(m-j)!} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{m!} \sum_{j=0}^m \frac{m!}{j!(m-j)!} \lambda_1^j \lambda_2^{m-j} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{m!} (\lambda_1 + \lambda_2)^m, \end{aligned}$$

where the last equality is from the binomial theorem. Thus,  $X_1 + X_2$  is Poisson with parameter  $\lambda_1 + \lambda_2$ .

A proof by induction can be used to extend to a sum of  $n$  Poisson random variables.  $\square$

## B.7 The Uniform Distribution

**Definition B.2** A random variable  $X$  has a uniform distribution on the interval  $[a, b]$  (for  $a < b$ ) if its pdf is

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{otherwise} \end{cases}. \quad (\text{B.17})$$

We denote this by  $X \sim \text{Unif}[a, b]$  ||

We can also define the uniform distribution over the intervals  $(a, b)$ ,  $(a, b]$ , and  $[a, b)$  by making the appropriate adjustment in the pdf.

**Theorem B.7** Let  $X \sim \text{Unif}[a, b]$ . Then

$$\text{E}[X] = \frac{a+b}{2}, \quad \text{Var}[X] = \frac{(a-b)^2}{12}. \quad (\text{B.18})$$

**Proposition B.1** Let  $X \sim \text{Unif}[0, \theta]$ . Then  $X/\theta \sim \text{Unif}[0, 1]$ . □

*Proof.* Exercise.

## B.8 The Exponential Distribution

**Definition B.3** A random variable  $X$  has the exponential distribution with parameter  $\lambda > 0$  if its pdf is

$$f(x) = \lambda e^{-\lambda x} \quad \text{for } x \geq 0. \quad (\text{B.19})$$

We will write  $X \sim \text{Exp}(\lambda)$ . ||

Another common parameterization is

$$f(x) = \frac{1}{\rho} e^{-x/\rho} \quad \text{for } x \geq 0, \quad (\text{B.20})$$

with  $\rho = 1/\lambda > 0$ . Here  $\rho$  is a *scale parameter* – doubling it corresponds to doubling all values – and is equal to the mean of the distribution.

**Theorem B.8** Let  $X$  have an exponential distribution with  $\lambda > 0$ . Then

$$\mathbb{E}[X] = \frac{1}{\lambda} = \rho, \quad \text{Var}[X] = \frac{1}{\lambda^2} = \rho^2. \quad (\text{B.21})$$

The exponential distribution is used to model *waiting times*, that is, time between the occurrence of successive events, such as the time between groups of customers arriving at a fast food restaurant. It is notable for being the only distribution with the *memoryless property*, that is,

$$P(X \geq t + h | X \geq h) = P(X \geq t). \quad (\text{B.22})$$

The proof is an exercise.

The parameter  $\lambda$  is a *rate parameter* – it represents how quickly events occur. Doubling  $\lambda$  cuts the expected time until the next arrival in half.

If  $X$  represents the time in minutes until an event occurs, then Equation (B.22) says that the probability the event will not occur within the next  $t$  minutes does not depend on how long you have already waited. If the current time is  $h$  (and the event has not occurred yet), the probability that the event will wait an additional  $t$  minutes is the same as the probability back at time  $h = 0$ .

Note that the times between individuals (opposed to groups) arriving at a fast food restaurant would not be memoryless. If someone just walked in the door, there is a higher than normal chance that someone else will arrive within the next 10 s (namely a friend or family member).

For some calculations with exponential variables, it is useful to *standardize* to rate 1 (and mean and variance 1).

**Proposition B.2** Let  $X$  be an exponential random variable with parameter  $\lambda$ . Then  $Y = \lambda X$  is exponential with parameter 1.

*Proof.* Let  $X \sim \text{Exp}(\lambda)$  with pdf  $f_X(x) = \lambda e^{-\lambda x}$ . Then, the cumulative distribution function (cdf) is  $F_X(x) = 1 - e^{-\lambda x}$ .

Let  $F_Y(x)$  and  $f_Y(x)$  denote the cdf and pdf, respectively, of  $Y = \lambda X$ . Then

$$F_Y(x) = P(Y \leq x) = P(\lambda X \leq x) = P\left(X \leq \frac{x}{\lambda}\right) = F_X\left(\frac{x}{\lambda}\right).$$

Thus,

$$F'_Y(x) = f_Y(x) = F'_X\left(\frac{x}{\lambda}\right) \frac{1}{\lambda} = e^{-x},$$

so  $Y \sim \text{Exp}(1)$ . □

## B.9 The Gamma Distribution

**Definition B.4** The gamma function is defined, for  $r > 0$ , by

$$\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} dx. \quad (\text{B.23})$$

**Theorem B.9**

1. For  $r > 1$ ,

$$\Gamma(r) = (r - 1)\Gamma(r - 1). \quad (\text{B.24})$$

2. For positive integer  $n$ ,

$$\Gamma(n) = (n - 1)! \quad (\text{B.25})$$

3.  $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}. \quad (\text{B.26})$

The proof is an exercise.

**Definition B.5** A continuous random variable  $X$  has the gamma distribution with parameters  $r > 0$ ,  $\lambda > 0$  if its pdf is given by

$$f(x) = \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x} \quad \text{for } x \geq 0. \quad (\text{B.27})$$

We denote this by  $X \sim \text{Gamma}(r, \lambda)$ .

Some examples are shown in Figure B.1. ||

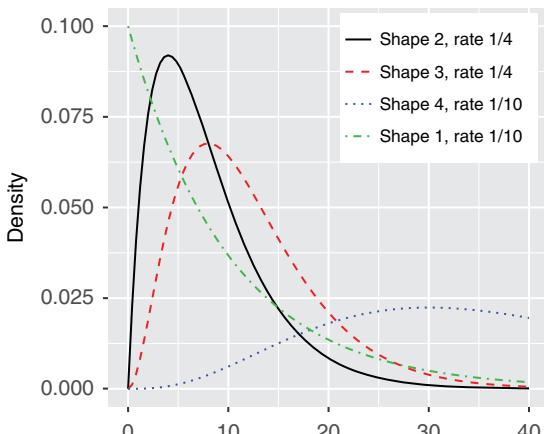
The  $r$  and  $\lambda$  are called *shape* and *rate* parameters respectively. Another common parameterization is to let  $\rho = 1/\lambda$

$$f(x) = \frac{1}{\rho^r \Gamma(r)} x^{r-1} e^{-x/\rho} \quad \text{for } x \geq 0. \quad (\text{B.28})$$

$\rho$  is a *scale* parameter.

The exponential distribution is a special case of the gamma distribution. That is  $X \sim \text{Exp}(\lambda)$  is the same as  $X \sim \text{Gamma}(1, \lambda)$ .

**Figure B.1** Densities of gamma distribution.



By using the substitution  $u = \lambda x$ , we find

$$\int_0^\infty x^{r-1} e^{-\lambda x} dx = \frac{\Gamma(r)}{\lambda^r}, \quad (\text{B.29})$$

so that Equation (B.27) integrates to 1.

**Theorem B.10** Let  $X$  be a gamma random variable with parameters  $r, \lambda$ . Then

$$\mathbb{E}[X] = \frac{r}{\lambda}, \quad \text{Var}[X] = \frac{r}{\lambda^2}. \quad (\text{B.30})$$

*Proof.*

$$\begin{aligned} \mathbb{E}[X] &= \int_0^\infty x \frac{1}{\Gamma(r)} \lambda^r x^{r-1} e^{-\lambda x} dx \\ &= \frac{\lambda^r}{\Gamma(r)} \int_0^\infty x^r e^{-\lambda x} dx \\ &= \frac{\lambda^r}{\Gamma(r)} \cdot \frac{\Gamma(r+1)}{\lambda^{r+1}} && \text{by Equation (B.29)} \\ &= \frac{r}{\lambda} && \text{by Theorem B.9.} \end{aligned}$$

We leave the proof of the variance as an exercise.  $\square$

The proof to the following proposition is an exercise (Exercise B.8).

**Proposition B.3** Let  $X$  be a gamma random variable with parameters  $r, \lambda$  and suppose  $c > 0$  is a constant. Then  $cX$  is also a gamma random variable with parameters  $r, \lambda/c$ .

We compute the moment generating function of a gamma random variable (see Section A.8).

$$\begin{aligned} M(t) &= \int_0^\infty e^{tx} f(x) dx \\ &= \frac{\lambda^r}{\Gamma(r)} \int_0^\infty x^{r-1} e^{-(\lambda-t)x} dx \\ &= \frac{\lambda^r}{\Gamma(r)} \cdot \frac{\Gamma(r)}{(\lambda-t)^r} && \text{by Equation (B.29)} \\ &= \left( \frac{\lambda}{\lambda-t} \right)^r. \end{aligned} \quad (\text{B.31})$$

for  $t < \lambda$ .

Thus, we can prove the following.

**Theorem B.11** Let  $X_1, X_2, \dots, X_n$  be independent random variables with  $X_i \sim \text{Gamma}(r_i, \lambda)$ ,  $i = 1, 2, \dots, n$ . Then the sum  $X = \sum_{i=1}^n X_i$  is also a gamma random variable with parameters  $r_1 + r_2 + \dots + r_n$  and  $\lambda$ .

*Proof.* Let  $M_i(t)$  denote the mgf of  $X_i$ ,  $i = 1, 2, \dots, n$ , and  $M(t)$  the mgf of the sum  $X$ . Then,

$$M(t) = \prod_{i=1}^n M_i(t) = \prod_{i=1}^n \left( \frac{\lambda}{\lambda - t} \right)^{r_i} = \left( \frac{\lambda}{\lambda - t} \right)^{r_1 + r_2 + \dots + r_n},$$

for  $t < \lambda$  (Theorem A.14). We see that  $M(t)$  is the mgf of a gamma distribution with parameters  $r_1 + r_2 + \dots + r_n$  and  $\lambda$ . Hence,  $X = X_1 + X_2 + \dots + X_n \sim \text{Gamma}(r_1 + r_2 + \dots + r_n, \lambda)$   $\square$

## B.10 The Chi-Square Distribution

**Definition B.6** A continuous random variable  $X$  has a chi-square distribution with  $m > 0$  degrees of freedom if its pdf is

$$f(x) = \frac{x^{m/2-1} e^{-x/2}}{2^{m/2} \Gamma(m/2)} \quad \text{for } x \geq 0. \quad (\text{B.32})$$

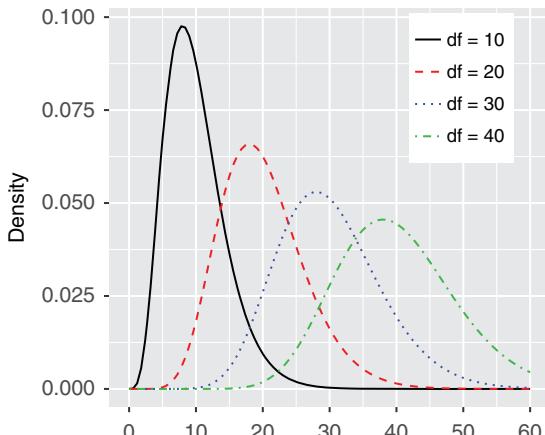
We write  $X \sim \chi_m^2$ .

Some examples are shown in Figure B.2.  $\parallel$

**Theorem B.12** Let  $X$  be a random variable following a chi-square distribution with  $m$  degrees of freedom. Then

$$\mathbb{E}[X] = m, \quad \text{Var}[X] = 2m. \quad (\text{B.33})$$

**Figure B.2** Densities for the chi-square distribution.



The chi-square distribution with  $m$  degrees of freedom is a special case of the gamma distribution, with  $r = m/2$  and  $\lambda = 1/2$ . Thus, it follows from Theorem B.11 that

**Theorem B.13** Let  $X_1, X_2, \dots, X_k$  be independent chi-square random variables with degrees of freedom  $m_1, m_2, \dots, m_k$ , respectively. Then  $X = X_1 + X_2 + \dots + X_k$  is a chi-square random variable with  $m_1 + m_2 + \dots + m_k$  degrees of freedom.

**Theorem B.14** Let  $Z$  be a random variable from a standard normal distribution. Then  $Z^2$  has a chi-square distribution with 1 degree of freedom.

*Proof.* Let  $f_{Z^2}$  denote the pdf for  $Z^2$ .

$$\begin{aligned} \int_{-\infty}^x f_{Z^2}(y) dy &= P(Z^2 \leq x) \\ &= P(-\sqrt{x} \leq Z \leq \sqrt{x}) \\ &= P(Z \leq \sqrt{x}) - P(Z \leq -\sqrt{x}) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{x}}^{\sqrt{x}} e^{-y^2/2} dy. \end{aligned}$$

Differentiate both sides with respect to  $x$ . Then

$$\begin{aligned} f_{Z^2}(x) &= \frac{1}{\sqrt{2\pi}} \left( e^{-x/2} \frac{d}{dx} \sqrt{x} - e^{-x/2} \frac{d}{dx} (-\sqrt{x}) \right) \\ &= \frac{1}{\sqrt{2\pi} \sqrt{x}} e^{-x/2}, \end{aligned}$$

which is the pdf for the  $\chi^2$  distribution with 1 degree of freedom.  $\square$

Thus, Theorem B.14 together with Theorem B.13 gives us Theorem B.15.

**Theorem B.15** Let  $Z_1, Z_2, \dots, Z_k$  be an independent random sample from the standard normal distribution. Then  $\sum_{i=1}^k Z_i^2$  has a  $\chi^2$  distribution with  $k$  degrees of freedom.

Recall that for random variables  $X_1, X_2, \dots, X_n$ , we define the sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

**Theorem B.16** Let  $X_1, X_2, \dots, X_n$  be a random sample from  $N(\mu, \sigma^2)$ . Then

1.  $\bar{X}$  and  $S^2$  are independent random variables.
2.  $(n - 1)S^2/\sigma^2$  has a  $\chi^2$  distribution with  $n - 1$  degrees of freedom.

*Proof.* We present only the proof of (2). We have  $(X_i - \mu)/\sigma \sim N(0, 1)$ ,  $i = 1, 2, \dots, n$ , so by Theorem B.15, we have

$$\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 = \sum_{i=1}^n \left( \frac{X_i - \mu}{\sigma} \right)^2 \sim \chi_n^2.$$

Now, recalling that  $\sum_{i=1}^n (X_i - \bar{X}) = 0$ ,

$$\begin{aligned} \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 &= \frac{1}{\sigma^2} \sum_{i=1}^n [(X_i - \bar{X}_i) + (\bar{X}_i - \mu)]^2 \\ &= \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 + \left( \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right)^2. \end{aligned}$$

Let  $U = (1/\sigma^2) \sum_{i=1}^n (X_i - \mu)^2$ ,  $V = (1/\sigma^2) \sum_{i=1}^n (X_i - \bar{X})^2$ ,  $W = ((\bar{X} - \mu)/(\sigma/\sqrt{n}))^2$ . Then by Theorems B.14 and B.15,  $U$  and  $W$  are chi-square random variables, with  $n$  and 1 degrees of freedom, respectively. It is a fact that their mgf's are  $M_U(t) = 1/(1 - 2t)^{n/2}$  and  $M_W(t) = 1/(1 - 2t)^{1/2}$ , respectively.

Now,  $U = V + W$ , and since  $V$  and  $W$  are independent, by Theorem A.14 we have  $M_U(t) = M_V(t)M_W(t)$ , or

$$\begin{aligned} M_V(t) &= \frac{M_U(t)}{M_W(t)} \\ &= \frac{(1 - 2t)^{-n/2}}{(1 - 2t)^{-1/2}} \\ &= (1 - 2t)^{-(n-1)/2}, \end{aligned}$$

which is the mgf of a chi-square distribution with  $n - 1$  degrees of freedom.

Since,

$$\begin{aligned} V &= \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 \\ &= \frac{n-1}{n-1} \cdot \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 \\ &= \frac{n-1}{\sigma^2} \cdot \frac{1}{(n-1)} \sum_{i=1}^n (X_i - \bar{X})^2 \\ &= \frac{n-1}{\sigma^2} S^2, \end{aligned}$$

this completes the proof. □

## B.11 The Student's $t$ Distribution

**Definition B.7** A random variable has a  $t$  distribution with  $k$  degrees of freedom ( $k > 0$ ) if its pdf is

$$f(x) = \frac{\Gamma(k+1)/2}{\Gamma(k/2)\sqrt{k\pi}} \left(1 + \frac{x^2}{k}\right)^{-(k+1)/2} \quad \text{for } -\infty < x < \infty.$$

||

**Theorem B.17** Let  $Z$  be a standard normal random variable and let  $W$  denote a chi-square distribution with  $k$  degrees of freedom, with  $Z$  and  $W$  independent. Then the random variable defined by

$$T = \frac{Z}{\sqrt{W/k}}$$

has a  $t$  distribution with  $k$  degrees of freedom.

*Proof.* Let  $f_T(t), f_Z(z), f_W(w)$  denote the pdf's of  $T, Z$ , and  $W$ , respectively. If  $f(z, w)$  denotes the joint pdf of  $Z$  and  $W$ , then by independence,  $f(z, w) = f_Z(z)f_W(w)$ . Thus,

$$\begin{aligned} P(T \leq t) &= P\left(\frac{Z}{\sqrt{W/k}} \leq t\right) = P(Z \leq t\sqrt{W/k}) \\ &= \int_0^\infty \int_{-\infty}^{t\sqrt{w/k}} f_Z(z)f_W(w) dz dw. \end{aligned}$$

By the fundamental theorem of calculus,  $d/dt P(T \leq t) = f_T(t)$ , so

$$\begin{aligned} f_T(t) &= \int_0^\infty \frac{d}{dt} \int_{-\infty}^{t\sqrt{w/k}} f_Z(z)f_W(w) dz dw \\ &= \int_0^\infty f_Z\left(t\sqrt{\frac{w}{k}}\right) \sqrt{\frac{w}{k}} f_W(w) dw \\ &= \int_0^\infty \frac{\sqrt{\frac{w}{k}} e^{-\frac{t^2 w}{2k}}}{\sqrt{2\pi}} \frac{w^{\frac{k}{2}-1} e^{-\frac{w}{2}}}{2^{\frac{k}{2}} \Gamma(k/2)} dw \\ &= \frac{1}{\sqrt{2\pi} 2^{k/2} \Gamma(k/2) \sqrt{k}} \int_0^\infty w^{\frac{k}{2}-\frac{1}{2}} e^{-w\left(\frac{1}{2}+\frac{t^2}{2k}\right)} dw. \end{aligned}$$

Now, make a  $u$ -substitution,  $u = \omega(\frac{1}{2} + t^2/(2k))$  to get

$$\begin{aligned} f_T(t) &= \frac{1}{\sqrt{2\pi} 2^{k/2}\Gamma(k/2)\sqrt{k}} \int_0^\infty \frac{u^{\frac{k-1}{2}} e^{-u}}{(1/2 + t^2/(2k))^{\frac{k-1}{2}}(1/2 + t^2/(2k))} du \\ &= \frac{1}{\sqrt{2\pi} 2^{k/2}\Gamma(k/2)\sqrt{k}(1/2 + t^2/(2k))^{(k+1)/2}} \int_0^\infty u^{\frac{k+1}{2}-1} e^{-u} du \\ &= \frac{\Gamma((k+1)/2)}{\sqrt{2\pi} 2^{k/2}\Gamma(k/2)\sqrt{k}(1/2 + t^2/(2k))^{(k+1)/2}} \\ &= \frac{\Gamma((k+1)/2)}{\sqrt{\pi k} \Gamma(k/2)(1 + t^2/k)^{(k+1)/2}}, \end{aligned}$$

which is the desired pdf.  $\square$

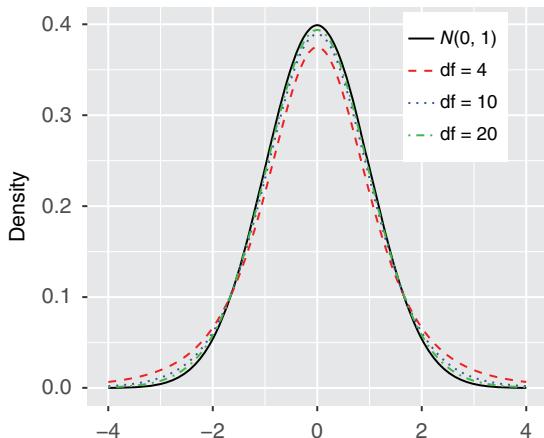
The pdf for a  $t$  distribution is bell shaped and symmetric about 0 with heavier tails than the standard normal. As  $k$  tends toward infinity, the density of the  $t$  distribution tends toward the density of the standard normal. Some examples are shown in Figure B.3.

**Theorem B.18** Let  $T$  denote a random variable with a  $t$  distribution with  $k$  degrees of freedom. Then

$$E[T] = 0 \quad \text{Var}[T] = k/(k-2), \quad k > 2. \quad (\text{B.34})$$

**Theorem B.19** Let  $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ . Then the random variable  $T = (\bar{X} - \mu)/(S/\sqrt{n})$  has a  $t$  distribution with  $n - 1$  degrees of freedom.

**Figure B.3** Densities for the  $t$  distribution.



*Proof.* Let  $X_1, X_2, \dots, X_n \sim N(\mu, \sigma^2)$ . Then  $\bar{X} \sim N(\mu, \sigma^2/n)$  so  $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n}) = \sqrt{n}(\bar{X} - \mu)/\sigma$  is a standard normal random variable. From Theorem B.16, we know that  $W = (n-1)S^2/\sigma^2$  is a chi-square random variable with  $n-1$  degrees of freedom and that  $Z$  and  $W$  are independent.

Thus, by Theorem B.17,

$$T = \frac{Z}{\sqrt{W/(n-1)}} = \frac{\sqrt{n}(\bar{X} - \mu)/\sigma}{\sqrt{((n-1)S^2/\sigma^2)/(n-1)}} = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has a  $t$  distribution with  $n-1$  degrees of freedom.  $\square$

## B.12 The Beta Distribution

**Definition B.8** A random variable  $X$  has a beta distribution with parameters  $\alpha > 0$  and  $\beta > 0$  if its pdf is

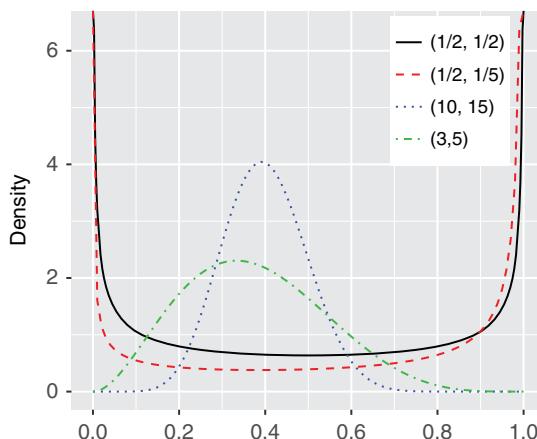
$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \quad \text{for } 0 \leq x \leq 1. \quad (\text{B.35})$$

We will denote this by  $X \sim \text{Beta}(\alpha, \beta)$ .

Some examples are shown in Figure B.4.  $\parallel$

Note that

$$\begin{aligned} \Gamma(\alpha)\Gamma(\beta) &= \int_0^\infty u^{\alpha-1}e^{-u} du \int_0^\infty v^{\beta-1}e^{-v} dv \\ &= \int_0^\infty \int_0^\infty u^{\alpha-1}v^{\beta-1}e^{-(u+v)} du dv. \end{aligned}$$



**Figure B.4** Densities for the Beta distribution  $(\alpha, \beta)$ .

Now, make the substitution  $x = u/(u + v)$  and  $y = u + v$ . Then  $u = xy$  and  $v = (1 - x)y$ . Computing the Jacobian for this transformation and making the appropriate changes in the limits of integration, we find

$$\begin{aligned}\Gamma(\alpha)\Gamma(\beta) &= \int_0^1 \int_0^\infty x^{\alpha-1}(1-x)^{\beta-1}y^{\alpha+\beta-1}e^{-y} dy dx \\ &= \Gamma(\alpha + \beta) \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx.\end{aligned}$$

Thus, the function in Equation (B.35) integrates to 1.

**Theorem B.20** Let  $X$  have a beta distribution,  $X \sim \text{Beta}(\alpha, \beta)$ . Then

$$E[X] = \frac{\alpha}{\alpha + \beta}, \quad \text{Var}[X] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \quad (\text{B.36})$$

The proof is left as an exercise.

## B.13 The F Distribution

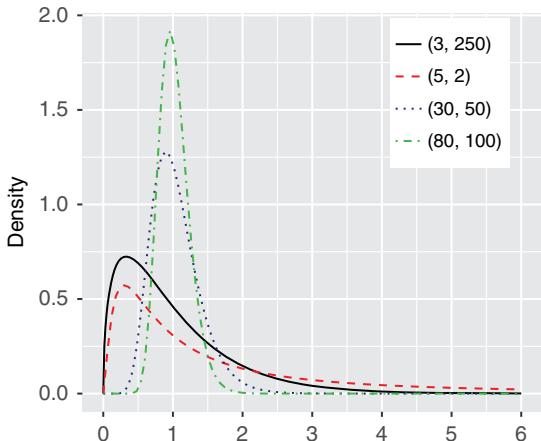
**Definition B.9** Let  $Y$  and  $W$  be independent chi-squared random variables with  $m$  and  $n$  degrees of freedom, respectively. The random variable

$$X = \frac{Y/m}{W/n} \quad (\text{B.37})$$

is called an  $F$  distribution with  $m$  and  $n$  degrees of freedom. We denote this by  $X \sim F_{m,n}$ .

Some examples are shown in Figure B.5. ||

**Figure B.5**  $F$  distribution with  $(m, n)$  degrees of freedom.



**Theorem B.21** The pdf of an  $F$  distribution with  $m, n$  degrees of freedom is

$$f(x) = \frac{\Gamma((m+n)/2)m^{m/2}n^{n/2}}{\Gamma(m/2)\Gamma(n/2)} \cdot \frac{x^{m/2-1}}{(mx+n)^{(m+n)/2}} \quad \text{for } x > 0. \quad (\text{B.38})$$

*Proof.* Let  $f_Y(y) = \frac{1}{2^{m/2}\Gamma(m/2)}y^{m/2-1}e^{-y/2}$  and  $f_W(w) = \frac{1}{2^{n/2}\Gamma(n/2)}w^{n/2-1}e^{-w/2}$  denote the pdf's for  $Y$  and  $W$ , respectively. Then by independence,

$$\begin{aligned} F_X(x) &= P\left(\frac{Y/m}{W/n} \leq x\right) = P(Y \leq \frac{m}{n}xW) \\ &= \int_0^\infty \int_0^{mwx/n} \frac{w^{n/2-1}e^{-w/2}}{2^{m/2}\Gamma(m/2)} \times \frac{y^{m/2-1}e^{-y/2}}{2^{n/2}\Gamma(n/2)} dy dw \\ &= \frac{1}{2^{\frac{m}{2}+\frac{n}{2}}\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \int_0^\infty \int_0^{mwx/n} w^{\frac{n}{2}-1}y^{\frac{m}{2}-1}e^{-\frac{w}{2}-\frac{y}{2}} dy dw. \end{aligned}$$

Now differentiate with respect to  $x$  and then set  $z = (mx/(2n) + 1/2)w$ :

$$\begin{aligned} f_X(x) &= \frac{1}{2^{m/2+n/2}\Gamma(m/2)\Gamma(n/2)} \int_0^\infty w^{n/2-1}(mwx/n)^{m/2-1}e^{-mwx/(2n)-w/2} \frac{mw}{n} dw \\ &= \frac{1}{2^{n/2+m/2}\Gamma(m/2)\Gamma(n/2)} \left(\frac{m}{n}\right)^{m/2} x^{m/2-1} \int_0^\infty \frac{z^{m/2+n/2-1}}{(mx/(2n) + 1/2)^{m/2+n/2-1}} e^{-z} \\ &\quad \times \frac{dz}{(mx/(2n) + 1/2)} \\ &= \frac{1}{2^{m/2+n/2}\Gamma(m/2)\Gamma(n/2)} \left(\frac{m}{n}\right)^{m/2} x^{m/2-1} \frac{2^{m/2+n/2}n^{m/2+n/2}}{(mx+n)^{m/2+n/2}} \\ &\quad \times \int_0^\infty z^{m/2+n/2-1}e^{-z} dz \\ &= \frac{m^{m/2}n^{n/2}}{\Gamma(m/2)\Gamma(n/2)} \frac{x^{m/2-1}}{(mx+n)^{(m+n)/2}} \Gamma\left(\frac{m+n}{2}\right). \end{aligned}$$
□

**Theorem B.22** Let  $X$  be a random variable from an  $F$  distribution with  $m$  and  $n$  degrees of freedom,  $X \sim F_{m,n}$ . Then

$$E[X] = \frac{n}{n-2}, \quad n > 2, \quad (\text{B.39})$$

$$\text{Var}[X] = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}, \quad n > 4. \quad (\text{B.40})$$

We leave the proof of the expectation as an exercise (Exercise B.14).

## Exercises

- B.1** A box contains 30 red balls and 20 blue balls.
- If you draw 15 balls at random without replacement, what is the probability that you draw 8 red balls and 7 blue balls?
  - If you draw 15 balls at random without replacement, how many red balls do you expect?
- B.2** You draw 100 values at random from the exponential distribution with  $\lambda = 1$ . Find the probability that 30 of these values fall in the interval  $[0, 0.25]$ , 30 fall in  $(0.25, 0.75]$ , 22 fall in  $(0.75, 1.25]$ , and the rest fall in  $(1.25, \infty)$ .
- B.3** According to [www.aabb.org](http://www.aabb.org), the distribution of blood types in the United States is 48% are of type O, 37% are of type A, 11% of type B, and 4% are of type AB. If you select 45 people at random,
  - what is the probability that the distribution of their blood types will be 20 of type O, 15 of type A, 6 of type B, and 4 of type AB?
  - what is the expected number of people in your sample with type A blood?
- B.4** Prove Proposition B.1.
- B.5** Prove the memoryless property for the exponential distribution (Equation B.22).
- B.6** Prove Theorem B.9.
- B.7** Find the variance of the gamma distribution with parameters  $r$  and  $\lambda$ .
- B.8** Prove Proposition B.3
- B.9** Verify Theorem B.13.
- B.10** Suppose  $X_1 \sim \chi^2_5$ ,  $X_2 \sim \chi^2_4$ , and  $X_3 \sim \chi^2_8$ . Let  $W = X_1 + X_2 + X_3$ . Find  $P(W \leq 10)$ .
- B.11** Let  $X \sim \text{Gamma}(2, \lambda)$ . Show that  $Y = 2\lambda X$  has a chi-square distribution with 4 degrees of freedom.

- B.12** Let  $X_1, X_2, \dots, X_n \sim N(\mu, \sigma^2)$  and define  $S^2 = 1/(n-1) \sum_{i=1}^n (X_i - \bar{X})^2$ . Prove that

$$\text{Var}[S^2] = \frac{2\sigma^4}{n-1}.$$

- B.13** Prove Theorem B.20.

- B.14** Prove that the expected value of  $X \sim F_{m,n}$  is  $n/(n-2)$  for  $n > 2$ . Hint: Suppose  $W \sim \chi_n^2$ . Show that  $E[1/W] = 1/(n-2)$ ,  $n > 2$ .

- B.15** Let  $X \sim F_{m,n}$ . Show that  $1/X \sim F_{n,m}$ .

- B.16** Let  $X \sim F_{m,n}$ . Define  $Y = mX/(mX + n)$ . Show that  $Y$  has a beta distribution with  $\alpha = m/2$  and  $\beta = n/2$ .

## Appendix C

### Distributions Quick Reference

Discrete

Probability mass function	Mean and variance	Moment generating function $M_x(t)$
<b>Bernoulli</b> $f(x; p) = p^x(1-p)^{1-x}$ $x = 0, 1, 0 \leq p \leq 1$	$\mu = p$ $\sigma^2 = p(1-p)$	$(1-p) + pe^t$
<b>Binomial</b> $f(x; n, p) = \binom{n}{x} p^x(1-p)^{n-x}$ $x = 0, 1, 2, \dots, n, 0 \leq p \leq 1, n \in \mathbf{Z}^+$	$\mu = np$ $\sigma^2 = np(1-p)$	$[pe^t + (1-p)]^n$
<b>Geometric</b> $f(x; p) = (1-p)^{x-1}p$ $x = 1, 2, 3, \dots, 0 < p \leq 1$	$\mu = \frac{1}{p}$ $\sigma^2 = \frac{1-p}{p^2}$	$\frac{pe^t}{1 - (1-p)e^t}$ $t < -\ln(1-p)$
<b>Hypergeometric</b> $H(x; n, M, N) = \frac{\binom{M}{x} \binom{N}{n-x}}{\binom{M+N}{n}}$ $x = 0, 1, \dots, n, n, M, N \in \mathbf{Z}^+$ $\max\{0, n-N\} \leq x \leq \min\{M, n\}$	$\mu = \frac{nM}{M+N}$ $\sigma^2 = \frac{nMN(M+N-n)}{(M+N)^2(M+N-1)}$	Omitted
<b>Negative binomial</b> $f(x; r, p) = \binom{x-1}{r-1} p^r(1-p)^{x-r}$ $x = r, r+1, r+2, \dots, 0 < p \leq 1, r \in \mathbf{Z}^+$	$\mu = \frac{r}{p}$ $\sigma^2 = \frac{r(1-p)}{p^2}$	$\left[ \frac{pe^t}{1 - (1-p)e^t} \right]^r$ $t < -\ln(1-p)$

<b>Poisson</b>		
$f(x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$	$\mu = \lambda$	$e^{i(\rho^2 - 1)}$
$x = 0, 1, 2, \dots, \lambda > 0$	$\sigma^2 = \lambda$	
<b>Beta</b>		
$f(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$ $0 < x < 1, \alpha > 0, \beta > 0$	$\mu = \frac{\alpha}{\alpha + \beta}$ $\sigma^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$	$1 + \sum_{k=1}^{\infty} \left( \prod_{s=0}^{k-1} \frac{\alpha + s}{\alpha + \beta + s} \right) \frac{t^k}{k!}$
<b>Cauchy</b>		
$f(x; \alpha, \beta) = \frac{\beta/\pi}{(x - \alpha)^2 + \beta^2}$ $-\infty < x < \infty, -\infty < \alpha < \infty, \beta > 0$	$\mu$ does not exist $\sigma^2$ does not exist	Does not exist
<b>Chi-square</b>		
$f(x; m) = \frac{1}{2^{m/2}\Gamma(m/2)} x^{(m-2)/2} e^{-x/2}$ $x \geq 0, m > 0$	$\mu = m$ $\sigma^2 = 2m$	$(1 - 2t)^{-m/2}$ $t < 1/2$
<b>Exponential</b>		
$f(x; \lambda) = \lambda e^{-\lambda x}$ $x \geq 0, \lambda > 0$	$\mu = \frac{1}{\lambda}$ $\sigma^2 = \frac{1}{\lambda^2}$	$(1 - t/\lambda)^{-1}$ $t < \lambda$
<b>F</b>		
$f(x; m, n) = \frac{\Gamma(\frac{m+n}{2}) m^{m/2} n^{n/2}}{\Gamma(\frac{m}{2}) \Gamma(\frac{n}{2})} \frac{x^{m/2-1}}{(mx + n)^{(m+n)/2}}$ $x > 0, m > 0, n > 0$	$\mu = \frac{n}{n-2}$ $\sigma^2 = \frac{2n^2(m+n-2)}{m(n-2)(n-4)}$	Does not exist

(Continued)

(Continued)

Probability mass function	Mean and variance	Moment generating function $M_x(t)$
<b>Gamma</b>		
$f(x; r, \lambda) = \frac{1}{\Gamma(r)} \lambda^r x^{r-1} e^{-\lambda x}$ $x \geq 0, r > 0, \lambda > 0$	$\mu = r/\lambda$ $\sigma^2 = r/\lambda^2$	$(1 - t/\lambda)^{-r}$ $t < \lambda$
<b>Normal</b>		
$f(x; \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$ $-\infty < x < \infty, \mu > 0, \sigma > 0$	$\mu = \mu$ $\sigma^2 = \sigma^2$	$e^{\mu t + \sigma^2 t^2/2}$
<b>Student's <math>t</math></b>		
$f(x; k) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{\pi k} \Gamma(\frac{k}{2})} (1 + x^2/k)^{-(k+1)/2}$ $-\infty < x < \infty, k \in \mathbb{Z}^+$	$\mu = 0$ $\sigma^2 = \frac{k}{k-2}$	Does not exist
<b>Uniform</b>		
$f(x; a, b) = \frac{1}{b-a}$ $a < x < b, -\infty < a < b < \infty$	$\mu = \frac{a+b}{2}$ $\sigma^2 = (b-a)^2/12$	$\frac{e^{bt} - e^{at}}{(b-a)t}$
<b>Weibull</b>		
$f(x; k, \lambda) = \frac{k \cdot x^{k-1}}{\lambda^k} e^{-(x/\lambda)^k}$ $x > 0, k > 0, \lambda > 0$	$\mu = \lambda \Gamma(1+1/k)$ $\sigma^2 = \lambda^2 (\Gamma(1+2/k) - (\Gamma(1+1/k))^2)$	$\sum_{n=0}^{\infty} \Gamma\left(1 + \frac{n}{k}\right) \frac{\lambda^n}{n!}$ $k \geq 1$

## Problem Solutions

### Solutions for Chapter 1

- 1.1** (a) Population: high school students. Sample: 1500 high school students. Statistic 47%. (b) Population: people living in the US. Parameter: 9.6%.
- 1.2** (a) Population of all offenders convicted of a misdemeanor or felony who were released in 2010. (b) Parameter.
- 1.4** (a) Patients played Tetris or kept a log. (b) No. (c) Yes, treatments were assigned randomly. (d) No, patients were recruited.
- 1.6** (a) Observational study. (b) No, we cannot establish causation. The participants were not randomized into two groups. (c) No, we cannot generalize to all urban American adolescents.
- 1.7** (a) Experiment. (b) Yes, participants were randomly assigned to treatments. (c) No, participants were recruited.
- 1.9**  $n/N$ .

### Solutions for Chapter 2

- 2.1**  $\ln(13.7143) = 2.6184 \neq 2.3877$ , but  $\log(15) = 2.70805 = m'$ .
- 2.3** (a) No. For example consider  $1 < 3 < 5$  with  $f(x) = x^2$ . (b) No. For example consider  $1 < 3 < 6$  with  $f(x) = (x - 3)^2$ . (c)  $f$  is linear. (d)  $f$  is increasing (or decreasing) and  $n$  is odd, or  $f$  is linear.

- 2.5** (a) Favor: 1385, Oppose: 808. (b) The `summary` command indicates that there were many non-responses. The `table` command does not give any indication that there were non-responses. (d) 74.3% of those who think the courts are not harsh enough favor the death penalty. 40.9% of those who think the courts are too harsh favor the death penalty.
- 2.9** (a) True.  $(w_1 + w_2 + \dots + w_n)/n = (n\bar{x} + n\bar{y})/n = \bar{x} + \bar{y}$ . (b) True.
- 2.11** Hint: Show that the pdf is symmetric about  $\theta$ :  $f(\theta - x) = f(\theta + x)$ .
- 2.12** (a) 1.085191, 14.3069. (b) -16.00965, 66.00965. (c) 46.58367.
- 2.13** (a)  $q_{0.05} = 0.0512/\lambda$ . (b) `qexp(0.05, 4)`.
- 2.16** No, the discrete nature of this random variable results in  $P(X \leq 1) = 0.007$  and  $P(X \leq 2) = 0.035$ .
- 2.19** (a) 3. (b) 4. (c) 6.
- 2.20** Around 60% versus 32%.

## Solutions for Chapter 3

- 3.1** (a)  $\bar{X}_t - \bar{X}_c = 11 - 7 = 4$ . (c)  $t = 4$  = observed statistic,  $P$ -value =  $2/10 = 0.2$ . (d) 0.2.
- 3.2** If you assume your professor is telling the truth, then the chance that a randomly drawn marble is red is  $1/10\,000$ . Either you are extremely lucky, or else there is reason to suspect your professor is wrong! With a  $1/10$  draw, it would be harder to tell your professor she is wrong!
- 3.5** (a) Observed test statistic 5.886. One simulation gives a  $P$ -value = 0.0002 so we conclude that the difference in mean delay times between the two carriers is statistically discernible. (b) Observed test statistic -5.663. One simulation gives  $P$ -value = 0.0002, so the difference in mean delay times between May and June is statistically discernible.
- 3.7** Code provided at <https://github.com/lchihara/MathStatsResamplingR>.
- 3.9** (a) The difference in proportions is statistically discernible. ( $P$ -value = 0.017). (b) The ratio of the variances is not 1 ( $P$ -value = 0.0374).

- 3.12** (b) Mean number of strikeouts in away games: 7.31; mean number of strikeouts in home games: 6.95. (c) Observed test statistic is 0.358.  $P$ -value = 0.39.
- 3.13** (b) No. The observations are not independent. Several actors and actresses have won more than once (for example, Anthony Hopkins and Francis McDormand).
- 3.15** 62.1% versus 50.8%.  $P$ -value = 0.284, so we cannot rule out chance variability.
- 3.17** With a two-sided  $P$ -value = 0.007, evidence supports the hypothesis that the ointment and soap treatments are effective.
- 3.20** (a) Vanilla and chocolate ice cream from the same manufacturer will not be independent (for instance, the two flavors will most likely use the same cream). (b) Vanilla 191.4 (58.6), Chocolate 198.7 (63.1). (c) One-sided  $P$ -value = 0.0006. Data supports claim that chocolate ice will have more of an impact on your waistline than vanilla.
- 3.22** (a) Matched pairs. (b) Left: 32.01 (2.40); right: 32.42 (2.56). (c) Two-sided  $P$ -value = 0.146 which suggests that mean pupillary distance (PD) measurements are the same for each eye.
- 3.23** Age is a possible confounder. Other answers possible.
- 3.26** Collider.

## Solutions for Chapter 4

- 4.1** There are 20 possible sets of size 3. The mean of the medians is 5.7. The median of the population is 5.5.
- 4.3** (a) {6, 8, 8, 9, 10, 10, 10, 11, 11, 12, 13, 14}. (b) No. (c)  $E[X] + E[Y] = 10.25$ .
- 4.9** 0.001.
- 4.11** 0.332.
- 4.13** 0.022.

**4.15** 91.**4.17** (a)  $N(15 - 2 \cdot 4, 3^2 + 2^2) = N(7, 5^2)$ . (b) Use `X <- rnorm(10^4, 15, 3)`, `Y <- rnorm(10^4, 4, 2)`. (c) 0.726.**4.19** (a)  $W = \bar{X} + \bar{Y} \sim N(36, 3.11^2)$ . (c) 0.901.**4.21** See Theorem B.15.**4.23** Simulated answers should be close to: (b) 10, 5. (c) 0.5.**4.25** (a) 13.86.**4.27** (a)  $f_{\max}(x) = 8(1/x^2 - 1/x^3)$ . (b) 1.545.**4.29**  $120e^{-12x}(1 - e^{-12x})^9$ .**4.30** (b) 0.0057.**4.33**  $P(X = k) = 30^k e^{-30}/k!$ ,  $k = 0, 1, \dots$ **4.35** (a) Yes. (c) 4.471.**4.37** (a) 0.849.

## Solutions for Chapter 5

**5.1** (a) and (c) are bootstrap samples.**5.2** Answers will vary.**5.3**  $3^3$  samples.**5.6** (a)  $N(36, 8^2/200)$ . (c) Bootstrap distribution using sample in (b): mean 35.91, se 0.53.**5.8** For odd  $n$ , median will be one of the sample points. Thus for small  $n$ , there will be only  $n$  possible values for the median, so the sampling distribution is much more “granular” than when  $n$  is even. As  $n$  increases, this becomes less apparent.**5.10** About (13.68, 23.23). Bias is about 10.8%.

- 5.13** (a) Bell-shaped, mean at about 5.196, SD at about 1.43. (b) (2.33, 8). (c) For the bootstrap distribution, we sample with replacement from the original samples – men from the men, and women from the women. With the permutation distribution, we sample assuming there is no difference in the means; in particular, we sample without replacement from the pooled data.
- 5.14** (b) bootstrap mean: 2.34, bootstrap SE 0.747. We are 95% confident that on average, the length of commercials on basic channels are from 0.89 to 3.84 min longer than on extended cable. (c) 0.0044, 0.0059, less than 0.6% of the standard error. (d)  $P$ -value = 0.0055 which is small enough to indicate that more than chance variability explains the difference in mean times.
- 5.17** (b) 0.149. (c) (0.126, 0.172).
- 5.19** ( $-0.075$ ,  $0.225$ ). Since the interval contains 0, we cannot rule out the possibility that the proportion of newborn baby girls in Alaska who weigh less than 2747 g is the same as the proportion of newborn baby girls in Wyoming who weigh less than 2747 g.
- 5.21** ( $-0.074$ ,  $-0.014$ ). We are 95% confident that the percentage of patients in the treatment group who will acquire the infection is from 1.4% to 7.4% lower than the percentage of patients in the control group.
- 5.23** (3.49, 11.49). (This is paired data so you should perform the bootstrap on the difference.)
- 5.25** (c) Bootstrap mean 1.63, SE 0.319. Right-skewed. (d) (1.17, 2.4). (e) 0.159.
- 5.27** (b) (1.46, 2.88). With 95% confidence, the mean protein content of meat dishes is from 1.46 to 2.88 times greater than the meat protein content of vegetarian dishes. (c) bias 0.045; 0.124 or over 12% of the bootstrap standard error.

## Solutions for Chapter 6

**6.3**  $\hat{\theta} = 4 / \left( \sqrt{5} + 3 + 3 + \sqrt{10} \right) = 0.351.$

**6.5** (a)  $\hat{\mu} = \bar{X}$ .  $\hat{\sigma} = \sqrt{(1/n) \sum_{i=1}^n (X_i - \mu)^2}.$

**6.7** Does not exist.

**6.9**  $\hat{N} = 9$ .

**6.11**  $5/\ln(576) = 0.7866$ .

**6.13**  $\hat{\lambda} = (n+m)/\left(\sum_{i=1}^n X_i + 2 \sum_{j=1}^m Y_j\right)$ .

**6.15** (a)  $\hat{\alpha} = n/\left(\sum_{i=1}^n x_i^\beta\right)$ . (b)  $n/\alpha = \sum_{i=1}^n x_i^\beta$  and  $\alpha \sum_{i=1}^n X_i^\beta \ln(x_i) - n/\beta = n \sum_{i=1}^n (\ln(x_i))$ .

**6.17**  $\hat{r} = 22.88, \hat{\lambda} = 3.138$ .

**6.19** Shape = 0.917, scale = 17.344.

**6.21** (a)  $\bar{X}/(\bar{X} - 2)$ .

**6.23**  $\hat{p} = 0.811, \hat{n} = 8.88$ .

**6.26**  $E[\bar{X}] = \theta + 1$ , so  $\bar{X}$  is biased.

**6.28**  $\sum_{i=1}^n a_i = 1$ .

**6.32** (a)  $-\sigma^2/n$ . (b)  $(\sigma^4/n^2) \cdot 2(n-1)$ . (c)  $2(n-1)\sigma^4/n^2 + (\sigma^2/n)^2$ .

**6.34**  $\text{Var}[\hat{\theta}_1] = \theta^2$ ,  $\text{Var}[\hat{\theta}_2] = (1/4)(2\theta^2) = (1/2)\theta^2$  and  $\text{Var}[\hat{\theta}_3] = (1/9)(\theta^2 + 2^2\theta^2) = (5/9)\theta^2$ , so  $\hat{\theta}_2$  is the most efficient and  $\hat{\theta}_1$  is the least efficient.

**6.37** The two curves approach each other.

**6.39** (a)  $2/(3\theta)$ . (b) Bias:  $-17/(27\theta)$ , MSE  $589/(1458\theta^2)$ . (c)  $27 T/10$ , where  $T = X_1/9 + X_2/9 + X_3/3$ .

**6.40** (d)  $\theta^2/((6n+1)(3n+1))$ .

**6.41** (a)  $f_{\max}(x) = (n\alpha x^{\alpha n-1})/(\beta^{\alpha n})$ .

**6.43** (d)  $(\pi - 4)/(4\lambda)$ .

**6.47** Is consistent.

## Solutions for Chapter 7

- 7.1** (a) Population mean here is not random. The probability that it is contained in the interval is 0 or 1. (c) Not 95% of the time, but 95% of the confidence intervals generated by each sample will contain true mean. (e) Each sample gives rise to a different confidence interval and 95% of these intervals will contain the true mean.
- 7.3** (a) (201.8, 218.2). (b) 97. (c) 166.
- 7.5**  $4n$ .
- 7.7** (a) (2.54, 3.64). (b) Skewness and/or outliers.
- 7.8** 118.01.
- 7.12** (28.34, 33.53) cm.
- 7.14** (b)  $(-533.61, -83.294)$ .
- 7.16** (b)  $(11.47, \infty)$ . We are 95% confident that on average, seedlings planted in fertilized plots grow at least 11.5 cm more than seedlings grown in non-fertilized plots.
- 7.18** (a) One weightlifter weighs over 300 lb! (b)  $(126.9, 163.5)$ . (c) Yes, the right endpoint decreases by around 20 lb.
- 7.21** This is a matched pairs setting! On average, chocolate ice cream has at least 3.84 more calories than vanilla.
- 7.23** Unbalanced sample sizes, pooled CI's are not capturing true mean difference at 95%. Balanced case, pooled CI does better.
- 7.24**  $[5.18, \infty)$ .
- 7.25** (a)  $(0.647, 0.705)$ . (b) Yes, the 95% confidence interval contains 50%  $(0.43, 0.51)$ .
- 7.27** (a) 1064. (b) 968.
- 7.29** (a)  $(0.07, 0.13)$ .

- 7.31** (b)  $(52.87, 103.29)$ . (c)  $(54.78, 105.01)$ . Bootstrap  $t$ :  $(56.84, 112.15)$ . Report the bootstrap  $t$ .
- 7.33** (a) 1497 non-smokers, 90 smokers. (c)  $(-27.68, 190.69)$ . Percentile:  $(-24.79, 187.13)$ . Bootstrap  $t$ :  $(-27.46, 190.27)$ . They are all roughly the same (note that 0 is in each interval). Report formula  $t$  or bootstrap  $t$ .
- 7.42** Yes.  $(\bar{X} - qS/\sqrt{n}, \bar{X} + qS/\sqrt{n})$ , where  $q$  denotes the 0.975-quantile of the  $t$  distribution with  $n - 1$  degrees of freedom.
- 7.43**  $(0.474/(2X), 11.143/(2X))$ .
- 7.48**  $(86.16, 559.21)$ .
- 7.52** (b) Hint: What property does  $x, x^3, e^x$  share?
- 7.53** (a)  $r/(n\lambda^2)$ . (b)  $f(x) = x^2$ , so  $f'(\mu) = 2\mu$ .  $\text{Var}[\bar{X}^2] \approx (2\mu)^2 r/(n\lambda^2)$ . (c)  $f(x) = \log(x)$ , so  $f'(\mu) = 1/\mu$ .  $\text{Var}[\bar{X}^2] \approx (1/\mu)^2 r/(n\lambda^2)$ . (d)  $f(x) = \sqrt{x}$ , so  $f'(\mu) = 1/(2\sqrt{\mu})$ .  $\text{Var}[\bar{X}^2] \approx (1/(4\mu))r/(n\lambda^2)$ . (e) Pick sample size  $n$  and parameters  $r$  and  $\lambda$ . Generate  $N$  samples, say  $N = 10^4$ ; for each compute  $\bar{x}$ ,  $\bar{x}^2$ ,  $\log \bar{x}$ , and  $\sqrt{\bar{x}}$ . Compute the variance of each set of  $N$  statistics, and compare. Double-check by repeating with other  $r$ ,  $\lambda$ , and  $n$ .
- 7.54** (a)  $\text{SE} = 0.49$ ,  $\bar{y}/\bar{x} = 1.96$ . (c) Using script given in text, 0.474.
- 7.55** (b)  $(0.156, 1.193)$ . (d)  $(0.192, 1.110)$ .

## Solutions for Chapter 8

- 8.1**  $P\text{-value} = 0.237$ , so we conclude that mean calcium levels are the same.
- 8.3**  $P\text{-value} = 0.007$ , so we conclude that on average, body temperatures are higher for children in Sodor.
- 8.5** (b) With  $t$ -test,  $P\text{-value} = 0.0226$ , so we would conclude that on average, walleye do weigh less than 2.5 lb. (c)  $P\text{-value} = 0.0542$ ; cannot rule out that average weight is the same.
- 8.7**  $t = 2.5991$ ,  $\text{df} = 13.85$ ,  $P\text{-value} = 0.0215$ .
- 8.9**  $P\text{-value} = 0.001$ ; with 95% confidence, we conclude that ales have 7.1 more calories on average than lagers.

**8.13** (a) Matched pairs. (c) One-sample  $t$ -test on difference:  $P$ -value = 0.132.  
 (d) Two-sample  $t$ -test:  $P$ -value = 0.122.

**8.15**  $P$ -value = 0.028. Evidence supports the hypothesis that the proportion of sex workers in Bamako who are HIV positive is greater than 0.391.

**8.17** Yes, statistically discernible,  $P$ -value = 0.041.

**8.18** (b) The percentage of patients on 30% inspired oxygen who got an infection was between 0.8% and 11.2% higher than the percentage of patients on 80% inspired oxygen. (c) We do not know if 30% inspired oxygen is better than no treatment at all.

**8.21**

		Two years later	
		No	Yes
Study start	No	$x$	28
	Yes	23	$y$

where  $x + y = 139$ . Since we are only interested in the children who switched, we do not need to know the specific  $x$  and  $y$  values. (Note though that we cannot compute the proportion of “no”s at the start of the study, or 2 years later.) If  $X \sim \text{Binom}(51, 0.5)$ , we want  $2P(X \geq 28)$ , which is approximately 0.576, so the intervention does not appear to be effective in changing the children’s eating of vegetables.

**8.23** (a) 47/655, 28/655. (b) *Hints:* This is paired data! To get a table, first create binary vectors. `mobile <- ifelse(MobileAds$m.cpc_post==0, 0, 1)`  $P$ -value = 0.014

**8.24** (a) 88/655, 114/655.

**8.27** (a) Type I error: we conclude that mean arsenic levels in the community are higher than 10 ppb even though it is in fact less than or equal to 10 ppb. The community may take unnecessary and expensive measures to lower the arsenic levels. Type II error: we conclude that mean arsenic levels in the community are less than or equal to 10 ppb even though in reality they are actually higher than 10 ppb. Community is drinking unsafe water and not taking any action.  
 (b) Not necessarily, since the  $t$  distribution has fatter tails.

**8.30**  $n \geq 30$ .

**8.32**  $n \geq 15$ .

**8.34**  $X \leq 22$  or  $X \geq 38$ .

**8.36** (a) 0.04. (b) 0.473.

**8.38** (a) 0.0497. (b) 0.548.

**8.40**  $1 - (3/4)^{\theta+1}$ .

**8.42**  $1 - \beta = P\left(Z < q_1 + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}}\right) + P\left(Z < q_2 - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right)$ .

**8.44** 0.003.

**8.48** (b) 0.224.

## Solutions for Chapter 9

**9.1**  $-1/100$ .

**9.3** 133.

**9.5** 87.

**9.9** (a) 0.49. (c) 0.91.

**9.12** Hint: Equation (9.12).

**9.13** (a) weight =  $20.078 + 1.607\text{height}$ . (b) 116.498. (c)  $R^2 = 0.5625$ ; 56% of the variability in weight is explained by the model.

**9.17** (a) Increasing variance. (c) Indicates that a linear model is appropriate.

**9.19** (b) Kills =  $1.736 + 0.947\text{Assts}$ . For every additional assist per set, there is associated a nearly one additional kill. About 93.7% of the variability in kills per set is explained by this model. (c) Yes, a straight-line model is appropriate.

**9.21** (a) 0.349. (b) Weight =  $-2379.69 + 148.995 \cdot \text{Gestation}$ . (c)  $R^2 = 0.122$ . (d) The problem is that  $\sigma$  is not constant for different gestation lengths (in part because there are so few data points at 42 weeks).

**9.23** (0.546, 1.198).

- 9.25** (b) Level =  $-3280 + 1.83\text{Year}$ . (c) No, there appears to be serial correlation.
- 9.27** 0.553; (0.368, 0.751).
- 9.29** (b) 0.23. (c)  $(-0.13, 0.54)$  which contains 0. (d) Yes independent,  $P\text{-value} = 0.151$ .
- 9.38** (a)  $\ln(\hat{p}/(1 - \hat{p})) = 2.905 - 0.037\text{Temperature}$ . (b)  $e^{-0.037(-5)} = 1.203$ . A 5 degree decrease in temperature increases the odds of an O-ring incident by a (multiplicative) factor of 1.203. Or, there is a 20% increase in the odds of an O-ring incident for every 5 degree drop in temperature. (d) The range of temperatures in the data set is from 53 to 81 degrees. We are extrapolating quite a bit at temperature 33 degrees.
- 9.40** (a)  $\ln(\hat{p}/(1 - \hat{p})) = -2.23 + 0.29\text{Hits}$ . (b)  $e^{0.29} = 1.34$ . Each additional hit increases the odds of winning by 34%. (d)  $(0.86, 0.98)$ . Answers will vary.

## Solutions for Chapter 10

- 10.1** (b) Test of independence. (d) Gender and recidivism are not independent ( $c \approx 28.1$ ).
- 10.3** (c)  $c = 10.408$ ,  $P\text{-value} = 0.001$ .
- 10.4** (b)  $c = 30.216$ ,  $df = 4$ .  $P\text{-value} = 0.0004$ . (c) Race/ethnicity and belief about morality are not independent.
- 10.5** (a) Independence. (b)  $P\text{-value} = 0.0725$ . Not enough evidence to support dependence. (c) Fisher's exact test  $P\text{-value} = 0.0405$ , permutation test  $P\text{-value} = 0.0389$ . (d) Class and preference for Instagram or Snapchat are not independent.
- 10.7** (a) This is a test of homogeneity. Let  $\pi_{ij}$  denote the proportion of fish from region  $i$  ( $i = 1, 2, 3$ ) and  $j$  rays, ( $j = \geq 36, 35, 34, 33, 32, \leq 31$ ). Then  $H_0$ :  $\pi_{1j} = \pi_{2j} = \pi_{3j}$  versus  $H_A$ : at least one pair of proportions not the same. (b)  $c = 12.803$ ,  $df = 10$ ,  $P\text{-value} = 0.23$ . We conclude there is not enough evidence to support the hypothesis that fin ray counts differ across the regions.

- 10.9** (a) Test of homogeneity. (b)  $c = 13.4621$ ,  $\text{df} = 1$ ,  $P\text{-value} = 0.0002$ . The difference between the two airlines is statistically discernible.
- 10.13** (a) Expected counts: 3.43, 11.57, 4.57, 15.42. Two are less than 5. (b)  $\binom{20}{2} \binom{15}{6} / \binom{35}{8} = 0.0404$ . (c)  $\sum_{k=0}^2 \binom{20}{k} \binom{15}{8-k} / \binom{35}{k} = 0.0461$ . (d) If the selection were completely random, then the chance of obtaining a committee consisting of 2 or fewer seniors is only 4.6%, which gives (mild) evidence that there is cause for suspicion.
- 10.15**  $c = 8.58$ ,  $\text{df} = 5 - 1 = 4$ ,  $P\text{-value}$  is 0.074. Yes, it is plausible that these numbers are drawn from a distribution with  $\text{pdf } f(x) = 2/x^3$ ,  $x \geq 1$ .
- 10.18** (a) Using the intervals: (0, 17], (17, 22], (22, 28], (28, 33], (33, 44], then  $c = 28.388$ ,  $\text{df} = 5 - 0 - 1 = 4$ ,  $P\text{-value} = 0$ . Test statistics may vary depending on the cut-offs used for the intervals. Conclude that the data are not consistent with a  $N(25, 10^2)$  distribution.
- 10.20** (a) 0.497. (b) Plausible that data comes from geometric distribution ( $c = 6.33$ ,  $\text{df} = 4$ ).
- 10.22**  $c = 3.632$ ,  $\text{df} = 4$ ,  $P\text{-value} = 0.458$ . The numbers appear to be drawn randomly.  $c$  will vary depending on the intervals used.
- 10.24** Recall that the estimated parameters are:  $\text{shape} = 0.917$ ,  $\text{scale} = 17.344$ .  $c = 14.217$ ,  $P\text{-value} = 0.047$ , so there is marginal evidence that times between successive earthquakes are not consistent with a Weibull distribution.
- 10.26**  $m \leq 4$  or  $m \geq 18$ .
- 10.30** The odds of a short pupil being bullied are 2.43 times higher than the odds of a not short pupil to be bullied.

## Solutions for Chapter 11

- 11.1** (a) Posterior column: 0.954; 0.5828; 0.3221. (b) After observing 4 wins out of 5, you now believe that there is a 32.2% chance that your long term probability of winning is 0.6. (c) 0.5; 0.5227.
- 11.3** (b) After observing 4 wins out of 7, you now believe that there is a 2.3% chance that your probability of winning is 0.2. (c) 0.5, 0.55.

**11.5** (a) (0.229, 0.272). (b) (0.253, 0.274). (c) 0.00016. (d) (0.247, 0.282).

**11.8** 0.55; 0.0012.

**11.10**  $\mu | \mathbf{x} \sim N(0.829, 0.0323^2)$ .

**11.12** (a)  $\mu | \mathbf{x} = 19.53$ ,  $\sigma | \mathbf{x} = 1.4797$ , precision 0.457. (c)  $\mu | \mathbf{x} = 19.14$ ,  $\sigma | \mathbf{x} = 1.531$ , precision 0.4267.

**11.14** (a) Not conjugate, so computations are difficult. (b) Is non-zero for  $\theta < 0$  and  $\theta > 1$ , values which are impossible.

**11.16** (b) Pareto distribution with parameters  $\alpha + N, \beta$ , (with  $\theta > \max\{\beta, X_1, X_2, \dots, X_N\}$ ). (c) 0.1748.

**11.18** (a)  $\theta^n \exp(-\theta \sum_{i=1}^n X_i)$ . (c) Gamma with parameters 14, 17.

## Solutions for Chapter 12

**12.1** (b)  $P\text{-value} = 5.5 \times 10^{-9}$ . The mean lengths are not the same across species.

**12.3** (b)  $P\text{-value} = 0.023$  so average age is not the same across races.

**12.4** (a) Delay times are not normally distributed. (b)  $P\text{-value} = 0.01$ . Mean flight delay times are not the same across the days of the week.

## Solutions for Chapter 13

**13.1** (a) The female distribution is centered about 9.3 and male distribution about 14.5, the means of the original data. The male distribution has a larger standard deviation, 1.16 versus 0.92; these numbers match  $s/\sqrt{n}$  for each dataset. Both smoothed bootstrap distributions are very close to normal. (d) Both are very close to normal, with the same mean  $-5.2$  (for female minus male). The smoothed bootstrap standard error is larger, 1.48 compared to 1.44. For comparison the usual formula standard error is 1.48.

**13.3** The distribution is close to normal, with slight positive skewness. The mean is 7.30, and standard deviation 6.9. A 95% bootstrap percentile interval is (6.01, 8.70).

- 13.7** (a) 0.9046; a confidence interval based on  $10^6$  replications, would be  $0.9046 \pm 2 \cdot 0.000125$ . (b) 0.9045242 with absolute error  $< 1e - 14$ .
- 13.11** Using  $N = 10^6$  gives approximately: (a)  $0.597 \pm 2 \cdot 0.00037$ . (b)  $0.597 \pm 2 \cdot 0.00025$ . (d) The second one gives more accurate estimates. The corresponding graph is flatter, the  $h$  values do not vary as much.
- 13.18**  $\hat{\lambda} = 0.29$ ,  $\hat{\beta} = 6.895$ .

## Solutions for Appendix B

**B.1** (a) 0.201; (b) 16.65.

**B.2** 0.000022.

**B.3** (a) 0.00138; (b) 16.65.

**B.10** `pchisq(10, 17)`.

## Bibliography

- Adrian, A., B. Holmes, and B. Stallsmith (2012). Impact of a gill parasite upon the minnow, *Notropis telescopus*. *Southeastern Naturalist* 11, 35–42.
- Agresti, A. (2012). *Categorical Data Analysis* (third ed.). New York: Wiley.
- Agresti, A. and B. Caffo (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *The American Statistician* 54(4), 280–288.
- Agresti, A. and B. Coull (1998). Approximate is better than “exact” for interval estimation of binomial proportions. *The American Statistician* 52(2), 119–126.
- Ahmed, A. and M. Hammarstedt (2009). Detecting discrimination against homosexuals: evidence from a field experiment on the internet. *Economics* 76(303), 588–597.
- Barnsley, R., A. Thompson, and P. Legault (1992). Family planning: football style. The relative age effect in football. *International Review for the Sociology of Sport* 27, 77–87.
- Bellman, R. (1966). *Adaptive Control Processes: A Guided Tour*. Princeton University Press.
- Benkouiten, S., D. Rezak, S. Badiaga, A. Veracx, R. Giorgi, D. Raoult, and P. Brouqui (2014). Effect of permethrin-impregnated underwear on body lice in sheltered homeless persons: a randomized controlled trial. *Journal of the American Medical Association Dermatology* 150, 273–279.
- Bickel, P. and K. Doksum (2001). *Mathematical Statistics* (second ed.), Volume 1. Upper Saddle River, New Jersey: Prentice-Hall.
- Bode, L., J. Kluytmans, H. Wertheim, D. Bogaers, C. Vandebroucke-Grausl, et al. (2010). Preventing surgical-site infections in nasal carriers of *Staphylococcus aureus*. *New England Journal of Medicine* 362, 9–17.
- Bombana, H., S. Bogstrand, H. Gjerde, et al. (2021). Use of alcohol and illicit drugs by trauma patients in Sao Paulo, Brazil. *Injury* 53, 30–36.
- Box, G. (1953). Non-normality and tests on variances. *Biometrika* 40, 318–335.
- Brashares, J., P. Arcese, M. Sam, P. Coppolillo, A. Sinclair, and A. Balmford (2004). Bushmeat hunting, wildlife declines, and fish supply in West Africa. *Science* 306(5699), 1180–1183.

- Brown, L., T. Cai, and A. DasGupta (2001). Interval estimation for a binomial proportion. *Statistical Science* 16(2), 101–133.
- Camill, P., L. Chihara, B. Adams, C. Andreassi, A. Barry, K. S. J. Limmer, M. Mandell, and G. Rafert (2010). Early life history transitions and recruitment of *Picea mariana* in thawed boreal permafrost peatlands. *Ecology* 2, 448–459.
- Carlin, B. and T. Louis (2009). *Bayesian Methods for Data Analysis* (third ed.). Boca Raton, FL: Chapman and Hall/CRC.
- Casella, G. and R. Berger (2001). *Statistical Inference*. Belmont, CA: Duxbury Press.
- Chan, K. (2008). Chinese children's perceptions of advertising and brands: an urban rural comparison. *Journal of Consumer Marketing* 25(2), 74–84.
- Chan, D., R. Ge, O. Gershony, T. Hesterberg, and D. Lambert (2010). Evaluating online ad campaigns in a pipeline: causal models at scale. In *KDD '10: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, pp. 7–16. ACM.
- Chance, B. and A. Rossman (2005). *Investigating Statistical Concepts, Applications, and Methods* (first ed.). Belmont, CA: Duxbury.
- Cochran, W. (1954). Some methods for strengthening the common  $\chi^2$  tests. *Biometrics* 10, 417–451.
- Collaborators, D. T., W. M. R. Collaborative, et al. (2017). Dexamethasone versus standard treatment for postoperative nausea and vomiting in gastrointestinal surgery: randomised controlled trial (dreams trial). *British Medical Journal* 357.
- Collett, D. (2003). *Modelling Binary Data* (second ed.). Chapman and Hall/CRC.
- DASL (1996). Data and Stories Library. <http://lib.stat.cmu.edu/DASL/> DataArchive.html. Carnegie Mellon University Statistics Department.
- Davison, A. and D. Hinkley (1997). *Bootstrap Methods and their Application*. Cambridge: Cambridge University Press.
- De Ridder, E., R. Pinxten, and M. Eens (2000). Experimental evidence of a testosterone-induced shift from paternal to mating behavior in a facultatively polygynous songbird. *Behavioral Ecology and Sociobiology* 49, 24–30.
- de Valpine, P., C. Paciorek, D. Turek, N. Michaud, C. Anderson-Bergman, F. Obermeyer, C. Wehrhahn Cortes, A. Rodriguez, D. Temple Lang, and S. Paganin (2021). *NIMBLE: MCMC, Particle Filtering, and Programmable Hierarchical Modeling*. R package version 0.12.1.
- Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B* 57, 1–38.
- Deng, A., Y. Xu, R. Kohavi, and T. Walker (2013). Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, pp. 123–132.
- Dobson, A. (2002). *An Introduction to Generalized Linear Models* (first ed.). Florida: Chapman and Hall/CRC.

- Draper, N. and H. Smith (1998). *Applied Regression Analysis* (third ed.). New York: Wiley.
- Edgeworth, F. (1885). Methods of statistics. *Journal of the Statistical Society of London Jubilee*, 181–217.
- Efron, B. and R. Tibshirani (1993). *An Introduction to the Bootstrap*. New York, NY: Chapman & Hall.
- Fan, R., T. Hesterberg, Y. Liu, and L. Zhang (2018). Methods for measuring brand lift of online Ads. In *JSM Proceedings, Statistics in Marketing Section*, Alexandria, VA, pp. 891–905. American Statistical Association.
- Fisher, R. A. (1922). On the interpretation of Chi-square from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society* 85, 87–94.
- Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.
- Gelman, A., J. Carlin, H. Stern, and R. Rubin (2010). *Bayesian Data Analysis* (second ed.). Florida: Chapman and Hall/CRC.
- Ghahramani, S. (2004). *Fundamentals of Probability with Stochastic Processes* (third ed.). New Jersey: Prentice Hall.
- Gilks, W., S. Richardson, and D. Spiegelhalter (1996). *Markov Chain Monte Carlo in Practice* (first ed.). London: Chapman and Hall/CRC.
- Goodman, L. (1952). Serial number analysis. *Journal of the American Statistical Association* 47, 622–634.
- Gregory, P. and R. Tasto (1976). Results of the Jack Mackerel subpopulation discrimination feasibility study. Technical report, California Department of Fish and Game. [http://aquacomm.fcla.edu/84/1/Marine\\_Resources\\_Administrative\\_Report\\_No\[1\]\\_76-2.pdf](http://aquacomm.fcla.edu/84/1/Marine_Resources_Administrative_Report_No[1]_76-2.pdf).
- Greif, R., O. Akca, E.-P. Horn, A. Kurz, and I. Sessler (2000). Supplemental perioperative oxygen to reduce the incidence of surgical-wound infection. *The New England Journal of Medicine* 342, 161–167.
- Griffith, G., T. Morris, M. Tudball, et al. (2020). Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nature Communications* 11, 1–12.
- Guinn, C., S. Baxter, W. Thompson, F. Frye, and C. Kopec (2002). Which fourth-grade children participate in school breakfast and do their parents know it? *Journal of Nutrition Education and Behavior* 34, 159–165.
- Hasumi, T., T. Akimoto, and Y. Aizawa (2009). The Weibull-log Weibull distribution for interoccurrence times of earthquakes. *Physics A: Statistical Mechanics and its Applications* 388, 491–498.
- Hershfield, D. (1971). The frequency of dry periods in Maryland. *Chesapeake Science* 12, 72–84.
- Hesterberg, T. (1988). *Advances in Importance Sampling*. Ph. D. thesis, Statistics Department, Stanford University.
- Hesterberg, T. (1991). Importance sampling for Bayesian estimation. In A. Buja and P. Tukey (Eds.), *Computing and Graphics in Statistics*, Volume 36 of *Volumes in Mathematics and Its Applications*, pp. 63–75. Springer-Verlag, Institute for Mathematics and Its Applications.

- Hesterberg, T. (1995). Weighted average importance sampling and defensive mixture distributions. *Technometrics* 37(2), 185–194.
- Hill, F., I. Mammarella, A. Devine, S. Caviola, M. Passolunghi, and Szűcs (2016). Maths anxiety in primary and secondary school students: gender differences, developmental changes and anxiety specificity. *Learning and Individual Differences* 48, 45–53.
- Iyadurai, L., S. Blackwell, R. Meiser-Stedman, P. Watson, M. Bonsall, R. Geddes, A. Nobre, and E. Holmes (2017). Preventing intrusive memories after trauma via a brief intervention involving Tetris computer game play in the emergency department: a proof-of-concept randomized controlled trial. *Molecular Psychiatry* 23, 674–682.
- Johnson, N. J. (1978). Modified t tests and confidence intervals for asymmetrical populations. *Journal of the American Statistical Association* 73, 536–544.
- Justus, C. G., W. R. Hargraves, A. Mikhail, and D. Gruber (1978). Methods for estimating wind speed frequency distributions. *Journal of Applied Meteorology* 17, 350–353.
- Kutner, M., C. Nachtsheim, J. Neter, and W. Li (2005). *Applied Linear Statistical Models* (fifth ed.). New York: McGraw Hill.
- Latter, O. (1902). An enquiry into the dimensions of the Cuckoo's egg and the relation of the variations to the size of the eggs of the foster-parent, with notes on coloration. *Biometrika* 1(2), 164–176.
- Llargues, E., R. Franco, A. Recasens, A. Nadal, M. Vila, M. Pérez, J. Manresa, I. Recasens, G. Salvador, J. Serra, E. Roure, and C. Castells (1979). Assessment of a school-based intervention in eating habits and physical activity in school children; the AVall study. *Journal of Epidemiology and Community Health* 65(10), 896–901.
- Lohr, S. (1991). *Sampling: Design and Analysis* (second ed.). Belmont, CA: Duxbury Press.
- Mahmud, M. A., M. Spigt, A.M. Bezabih, I. L. Pavon, G.-J. Dinant, and R. B. Velasco (2015). Efficacy of handwashing with soap and nail clipping on intestinal parasitic infections in school-aged children: a factorial cluster randomized controlled trial. *PLoS medicine* 12(6), e1001837.
- McCullagh, P. and J. Nelder (1989). *Generalized Linear Models* (second ed.). Florida: Chapman and Hall/CRC.
- McCullough, B. D. (2000). Is it safe to assume that software is accurate. *International Journal of Forecasting* 16, 349–357.
- McLachlan, G. and T. Krishnan (1997). *The EM Algorithm and Extensions* (first ed.). New York: Wiley.
- Miao, W. and P. Chiou (2008). Confidence intervals for the difference between two means. *Computational Statistics and Data Analysis* 52, 2238–2248.
- Moser, B. and G. Stevens (1992). Homogeneity of variance in the two-sample means test. *The American Statistician* 46, 19–21.

- Mukamal, K., L. Kuller, A. Fitzpatrick, W. Longstreth, M. Mittleman, and D. Siscovich (2003). Prospective study of alcohol consumption and risk of dementia in older adults. *Journal of the American Medical Association* 289, 1405–1413.
- Newcombe, R. (1998). Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in medicine* 17, 857–872.
- Pan, W. (2009). Approximate confidence intervals for one proportion and difference of two proportions. *Computational Statistics and Data Analysis* 40, 143–157.
- Pandey, S., S. Poudel, A. Gaire, R. Poudel, P. Subedi, J. Gurung, R. Sharma, and J. Thapa (2021). Knowledge, attitude and reported practice regarding donning and doffing of personal protective equipment among frontline healthcare workers against COVID-19 in Nepal: a cross-sectional study. *PLOS Global Public Health* 1(11), e0000066.
- Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. New York: Cambridge University Press.
- Pearl, J., M. Glymour, and N. Jewell (2016). *Causal Inference in Statistics: A Primer*. New York: Wiley.
- Pearson, K. (1900). On a criterion that a given system of deviations from the probably in the case of a correlated system of variables is such that it can be reasonably be supposed to have Arisen from random sampling. *Philosophical Magazine Series 5* 50, 157–175.
- Pearson, K. (1922). On the  $\chi^2$  goodness of fit. *Biometrika* 14, 186–191.
- Pearson, K. (1923). Further note on the  $\chi^2$  goodness of fit. *Biometrika* 14, 418.
- Pejtersen, J. H. (2020, 05). The effect of monetary incentive on survey response for vulnerable children and youths: a randomized controlled trial. *PLoS ONE* 15(5), 1–12.
- Pitman, J. (1993). *Probability*. New York: Springer.
- Primack, B., E. Douglas, and K. Kraemer (2010). Exposure to cannabis in popular music and cannabis use among adolescents. *Addiction* 105, 515–523.
- Resnick, B. (2019, March). Statistical significance and p-values explained. [posted online; 22-March-2019].
- Ricker, W. (1973). Linear regressions in fishery research. *Bulletin of Fisheries Research Board of Canada* 30, 409–433.
- Ricker, W. (1975). Computation and interpretation of biological statistics of fish populations. *Bulletin of Fisheries Research Board of Canada* 191, 382.
- Robert, C. and G. Casella (2004). *Monte Carlo Statistical Methods* (second ed.). New York: Springer.
- Robert, C. and G. Casella (2010). *Introducing Monte Carlo Methods with R*. New York: Springer.
- Romesburg, H. and K. Marshall (1979). Fitting the geometric distribution to capture frequency data. *Journal of Wildlife Management* 43, 79–84.

- Ronay, R. and W. von Hippel (2010). The presence of an attractive woman elevates testosterone and physical risk taking in young men. *Social Psychological and Personality Science* 1, 57–64.
- Ross, S. (2009). *A First Course in Probability* (eighth ed.). New Jersey: Prentice Hall.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66(5), 688–701.
- Samaha, F. F., N. Iqbal, P. Seshadri, et al. (2009). A low-carbohydrate as compared with a low-fat diet in severe obesity. *New England Journal of Medicine* 348(21), 2074–2081.
- Sandler, R. S., S. Halabi, J. A. Baron, B. S. Bandinger, et al. (2003). A randomized trial of aspirin to prevent colorectal adenomas in patients with previous colorectal cancer. *New England Journal of Medicine* 348(10), 883–890.
- Scheaffer, R. and L. Young (2010). *Introduction to Probability and its Applications* (third ed.). Boston, MA: Brooks/Cole, Cengage Learning.
- Scheffe, H. (1970). Practical solutions of the Behrens-Fisher problem. *Journal of the American Statistical Association* 65, 1501–1508.
- Schenker, N. and J. Gentleman (2001). On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician* 55, 182–186.
- Seguro, J. V. and T. W. Lambert (2000). Modern estimation of the parameters of the Weibull wind speed distribution for wind energy analysis. *Journal of Wind Engineering and Industrial Aerodynamics* 85, 75–84.
- Siegfried, D., D. Linville, and D. Hille (2010). Analysis of nest sites and the resplendent quetzal (*pharomachrus mocinno*): relationship between nest and snag heights. *Wilson Journal of Ornithology* 122, 608–611.
- Smith, S., N. Weissman, C. Anderson, M. Sanchez, E. Chuang, et al. (2010). Multicenter, placebo-controlled trial of lorcasirten for weight management. *New England Journal of Medicine* 363, 245–256.
- Smith, T., P. Mardsen, M. Hout, and J. Kim (1972–2014). General social survey. Sponsored by the NSF.
- Stuart, A. and K. Ord (2009). *Kendall's Advanced Theory of Statistics, Volume 1: Distribution theory* (sixth ed.). New York: Wiley.
- Suess, E. and B. Trumbo (2010). *Introduction to Probability Simulation and Gibbs Sampling with R*. New York: Springer.
- Ten Hwang, Y., S. Larivière, and F. Messier (2005). Evaluating body condition of striped skunks using non-invasive morphometric indices and bioelectrical impedance analysis. *Wildlife Society Bulletin* 33(1), 195–203.
- Thomas, D., J. Apps, R. Hoffmann, M. McCrea, and T. Hammeke (2015). Benefits of strict rest after acute concussion: a randomized controlled trial. *Pediatrics* 135, 213–223.
- Tiampo, K., D. Weatherley, and S. Weinstein (2008). *Earthquakes: Simulations, Sources and Tsunamis*. Pageoph Topical Volumes. Birkhauser.

- Tippett, L. H. C. (1952). *The Methods of Statistics* (fourth ed.). New York: Wiley.
- Trafimow, D. and M. Marks (2015). Editorial. *Basic and Applied Social Psychology* 37(1), 1–2.
- Tukey, J. (1977). *Exploratory Data Analysis* (first ed.). Addison Wesley.
- van Linschoten, R., M. van Middelkoop, M. Y. Berger, E. M. Heintjes, J. A. Verhaar, S. P. Willemsen, B. W. Koes, and S. M. Bierma-Zeinstra (2009). Supervised exercise therapy versus usual care for patellofemoral pain syndrome: an open label randomised controlled trial. *BMJ* 339.
- Verzani, J. (2010). *UsingR: data sets for the text “Using R for Introductory Statistics”*. R package version 0.1-12.
- Voss, L. and J. Mulligan (2000). Bullying in school: are short pupils at risk? Questionnaire study in a cohort. *British Medical Journal* 320, 612–613.
- Wang, C., C. Schmid, P. Hibberd, R. Kalish, R. Roubenoff, R. Rones, and T. McAlindon (2009). Tai Chi is effective in treating knee osteoarthritis: a randomized controlled trial. *Arthritis Care and Research* 61, 1545–1553.
- Wardrop, R. (1995). *Statistics: Learning in the Presence of Variation* (first ed.). Dubuque, IA: W. C. Brown Publishers.
- Wasserstein, R. and N. Lazar (2016). The ASA’s statement on P-values: context, process, and purpose. *The American Statistician* 70(2), 129–133.
- Wasserstein, R., L. Schirm, and N. Lazar (2019). Moving to a world beyond “ $p < 0.05$ ”. *The American Statistician* 73, 1–19.
- Weisberg, S. (2005). *Applied Linear Regression* (third ed.). New York: Wiley.
- Weisser, D. (2003). A wind energy analysis of Grenada: an estimation using the ‘Weibull’ density function. *Renewable Energy* 28, 1803–1812.
- Westman, E., W. Yancy, J. Mavropoulos, M. Marquart, and J. McDuffie (2008). The effect of a low-carbohydrate, ketogenic diet versus a low-glycemic index diet on glycemic control in type 2 diabetes mellitus. *Nutrition and Metabolism* 5. BioMed Central, The Open Access Publisher, <https://doi.org/10.1186/1743-7075-5-36>.
- Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* 22, 209–212.
- Witmer, J. (2019). Editorial. *Journal of Statistics Education* 27(3), 136–137.
- Wu, C. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics* 11, 95–103.
- Xu, P., X. Song, W. Wang, F. Wang, L. Cao, and Q. Liu (2012). Seroprevalence of *Toxoplasma gondii* infection in chickens in Jinzhou, northeastern China. *Journal of Parasitology* 98, 1300–1301.
- Yang, N., M. Mu, J. Hu, W. Gao, S. Yang, and J. He (2013). Seroprevalence of *Toxoplasma gondii* infection in pet dogs in Shenyang, northeastern China. *Journal of Parasitology* 99, 176–177.
- Zhou, J., E. Erdem, G. Li, and J. Shi (2010). Comprehensive evaluation of wind speed distribution models: a case study for North Dakota sites. *Energy Conversion and Management* 51, 1449–1458.



## Index

### **a**

- Acceptance region 274
- Agresti-Coull interval 214–215
- Alpha level 264
  - Significance Level 277
- Alternative hypothesis 46
- ANOVA 429
  - permutation test approach 438
- autocorrelation 192
- Average treatment effect 458

### **b**

- Barchart 21
- Bayes 399–421
  - Bayes Theorem 400
  - computational issues 462–464
  - importance sampling 478–482
- Bellman, Richard 463
- Bernoulli distribution 505–506
- Beta distribution 522–523
  - as prior for binomial data 410
- Bias 12, 126–129, 161
- Binomial distribution 505–506
- Blinding 12
- Bonferroni correction 280
- Bootstrap 103
  - accuracy 127, 131–136
  - bias 112, 127
  - bias for parametric bootstrap 451
  - bootstrap distributions and sampling distributions 107

### **bootstrapping idea, the** 105

- center, bias, spread, shape 107
- distribution too narrow 447
- estimate cdf with ecdf 111
- how many resamples? 136
- parametric 449–452
- percentile interval 115–116
- vs. permutation test 119
- ratio of means 123
- regression 342
- relative risk 128
- sample rows of the data 342
- single population 104
- smoothed 444–449
- standard error 105
- stratified 453
- t* confidence interval 217–224
- t*-test 250–251
- two populations 116, 117
- variation of bootstrap distribution 131, 136

### **Z interval** 241

### **Boxplot** 27

### **c**

- $\chi_m^2$  372
- Case studies
  - Beer and hot wings 9, 35, 36, 50–54, 59, 140, 360, 488
  - Birth weights of babies 2, 29, 74, 103, 191, 220, 252, 362

- Case studies (*contd.*)
- Black spruce seedlings 10, 41, 72, 233, 234, 296, 309–328
  - Bushmeat 346–350
  - Flight delays 1, 21, 40, 71, 72, 95, 144, 232, 234, 238, 392, 440
  - General Social Survey (GSS) 7, 40, 213, 367–372, 375, 392
  - Google mobile ads 13, 41, 175, 238, 289, 301
  - Recidivism 4, 40, 63, 72, 74, 95
  - Verizon repair times 3, 55–58, 60–62, 120, 123, 220
  - Wind energy 156, 166, 449
- Cauchy distribution 42, 154, 171, 284
- Cause and effect 11, 67
- Cell means model 430
- Census 8
- Center 23, 25
- Central Limit Theorem 85
- accuracy 92
  - binomial data 88
  - finite population 93
- Chebyshev's Inequality 501
- Chi-square
- test of goodness-of-fit 382–387
  - test of homogeneity 380–382
  - test of independence 371–380
  - test statistic 369, 374
- Chi-square distribution 332, 371, 517
- sum of chi-square random variables 518
- Coefficient of variation 140
- Conditioning 379
- Confidence interval
- Agresti-Coull 214
  - and hypothesis tests 273
  - bootstrap  $t$  vs. formula  $t$  223
  - for difference of two proportions 215–216
  - for difference of two means 195–200
- in general 201–205
  - for one mean ( $\sigma$  known) 183–188
  - for one mean ( $\sigma$  unknown) 188–195
  - for one proportion 211–215
  - one sample proportion exact interval 275
  - one-sided 209–211
  - properties 224
  - score 212
  - $t$  confidence interval 190, 219, 220
  - Wald 214
  - $z$  187
- Confounding 68
- Conjugate family 410
- Consistency 169–171
- Contingency table 22, 367
- Continuity correction 90–92, 249–250, 257–258
- Yates 377
- Control group 12
- Control variates 455–462
- Convergence in probability 170
- Correlation 313
- bootstrap 342
  - coefficient 313
  - permutation test of independence 345–346
  - sample 315
- Covariance 309–313
- Cramer-Rao Inequality 166
- Credible interval 411
- Critical
- region 265, 272
  - value 265
- Cumulative distribution function 494
- Curse of dimensionality 463

**d**

Data sets

- Alelager 296, 363
- Bangladesh 112, 139, 219, 222, 237, 250, 488

- BookPrices 143  
 Cafeteria 73, 144  
 Cereals 391  
 Challenger 365  
 ChiMarathonMen 43, 439  
 Cuckoos 439  
 Diving2017 66, 143, 201  
 Eyes 75, 297, 364  
 Fatalities 350–357  
 FishMercury 140  
 Girls2004 141, 233  
 Groceries 75, 142, 234  
 IceCream 75, 143, 235  
 ILBoys 429, 436  
 Illiteracy 360, 363  
 Lottery 396  
 MathAnxiety 141, 364  
 Maunaloa 363  
 MnGroundwater 139, 141, 238  
 Nasdaq 295  
 NBA1617 359  
 Olympics2012 233, 234, 358  
 Oscars 73  
 Phillies2009 73, 366, 395  
 Quakes 178, 488  
 Quetzal 296, 360  
 RangersTwins2016 358  
 Salaries 296  
 Service 178  
 Skateboard 117  
 Skating2010 329, 335, 339, 346  
 Starcraft 439  
 Titanic 366  
 TV 140  
 TXBirths2004 238  
 Volleyball2009 361  
 wafers 436  
 Walleye 295, 363  
 Watertable 365  
 Data snooping 279  
 Degrees of freedom  
   chi-square 372  
 Welch's approximation 196, 242  
 Delta method 226–230  
 Density  
   conditional 496  
   marginal 496  
 Density estimate  
   kernel 444  
 Distribution, see specific  
   distributions  
 Dot plot 24
- e**
- Edgeworth approximation 92  
 Edgeworth, Francis, Y. 277  
 Effect size 271  
 Efficiency 164–167  
   relative 471, 472  
 Efron, Bradley 136  
 EM algorithm 483–488  
 Empirical cumulative distribution  
   function 34  
 Estimate 150  
 Estimator 78, 150  
 European stock option 470–472,  
   474  
 Exhaustive calculation 51  
 Expected value 494  
 Experiment 11  
 Experimental units 11  
 Exploratory data analysis 21  
 Exponential distribution 513–514  
   rate parameter 514  
   scale parameter 513
- f**
- F* statistic 434  
*F* distribution 434, 523–524  
 Fisher information 166, 173  
 Fisher's exact test 378–379  
 Fisher, Ronald, A. 264, 277, 372

Fitted value 318  
 Five-number summary 26

***g***

*G* statistic 388  
 Galton, Francis 321  
 Gamma distribution 514–517  
     sum of gamma random variables 517  
 Geometric distribution 508–509  
 German tank problem 204, 241  
 Goodness-of-fit test  
     all parameters known 382–385  
     some parameters estimated 385–387

Google Analytics Content Experiments 280, 418

***h***

Histogram 23  
 Homogeneity, test of 380–382  
 Human evaluation 455  
 Hypergeometric distribution 378, 510–511  
 Hypothesis  
     composite 281  
     simple 281  
 Hypothesis test 46  
     one sample mean *t*-test 246–247  
     one sample mean *z* test 245–246  
     one sample proportion *z* test 248–250  
     two sample means *t*-test 252–255  
     two sample proportions *z* test 255–261  
     two-sided 46

***i***

iid 499  
 Importance sampling 468–482  
     design distribution 470  
     importance function 470  
     target distribution 470

Improper prior 407  
 Independence  
     chi-square test of 371–380  
     permutation test of 369  
     of random variables 495

Indicator function 476  
 Inference 8  
 Interquartile range 26

***j***

Joint density 495

***k***  
 Kurtosis 38

***l***

Law of Averages 500  
 Law of Total Probability 493

Likelihood  
     complete data 485  
     Bayesian 400  
     function 148, 149, 151  
     generalized ratio test 285–288  
     incomplete data 485  
     ratio and chi-square 388–389  
     ratio test 281–285  
     ratio test statistic 282, 285

Linear model  
     assumptions 330  
     conditions 330, 340–342  
     inference for response 336–340  
     inference for slope and intercept 333–336

Location family 205

Location parameter 205

Location-scale family 208

Logistic regression 350–357  
     inference 355–357

***m***

Margin of error 187  
 Matched pairs 66–67, 122  
     median 143

- paired *t*-test 201, 254  
proportions test 259  
Maximum likelihood estimate 149,  
151  
Maximum likelihood estimation  
asymptotic bias 163  
asymptotic normality 173  
continuous variables 150–155  
discrete variables 148–150  
multiple parameters 155–158  
unbiased estimator 161–163  
McNemar's test 259  
Mean 25  
midmean 25  
of a sample 498  
trimmed 25  
Mean absolute deviation 26  
Mean square error 167–169, 434  
Mean square for treatments 434  
Median 25  
Method of moments 158–160  
Moment 501  
central 38, 501  
Moment generating function 502  
Monte Carlo  
integration 466  
sampling 130  
Multinomial distribution 373,  
506–508  
Multiple testing 279
- n**  
Natural variability 22  
Negative binomial distribution  
509–510  
Neyman–Pearson Lemma 282  
Normal distribution 85, 497  
MLE for multiple parameters 155  
as prior for continuous data 415  
sum of squared normal random  
variables 518  
sums of normal random variables  
499
- Normal quantile plot 29, 31  
Null distribution 48, 245  
Null hypothesis 46
- o**  
Observation 1  
Observational study 11  
Odds 351  
Odds ratio 352, 378, 397  
Order statistics 83, 100  
Outlier 192, 248, 324
- p**  
*P*-value 47, 245, 277  
vs. critical region 272–273  
one-sided vs two-sided 58  
for permutation test 51  
two-sided 59  
Parameter 6  
Pearson, Karl 369, 372  
Permutation distribution 52  
Permutation test 50, 119  
correlation 345–346  
conditions 62  
of independence 369–372  
two sample 52  
Placebo 12  
Plug-in principle 109  
Poisson distribution 150, 511–513  
sum of Poisson random variables  
512  
Population 5  
finite 5  
infinite 5  
standard deviation 26  
variance 26  
Post-stratification 453  
Posterior distribution 400, 407  
Power 267–272  
Precision 416  
Predicted value 318  
Prediction interval 338

- Prior distribution 400  
 conjugate family 410  
 flat 410  
 improper 407  
 noninformative 410, 416
- Probability density function 494
- Probability mass function 493
- q**  
 $q_p$  29  
 Quantile 29  
 Quantile-quantile plot 33
- r**  
 $\rho(X, Y)$  313  
 $r$  315  
 $r^2$  323  
 R-square 323  
 Random assignment 12  
 Random variable  
   continuous 493  
   discrete 493  
 Range 26  
 Reference distribution 48, 245  
 Regression  
   least-squares 319  
   logistic 350  
   models 321  
   multiple 328  
   toward the mean 320–321  
 Relative risk 129, 239  
 Replication crisis 276  
 Resample  
   permutation 51  
 Residual 323  
   conditions in regression 341–342  
   plot 323  
   standard error  $S$  332
- s**  
 $S$  162, 332  
 $ss_x, ss_y, ss_{xy}$  318  
 $s$  26
- Sample 5  
 finite population 5  
 probability 8  
 random 5  
 with replacement 5  
 without replacement 5
- Sample standard deviation 26  
 as estimate of  $\sigma$  188  
 $S^2$  is unbiased estimate of  $\sigma^2$  162
- Sampling  
 cluster 9  
 multi-stage 9  
 optimal stratified 454  
 stratified 9, 452–455
- Sampling distribution 77, 78  
 by calculating 82  
 of the maximum of a sample 82  
 of the minimum of a sample 82  
 by simulation 79, 81
- Sampling frame 8
- Scale family 207
- Scale parameter 207
- Scatter plot 36  
 Scatter plot smooth 326
- Sequential data 417
- Sidak correction 279
- Signal-to-noise 6
- Skewness 38  
 and  $t$  confidence intervals 194, 198, 224
- Smoothing spline 326
- Spread 23
- Standard deviation 26  
 of a random variable 494
- Standard error 77, 78  
 residual 332
- Statistic 6
- Statistical practice 13, 174–176, 289–294
- Statistically discernible 47  
 Lack of 278
- Statistical significance 277  
 vs. practical importance 277

Strong law of large numbers 500  
 Student's *t* distribution, *see t*  
     distribution 189  
 Subjects 11  
 Sum of squares 318  
     error 431  
     total 431  
     treatment 431  
 Surveys 8

**t**

*t* distribution 189, 334, 520–522  
     noncentral 272, 305  
*t* statistic  
     one sample mean 189, 247  
     two sample means 196, 254  
 Test statistic 47  
     choice of 54  
 Training data 455  
 Transformation invariance 171–173,  
     225, 230  
 Translated exponential distribution  
     474  
 Treatments 11  
 Tukey, John 21  
 Type I error 261–267  
 Type II error 261, 267–272

**u**

Ulam, Stanislaw 130  
 Unbiased  
     estimator 161–163  
     statistic 127  
 Undercoverage 8  
 Uniform distribution 152, 153, 513  
     density of the maximum 83  
     density of the minimum 83

**v**

Variable 1  
     categorical 21  
     lurking 496  
 Variance  
     pooled 200, 255  
     of a random variable 494  
 Verizon 125  
 von Neumann, John 130

**w**

Weibull distribution 156, 449  
 Weights 175–176  
 Welch's approximation 196, 242  
 Wilson interval 214



# **WILEY END USER LICENSE AGREEMENT**

Go to [www.wiley.com/go/eula](http://www.wiley.com/go/eula) to access Wiley's ebook EULA.