

Margin of Error

Statistics for a World That Needs Better Decisions

Vivek H. Patil, Ph.D.

March 2026

Table of contents

Margin of Error	1
Statistics for a World That Needs Better Decisions	1
Preface	5
About the Author	5
Preface	6
How to Use This Book	8
Chapter Structure	8
The Companion Website	8
Errata and Feedback	9
For Instructors	9
For Students	9
What to Expect	9
A Note on AI Tools	10
I Foundations	11
1 Why Statistics Matters Now	13
1.1 The Water Was Poisoned	13
1.2 What Statistics Is (And What It Is Not)	16
1.2.1 Statistics Is Not Math (Exactly)	17

1.2.2	The Three Big Questions	17
1.2.3	Statistics in Everyday Decisions	18
1.3	Types of Data: The Raw Material	19
1.3.1	Variables	19
1.3.2	Categorical Variables	19
1.3.3	Numerical Variables	20
1.3.4	Identifying Variable Types in Practice	20
1.3.5	Why This Matters	21
1.3.6	The “It Has Numbers So It Must Be Numerical” Trap	22
1.4	Where Data Comes From (And Why You Should Care)	23
1.4.1	Populations and Samples	24
1.4.2	Observational Studies vs. Experiments	24
1.4.3	A Business Example	25
1.5	Data Quality: Garbage In, Garbage Out	26
1.5.1	Common Data Quality Problems	26
1.5.2	A Cautionary Tale: The Spreadsheet Error That Changed History	27
1.5.3	The First Rule of Data Analysis	27
1.6	The AI Question	28
1.6.1	What AI Can and Cannot Do	30
1.6.2	Using AI as a Learning Tool (Not a Crutch)	31
1.7	Reading the News Statistically	31
1.8	Looking Ahead	32
1.9	Further Reading and References	32
1.10	Key Terms	33
1.11	Exercises	34
1.11.1	Check Your Understanding	34
1.11.2	Apply It	35
1.11.3	Think Deeper	37

2	Asking Good Questions: Research Design	39
2.1	The Experiment You Never Agreed To	39
2.2	Two Kinds of Studies	40
2.2.1	Observational Studies	41
2.2.2	Experiments	42
2.2.3	The Spectrum Between Observation and Experiment	43
2.3	How to Choose Your Sample	44
2.3.1	Simple Random Sampling	44
2.3.2	Stratified Sampling	45
2.3.3	Cluster Sampling	46
2.3.4	Systematic Sampling	46
2.3.5	Convenience Sampling and Its Problems	47
2.4	Bias: When Data Leads You Astray	49
2.4.1	Selection Bias	49
2.4.2	Response Bias	50
2.4.3	Non-Response Bias	51
2.4.4	Survivorship Bias	51
2.5	Confounding: The Hidden Third Variable	52
2.5.1	More Examples of Confounding	54
2.5.2	Why Confounding Makes Causation Hard	55
2.6	When Can We Claim Causation?	56
2.6.1	The Experimental Path	56
2.6.2	The Observational Path	57
2.7	The Ethics of Research Design	58
2.7.1	Informed Consent	58
2.7.2	Power and Data Collection	59
2.7.3	Ethical Obligations at Every Stage	60
2.8	Putting It All Together: Evaluating a Study	62
2.9	Looking Ahead	63

2.10	Key Terms	63
2.11	Further Reading and References	65
2.12	Exercises	66
2.12.1	Check Your Understanding	66
2.12.2	Apply It	67
2.12.3	Think Deeper	69
3	Summarizing Data with Numbers	71
3.1	Two Economies, One Country	71
3.2	Measures of Center	73
3.2.1	The Mean	73
3.2.2	The Median	73
3.2.3	The Mode	74
3.2.4	When to Use Which	74
3.3	Why Measures of Center Are Not Enough	76
3.4	Measures of Spread	77
3.4.1	The Range	77
3.4.2	The Interquartile Range (IQR)	77
3.4.3	Variance and Standard Deviation	78
3.4.4	Population vs. Sample Notation	80
3.4.5	Interpreting the Standard Deviation	80
3.5	Percentiles, Quartiles, and the Five-Number Summary	81
3.5.1	Percentiles	81
3.5.2	Quartiles	82
3.5.3	The Five-Number Summary	82
3.5.4	Describing the Income Distribution with Percentiles	85
3.6	Outliers	86
3.6.1	What Is an Outlier?	86
3.6.2	Detecting Outliers with the IQR Method	87

3.6.3	The Ethics of Outlier Removal	88
3.6.4	How Outliers Affect Summary Statistics	89
3.7	Putting It All Together	89
3.8	A Note on Income Data	92
3.9	Choosing the Right Summary Statistics	92
3.10	Looking Ahead	93
3.11	Key Terms	93
3.12	Further Reading and References	94
3.13	Exercises	95
3.13.1	Check Your Understanding	95
3.13.2	Apply It	96
3.13.3	Think Deeper	98
4	Summarizing Data with Pictures	101
4.1	The Maps That Showed What Everyone Already Knew	101
4.2	Principles of Effective Visualization	104
4.3	Histograms and Density Plots	105
4.3.1	Histograms	105
4.3.2	Density Plots	106
4.4	Box Plots	108
4.5	Bar Charts	110
4.5.1	When to Use Bar Charts	111
4.5.2	Dos and Don'ts	111
4.5.3	A Special Case: Pie Charts	112
4.6	Scatter Plots	113
4.7	Correlation	116
4.7.1	Pearson's Correlation Coefficient	117
4.7.2	Interpreting Correlation	118
4.7.3	Building Your Correlation Intuition	119

4.8	The Correlation-Causation Trap	120
4.8.1	Spurious Correlations	121
4.8.2	Why We Keep Falling for It	122
4.9	Distortion as an Ethical Failure	123
4.9.1	Truncated Axes	123
4.9.2	Cherry-Picked Time Ranges	124
4.9.3	Manipulated Aspect Ratios	125
4.9.4	Misleading Area and Volume Comparisons	126
4.9.5	The Dual-Axis Trap	126
4.10	Putting It All Together	129
4.11	Looking Ahead	130
4.12	Key Terms	130
4.13	Further Reading and References	131
4.14	Exercises	133
4.14.1	Check Your Understanding	133
4.14.2	Apply It	134
4.14.3	Think Deeper	137

II Probability and Inference 139

5 Probability: Thinking About Chance 141

5.1	The Test Said Yes	141
5.2	What Probability Means	143
5.2.1	The Language of Probability	143
5.3	Basic Probability Rules	144
5.3.1	The Complement Rule	144
5.3.2	The Addition Rule	145
5.3.3	The Multiplication Rule and Independence	146
5.4	Conditional Probability and Two-Way Tables	148

5.4.1	Conditional Probability	148
5.4.2	Working with Two-Way Tables	148
5.4.3	Testing for Independence with Two-Way Tables	150
5.5	Bayes' Theorem	151
5.5.1	The Problem of Inverse Probability	151
5.5.2	Why Natural Frequencies Work Better	152
5.5.3	The Tree Diagram Approach	154
5.5.4	What Happens When the Base Rate Changes	155
5.6	Random Variables and Expected Value	158
5.6.1	Random Variables	158
5.6.2	Probability Distributions	158
5.6.3	Expected Value	159
5.6.4	Variance and Standard Deviation of a Random Variable	160
5.7	The Binomial Distribution	161
5.7.1	When Binomial Applies	161
5.7.2	The Binomial Formula	161
5.7.3	A Worked Example	162
5.7.4	Mean and Standard Deviation of the Binomial	163
5.7.5	Building Intuition	163
5.7.6	When the Binomial Does Not Apply	165
5.8	Tying It Together	165
5.9	Looking Ahead	166
5.10	Further Reading and References	166
5.11	Key Terms	168
5.12	Exercises	169
5.12.1	Check Your Understanding	169
5.12.2	Apply It	170
5.12.3	Think Deeper	172

6	The Normal Distribution and the Central Limit Theorem	175
6.1	Every Baby Gets Weighed	175
6.2	The Normal Distribution	176
6.2.1	Shape, Center, and Spread	176
6.2.2	The Density Curve	177
6.3	Z-Scores	178
6.3.1	Standardizing Values	178
6.3.2	Why Z-Scores Matter	179
6.3.3	From Z-Scores to the Standard Normal	181
6.4	The Empirical Rule	181
6.5	Normal Probability Calculations	183
6.5.1	Using Z-Tables	183
6.5.2	Common Probability Questions	183
6.5.3	Finding Values from Probabilities (Working Backward)	184
6.6	Assessing Normality	185
6.6.1	QQ Plots	185
6.6.2	The Shapiro-Wilk Test	187
6.7	From Individual Values to Sample Means	189
6.8	Sampling Distributions	190
6.8.1	The Concept	190
6.8.2	Properties of the Sampling Distribution	190
6.9	Standard Error vs. Standard Deviation	191
6.10	The Central Limit Theorem	193
6.10.1	The Statement	193
6.10.2	Why This Result Matters	193
6.10.3	How Large Is “Large Enough”?	195
6.10.4	A Visual Walkthrough	196
6.10.5	The CLT in Action	197
6.10.6	What the CLT Does Not Say	198

6.11 Putting It All Together	198
6.12 Looking Ahead	199
6.13 Further Reading and References	200
6.14 Key Terms	201
6.15 Exercises	202
6.15.1 Check Your Understanding	202
6.15.2 Apply It	202
6.15.3 Think Deeper	205
7 Confidence Intervals	207
7.1 The Night the Polls “Failed”	207
7.2 A Single Number Is Never Enough	208
7.3 The Logic of Confidence Intervals	209
7.4 Confidence Intervals for Means	210
7.4.1 The Standard Error	211
7.4.2 Why t Instead of z	211
7.4.3 Degrees of Freedom	212
7.4.4 A Worked Example	212
7.4.5 Conditions for the t -Interval	213
7.5 Confidence Intervals for Proportions	213
7.5.1 Where the Formula Comes From	214
7.5.2 The Formula	214
7.5.3 A Worked Example	215
7.5.4 Conditions for the Proportion Interval	215
7.5.5 A Practical Illustration	215
7.6 What “95% Confident” Actually Means	216
7.6.1 Why This Distinction Matters	218
7.6.2 Common Misinterpretations (and Corrections)	219
7.7 What Makes Confidence Intervals Wider or Narrower	220

7.7.1	Confidence Level	220
7.7.2	Sample Size	220
7.7.3	Variability in the Data	221
7.7.4	A Summary Table	222
7.8	Choosing the Right Sample Size	223
7.8.1	Sample Size for a Mean	223
7.8.2	Sample Size for a Proportion	224
7.9	Connecting Back to the Polls	225
7.10	Ethics of Sample Size	225
7.11	Putting It All Together	226
7.12	A Few Final Cautions	228
7.13	Looking Ahead	229
7.14	Further Reading and References	229
7.15	Key Terms	230
7.16	Exercises	231
7.16.1	Check Your Understanding	231
7.16.2	Apply It	232
7.16.3	Think Deeper	234
8	Hypothesis Testing	237
8.1	Did It Actually Work?	237
8.2	The Logic of Hypothesis Testing	238
8.2.1	Starting with a Question	238
8.2.2	Null and Alternative Hypotheses	239
8.2.3	The Courtroom Analogy	239
8.2.4	Where the Analogy Breaks Down	240
8.3	Test Statistics and P-Values	241
8.3.1	The Test Statistic	241
8.3.2	What Is a P-Value?	242

8.3.3	The Significance Level	243
8.3.4	What a P-Value Is NOT	244
8.4	Type I and Type II Errors	245
8.4.1	Type I Error (False Positive)	245
8.4.2	Type II Error (False Negative)	245
8.4.3	Power	245
8.4.4	Why Power Matters in Practice	246
8.5	Statistical Significance vs. Practical Significance	249
8.5.1	The Large-Sample Problem	249
8.5.2	Effect Sizes	249
8.6	One-Sample t-Test	251
8.6.1	When to Use It	251
8.6.2	Example	251
8.6.3	Confidence Intervals and Hypothesis Tests: Two Sides of the Same Coin .	253
8.6.4	Assumptions	253
8.7	Two-Sample t-Tests	253
8.7.1	Independent Samples t-Test	254
8.7.2	Paired Samples t-Test	256
8.7.3	Independent vs. Paired: How to Choose	257
8.8	When the Data Isn't Normal: Rank-Based Tests	258
8.9	Chi-Square Tests	259
8.9.1	Chi-Square Goodness of Fit Test	259
8.9.2	Chi-Square Test of Independence	261
8.9.3	Assumptions for Chi-Square Tests	263
8.10	One-Tailed vs. Two-Tailed Tests	263
8.11	The Hypothesis Testing Procedure, Summarized	264
8.12	Putting It All Together: The Opower Study Revisited	267
8.13	Common Misconceptions and Pitfalls	268
8.14	Looking Ahead	269

8.15	Further Reading and References	269
8.16	Key Terms	271
8.17	Exercises	273
8.17.1	Check Your Understanding	273
8.17.2	Apply It	274
8.17.3	Think Deeper	277
III	Comparing and Modeling	279
9	Comparing Groups	281
9.1	Are Emily and Greg More Employable Than Lakisha and Jamal?	281
9.2	The Problem with Running Multiple t-Tests	283
9.3	ANOVA: The Big Picture	284
9.4	The F-Statistic: Between vs. Within	285
9.4.1	Making It Precise	287
9.4.2	The ANOVA Table	289
9.4.3	A Worked Example	289
9.5	Assumptions of ANOVA	290
9.5.1	Independence	291
9.5.2	Normality	291
9.5.3	Equal Variances (Homogeneity of Variance)	291
9.6	Post-Hoc Tests: Finding Where the Differences Are	292
9.6.1	Tukey's Honestly Significant Difference (HSD)	292
9.6.2	Other Post-Hoc Options	293
9.7	Effect Sizes for Group Comparisons	294
9.7.1	Eta-Squared (η^2)	295
9.7.2	Partial Eta-Squared (η_p^2)	295
9.7.3	Cramér's V for Categorical Comparisons	296
9.8	When to Use Which Test: A Decision Framework	297

9.9	Beyond the Résumé Study: ANOVA in Practice	298
9.10	The Ethics of Comparing Groups	301
9.11	Connecting ANOVA to the Résumé Study	302
9.12	A Note on the Chi-Square Test	304
9.13	Pulling It All Together	305
9.14	Looking Ahead	306
9.15	Further Reading and References	307
9.16	Key Terms	308
9.17	Exercises	310
9.17.1	Check Your Understanding	310
9.17.2	Apply It	310
9.17.3	Think Deeper	312
10	Simple Linear Regression	315
10.1	The Scatter Plot That Launched a Thousand Debates	315
10.2	One Variable to Predict Another	317
10.3	The Regression Line	318
10.3.1	The Least Squares Criterion	319
10.4	Fitting the Line	320
10.4.1	The Slope	320
10.4.2	The Intercept	321
10.4.3	A Worked Example	322
10.5	R-Squared: How Much Does the Model Explain?	323
10.5.1	Interpreting R-Squared	324
10.5.2	Limitations of R-Squared	325
10.6	Is the Slope Real? Inference for Regression	325
10.6.1	Standard Error of the Slope	326
10.6.2	Testing Whether the Slope Is Zero	326
10.6.3	Confidence Interval for the Slope	327

10.6.4	Reading Regression Output	328
10.6.5	Statistical vs. Practical Significance in Regression	328
10.7	Residuals: What the Model Misses	329
10.7.1	Properties of Residuals	329
10.7.2	Signal, Noise, and the Population Model	330
10.8	The Assumptions Behind Regression	330
10.8.1	L: Linearity	330
10.8.2	I: Independence	331
10.8.3	N: Normality of Residuals	331
10.8.4	E: Equal Variance (Homoscedasticity)	332
10.9	Checking Assumptions with Residual Plots	332
10.9.1	Residuals vs. Fitted Values Plot	333
10.9.2	Normal QQ Plot	333
10.10	Prediction and the Danger of Extrapolation	335
10.10.1	Interpolation vs. Extrapolation	335
10.11	A Complete Example	336
10.12	Before You Regress: Examine Your Variables	338
10.13	Outliers and Influential Points	338
10.14	Correlation and Regression: Close Relatives	340
10.15	What Regression Cannot Do	340
10.16	Looking Ahead	342
10.17	Further Reading and References	342
10.18	Key Terms	343
10.19	Exercises	345
10.19.1	Check Your Understanding	345
10.19.2	Apply It	346
10.19.3	Think Deeper	347

11 Multiple Regression	349
11.1 The Question That Launched a Thousand Arguments	349
11.2 From One Predictor to Many	350
11.2.1 How the Coefficients Are Estimated	351
11.3 “Holding All Other Variables Constant”	352
11.4 Categorical Predictors and Dummy Variables	353
11.4.1 Categorical Variables with More Than Two Categories	353
11.4.2 Choosing the Reference Category	354
11.5 Interaction Effects	355
11.6 R-Squared vs. Adjusted R-Squared	355
11.7 Multicollinearity	357
11.7.1 What It Is	357
11.7.2 The Extreme Case	357
11.7.3 Detecting Multicollinearity with VIF	357
11.7.4 What To Do About It	358
11.8 When Regression Can and Cannot Support Causal Claims	359
11.8.1 The Problem of Confounding	359
11.8.2 The Catch	359
11.8.3 When Causal Claims Are Defensible	360
11.8.4 The Gold Standard	360
11.9 Omitted Variable Bias	361
11.9.1 The Intuition	361
11.9.2 The Formal Condition	362
11.9.3 The Direction of Bias	362
11.10 Simpson’s Paradox	363
11.10.1 The Berkeley Admissions Case	363
11.10.2 What Simpson’s Paradox Teaches Us	364
11.11 The Ethics of “Controlling Away” Discrimination	365
11.12 The Bias-Variance Trade-Off in Model Building	366

11.13 Putting It All Together with the Wage Gap	367
11.14 Assumptions and Diagnostics	368
11.15 A Note on Statistical Significance of Individual Predictors	370
11.16 Looking Ahead	371
11.17 Further Reading and References	371
11.18 Key Terms	372
11.19 Exercises	374
11.19.1 Check Your Understanding	374
11.19.2 Apply It	374
11.19.3 Think Deeper	377
12 Where Do You Go from Here?	379
12.1 The Map So Far	379
12.2 Patterns Over Time	380
12.3 Causation Gets Serious	382
12.4 When Prediction Is the Point	383
12.5 Updating Your Beliefs	385
12.6 The Responsibility You Carry	386
12.7 The Statistical Mindset	389
12.8 Where to Keep Learning	391
12.9 Parting Words	392
12.10 References	393
13 Interactive Apps	395
13.1 Chapter 1 — Why Statistics Matters Now	397
13.1.1 Variable Type Explorer	397
13.2 Chapter 2 — Asking Good Questions: Research Design	397
13.2.1 Sampling Explorer	397
13.2.2 Total Survey Error Explorer	398
13.3 Chapter 3 — Summarizing Data with Numbers	398

13.3.1 Datasaurus Explorer	398
13.4 Chapter 4 — Summarizing Data with Pictures	398
13.4.1 Correlation Game	398
13.5 Chapter 5 — Probability	398
13.5.1 Probability Rules Sandbox	398
13.5.2 Base Rate Lab	399
13.5.3 Binomial Explorer	399
13.6 Chapter 6 — Normal Distribution and the Central Limit Theorem	399
13.6.1 Normal Curve Explorer	399
13.6.2 Distribution Explorer	399
13.7 Chapter 7 — Confidence Intervals	399
13.7.1 Confidence Interval Simulator	399
13.8 Chapter 8 — Hypothesis Testing	400
13.8.1 Hypothesis Testing Playground	400
13.8.2 P-Hacking Simulator	400
13.9 Chapter 9 — Comparing Groups	400
13.9.1 ANOVA Visualizer	400
13.10 Chapters 10 & 11 — Simple and Multiple Regression	400
13.10.1 Regression Explorer	400
14 Datasets	403
14.1 Download	405
14.2 Loading Data in R	406
14.3 Data Sources	407
15 Get Margin of Error	409
15.0.1 Print Edition	411
15.0.2 Kindle Edition	411
15.1 What you get on this companion site	411

Appendices 413**Appendix A: Statistical Tables 413**

Standard Normal (Z) Table	413
t-Distribution Critical Values	415
Chi-Square Distribution Critical Values	416
F-Distribution Critical Values ($\alpha = 0.05$)	417

Appendix B: Dataset Descriptions 419

Chapter 1: Why Statistics Matters Now	419
flint-water-lead.csv	419
Chapter 2: Research Design	420
sampling-demo.csv	420
Chapter 3: Summarizing Data with Numbers	421
acs-household-income.csv	421
Chapter 4: Summarizing Data with Pictures	421
housing-redlining.csv	421
spurious-correlations.csv	422
Chapter 5: Probability	423
covid-testing.csv	423
Chapter 6: The Normal Distribution and CLT	424
birth-weights.csv	424
Chapter 7: Confidence Intervals	424
polling-data.csv	424
Chapter 8: Hypothesis Testing	425
energy-reports.csv	425
Chapter 9: Comparing Groups	426
resume-callbacks.csv	426
Chapter 10: Simple Linear Regression	427
education-spending.csv	427

Chapter 11: Multiple Regression	428
wage-gap.csv	428
Appendix C: Glossary of Key Terms	429
Appendix D: Guide to the Companion Resources	437
Interactive Applications	437
Datasets	438
Errata and Feedback	438
Instructor Resources	438

Margin of Error

Statistics for a World That Needs Better Decisions

Welcome to *Margin of Error*, a book about statistical thinking for business professionals, published by Margin of Error Media.

This book teaches statistical thinking through real stories, real data, and real ethical questions. It is designed for business students and professionals who need to reason under uncertainty and evaluate quantitative claims.

Start with the Preface for how to use this book, or jump to Chapter 1 to begin.

Companion tools: Explore the interactive applications, referenced from chapter callouts throughout the book.

Datasets: All datasets used in the book are available for download as CSV files.

Copyright

© 2026 Vivek H. Patil, Ph.D. All rights reserved.

Published by Margin of Error Media LLC
marginoferrormedia.com

No part of this book may be reproduced, distributed, or transmitted in any form or by any means without the prior written permission of the publisher, except for brief quotations in reviews or scholarly works with full attribution.

For permissions, licensing, or classroom adoption inquiries:
patilv@gmail.com

First edition: 2026

ISBN (print): [ISBN-PENDING]

ISBN (ebook): [ISBN-PENDING]

Preface

Vivek H. Patil, Ph.D.

Published by Margin of Error Media LLC

For every student who ever walked into a statistics course convinced they were “not a math person.”

For the educators who refuse to let a beautiful subject be reduced to formulas on a board.

And for anyone who has ever been misled by a number and wished they knew how to push back.

About the Author

Vivek H. Patil, Ph.D. has over two decades of experience in marketing, data analytics, and research methodology. His research integrates measurement theory and statistics with frameworks from economics, social psychology, and cognitive psychology to examine human behavior. He has authored or co-authored 25+ peer-reviewed articles in journals including the *Journal of Business Research*, *PLOS ONE*, *Journal of Marketing Analytics*, *Scientometrics*, and the *American Journal of Health Promotion*.

He holds a Ph.D. in Business (Marketing) from the University of Kansas, an M.Eng. in Software Systems from BITS Pilani, and a Master of Management Studies from BITS Pilani.

That grounding in both research and practice extends into his other work. He is the founder of VeloxPortfolio.com, an AI-powered tool that transforms resumes into portfolio websites, and the

publisher of books on data analytics and research methods through Margin of Error Media LLC. His teaching materials, datasets, interactive tools, and current projects are available at patilv.com.

Preface

This book started with a frustration.

Across more than two decades of teaching courses that draw heavily on data and statistical reasoning — marketing research, business analytics, data visualization, survey design, multivariate statistics — I have watched students arrive convinced that anything involving numbers will be painful, abstract, and irrelevant to their lives. Some of them are right, depending on the course. Too many approaches to statistics treat the subject as a sequence of formulas to memorize and procedures to execute. They drain the life out of a deeply useful and consequential way of thinking humans have developed.

Statistics is not a collection of tests. It is a way of reasoning under uncertainty. It is how we figure out whether a medical treatment works, whether a policy is helping or hurting, whether the pattern we see in data is real or a mirage. It is how a pediatrician in Flint, Michigan showed that children were being poisoned by their own water supply. It is how researchers demonstrated that identical resumes with different names received different callback rates. It is how epidemiologists tracked the spread of a virus and how election forecasters try, and sometimes fail, to tell us what is coming.

Every one of those stories involves people. Real people, with real consequences riding on whether the analysis was done well or done badly. That is the thread that runs through this book: data analysis is a responsibility, more than a skill.

Margin of Error is written for anyone who needs to reason carefully with data — students in undergraduate or graduate courses, working professionals across fields, or anyone who regularly encounters quantitative claims and wants to evaluate them more honestly. The book is rigorous enough to serve as a primary course text and accessible enough for the reader who has not taken a formal statistics course in years.

Three things make this book different from the textbook you might expect.

First, every chapter opens with a real story. Not a contrived example, not a hypothetical scenario, but an actual situation where statistics mattered for real people. The Flint water crisis. The

Facebook emotional contagion experiment. Redlining in American housing. Resume discrimination. Election polling. These stories are not decorations. They are the foundation. The statistical concept in each chapter emerges from the story, because that is how statistics works in practice: you start with a question that matters, and you figure out how to answer it with data.

Second, ethics is woven throughout, not siloed. There is no standalone chapter on ethics that students skip or instructors cut for time. Instead, every chapter includes moments where you confront the ethical dimensions of the method you are learning. Who collected this data? Who benefits from this analysis? Who might be harmed? When is it wrong to remove an outlier? What happens when people game the p -value until they get the result they want? These are not philosophical distractions. They are core statistical concerns.

Third, this book takes AI seriously without being an AI book. We are living through a moment when artificial intelligence can generate analyses, predictions, and recommendations at a scale no human could match. That makes statistical literacy more important, not less. AI can compute. It cannot think about what the computation means. Throughout this book, you will encounter “AI Reality Check” boxes that show how AI tools can get statistical reasoning wrong, and what a statistically literate person would catch. Your job is not to out-calculate a machine. Your job is to be the person who knows whether the calculation was the right one to run.

The book is software-agnostic. The main text teaches statistical thinking, not button-clicking. For hands-on work, the companion website at stats.marginofferrormedia.com hosts interactive applications that let you explore concepts visually, plus downloadable datasets for every analysis in the book.

I owe thanks to many people who shaped this book. My students, who over two decades have taught me what confuses them, what excites them, and what they need from a course built around data. My colleagues and collaborators, who create an environment where teaching and scholarship reinforce each other. And the researchers who collected and shared the data used throughout this book, sometimes at considerable personal effort, so that others could learn from real problems.

Most of all, I owe thanks to my family. My wife, Anvita, whose patience with a husband who talks about sampling distributions at dinner is a form of grace I do not deserve. And my daughters, Aashi and Avisha, whose curiosity sparks ideas, whose skepticism discounts the bad ones, and whose presence is a daily reminder to do better work. This book exists because they made the space for it.

Vivek H. Patil

How to Use This Book

Chapter Structure

Every chapter follows the same pattern:

- **Opening Story:** A real case where the chapter’s statistical concept played a central role. Read it. It sets up everything that follows.
- **Main Content:** The statistical ideas, explained in plain language first and then in formal terms. Formulas are introduced when they clarify thinking, not for their own sake.
- **Notes:** Short asides in boxes throughout the text. Historical context, interesting facts, or dry observations. They reward the reader who slows down.
- **Ethics Moments:** These are not interruptions. They are part of the method.
- **AI Reality Check:** Brief examples of how AI tools can stumble on statistical reasoning, and what a careful analyst would catch.
- **Try It Online:** Callouts that point you to interactive applications on the companion website. When you see one, open it.
- **Exercises:** Three tiers at the end of every chapter.
 - *Check Your Understanding:* Can you explain the concepts in your own words?
 - *Apply It:* Can you work through the analysis?
 - *Think Deeper:* Can you wrestle with what the numbers mean and whether the analysis was done responsibly?

The Companion Website

Visit stats.marginoferrormedia.com for:

- **Interactive Applications:** Tools that let you explore sampling, probability, confidence intervals, hypothesis testing, regression, and more — no installation required, everything runs in your browser.
- **Datasets:** Downloadable CSV files for every dataset used in the book.
- **Instructor Resources:** Under development. Contact patilv@gmail.com for the current state.

A complete, chapter-by-chapter list of the interactive applications used in the book is maintained at stats.marginoferrormedia.com/apps.html. Every “Try It Online” callout in the chapters links into that page, so if an app’s host ever moves, the book’s references stay valid.

Errata and Feedback

This book will improve over time, and I want your help. If you find an error, have a suggestion, or want to flag something that could be explained better, you can open an issue at github.com/patilv/moe-stats-issues or email me directly at patilv@gmail.com. Every piece of feedback is read, and corrections are incorporated into future editions.

For Instructors

The exercise sets are deliberately large so that you can vary assignments across sections and terms. The main text is software-agnostic; companion materials use R with the tidyverse, and instructors using Python or other environments can adapt the datasets and exercises without loss. Instructor materials including sample syllabi and additional assessment resources are under development. Contact patilv@gmail.com in the meantime.

For Students

If you are anxious about this material, you are in good company. Most people are. The single best thing you can do is work through the examples and exercises rather than just reading about them. Open the companion site. Run the code. Use the interactive applications. Statistics is not a spectator sport.

What to Expect

Over the next twelve chapters, we will build your statistical toolkit from the ground up.

Part I: Foundations (Chapters 1–4) covers the basics. You will learn to identify variable types, understand research design, summarize data with numbers and pictures, and start developing an eye for the difference between real patterns and noise.

Part II: Probability and Inference (Chapters 5–8) is the heart of the course. You will learn how probability works, why the normal distribution matters, what confidence intervals actually mean (it is not what most people think), and how hypothesis testing allows us to make decisions under uncertainty. This section also tackles the misuse of these tools head-on, including p -hacking, the replication crisis, and the difference between statistical significance and practical importance.

Part III: Comparing and Modeling (Chapters 9–12) puts the tools to work. You will compare groups using ANOVA, model relationships using regression, and examine when statistical analyses can and cannot support causal claims. The final chapter points toward where statistics goes from here, including brief introductions to time series, causal inference, machine learning, and Bayesian thinking.

Every dataset used in the exercises is available on the companion website. Some familiarity with R will help when working through the analyses, though the book is written so the statistical reasoning stands on its own. If you have never written a line of code, plan to work slowly with the early datasets and give yourself time to get comfortable before the methods get more involved.

Throughout, you will encounter real stories, real questions, and data drawn from real sources where possible. Where a concept is better illustrated with a constructed or simulated dataset, that is noted clearly in the text.

A Note on AI Tools

This book is my own work. I used AI tools at various points to support that work: to assist with building the technical infrastructure behind the companion website and interactive applications, to review and revise drafts, and for targeted research assistance. The primary tool was Claude, developed by Anthropic. Gemini, developed by Google, served as a secondary review tool in limited instances.

Every analytical position reflects my own judgment, shaped by two decades of work in research, teaching, and data analysis. The responsibility for what appears here, including any errors, is entirely mine.

I mention this not as a disclaimer but as a matter of principle. The norms around AI use in scholarly work are still being worked out, and I would rather be transparent about my process than pretend the tools I used do not exist.

Part I

Foundations

Chapter 1

Why Statistics Matters Now

1.1 The Water Was Poisoned

In April 2014, the city of Flint, Michigan switched its water supply from the Detroit Water and Sewerage Department to the Flint River. The decision was a cost-cutting measure. Officials assured residents the water was safe.

Within weeks, residents noticed the water looked wrong. It smelled wrong. It tasted wrong. Some reported rashes after bathing. Others watched their hair fall out. They complained, and they were told the water met all federal standards.

They were being misled, though whether the deception was deliberate or negligent would become its own debate. What was not debatable was the chemistry. The Flint River water was highly corrosive, and the city had failed to apply the corrosion-control treatment required by federal law. As untreated river water flowed through Flint's aging lead pipes, it leached lead, a potent neurotoxin, directly into the drinking water of roughly 100,000 people. Among the most vulnerable were children, for whom even small amounts of lead exposure can cause irreversible damage to cognitive development.

The story of how this crisis was uncovered is a story about statistics.

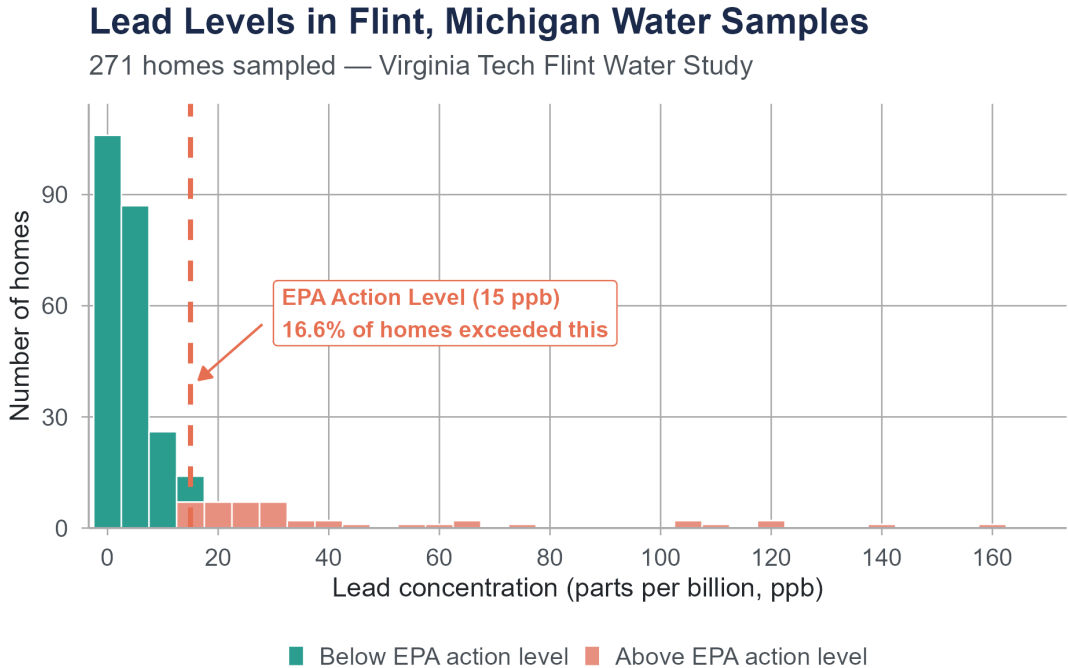
Marc Edwards, an environmental engineering professor at Virginia Tech, had a track record of investigating water contamination. When Flint residents reached out to him, his team organized a citizen science project. They sent water-testing kits to about 300 Flint households and analyzed 271 returned samples. Among the first-draw samples (water collected directly from the tap before any flushing), 16.6% of homes exceeded the EPA action level of 15 parts per billion (ppb), seven

came in above 100 ppb, and the highest first-draw reading was 158 ppb. In a separate, widely reported individual case, Edwards's team measured lead at 13,200 ppb in stagnated water at one Flint resident's home — a level so high that the EPA classifies it as hazardous waste.

At the same time, Dr. Mona Hanna-Attisha, a pediatrician at Hurley Medical Center in Flint, was growing concerned. She obtained blood lead level data for children in Flint from before and after the water switch. Using statistical analysis, she compared the percentage of children with elevated blood lead levels (above 5 micrograms per deciliter) across the two time periods. Her analysis showed that the percentage of children with elevated blood lead levels had roughly doubled after the water source change, and in some zip codes the increase was even more pronounced.

When Hanna-Attisha presented her findings, state officials publicly dismissed them. They said her data was wrong, her methods were flawed, her conclusions were irresponsible. She went back to her data, reconfirmed her analysis, and stood firm. Within weeks, the state was forced to acknowledge what the data made undeniable: the water in Flint was poisoning its children.

Two researchers. Two datasets. Two statistical analyses. That is what it took to force the truth into the open against an establishment that preferred to deny it.



Virginia Tech Flint Water Study (flintwaterstudy.org). First-draw samples, August–September 2015.

Figure 1.1: Distribution of lead levels in sampled Flint homes, showing a right-skewed pattern with 16.6% of homes (45 of 271) exceeding the EPA action level of 15 parts per billion. Source: Virginia Tech Flint Water Study (2015).

This is what statistics is for.

But notice what statistics required in Flint. It was not enough to have data. Residents had data in the most visceral sense, they could see and smell and taste that something was wrong. What they lacked was the ability to translate their experience into evidence that institutions would be forced to take seriously. Edwards needed a systematic sampling plan that could withstand scrutiny, not just a few alarming numbers but a distribution of lead levels across hundreds of homes that painted an undeniable picture. Hanna-Attisha needed a comparison framework, before-and-after data analyzed with methods that controlled for alternative explanations. Both needed to present their findings in ways that could survive hostile examination by officials who had every incentive to discredit them.

This is a pattern you will see throughout this book. The statistical tools themselves are not complicated. A mean is a mean. A comparison is a comparison. What makes the difference is knowing

which tool to use, knowing what questions to ask of the data, and knowing how to defend your conclusions when they are challenged. Those skills, the skills of statistical thinking, are what separate someone who can crunch numbers from someone who can change minds with evidence.

1.2 What Statistics Is (And What It Is Not)

If your previous experience with statistics involves memorizing formulas, plugging numbers into a calculator, and hoping the answer matches the back of the book, consider a different frame. That approach, the mechanical one, misses the point of the discipline entirely. It turns statistics into a set of rituals: do this step, then this step, then report this number. Rituals are easy to follow and easy to forget. What stays with you is understanding.

Statistics is a way of reasoning under uncertainty. The world is complicated. Information is incomplete. People disagree about causes and effects. Data has noise. And yet, decisions must be made. Should this drug go to market? Is this school program working? Did this marketing campaign increase sales, or did sales go up for a different reason? Are the children of Flint sicker than they should be?

We cannot answer these questions by intuition alone. Intuition is powerful, but it is also unreliable, subject to every cognitive bias cataloged by decades of psychology research. We cannot answer them by anecdote. One person's story, no matter how compelling, cannot tell us what is happening across a population. And we cannot answer them by collecting data and staring at it, because raw data without analysis is just a spreadsheet.

Statistics gives us a framework for moving from data to defensible conclusions. It provides tools for summarizing what we observe, quantifying how confident we should be, testing whether patterns are real or accidental, and making predictions about what might happen next. It is, in a sense, the grammar of evidence. Just as knowing grammar helps you write clearly and spot bad arguments, knowing statistics helps you think clearly about data and spot bad analyses.

Here is what statistics is not: a set of mechanical procedures that produce truth. No statistical test can tell you what to believe. No p -value can decide whether a policy is good. No regression coefficient can settle a moral argument. Statistics provides evidence, and evidence must be interpreted by people who understand both the methods and the context. The methods are this book's job. The context is yours.

1.2.1 Statistics Is Not Math (Exactly)

Students often arrive in a statistics course expecting it to feel like a math class. There will be formulas, yes. There will be numbers. You will occasionally need to square things and take square roots. But statistics is not a branch of mathematics in the way that calculus or linear algebra is. Mathematics deals in certainty: given these axioms, this theorem follows. Statistics deals in uncertainty: given this data, here is what we can reasonably conclude, and here is how confident we should be about that conclusion.

This distinction matters because it means that two competent statisticians can look at the same data and reach different conclusions, not because one of them made an arithmetic error, but because they made different judgment calls about what method to use, what variables to include, how to handle missing data, or what level of evidence to require. These are not failures of the discipline. They are features of reasoning about a world that is irreducibly uncertain.

If you find the formulas in this book intimidating, I have good news: the formulas are tools, not the point. The point is the reasoning. A student who understands *why* we compute a confidence interval and *what it means* but needs to look up the formula is in far better shape than a student who can compute the interval flawlessly but has no idea what to do with the result.

1.2.2 The Three Big Questions

Almost every statistical analysis, from the simplest bar chart to the most complex machine learning model, is trying to answer one of three questions:

1. **What happened?** This is description. How many customers did we have last quarter? What was the average income in this zip code? What percentage of patients experienced side effects? Descriptive statistics organizes and summarizes data so we can see patterns.
2. **Is it real?** This is inference. We observed a difference between two groups, but could that difference have arisen by chance? We see a trend in the data, but would the trend hold up with new data? Inferential statistics helps us move from what we see in a sample to conclusions about a larger population, and it quantifies the uncertainty involved.
3. **What will happen?** This is prediction. Given what we know about past customers, which new prospects are likely to buy? If we increase advertising spending, what is our best estimate of the effect on sales? Based on a patient's symptoms and test results, what is the probability of a particular diagnosis? Predictive modeling uses patterns in existing data to forecast outcomes. This is where statistics overlaps most with machine learning, and where the distinction

between correlation and causation becomes most consequential, because a prediction based on a spurious pattern will fail the moment conditions change.

These three questions are not always cleanly separable. A good analysis often involves all three. But keeping them in mind will help you understand why we use the tools we use throughout this book.

1.2.3 Statistics in Everyday Decisions

You may think statistics is something that happens in research labs and corporate boardrooms, and it does, but it also happens every time you make a decision under uncertainty, which is to say, constantly.

When you check the weather forecast and see “70% chance of rain,” you are consuming a statistical prediction. That number came from a model that analyzed atmospheric data, compared current conditions to historical patterns, and produced a probability. You then make a judgment call: bring an umbrella, move the picnic indoors, or take your chances. The model does not make the decision. You do. But the model gives you something better than guessing.

When you read that a new medication “reduces the risk of heart attack by 50%,” that sounds dramatic. But statistics teaches you to ask: 50% of what? If the baseline risk was 2 in 1,000, a 50% reduction means it dropped to 1 in 1,000. That is a difference of one person per thousand, which might not change your decision about whether to take a pill with unpleasant side effects every day for the rest of your life. The percentage sounds large. The absolute numbers tell a different story. Learning to ask the right follow-up questions is a core statistical skill.

When a company advertises that “9 out of 10 dentists recommend” their toothpaste, a statistically literate person wants to know: 9 out of 10 dentists surveyed about what? Were they asked to choose between this brand and nothing at all, or between this brand and its competitors? How were these dentists selected? Were they paid? The claim is technically a statistic, but without context, it communicates almost nothing.

When your social media feed shows you an article claiming that people who own dogs live 24% longer, statistics gives you the vocabulary to ask whether this was an experiment (it was not, because researchers did not randomly assign people to own dogs) and whether there might be confounding variables (dog owners may be wealthier, more active, less likely to live alone, all of which independently predict longer life). The dogs might not be doing the work. They might just be along for the ride.

These examples share a common thread: someone is presenting a number, and that number is supposed to influence what you think or what you do. Statistical literacy is the ability to evaluate whether the number deserves that influence. It is a form of self-defense.

1.3 Types of Data: The Raw Material

Before you can analyze data, you need to understand what kind of data you have. This sounds simple, and in a sense it is, but getting it wrong leads to analyses that do not make sense. You would not try to calculate the average zip code, but people do try to calculate the average of survey ratings on a 1-to-5 scale and then argue about whether a 3.7 means something different from a 3.4, which turns out to be a more interesting question than it first appears.

1.3.1 Variables

A **variable** is any characteristic that can vary across the people, things, or events you are studying. Height is a variable. Gender is a variable. Number of purchases last month is a variable. Whether a customer churned is a variable.

The cases, or units of observation, are the people, things, or events themselves. In a dataset of 500 customers, each customer is a case (one row in the spreadsheet), and the columns, things like age, purchase amount, satisfaction rating, region, are the variables. This rows-and-columns structure is so fundamental that it is worth pausing to make sure it is clear before we go further.

1.3.2 Categorical Variables

A **categorical variable** places each case into one of a set of groups or categories.

Nominal variables have categories with no natural ordering. Examples: blood type (A, B, AB, O), state of residence, brand of phone, industry classification. You can count how many cases fall into each category and compute proportions, but calculating a mean makes no sense.

Ordinal variables have categories with a natural order, but the distances between categories are not necessarily equal. Examples: education level (high school, bachelor's, master's, doctorate), customer satisfaction (very dissatisfied, dissatisfied, neutral, satisfied, very satisfied), pain rating on a 1-to-10 scale. You can say one category is higher than another, but you cannot confidently

say the gap between “satisfied” and “very satisfied” is the same size as the gap between “neutral” and “satisfied.”

i Note

The question of whether to treat ordinal data as numerical (computing means and standard deviations) or categorical (computing frequencies and proportions) generates one of the longest-running debates in statistics. In practice, researchers often treat ordered rating scale responses (such as 1-to-5 or 1-to-7 agreement scales) as numerical. Whether this is appropriate depends on context. We will revisit this in Chapter 9 when choosing between ANOVA and nonparametric methods for ordinal outcomes.

Some disciplines use a four-level measurement framework developed by S.S. Stevens (1946): nominal, ordinal, interval, and ratio. Interval and ratio are both continuous numerical variables; the difference is whether a true zero exists. Ratio scales (income, height) support statements like “twice as much.” Interval scales (temperature in Fahrenheit, IQ scores) do not. For selecting statistical methods, this distinction rarely changes what you do. If you encounter Stevens’ framework in other courses or in the research literature, the two systems are compatible.

1.3.3 Numerical Variables

A **numerical variable** (also called quantitative) represents a measured quantity where arithmetic operations are meaningful.

Discrete numerical variables take on countable values, usually integers. Examples: number of children in a household (0, 1, 2, 3, ...), number of defective items in a shipment, number of website visits.

Continuous numerical variables can take any value within a range, including decimals. Examples: height (5.74 feet), temperature (98.6 degrees), time to complete a task (43.7 seconds), revenue (\$1,247,893.41). In practice, all measurements are rounded to some level of precision, but the underlying concept is that the variable could, in principle, take infinitely many values.

1.3.4 Identifying Variable Types in Practice

Let’s apply this framework to a real situation. Imagine a hospital tracking patient data. For each patient, the hospital records:

- **Patient ID** (like MRN-00482): Nominal. It is a label, nothing more. The fact that it contains numbers does not make it numerical.
- **Age**: Continuous numerical. A patient could be 42.7 years old, even if the hospital rounds to whole years.
- **Blood pressure** (systolic, in mmHg): Continuous numerical.
- **Number of previous hospitalizations**: Discrete numerical. You can have 0, 1, 2, or 3 previous stays, but not 2.7.
- **Primary diagnosis** (diabetes, pneumonia, heart failure): Nominal. There is no inherent ordering among diseases.
- **Pain level** (1 to 10 scale): Ordinal. The distances between levels are not guaranteed to be equal. The jump from 2 to 3 might feel very different from the jump from 8 to 9.
- **Insurance type** (private, Medicare, Medicaid, uninsured): Nominal.
- **Discharge status** (discharged home, transferred, deceased): Nominal.

Notice how a single dataset can contain all four types of variables. Getting the classification right is not academic. If someone computes the “average insurance type,” the software might produce a number, but that number is nonsense. If someone treats pain level as continuous and runs a sophisticated regression, the results might be interpretable, or they might be misleading, depending on whether the gaps between pain levels are roughly equal in the population being studied.

Try It Online

Open the **Variable Type Explorer** on the companion website. Classify variables from a hospital dataset as nominal, ordinal, discrete, or continuous, and get instant feedback. Then switch to the “What Goes Wrong” tab to see what happens when you apply the wrong summary statistic to the wrong type of variable, like computing the mean of insurance types or a histogram of patient IDs. Try the “Numbers Trap” tab to test whether you can spot variables that contain digits but are not numerical. A quiz mode lets you build speed and accuracy. Getting this classification right is the first decision in any analysis.

1.3.5 Why This Matters

The type of variable determines which statistical methods are appropriate. You summarize categorical data with counts, proportions, and bar charts. You summarize numerical data with means, standard deviations, and histograms. You compare groups of categorical data with chi-square tests

and groups of numerical data with t -tests or ANOVA. Choosing the wrong method for your variable type produces output that looks like an answer but is not one.

Understanding what kind of variable you are working with is a useful habit in data analysis. Treating it as a default starting point, like reading the ingredients before you start cooking, will save you from a surprising number of wrong turns.

Types of Variables

Variable type determines which statistical methods are appropriate

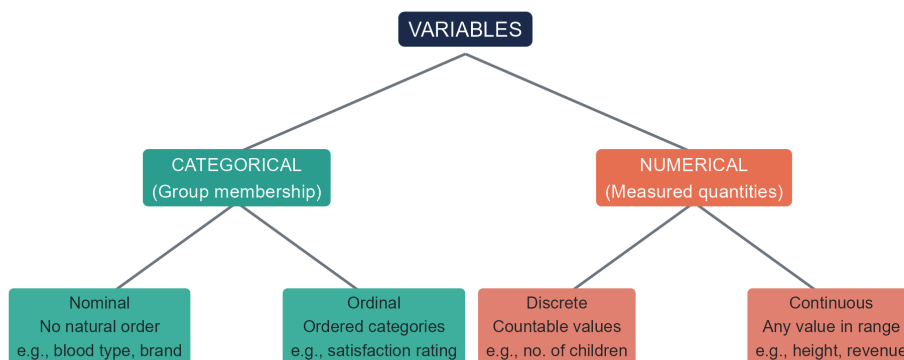


Figure 1.2: A taxonomy of variable types, showing how variables divide into categorical (nominal and ordinal) and numerical (discrete and continuous) types.

1.3.6 The “It Has Numbers So It Must Be Numerical” Trap

One of the most common mistakes beginners make is assuming that any variable containing numbers is numerical. This seems logical, but it is wrong often enough to be worth highlighting.

Zip codes are numbers. But what does an “average zip code” of 48,203.7 mean? Nothing. Zip codes are labels assigned to geographic areas, and the fact that they happen to be numeric is an accident of the postal system. Jersey numbers in basketball are numbers, but calculating that the average jersey number on a team is 23.4 is meaningless. Social Security numbers, credit card numbers, phone numbers, account numbers, all of these contain digits but are not numerical variables. No arithmetic operation on them produces useful information.

Going the other direction, some variables that look categorical have a numerical interpretation lurking underneath. A student's class year (freshman, sophomore, junior, senior) is ordinal, but it also corresponds to years of college completed (roughly 1, 2, 3, 4), and in some analyses treating it as numerical is reasonable. An ordered rating scale of 1 to 5 is technically ordinal, but if you are averaging across thousands of responses, treating it as numerical often produces results that are practical and defensible, even if purists object. The key is to make the choice deliberately, with awareness of the assumption you are making, rather than by default.

The safest habit is to ask yourself two questions whenever you encounter a variable: "Would it make sense to add or average these values?" and "Do the numbers represent a real quantity, or are they just convenient labels?" If the answer to the first question is no, or the answer to the second is "just labels," you are dealing with a categorical variable regardless of what the data looks like. (The "Numbers Trap" tab in the Variable Type Explorer app lets you practice spotting these cases across different datasets.)

1.4 Where Data Comes From (And Why You Should Care)

Data does not appear out of thin air. Someone decides what to measure, how to measure it, whom to measure, and when. Each of those decisions shapes what the data can and cannot tell you. Understanding the origin of your data is not a preliminary step you rush through to get to the interesting analysis. It *is* the analysis, or at least the part that determines whether the analysis means anything.

Consider a simple example. A manager looks at quarterly sales data and notices that revenue is higher in quarters when the company ran social media campaigns. They conclude that social media drives revenue. But they have not considered that the company tends to run social media campaigns during the holiday season, when revenue is naturally higher anyway. The data is accurate. The math is correct. But the conclusion is wrong because they did not ask a basic question about where the pattern came from.

Every dataset tells two stories. There is the story in the data: the patterns, the trends, the relationships between variables. And there is the story *behind* the data: how it was collected, why it was collected, what was included, and what was left out. The second story is at least as important as the first, and it is the one most people ignore.

1.4.1 Populations and Samples

The **population** is the entire group you want to know about. All registered voters in the United States. All customers who purchased from your website this year. All children in Flint who were exposed to contaminated water.

A **sample** is a subset of the population that you actually observe. We rarely have data on an entire population. Instead, we collect data from a sample and use it to draw conclusions about the population. This leap, from sample to population, is the central challenge of statistical inference, and we will spend much of this book learning how to do it well and understanding when it goes wrong.

The quality of your conclusions depends entirely on the quality of your sample. A biased sample produces biased conclusions, no matter how sophisticated your analysis. We will cover sampling methods in detail in Chapter 2.

Consider a concrete example. Suppose a university wants to know what percentage of its 15,000 students support a tuition increase to fund a new recreation center. Surveying all 15,000 students is impractical, so the university surveys 400 students instead. If those 400 are selected using simple random sampling (meaning every student has an equal chance of being selected and selections are made independently), then the percentage who support the increase in the sample should be close to the percentage in the full student body. How close? That depends on the sample size and the variability in opinions, and we will learn to quantify “how close” precisely when we study confidence intervals in Chapter 7.

But if the survey is conducted by setting up a table outside the existing recreation center, the sample is no longer random. Students who already use the rec center, and therefore benefit most from an upgrade, are overrepresented. The result will overestimate support. The math might be perfect. The conclusion will still be wrong, not because of a computational error but because of a sampling error. This is the kind of error that no amount of statistical sophistication can fix after the fact.

1.4.2 Observational Studies vs. Experiments

In an **observational study**, researchers collect data by observing participants as they naturally are. They do not intervene or assign treatments. Most business and social science research is observational: surveys, transaction records, census data, web analytics.

In an **experiment**, researchers deliberately assign treatments to participants and then compare outcomes. The gold standard is a **randomized controlled experiment**, where participants are

randomly assigned to a treatment group or a control group. Random assignment is a deceptively simple idea with profound consequences. By letting chance determine who receives the treatment and who does not, you ensure that, on average, the two groups are similar in every way except the treatment. Any pre-existing differences, age, income, health, personality, motivation, all of them, get distributed approximately equally between the groups. This means that if the treatment group does better than the control group, the most plausible explanation is the treatment, because nothing else systematically differs between the groups.

The distinction matters for one reason in particular: **causation**. Experiments can support causal claims. Observational studies, in general, cannot. If you observe that people who exercise more also weigh less, you have found an association. You have not shown that exercise causes weight loss, because people who exercise might differ from those who do not in a dozen other ways, diet, genetics, occupation, socioeconomic status, motivation, access to healthy food, that also affect weight. The people who choose to exercise might be exactly the kind of people who would weigh less regardless. Without random assignment, you cannot disentangle the effect of exercise from the characteristics of people who exercise.

This is a foundational idea in the book, and we will return to it again and again: correlation does not imply causation. You have heard this before. By the end of this book, you will understand exactly why it is true and what it takes to move beyond correlation to causal claims.

1.4.3 A Business Example

Suppose a retail company notices that customers who use its mobile app spend 35% more per year than customers who do not. The marketing team proposes spending \$2 million to acquire more app users, reasoning that getting people onto the app will increase their spending.

Do you see the problem? The data shows an association between app usage and spending. But customers who downloaded the app may already have been the company's most engaged and highest-spending customers. They did not spend more *because* of the app. They downloaded the app *because* they were already committed to the brand. Spending \$2 million to push the app on less-engaged customers might produce no increase in revenue at all.

To determine whether the app actually causes higher spending, the company would need an experiment: randomly assign some new customers to receive a prompt to download the app and others to receive no prompt, then compare spending over the next year. Only that kind of design can disentangle the effect of the app from the characteristics of the people who voluntarily use it.

This is not a hypothetical. Companies make this mistake constantly. They observe that their best customers behave in a certain way, invest heavily in getting other customers to mimic that behavior, and then wonder why the results are disappointing. The answer, almost always, is confounding. The behavior was a symptom of engagement, not a cause of it.

1.5 Data Quality: Garbage In, Garbage Out

The most elegant statistical analysis in the world is worthless if the data feeding it is flawed. This principle is old enough to have a cliché: garbage in, garbage out. But it is worth dwelling on, because in practice, data quality problems are far more common and far more damaging than choosing the wrong statistical test.

1.5.1 Common Data Quality Problems

Missing data. Values are missing because a survey respondent skipped a question, a sensor malfunctioned, a record was lost. The pattern of missingness matters. If data is missing at random, some methods can handle it. If data is missing for a reason related to the variable itself (e.g., people with high incomes are less likely to report their income), the missingness introduces bias.

Measurement error. The value recorded does not match the true value. Self-reported data is particularly prone to this. People misremember, exaggerate, or shade the truth. But measurement error also affects supposedly objective data, instruments can be miscalibrated, definitions can be inconsistent, data entry can introduce typos.

Selection bias. The sample does not represent the population. Online surveys miss people without internet access. Customer satisfaction surveys miss customers who already left. Hospital data overrepresents people who sought treatment, missing those who were sick but did not see a doctor.

Confounding variable. A third variable influences both the variable you think of as the cause and the variable you think of as the effect, creating a spurious association. We will spend considerable time on this problem.

Survivorship bias. You analyze data only on the things that survived some selection process, ignoring those that did not. During World War II, the mathematician and statistician Abraham Wald famously demonstrated this error when the U.S. military studied bullet holes on returning bombers and proposed armoring the wrong areas. We will tell this story in full in Chapter 2. The key insight is that your data only includes survivors, and that applies far beyond military aviation.

When you study successful companies, you miss the companies that tried the same strategies and failed. When you study long-lived people, you miss those who died young. When you study companies that IPO'd, you miss the startups that never made it.

Data entry and processing errors. Someone typed 10000 instead of 1000. A column was mislabeled. A decimal point ended up in the wrong place. A data merge went wrong and matched the wrong records. These errors are mundane and disturbingly common.

1.5.2 A Cautionary Tale: The Spreadsheet Error That Changed History

In 2010, two Harvard economists, Carmen Reinhart and Kenneth Rogoff, published an influential paper arguing that countries whose national debt exceeded 90% of GDP experienced sharply lower economic growth. The finding was cited by policymakers around the world to justify austerity measures, spending cuts that affected millions of people during the aftermath of the global financial crisis.

Three years later, a graduate student named Thomas Herndon tried to replicate their results for a class assignment. He could not. When he obtained the original spreadsheet, he discovered several problems. One was a simple coding error: a formula in Microsoft Excel had excluded five countries from an average because the cell range was wrong. Another was a questionable decision about which data points to include and which to exclude. When the errors were corrected and the excluded data was added back, the dramatic growth cliff at 90% debt largely disappeared.

The original finding was not fabricated. The economists had not committed fraud. They had made a spreadsheet mistake that anyone could make and a subjective data selection decision that happened to point in a particular direction. But the policy consequences were enormous. Governments had used the 90% threshold to justify austerity programs that cut public services, reduced government employment, and slowed economic recovery. This episode is a reminder that even careful, experienced researchers can make errors that have consequences far beyond the spreadsheet, and replication and scrutiny are not attacks on researchers, but essential features of how knowledge gets corrected over time.

1.5.3 The First Rule of Data Analysis

Before you compute a single summary statistic, look at your data. Examine its structure. Check for oddities. Ask how it was collected. Ask who collected it and why. Ask what might be missing.

This sounds tedious, and it can be. But skipping it is how conclusions go wrong. The Reinhart-Rogoff episode is one high-profile example, but the same dynamic plays out constantly in business analytics, public health, and policy research: a processing error that goes undetected because no one looked carefully at the data before running the analysis.

Ethics Moment

Data quality is an ethical issue, more than a technical one. When Flint officials reported water testing results, they used sampling methods that systematically avoided the worst-affected homes, a practice sometimes called “strategic sampling.” The data they published was not technically fabricated. It was strategically collected to produce a predetermined conclusion. This is one of the most dangerous forms of data manipulation because the numbers look clean. Being literate about data quality is your first defense against being misled, and your first obligation if you are the one producing the analysis.

1.6 The AI Question

You are learning statistics at a peculiar moment in history. Artificial intelligence systems can now generate analyses, build predictive models, write code, produce visualizations, and summarize findings faster than any human. If a machine can do all that, why should you spend a semester learning to do it yourself?

The answer is that computation is the easy part. It always has been. Even before AI, a calculator could compute a mean, and software could run a regression. The hard parts, the parts that matter, have always been the parts that require judgment:

- Is this the right question to ask?
- Is this data appropriate for answering it?
- Does this analysis method match the structure of the data?
- Are the assumptions of the method satisfied?
- What do the results actually mean in context?
- Who is affected by these findings, and what are the consequences of getting it wrong?

AI cannot reliably answer any of these questions. It can produce confident-sounding responses to all of them, which is arguably worse than not answering at all. A machine that generates a wrong answer with no hesitation is more dangerous than a machine that admits it does not know.

Consider what happened when Flint officials insisted the water was safe. The data they cited was real. The summaries were calculated correctly. The problem was not computation, it was judgment. They chose what to measure (lead levels at carefully selected homes), how to measure it (flushing the tap before testing, which reduced lead readings), and how to interpret the results (within federal limits, therefore safe). Every one of those choices required human judgment, and every one was wrong in ways that hurt real people.

Statistical literacy is the ability to evaluate these judgments. In an age when AI can produce endless analyses, the person who can tell the difference between a good analysis and a bad one has never been more valuable.

AI Reality Check

Ask a large language model how to summarize the “typical” amount customers spend per order at an e-commerce store. Most will recommend the mean, explaining that it “uses all the data” and is “more informative.” What they routinely fail to mention is that transaction amounts are almost always right-skewed: most customers place modest orders, while a small number of large purchases pull the mean well above what a typical customer actually spends. If the median order is \$42 but a handful of bulk orders push the mean to \$87, reporting the mean implies a “typical customer” who does not exist in your data. A manager who uses that number to set pricing expectations or design loyalty tiers is building strategy on a distorted picture. The calculation is correct. The choice of which calculation to run requires the kind of contextual judgment that AI does not reliably supply.

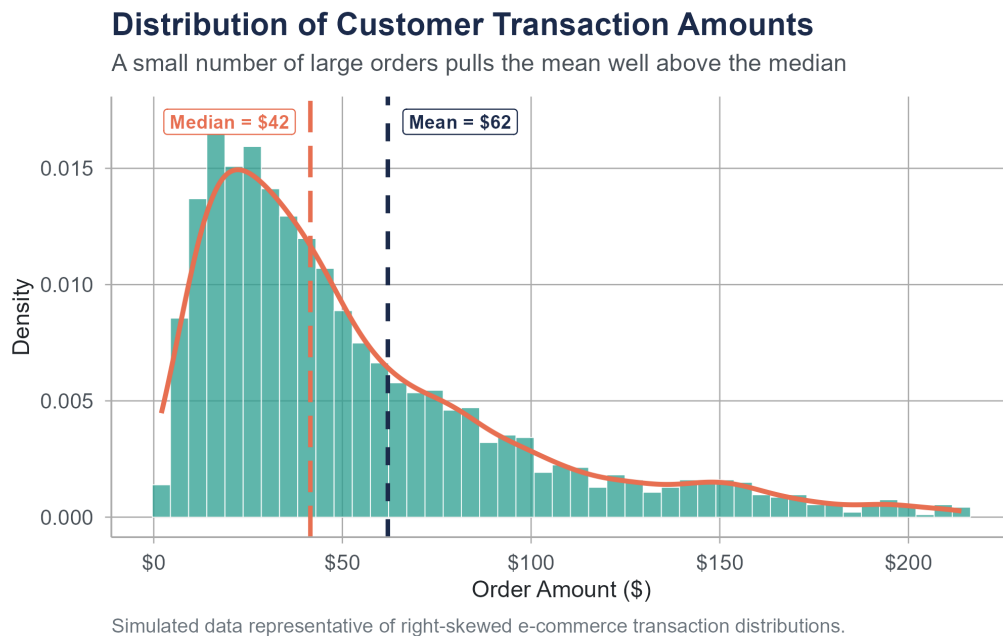


Figure 1.3: Simulated e-commerce transaction amounts showing the mean pulled far to the right of the median by a small number of large orders. Source: Simulated data.

1.6.1 What AI Can and Cannot Do

Let's be specific about what current AI tools handle well and where they fall short in the context of statistics:

AI can perform calculations quickly and accurately. It can run statistical tests, fit models, and generate visualizations if given clean data and clear instructions. It can look up formulas and explain procedures. For routine computation, AI is a useful assistant.

AI struggles with determining whether the assumptions of a test are met, understanding the context behind data, recognizing when results are technically correct but substantively misleading, handling ambiguous or incomplete data, and making judgment calls about what analysis to run in the first place.

AI fails at ethical reasoning about data use, understanding the social consequences of an analysis, and knowing what it does not know. When an AI tool encounters a question it cannot answer, it often answers anyway, with unearned confidence.

Your goal in this book is not to become a calculator. It is to become a thinker, someone who knows which questions to ask, which methods to apply, what the results mean, and whether the analysis was done with integrity. These are human skills. They are the skills this book teaches.

1.6.2 Using AI as a Learning Tool (Not a Crutch)

Used thoughtfully, AI tools can support your learning. If you are struggling with a concept, asking an AI to explain it a different way can sometimes unblock your understanding faster than re-reading the same paragraph five times. If you have written R code that produces an error, AI can often identify the syntax mistake.

But there is a line between using AI as a tutor and using it as a shortcut. If you paste an exercise into an AI tool and submit whatever it produces, you will get a grade for that assignment, but you will not get the learning. And the learning is the product you are paying for. Employers are not hiring you for your ability to copy-paste prompts. They are hiring you for your ability to look at data and figure out what it means, what it does not mean, and what to do about it.

Here is a useful rule of thumb: if you cannot explain your answer without referring to what the AI told you, you do not yet understand it. Use AI to help you *get* to understanding, not to bypass it.

1.7 Reading the News Statistically

Once you develop statistical thinking, you will start noticing how numbers are used, and misused, in the media, in advertising, in political arguments. This can be both empowering and maddening.

A headline screams: “New Study Finds Chocolate Reduces Risk of Heart Disease.” The statistically literate reader asks: was this an experiment or an observational study? If observational (and it almost certainly was), then confounders could explain the association. Perhaps people who eat moderate amounts of chocolate also have other lifestyle habits that promote heart health. What was the sample size? Who funded the study? Was the study published in a peer-reviewed journal, or just announced in a press release? What was the effect size? Beyond whether the result passed a statistical test, was the reduction in risk large enough to be practically meaningful?

A politician claims: “Crime is up 15% under the current administration.” The statistically literate listener asks: up 15% compared to what baseline? If crime dropped dramatically during an unusual period (say, a pandemic when people stayed home) and then partially returned to normal, a 15% increase from the abnormal low might still represent a historically normal level. What category

of crime is being measured? Is it violent crime, property crime, or all crime? Are we looking at reported crimes or arrests or convictions? Each produces different numbers.

A company advertises: “Clinically proven to reduce wrinkles.” The statistically literate consumer asks: established by what kind of study? How many participants? Compared to what, a placebo, a competitor, or nothing? What does “reduce wrinkles” mean operationally, and who judged the reduction?

These are not rhetorical questions. They are the exact questions that statistical training teaches you to ask. By the end of this book, they will become second nature.

1.8 Looking Ahead

Chapter 1 has introduced the vocabulary and orientation the rest of this book builds on: what statistics is for, how variables are classified, where data comes from, and why the difference between an observational study and an experiment matters before a single formula is written. The next chapter goes deeper on the collection side. How are samples actually drawn, and what can go wrong? What makes a study design capable of supporting a causal claim, and what stops most studies from getting there? And how do the choices researchers make before collecting a single data point determine whether every subsequent analysis is honest or misleading? Those questions are the subject of Chapter 2.

1.9 Further Reading and References

The following works are cited in this chapter or provide valuable additional context for the ideas covered here.

On the Flint water crisis: Hanna-Attisha, M., LaChance, J., Sadler, R.C., & Champney Schnepf, A. (2016). Elevated blood lead levels in children associated with the Flint drinking water crisis: A spatial analysis of risk and public health response. *American Journal of Public Health*, 106(2), 283–290.

Virginia Tech Flint Water Study (2015). Lead testing results from Flint, Michigan. Data and technical reports available at flintwaterstudy.org.

Hanna-Attisha, M. (2018). *What the eyes don't see: A story of crisis, resistance, and hope in an American city*. One World. (The full account of the Flint crisis in Hanna-Attisha's own words, accessible to a general audience.)

On measurement scales: Stevens, S.S. (1946). On the theory of scales of measurement. *Science*, 103(2684), 677–680. (The original paper introducing the nominal/ordinal/interval/ratio framework.)

On data quality and the Reinhart-Rogoff episode: Reinhart, C.M., & Rogoff, K.S. (2010). Growth in a Time of Debt. *American Economic Review*, 100(2), 573–578. (The original paper.)

Herndon, T., Ash, M., & Pollin, R. (2014). Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Cambridge Journal of Economics*, 38(2), 257–279. (The replication study that uncovered the spreadsheet error.)

On survivorship bias (treated more fully in Chapter 2): Mangel, M., & Samaniego, F.J. (1984). Abraham Wald's work on aircraft survivability. *Journal of the American Statistical Association*, 79(386), 259–267. (Reprints and contextualizes Wald's declassified 1943 memorandum alongside a biographical account.)

For further reading on statistical thinking: Spiegelhalter, D. (2019). *The art of statistics: Learning from data*. Pelican. (A readable, example-driven introduction to statistical reasoning with a similar spirit to this book.)

Wheelan, C. (2013). *Naked statistics: Stripping the dread from the data*. W.W. Norton. (An accessible, conversational treatment of core statistical concepts for a general audience.)

1.10 Key Terms

- **Statistics:** The science of collecting, organizing, analyzing, and interpreting data to make informed decisions under uncertainty.
- **Variable:** A characteristic that can take different values across cases or observations.
- **Categorical variable:** A variable whose values represent group membership. Includes nominal (no natural order) and ordinal (natural order) types.
- **Numerical variable:** A variable whose values represent measured quantities. Includes discrete (countable) and continuous (any value in a range) types.
- **Population:** The entire group of interest.

- **Sample:** A subset of the population from which data is collected.
 - **Observational study:** A study in which researchers observe participants without intervening or assigning treatments.
 - **Experiment:** A study in which researchers deliberately assign treatments to participants and compare outcomes.
 - **Randomized controlled experiment:** An experiment where participants are randomly assigned to treatment and control groups.
 - **Causation:** The relationship in which one variable directly produces a change in another. Establishing causation typically requires a randomized experiment or, in observational data, careful reasoning about confounders.
 - **Confounding variable:** A variable that influences both the explanatory variable and the response variable, creating a misleading association between them.
 - **Selection bias:** Systematic error arising from how participants are chosen for a sample or study.
 - **Missing data:** Values that are absent from a dataset, potentially introducing bias depending on the pattern and reason for missingness.
-

1.11 Exercises

1.11.1 Check Your Understanding

1. In the Flint water crisis, Dr. Hanna-Attisha compared blood lead levels in children before and after the water source change. Was this an observational study or an experiment? Explain your reasoning.
2. Identify the variable type (nominal, ordinal, discrete numerical, or continuous numerical) for each of the following:
 - a. A customer's zip code
 - b. The number of products in a shopping cart
 - c. A student's letter grade (A, B, C, D, F)
 - d. The time a customer spent on a website (in seconds)
 - e. A person's blood type
 - f. The number of siblings a person has
 - g. Temperature measured in Fahrenheit

- h. A hotel star rating (1 through 5 stars)
3. Explain the difference between a population and a sample. Why do we typically work with samples rather than populations?
4. A news article states: “People who eat breakfast every day earn 20% more than people who skip breakfast.” Does this mean eating breakfast causes higher earnings? What alternative explanations might account for this association?
5. Describe two ways that data quality problems could lead to incorrect conclusions. Use examples different from those given in the chapter.
6. What is the difference between a discrete numerical variable and a continuous numerical variable? Give one example of each that was not used in the chapter.
7. A company surveys its current customers about satisfaction and finds that 89% are satisfied. What kind of bias might affect this result, and in which direction?
8. Explain why the distinction between observational studies and experiments matters for making causal claims.
9. In your own words, explain what the phrase “garbage in, garbage out” means in the context of data analysis.
10. A researcher collects data on the relationship between ice cream sales and drowning deaths and finds a strong positive association. More ice cream sales coincide with more drownings. Should the researcher conclude that ice cream causes drowning? What is a more plausible explanation?

1.11.2 Apply It

(See Appendix B for complete variable descriptions for all datasets used in these exercises.)

For the following problems, use the Flint water dataset available on the companion website ([datasets/flint-water-lead.csv](#)). This dataset contains first-draw lead testing results from the Virginia Tech Flint Water Study, collected August–September 2015. It includes lead levels in parts per billion for 271 Flint homes, cleaned from the original source file (`Flint-Samples-FINAL.xlsx`, available at [flintwaterstudy.org](#)) using `flint-data-cleaning.R`. Both the dataset and cleaning script are available on the companion website.

1. Load the dataset and identify each variable. Classify each as nominal, ordinal, discrete numerical, or continuous numerical. How many observations (cases) are in the dataset? How many variables? Note that the `notes` column flags two homes that were sampled twice; how should you handle these when interpreting your results?

2. Examine the `lead_ppb` variable. What is the highest value in the dataset? The lowest? Based on the EPA action level of 15 ppb, what proportion of homes in the sample exceed this threshold?
3. Examine the dataset for missing values. Note whether any are present. Then explain why a carefully curated research dataset might be more complete than data collected through routine record-keeping. What kinds of lead testing data might go missing in a less controlled setting, and what would be the consequences of that missingness for conclusions about the scope of the problem?
4. The dataset includes a variable for ward (geographic district within Flint). How many distinct wards are represented? Is this variable categorical or numerical? Explain.
5. Create a simple summary of lead levels by ward. You do not need formal statistical tests yet, just compute the count, mean, and median lead level for each ward. What do you notice?
6. Using any data source you can find (the U.S. Census Bureau's American Community Survey is a good starting point), find three variables about a city or county of your choice. Identify each variable's type. Describe one question you could potentially answer by analyzing the relationship between two of these variables.
7. A local coffee shop tracks the following for each transaction: date, time of day, drink ordered, size (small/medium/large), price paid, whether the customer used a loyalty card (yes/no), and the barista who made the drink. Classify each variable and identify which would be useful for answering the question: "Do loyalty card users spend more per visit?"
8. Find a news article that reports a statistical finding (e.g., "Study finds that X is associated with Y"). Identify: (a) the population of interest, (b) the sample used, (c) whether the study was observational or experimental, and (d) whether the article's headline implies causation. If it does, is that implication justified?
9. A company wants to know whether a new website design leads to more purchases. They show the new design to visitors from California and the old design to visitors from New York, then compare purchase rates. What is wrong with this approach? What would you recommend instead?
10. Consider a dataset that contains the following variables for a set of 1,000 employees: employee ID, department, years of experience, annual salary, performance rating (1-5), and whether the employee was promoted in the past year (yes/no). For each variable, identify its type. Then propose two specific questions that could be investigated using this dataset, and identify which variables you would need for each question.

1.11.3 Think Deeper

1. In the Flint case, state officials initially dismissed Dr. Hanna-Attisha's findings by questioning her methodology. Later, their own data confirmed her conclusions. What does this episode suggest about the relationship between statistical evidence and institutional power? How should a data analyst respond when their findings are politically inconvenient?
2. The chapter argues that statistical literacy is more important in the age of AI, not less. Do you agree? Can you think of a scenario where relying on an AI tool for statistical analysis without understanding the underlying methods could lead to harm?
3. The chapter mentions "strategic sampling," the practice of selecting samples in a way that produces a desired result. This is different from outright fabrication of data, but the effect can be similar. Should strategic sampling be treated as a form of scientific fraud? Who has the responsibility to catch it?
4. Suppose you are hired as a data analyst for a company that asks you to analyze customer data that was collected without customers' knowledge or explicit consent. The data is already collected. The analysis could help improve the product. What ethical considerations should guide your decision about whether to proceed?
5. Think about a decision that affects your community (a school policy, a local business decision, a government program). What data would you want to see before that decision is made? What kind of study (observational or experimental) would be most appropriate for gathering that data? What practical or ethical barriers might prevent the ideal study from being conducted?
6. The chapter distinguishes between three big questions statistics can answer: "What happened?", "Is it real?", and "What will happen?" Think of a real-world situation (in business, healthcare, education, or policy) where all three questions are relevant. Describe specifically what each question would look like in that context.

Chapter 2

Asking Good Questions: Research Design

2.1 The Experiment You Never Agreed To

In January 2012, roughly 689,000 Facebook users had their News Feeds quietly altered. For one week, some users saw fewer posts containing positive emotional words. Others saw fewer posts containing negative emotional words. The researchers, a team from Facebook’s Core Data Science group and Cornell University, wanted to know whether emotional content in a social media feed could influence a person’s own emotional state. They called it “emotional contagion.”

The results, published in the *Proceedings of the National Academy of Sciences* (PNAS) in June 2014, were modest. Users who saw fewer positive posts wrote slightly more negative posts themselves, and vice versa. The effect sizes were tiny. Under normal circumstances, a study with small effects and a narrow scope might have attracted a paragraph in a science news roundup and nothing more.

Instead, it became one of the most controversial research studies of the decade.

The problem was not the findings. The problem was the design.

Nearly 700,000 people had been enrolled in a psychological experiment without their knowledge. No one had asked them. No one had told them. No one had given them the option to say no. The researchers argued that Facebook’s Terms of Service, which users accept when they create an account, covered this kind of research. The Terms of Service did mention that Facebook might use data for “research.” But there is a wide gap between a company analyzing click patterns to

improve ad targeting and a company deliberately manipulating what people see in order to study changes in their emotional state. Most people, when they click “I Agree” on a terms-of-service page, are not imagining that they have just consented to a mood manipulation experiment.

The backlash was swift and broad. Privacy advocates pointed out that the study lacked meaningful informed consent. Psychologists noted that the study would never have been approved by a standard university institutional review board (IRB) without explicit consent from participants. The editor of PNAS added an “editorial expression of concern” to the published paper, a rare step, acknowledging concerns about the ethical procedures. Cornell’s IRB, which had reviewed the project, had determined that its own review was not required because Facebook had collected the data, not the Cornell researchers. This interpretation struck many as a convenient loophole rather than a principled decision. In the years since, the U.S. Department of Health and Human Services updated the Common Rule (the federal regulations governing human subjects research), with revisions that took effect in 2018. Those revisions extended protections more clearly to research involving data collected by third parties. The Facebook study helped catalyze that conversation, though regulatory gaps between academic and corporate research settings remain an active area of debate.

What makes the Facebook study such a useful starting point for a chapter on research design is that it illustrates, in a single case, most of the principles we need to discuss. The researchers had a question. They chose a method to answer it. They selected a sample. They made decisions about what to manipulate, what to measure, and whom to include. And they made those decisions with insufficient attention to the people affected by them.

Good research design is about more than getting the right answer. It is about asking the right question, choosing the right approach, protecting the people involved, and being transparent about every decision along the way. This chapter covers how to do all of that.

2.2 Two Kinds of Studies

In Chapter 1, we introduced the distinction between observational studies and experiments. That distinction is so central to everything that follows in this book that we need to treat it in much greater depth here.

Connecting to Research Methods

If you have taken a course in research methods, you may have encountered a different but related classification: studies are sometimes described as **exploratory**, **descriptive**, or **causal**. Exploratory studies aim to develop initial understanding of a phenomenon that has not been well studied; they generate questions rather than test them. Descriptive studies aim to characterize what exists: how common something is, how groups compare, what patterns are present. Causal studies ask whether one thing *produces* another.

These two frameworks are not in conflict. They operate at different levels. Exploratory and descriptive studies are almost always observational: the researcher is watching and measuring, not intervening. Causal studies are most credibly conducted as experiments, because only controlled manipulation with random assignment can rule out the alternative explanations that observational data leaves open. When you see the observational/experimental distinction in this chapter, you can read it as the *methodological* dimension of a question whose *purpose* might be descriptive or causal. Understanding both frameworks deepens your ability to evaluate what a study can and cannot claim.

2.2.1 Observational Studies

In an observational study, the researcher watches, measures, and records, but does not intervene. The world does what it does, and you take notes.

Consider a researcher who wants to understand the relationship between sleep and academic performance among college students. They survey 500 students, asking them how many hours of sleep they typically get per night and collecting their GPAs from university records. They then look at whether students who sleep more tend to have higher GPAs.

This is an observational study. The researcher did not assign anyone to a sleep schedule. They observed the students as they were and recorded the data.

Observational studies are everywhere. A hospital compares outcomes for patients who chose Surgery A versus Surgery B. An economist examines whether countries with higher minimum wages have lower poverty rates. A marketing analyst looks at whether customers who open promotional emails spend more than customers who do not. In each case, the researcher is observing existing differences between groups, not creating those differences.

The great advantage of observational studies is feasibility. They are often cheaper, faster, and more ethical than experiments. You cannot randomly assign people to smoke for 30 years to study the

health effects of tobacco. You cannot randomly assign children to poverty to study its effects on brain development. You cannot randomly assign countries to adopt different trade policies. For many of the most important questions in medicine, social science, and policy, observational studies are the only realistic option.

The great disadvantage is that observational studies, on their own, cannot establish causation. We will spend a full section on why later in this chapter. For now, the short version is that when you observe a difference between two groups that already existed before you showed up, you cannot be sure whether the difference in outcomes is caused by the factor you are studying or by some other way in which the groups differ.

2.2.2 Experiments

In an experiment, the researcher intervenes. They do more than observe the world as it is. They change something and watch what happens.

The defining feature of a true experiment is that the researcher controls the assignment of treatments. In the simplest case, participants are divided into two groups. One group receives the treatment of interest (the **treatment group**), and the other does not (the **control group**). Then the researcher compares outcomes between the two groups.

Suppose our sleep researcher wanted to move beyond observation. They recruit 100 students and randomly assign 50 of them to a “sleep coaching” program designed to increase their nightly sleep by one hour. The other 50 receive no intervention. At the end of the semester, they compare the GPAs of the two groups.

This is an experiment. The researcher created the difference between the groups by assigning the treatment. They did not merely observe a pre-existing difference.

Randomization is the ingredient that makes experiments so powerful. When participants are randomly assigned to treatment and control groups, any pre-existing differences between participants (their motivation, prior academic performance, stress levels, coffee consumption, everything) get distributed roughly equally across both groups. This means that if the treatment group ends up with higher GPAs at the end of the study, the most plausible explanation is the treatment itself, because the groups were comparable in every other respect at the start.

This is a strong claim, and it has limits that we will discuss. But it is the foundational logic of experimental design, and it is why **randomized controlled experiments** are often called the “gold standard” for establishing cause and effect.

2.2.3 The Spectrum Between Observation and Experiment

The line between observational studies and experiments is clearer in textbooks than in practice. Real research often falls somewhere in between.

Quasi-experiments are studies where the researcher exploits a “natural” assignment to treatment and control groups that was not random in the strict sense but might be close enough to be useful. For example, when a state raises its minimum wage and a neighboring state does not, researchers can compare employment trends in the two states before and after the change. Nobody randomly assigned states to different minimum wage policies, but the comparison has some of the logic of an experiment. Economists call this a “natural experiment” or a “difference-in-differences” design.

A/B tests are experiments conducted in digital environments. When a tech company shows half its users a blue “Buy Now” button and the other half a green one, then compares click-through rates, that is a randomized experiment. The Facebook emotional contagion study was, structurally, an A/B test. The ethical problem was not the experimental design itself but the absence of informed consent.

Retrospective studies look backward in time, using existing records to study past events. A researcher might examine medical records to compare outcomes for patients who received different treatments years ago. These are observational studies, but they often feel like experiments because the treatments are well-defined. The crucial difference is that the researcher did not control who got which treatment, so confounding is a serious concern.

Understanding where a study falls on this spectrum is essential for knowing how much confidence to place in its conclusions.

Note

The phrase “gold standard” for randomized controlled trials comes from medicine, where the randomized controlled trial has been the benchmark for evaluating treatments since the mid-twentieth century. The first widely recognized RCT was the 1948 British Medical Research Council trial of streptomycin for pulmonary tuberculosis, designed and led by Austin Bradford Hill. Before that, treatments were evaluated by clinical experience, case reports, and expert opinion, methods that sound reasonable but produced centuries of medical practices that ranged from useless to actively harmful. Bloodletting persisted for over 2,000 years in Western medicine partly because no one ran a controlled experiment to check whether it worked. Bradford Hill later articulated the criteria for causal inference from observational data that bear his name, discussed later in this chapter.

2.3 How to Choose Your Sample

We introduced sampling concepts in Chapter 1; here we examine them in the context of research design.

Imagine you want to know the average amount of student loan debt carried by undergraduates at a large university with 15,000 students. You cannot survey all 15,000. You need a sample. But which students should you include?

The answer to this question matters far more than most people realize. A badly chosen sample can produce results that are not merely imprecise but systematically wrong. And the way a sample goes wrong is often invisible in the final data, which makes it especially dangerous.

2.3.1 Simple Random Sampling

The most intuitive sampling method is the one you probably thought of first. Get a list of all 15,000 students, assign each one a number, and use a random number generator to select, say, 500 of them. Every student has an equal chance of being selected, and every possible sample of 500 students has an equal chance of being the one you draw.

This is **simple random sampling**, and it is the foundation on which most of the statistical methods in this book are built. When we talk about “random samples” without further qualification, this is what we mean.

Simple random sampling has a critical virtue. Because every member of the population has the same chance of being included, there is no systematic tendency for the sample to over-represent or under-represent any subgroup. Your sample might, by chance, include more seniors than freshmen, or more engineering majors than English majors. But there is no built-in bias pushing it in any particular direction. And as the sample gets larger, those random fluctuations shrink.

The practical challenge is that simple random sampling requires a complete list of the population, what statisticians call a **sampling frame**. For our student debt example, the university’s registrar presumably has such a list. But for many populations of interest, no such list exists. There is no master list of all Americans who have experienced food insecurity. There is no directory of every small business in the informal economy. When the sampling frame is incomplete or does not exist, simple random sampling is impossible, and researchers must get creative.

2.3.2 Stratified Sampling

Suppose you suspect that student loan debt varies a lot by class year, with seniors carrying more debt than freshmen simply because they have been borrowing longer. You want your sample to reflect the university's distribution across class years. You could trust simple random sampling to give you roughly the right proportions, but with only 500 students, you might end up with too few freshmen or too many juniors by chance alone.

Stratified sampling solves this by dividing the population into non-overlapping subgroups (called **strata**) and then drawing a simple random sample from each stratum separately. The most common form draws from each stratum in proportion to its size in the population; this is called **proportional stratified sampling**. If the university is 30% freshmen, 25% sophomores, 25% juniors, and 20% seniors, you draw 150 freshmen, 125 sophomores, 125 juniors, and 100 seniors, giving you exactly 500. Now your sample is guaranteed to match the population's composition by class year.

Stratification improves precision when the variable you stratify on is related to the variable you are measuring. If seniors do in fact carry more debt than freshmen, a stratified sample will give you a more accurate estimate of the overall average than a simple random sample of the same size, because you have eliminated one source of random variation.

You can stratify on more than one variable (for instance, class year and residency status, in-state vs. out-of-state), but the more strata you create, the more complex the logistics become. With four class years and two residency categories, you already have eight strata, and you need enough students in each one to produce reliable estimates.

Note

When proportional stratified sampling gives you too few observations in a small but important subgroup, **disproportional stratified sampling** may be more appropriate. In this approach, you intentionally over-sample from smaller strata to ensure adequate representation. A study of religious discrimination in hiring might use proportional sampling for large groups but deliberately over-sample members of small religious minorities, who would otherwise appear in too few numbers for meaningful analysis. After data collection, the over-sampled groups are statistically weighted back to their true population proportions for any estimates about the population as a whole. This approach is common in studies of rare conditions, small demographic groups, and under-studied populations where proportional sampling would produce cells too thin to analyze.

2.3.3 Cluster Sampling

Now suppose you do not have a list of all 15,000 students but you do have a list of all 600 course sections offered this semester, and you know that virtually every student is enrolled in at least one course. You could randomly select 30 course sections and survey every student in those sections.

This is **cluster sampling**. Instead of sampling individuals directly, you sample naturally occurring groups (clusters) and then include all members of the selected clusters.

Cluster sampling is often used for practical reasons. It is easier to visit 30 classrooms than to track down 500 individual students scattered across campus. Public health surveys often use cluster sampling, randomly selecting neighborhoods or clinics and then surveying everyone at the selected sites.

The downside is that cluster sampling is generally less precise than simple random sampling or stratified sampling for a given sample size. People within the same cluster tend to be more similar to each other than to people in other clusters. Students in the same course section might be more similar in major, class year, and interests than a random cross-section of the university. This means each cluster gives you less new information than you would get from the same number of randomly selected individuals.

2.3.4 Systematic Sampling

There is one more common method worth knowing. Suppose you have a list of all 15,000 students sorted alphabetically and you want a sample of 500. You calculate that 15,000 divided by 500 equals 30. You pick a random starting point between 1 and 30, say student number 17, and then select every 30th student after that. Student 17, student 47, student 77, and so on.

This is **systematic sampling**. It is simple to implement and produces a sample that is spread evenly across the list. As long as the list itself does not have a hidden pattern that aligns with your sampling interval (which is rare but not impossible), systematic sampling works well in practice.

One risk arises when the list has a periodic structure. If a list of apartment units alternates between corner units and interior units, and your sampling interval happens to match that pattern, you could end up with a sample that is all corner units or all interior units. In practice this is unusual, but worth checking.

 Try It Online

The **Sampling Explorer** lets you compare all four sampling methods side by side. Draw repeated samples using simple random, stratified, cluster, and systematic methods and watch how much the estimates vary across approaches. Try drawing 20 stratified samples and 20 simple random samples from the same population and compare how tightly the estimates cluster. The difference is immediate and visible.

2.3.5 Convenience Sampling and Its Problems

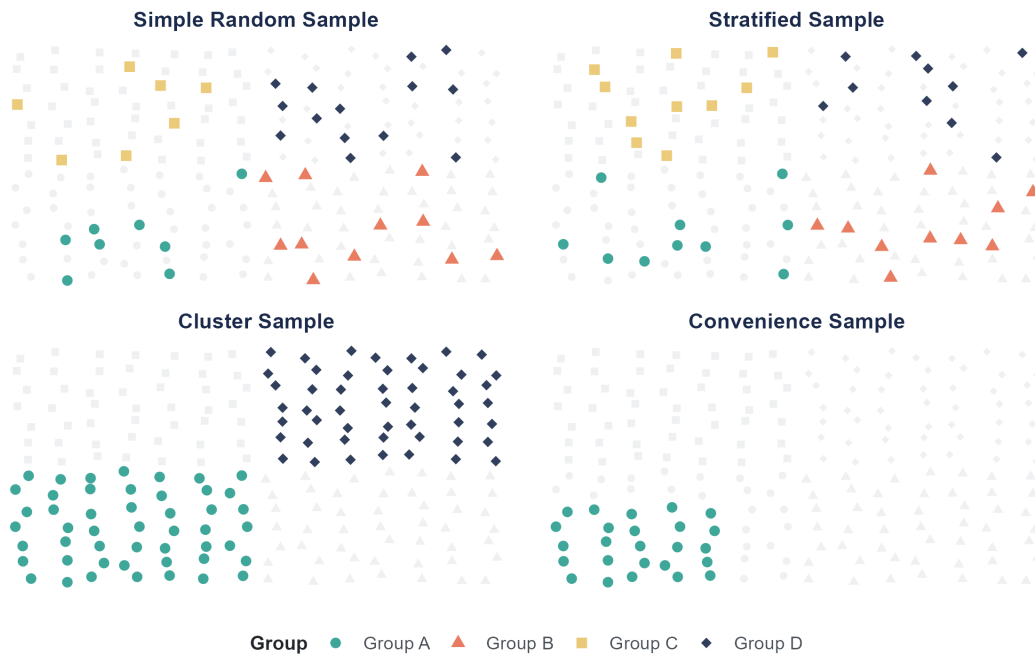
In practice, many studies use none of the methods described above. Instead, they rely on whoever is easy to reach, a practice called **convenience sampling**. A psychology professor studies the students in their introductory class. A political pollster surveys people who answer their phones. A health researcher recruits volunteers from a social media post.

Convenience samples are not inherently useless, but they are inherently limited. The people who are easy to reach are almost never representative of the broader population. Students in a psychology class are younger, more educated, and more likely to be from what researchers Henrich, Heine, and Norenzayan (2010) famously called WEIRD societies (Western, Educated, Industrialized, Rich, and Democratic), than the world population at large. Their widely cited paper in *Behavioral and Brain Sciences* argued that the overwhelming majority of published psychology research has been conducted on WEIRD participants, yet the resulting findings have routinely been described as universal features of human cognition, behavior, and social interaction. The implications are substantial. Findings about perception, memory, moral reasoning, social norms, and decision-making that appear stable and robust in Western undergraduate samples have, in many cases, not replicated in other cultural contexts, or have replicated only partially. If convenience sampling has systematically shaped a field's knowledge base for decades, then our understanding of "what people do" may really be a description of "what a specific and unrepresentative slice of humanity does." People who answer calls from unknown numbers differ from those who do not, though it is a distinction most of us have quietly settled for ourselves. Social media volunteers differ from people who are not on social media or who do not follow the researcher's account.

The problem is not that convenience samples exist. Many valuable studies have used them. The problem arises when researchers draw conclusions about a broad population from a convenience sample without acknowledging the limitations. If your study participants are all undergraduates at a single university, your findings apply to undergraduates at that university. They may or may not generalize to other people, and claiming otherwise requires an argument, not an assumption.

Comparing Sampling Methods

Same population, different ways of selecting a sample



Each panel highlights selected individuals (colored) from the full population (gray)

Figure 2.1: Four sampling methods applied to the same population, with selected individuals highlighted in color. Simple random sampling selects individuals at random, stratified sampling ensures proportional representation from each subgroup, cluster sampling includes entire groups, and convenience sampling selects individuals near a single location.

Try It Online

Open the **Sampling Explorer** on the companion website. Start with samples of 10, then increase to 50, 100, and 500, and watch the spread of estimates shrink. Try setting the population to have a rare subgroup (say, 5% of the total) and see how small samples can miss it entirely while larger samples capture it reliably.

2.4 Bias: When Data Leads You Astray

The word “bias” has a casual meaning (prejudice, unfairness) and a technical one. In statistics, **bias** refers to a systematic tendency for a method to produce results that are consistently too high, too low, or skewed in some particular direction. A biased method does more than make random errors. It makes errors that lean in the same direction, again and again.

Random error is like a dartboard where your throws scatter around the bullseye. Sometimes you are high, sometimes low, sometimes left, sometimes right, but on average you are close to the center. Bias is like a dartboard where the sights are off. Every throw lands in the same wrong spot, and throwing more darts does not fix the problem.

This distinction matters because increasing your sample size fixes random error but does not fix bias. A survey that systematically misses low-income households does not become representative just because you survey more people. It becomes a larger, more confidently wrong dataset.

Try It Online

The **Total Survey Error Explorer** maps every error type covered in this section as a clickable taxonomy tree. Select any node to see its definition, a practitioner example, and mitigation strategies. Each error is tagged with whether increasing sample size actually reduces it, a direct illustration of the distinction between random and systematic error. Use the List View to search across error types, or test yourself with the eight-question Self-Check quiz.

2.4.1 Selection Bias

Selection bias occurs when the process of selecting participants produces a sample that differs systematically from the population. We touched on this in Chapter 1. Here we dig deeper.

One famous historical example is the 1936 *Literary Digest* presidential poll. The magazine mailed mock ballots to about 10 million people, drawing its list from telephone directories, automobile registration records, magazine subscriber lists, and club and association rosters. It received approximately 2.4 million responses, an enormous sample by any standard. Based on those responses, the *Literary Digest* predicted that Alf Landon would defeat Franklin Roosevelt in a landslide, with Landon taking 57% of the popular vote. Roosevelt won 46 of 48 states, capturing approximately 61% of the vote.

What went wrong? Two problems compounded each other. First, the mailing list skewed heavily toward wealthier Americans: in 1936, during the Great Depression, telephones, cars, club memberships, and magazine subscriptions were all markers of relative affluence, and wealthier voters tended to favor Landon. But modern scholarship, notably Squire (1988), has established that this sampling bias alone cannot fully explain the magnitude of the error. The second and arguably more important problem was non-response bias. Of the 10 million ballots mailed, only about 24% came back. Landon supporters, more motivated by opposition to Roosevelt's New Deal, were far more likely to return their ballots. Roosevelt's supporters, who formed the larger share of the population, responded at much lower rates. The result was not merely a biased sample, it was a biased sample further filtered by differential non-response, compounding the original error. Meanwhile, George Gallup, using a much smaller but more carefully selected sample of about 50,000, correctly predicted Roosevelt's victory.

The lesson has not been fully learned. In 2016, many polls underestimated support for Donald Trump in key states, partly because their samples underrepresented voters without college degrees, a group that broke heavily for Trump. The problem was not sample size. National polls surveyed thousands of people. The problem was that the samples did not accurately reflect the composition of the electorate.

Selection bias can also emerge from the way data is generated. Hospital data contains information only about people who went to the hospital, which means it misses everyone who was sick but did not seek treatment. Online product reviews come from people motivated enough to write a review, who tend to be either very satisfied or very dissatisfied. Crime statistics reflect crimes that were reported, investigated, and documented, which may differ from the actual incidence of crime in ways that correlate with race, geography, and policing practices.

2.4.2 Response Bias

Even when you have a good sample, the answers people give may not be accurate. **Response bias** occurs when respondents systematically give answers that deviate from the truth.

People tend to over-report behaviors they consider socially desirable (exercising, voting, flossing) and under-report behaviors they consider undesirable (drinking, drug use, prejudice). This is called **social desirability bias**, and it affects self-reported data of all kinds.

The way a question is worded can also introduce bias. In a classic demonstration, Loftus and Palmer (1974) found that people estimated higher vehicle speeds when asked "How fast were the cars going when they *smashed* into each other?" compared to "How fast were the cars going

when they *contacted* each other?”, a difference of roughly 9 miles per hour on average (40.5 mph versus 31.8 mph), despite participants watching the same footage. Same event, different wording, different answers.

Leading questions, loaded terms, question order, and even the visual layout of a survey can all shape responses. A survey that asks “Do you support raising taxes to fund early childhood education?” will get different results than one that asks “Do you support raising taxes?” followed by a separate question about early childhood education. The survey designer’s choices are baked into the data, often invisibly.

2.4.3 Non-Response Bias

You mail out 1,000 surveys and get 200 back. Your response rate is 20%. The question that should keep you up at night is whether the 800 people who did not respond are different from the 200 who did.

Non-response bias occurs when the people who choose not to participate are systematically different from those who do. If dissatisfied customers are less likely to respond to a satisfaction survey (perhaps because they have already left), your results will overestimate satisfaction. If healthier people are more likely to participate in a health study (because they have the time and energy), your results will overestimate the health of the population.

Non-response bias has gotten worse over time as survey response rates have declined. In the 1970s, telephone surveys routinely achieved response rates of 70% or higher. Today, rates below 10% are common for cold-call surveys. When nine out of ten people decline to participate, the assumption that respondents represent the broader population becomes very hard to defend.

Researchers address non-response bias through follow-up contacts, incentive payments, and statistical weighting, adjusting the results to account for known differences between respondents and the population. None of these solutions is perfect, and all of them require assumptions about why people did or did not respond.

2.4.4 Survivorship Bias

During World War II, the U.S. military studied bombers that returned from combat missions to figure out where to add armor. The planes that came back had bullet holes concentrated in certain areas: the fuselage, the wings, the fuel system. The initial inclination was to reinforce those areas.

Abraham Wald, a mathematician working with the Statistical Research Group at Columbia University, pointed out the flaw in this reasoning. The military was only seeing the planes that had survived. The planes hit in the engines or cockpit had not made it back. The bullet holes on the surviving planes showed where a plane *could* afford to be hit. The missing data (the planes that did not return) indicated where the armor was actually needed. Wald's analysis, documented in a 1943 memorandum, redirected attention to the unobserved cases rather than the observed ones.

This is **survivorship bias**: the error of drawing conclusions from a sample that has been filtered by a survival process, while ignoring the cases that did not survive the filter. (Chapter 1 briefly noted this concept and the Wald example; here we treat it in full.)

Survivorship bias appears in many contexts. Business books study successful companies and extract lessons, without accounting for how many companies followed the same practices and failed. We read about college dropouts who became billionaires without accounting for the far larger number of dropouts who did not. Mutual fund performance statistics can appear impressive partly because funds that perform badly get closed and disappear from the database, a phenomenon that inflates apparent average returns.

The antidote to survivorship bias is to ask a specific question about every dataset you encounter. Who or what is missing from this data, and why?

Note

Abraham Wald's work on survivorship bias during World War II is one of statistics' most celebrated examples, but Wald himself was far more than this single insight. Born in 1902 in Cluj, then part of the Austro-Hungarian Empire (present-day Romania), Wald was a Jewish mathematician who fled to the United States after the Nazi annexation of Austria in March 1938. At Columbia, he produced foundational work in statistical decision theory and sequential analysis, the latter allowing researchers to analyze data as it accumulates rather than waiting for a predetermined sample size. This work remains central to how clinical trials are monitored today. Wald died in a plane crash in southern India in December 1950, at the age of 48.

2.5 Confounding: The Hidden Third Variable

This is a central concept in the book, and one that is often ignored in everyday reasoning.

Consider a study that finds that people who own pets live longer than people who do not. This is the same type of association we briefly flagged in Chapter 1 when discussing how to read news statistics; here we trace exactly how the reasoning can go wrong. The news headline writes itself. But before you rush to adopt a dog, think about what else might differ between pet owners and non-pet owners. Pet owners might tend to be wealthier (pets cost money). They might be more likely to have stable housing (many landlords do not allow pets). They might be more physically active (dogs need walks). All of these factors (wealth, housing stability, physical activity) are themselves associated with living longer.

A **confounding variable** (sometimes called a **confounder** or **lurking variable**) is a variable that is associated with both the explanatory variable and the outcome variable, creating a misleading association between them. If wealth is associated with both pet ownership and longevity, then wealth confounds the relationship between pet ownership and longevity. The apparent benefit of owning a pet might just be the benefit of having enough money and stability to own a pet.

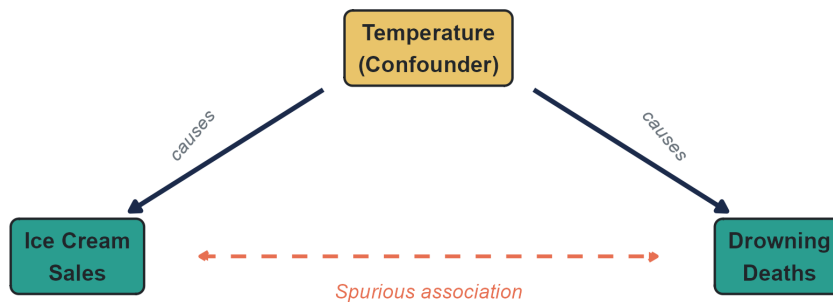
In a confounded relationship, there are three variables.

1. The **explanatory variable** (the one you think might be the cause). In our example, pet ownership.
2. The **response variable** (the outcome you are measuring). In our example, lifespan.
3. The **confounding variable** (the hidden third variable that is connected to both). In our example, wealth.

The confounding variable creates a pathway between the explanatory and response variables that has nothing to do with any direct causal connection. Even if pets had absolutely zero effect on lifespan, you would still see an association between pet ownership and longevity, because both are linked to wealth.

Confounding: When a Hidden Variable Drives Both

Temperature drives both ice cream sales and drowning deaths



Without accounting for temperature, one might wrongly conclude that ice cream causes drowning

Figure 2.2: A directed acyclic graph showing how temperature acts as a confounding variable. Temperature causes both ice cream sales and drowning deaths, creating a spurious association between the two. Without accounting for the confounder, the two appear related.

2.5.1 More Examples of Confounding

Confounding is not exotic. It is the default state of observational data.

Ice cream and drowning. There is a positive correlation between ice cream sales and drowning deaths. The confounder is temperature. Hot weather drives both ice cream consumption and swimming, and more swimming means more drownings. Ice cream does not cause drowning.

Shoe size and reading ability in children. Among a sample of children aged 5 through 12, shoe size is positively correlated with reading test scores. The confounder is age. Older children have bigger feet and better reading skills. Buying larger shoes will not improve your child's reading.

Number of firefighters at a fire and amount of property damage. More firefighters are dispatched to larger fires, which cause more damage. Sending fewer firefighters to a fire would not reduce the damage.

Hospital quality. A hospital that specializes in complex cancer surgeries might have a higher mortality rate than a community hospital that handles mostly routine procedures. Comparing their mortality rates directly is misleading because the complexity of cases, a confounding variable, differs substantially between them. This is a version of **Simpson's paradox**, where a trend that

appears in aggregated data reverses or disappears when the data is broken into subgroups. We will encounter this phenomenon in greater depth in later chapters; for now, note it as a reason why aggregated comparisons can mislead.

Education and income. People with more education tend to earn more money. But is this because education increases earning potential, or because people from wealthier families are more likely to attend college and would have earned more regardless? Family background is a confounder that makes the causal effect of education itself very difficult to isolate.

2.5.2 Why Confounding Makes Causation Hard

The fundamental problem with confounding is that there are always potential confounders you have not thought of. Even if you carefully measure and account for every confounding variable you can identify, there might be others you missed. In statistical language, there is always the possibility of unmeasured confounding or residual confounding.

This is why observational studies, no matter how large or carefully designed, generally cannot establish causation on their own. They can establish association. They can control for known confounders using statistical techniques (which we will explore in the regression chapters). They can make the case for causation more or less plausible. But they cannot eliminate the possibility that some unmeasured variable is responsible for the pattern in the data.

Randomized experiments solve this problem elegantly. When participants are randomly assigned to groups, confounders (both measured and unmeasured) are distributed roughly equally across groups. A randomized experiment on the effect of pet ownership on lifespan would randomly assign some people to own pets and others not to. (Setting aside the ethical and practical difficulties of such a study.) If the pet owners lived longer, you could be more confident that pets were the reason, because randomization would have balanced out wealth, housing stability, physical activity, and every other potential confounder, including ones you never thought to measure.

This is the deepest reason why the distinction between observational and experimental studies matters. It is not about methodology for its own sake. It is about whether confounders have been controlled, and therefore whether a causal interpretation is defensible.

2.6 When Can We Claim Causation?

Given everything we have discussed, when is it legitimate to say that X causes Y? This is a question we will revisit in later chapters, especially when we cover regression, but it is worth laying the groundwork now.

2.6.1 The Experimental Path

The most straightforward path to a causal claim is a well-designed randomized experiment. At its core, establishing causation requires three things. First, there must be a demonstrated relationship between the variables: X and Y must co-vary, meaning changes in X are associated with changes in Y. This is sometimes called **concomitant variation**. Second, the cause must precede the effect; X must come before Y in time. Third, the relationship must survive the elimination of alternative explanations: other variables that could plausibly produce the observed pattern must be ruled out.

In a well-designed experiment, all three conditions can be satisfied together. Random assignment creates the association, the manipulation establishes the time order, and the comparison between groups allows you to rule out alternatives. The strongest experiments also incorporate the following features.

A control group provides a baseline: without it, you cannot know whether the outcome changed because of the treatment or because of something else that happened during the study period.

Blinding prevents expectations from influencing results. When possible, participants do not know whether they are in the treatment or control group (single-blind), and neither do the researchers who assess outcomes (double-blind). Without blinding, knowing one's group assignment can change behavior in ways that contaminate the comparison.

Replication across multiple studies, ideally by different researchers in different settings, is what transforms a single finding into scientific consensus. The history of psychology offers many cautionary tales of effects that failed to replicate, so convergence across independent studies matters.

When all of these conditions are met, a causal claim is on strong footing. Few studies satisfy every condition perfectly, so the strength of a causal claim is always a matter of degree.

2.6.2 The Observational Path

For many important questions, experiments are not feasible. You cannot randomly assign people to smoke, to grow up in poverty, or to live in neighborhoods with different levels of air pollution. Does that mean we can never make causal claims about the effects of smoking, poverty, or pollution?

In practice, the scientific community does make causal claims based on observational data, but the bar is higher. The most famous example is smoking and lung cancer. The evidence that smoking causes cancer is overwhelming, but it comes almost entirely from observational studies. No randomized experiment has ever assigned humans to smoke for decades.

The case for causation from observational data typically rests on several converging lines of evidence, articulated by epidemiologist Sir Austin Bradford Hill in a 1965 lecture. The criteria that bear his name include the following considerations.

Consistency. The association has been found across multiple studies, in different populations, using different methods.

Strength. The association is large. Smokers are not 5% more likely to develop lung cancer. They are 15 to 30 times more likely, depending on the amount smoked.

Dose-response relationship. More exposure is associated with more effect. Heavier smokers develop cancer at higher rates than lighter smokers.

Temporal ordering. The cause precedes the effect. Smoking comes before cancer, not the other way around.

Biological plausibility. There is a credible mechanism. We understand how carcinogens in tobacco smoke damage lung tissue.

Elimination of alternative explanations. (This is a summarizing principle rather than one of Hill's original nine named criteria.) Known confounders have been accounted for, and the association persists.

These criteria are guidelines rather than a checklist. No single criterion is necessary or sufficient. But the more of them that are satisfied, the stronger the case for causation.

We will return to these ideas in Chapters 10 and 11, where regression models provide tools for statistically controlling for confounders. For now, the takeaway is this: experiments are the most direct route to causal claims. Observational studies can support causal claims when the evidence is strong, consistent, and resistant to alternative explanations. And a single observational study, no matter how large, is rarely enough on its own.

2.7 The Ethics of Research Design

The Facebook study was not the first time researchers caused harm by treating research design as a purely technical exercise. It was not even the worst case. But it serves as a useful anchor for a conversation about ethics that runs through every stage of the research process.

2.7.1 Informed Consent

The principle of **informed consent** means that people who participate in research should know they are participating, understand what the research involves, understand any risks, and have the genuine ability to decline without penalty.

This principle did not emerge from abstract philosophizing. It emerged from horror. The Nuremberg Code of 1947 was written in response to the medical experiments conducted by Nazi physicians on concentration camp prisoners during World War II. The Declaration of Helsinki, first adopted in 1964 by the World Medical Association and revised multiple times since, extended and formalized these protections specifically for medical research involving human participants. The Belmont Report, published in 1979 by the U.S. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, established the ethical framework that governs research in the United States to this day, built on three principles: respect for persons, beneficence, and justice.

In the Facebook study, informed consent was essentially absent. The researchers argued that the Terms of Service constituted consent, but terms of service are legal documents designed to protect the company, not research protocols designed to protect participants. Millions of people accept terms of service without reading them. Even among those who read them, very few would understand that they were consenting to participation in a psychological experiment.

The standard for informed consent in research is higher than the standard for agreeing to a commercial contract, and it should be. Research participants are not customers making a purchasing decision. They are people whose data, behavior, or bodies are being used to generate knowledge, often for the benefit of others. They deserve to know what they are contributing to and to have a real choice about whether to contribute.

2.7.2 Power and Data Collection

Who gets studied, and who does the studying? These questions sit at the center of an established body of scholarship in research ethics, feminist science studies, and critical data studies, not a peripheral concern but a recognized analytical framework for understanding how power shapes what knowledge gets produced.

The language we use in research reflects these power dynamics. For decades, people enrolled in studies were routinely called “subjects,” a word that carries an implicit relationship of authority over the studied person. The preferred term today is participant, which recognizes that people who take part in research have agency and rights, not just compliance and obligations. The shift in language is not cosmetic. It reflects a genuine shift in how the research community understands the relationship between researchers and the people they study.

In their book *Data Feminism*, Catherine D’Ignazio and Lauren Klein argue that data science (broadly, the practice of extracting knowledge and insight from data using statistical, computational, and domain expertise) is not neutral. The choices about what to measure, what categories to use, whose perspectives to center, and whose to ignore are shaped by the power structures of the institutions that collect and analyze data. A dataset is not a mirror of reality. It is a product of decisions, and those decisions reflect who had the power to make them. As they put it, in data work as in other domains, power imbalances shape what questions get asked, who gets to ask them, and whose lives become the data.

Cathy O’Neil’s *Weapons of Math Destruction* (2016) documents how this plays out in practice. Algorithms used for hiring, lending, college admissions, and criminal sentencing can systematically disadvantage the communities that have the least power to contest or even detect their operation, often while appearing to be objective because they are mathematical. The appearance of neutrality can obscure embedded bias in ways that human decision-makers’ biases cannot, precisely because the process is opaque.

Consider the case of Henrietta Lacks, an African American woman whose cancer cells were taken during a biopsy at Johns Hopkins Hospital in 1951 without her knowledge or consent. Those cells, known as HeLa cells, became one of the most important tools in medical research, contributing to developments in cancer treatment, vaccine development, virology, and genetics. The Lacks family did not learn about the cells until decades later and received no compensation while companies commercialized products developed using her tissue. The story, documented in Rebecca Skloot’s *The Immortal Life of Henrietta Lacks* (2010), illustrates how race, class, and institutional power shaped who was treated as a research participant and on whose terms.

The collection of data from marginalized communities without their consent or benefit is not only a historical issue. Joy Buolamwini and Timnit Gebru’s 2018 research, published as “Gender Shades” in the *Proceedings of Machine Learning Research*, demonstrated that commercial facial recognition systems performed substantially worse on darker-skinned faces than on lighter-skinned faces, with error rates for darker-skinned women up to 34 percentage points higher than for lighter-skinned men. Predictive policing algorithms trained on historical arrest data, which reflects decades of racially biased policing practices, produce outputs that concentrate police attention in communities of color, creating a self-reinforcing cycle. Data reflects the world that generated it, including the injustices in that world. Analyzing that data without acknowledging those injustices can reinforce them.

Ethics Moment

In 2016, the investigative journalism team at ProPublica (Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner) published an analysis of COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), a commercial risk assessment algorithm used across the United States to predict whether a defendant would re-offend before trial or sentencing. Their analysis found that the algorithm was roughly equally accurate overall for Black and white defendants, but the types of errors it made differed systematically. Black defendants who did not go on to re-offend were almost twice as likely as white defendants to be incorrectly classified as high risk. White defendants who did re-offend were more likely to be incorrectly classified as low risk. The company that produced COMPAS argued that its algorithm satisfied one established definition of fairness: equal predictive accuracy across groups. ProPublica argued that it violated another: equal false positive rates across groups. Subsequent academic research has established that these two definitions of fairness are mathematically incompatible: no algorithm can simultaneously satisfy both when base rates differ across groups. This means “fairness” in algorithmic systems is not a technical problem with a single correct solution. It is an ethical and political question about which errors matter more and who bears their cost, questions that statistics can illuminate but cannot resolve. For the original analysis, see “Machine Bias” at propublica.org.

2.7.3 Ethical Obligations at Every Stage

Ethics is not a box you check at the beginning of a study and then forget about. Ethical considerations arise at every stage of research design.

In choosing a question. Whose questions get asked? Whose problems get studied? When data science resources (defined earlier in this chapter) flow disproportionately toward problems that interest wealthy institutions, the problems of less powerful communities go unexamined.

In designing the study. Who is included and who is excluded? What categories are offered on a survey? If a survey asks about gender with only “male” and “female” as options, it erases non-binary respondents. If a clinical trial excludes pregnant women “for safety,” it means that drugs reach the market without evidence about their effects during pregnancy.

In collecting data. Are participants fully informed? Do they have actual power to decline? A worker asked to participate in a company-sponsored study may feel pressured even if participation is technically voluntary. A student asked by a professor to fill out a survey may worry about consequences for declining.

In analyzing data. Are the researchers transparent about their methods? Do they report results that are inconvenient for their hypothesis, or do they selectively present only the findings that tell the story they want? This is a form of bias that no sampling method can fix.

In reporting results. Who has access to the findings? Are the results communicated in ways that people can actually understand and use? Publishing a finding in an academic journal behind a paywall serves a very different constituency than making it available to the community that was studied.

D’Ignazio and Klein’s *Data Feminism* (MIT Press, 2020) is available as an open-access book at data-feminism.mitpress.mit.edu and provides a thorough, readable treatment of these themes. It is worth engaging with in full, not because it provides easy answers, but because it provides better questions.

AI Reality Check

Ask an AI tool to summarize a research study and it will almost always give you a confident answer. What it will rarely do, unless you ask, is tell you whether the study was observational or experimental, and what that distinction means for the conclusions.

This matters because AI tools routinely describe observational findings using causal language. “Researchers found that coffee reduces the risk of heart disease” is how an AI might summarize a study that actually found an association between coffee consumption and lower heart disease rates in a cohort of volunteers. The study did not manipulate coffee intake. It did not randomly assign anyone to drink or abstain. It cannot support the claim that coffee *re-*

duces anything. But the AI summary sounds like it can, because the AI does not distinguish between “is associated with” and “causes.”

The same problem extends to confounders. If you ask an AI to identify the key findings of a study, it will report the results as stated. It will not, on its own, ask whether important confounders were measured, whether the comparison groups were comparable, or whether the sampling method limits generalizability. These are the questions this chapter has trained you to ask, and they are precisely the questions that AI tools skip.

When you use AI to help you read or summarize research, always follow up with three questions: Was this study observational or experimental? What confounders might explain the result? How was the sample selected? If the AI’s summary does not address these, the summary is incomplete regardless of how polished it sounds.

2.8 Putting It All Together: Evaluating a Study

When you encounter a research study (whether in a news article, a journal paper, or a business report), here is a framework for evaluating its design.

- 1. What is the research question?** Is it clearly stated? Is it a question that data can actually answer?
- 2. Is it observational or experimental?** If observational, be cautious about causal claims. If experimental, check whether randomization was used and whether there was a proper control group.
- 3. How was the sample selected?** Was it a probability sample (simple random, stratified, cluster, systematic) or a convenience sample? Does the sampling frame match the population of interest?
- 4. How large is the sample?** Larger samples give more precise estimates, but size alone does not fix bias. A biased sample of 10,000 is worse than an unbiased sample of 500.
- 5. What biases might be present?** Think about selection bias, response bias, non-response bias, and survivorship bias. Consider who might be missing from the data and why.
- 6. What confounders might exist?** If the study claims an association between two variables, think about what third variables could be creating or inflating that association.
- 7. Were ethical standards followed?** Did participants give informed consent? Were vulnerable populations protected? Is the data being used in ways participants would reasonably expect?

8. Are the conclusions proportional to the evidence? A well-designed randomized experiment can support stronger conclusions than an observational study. A single study, no matter how well designed, provides weaker evidence than multiple studies with consistent findings.

This framework is not a formula. It is a habit of mind. The more you practice it, the better you will become at distinguishing studies that deserve your trust from studies that do not.

2.9 Looking Ahead

In Chapters 3 and 4, we will turn to the tools for describing what is in a dataset, using numbers and pictures. These tools may seem simpler than research design, but they depend on everything we have covered here. A beautiful histogram of a biased sample is still a picture of bias. A summary statistic calculated from confounded data summarizes the wrong thing. The methods in the next chapters are powerful, but only when applied to data that was collected thoughtfully.

Research design is not the glamorous part of statistics. It does not involve elegant formulas or impressive software. But it is the part that determines whether your analysis means anything at all.

2.10 Key Terms

- **Observational study:** A study in which the researcher observes and measures without intervening or assigning treatments.
- **Experiment:** A study in which the researcher deliberately assigns treatments to participants and compares outcomes.
- **Randomized controlled experiment:** An experiment in which participants are randomly assigned to treatment and control groups, considered the strongest design for establishing cause and effect.
- **Quasi-experiment:** A study that resembles an experiment but lacks true random assignment, often exploiting natural or policy-driven variation.
- **A/B test:** A randomized experiment, typically conducted in a digital environment, comparing two versions of something (a webpage, an email, an ad) to see which performs better.
- **Treatment group:** The group in an experiment that receives the intervention being studied.

- **Control group:** The group in an experiment that does not receive the intervention, serving as a baseline for comparison.
- **Blinding:** A design feature in which participants (single-blind) or both participants and researchers (double-blind) do not know who is in the treatment or control group.
- **Simple random sampling:** A sampling method in which every member of the population has an equal chance of being selected.
- **Sampling frame:** A complete list of the population from which a sample is drawn.
- **Stratified sampling:** A sampling method that divides the population into non-overlapping subgroups (strata) and samples from each stratum separately.
- **Cluster sampling:** A sampling method that randomly selects groups (clusters) from the population and includes all members of the selected clusters.
- **Systematic sampling:** A sampling method that selects every k th member from a list, starting at a randomly chosen point.
- **Convenience sampling:** A non-probability sampling method that uses whoever is easiest to reach.
- **Bias:** A systematic tendency for a method to produce results that deviate from the truth in a consistent direction.
- **Selection bias:** Bias arising from the way participants are chosen for a study, producing a sample that does not represent the population.
- **Response bias:** Bias arising from participants giving inaccurate answers, often due to social desirability, question wording, or question order.
- **Non-response bias:** Bias arising when people who do not participate in a study are systematically different from those who do.
- **Survivorship bias:** Bias arising from studying only the cases that passed through some selection or survival process, while ignoring those that did not.
- **Social desirability bias:** A specific form of response bias in which respondents over-report socially approved behaviors and under-report disapproved ones.
- **Confounding variable (confounder):** A variable that influences both the explanatory variable and the response variable, creating a misleading association between them.
- **Lurking variable:** Another term for a confounding variable, emphasizing that it may not be measured or even recognized.
- **Simpson's paradox:** A phenomenon in which a trend that appears in aggregated data reverses or disappears when the data is separated into subgroups.
- **Informed consent:** The principle that research participants should understand what a study involves, including any risks, and should freely agree to participate.

- **Institutional review board (IRB):** A committee that reviews research proposals to ensure they meet ethical standards for the protection of human participants.
 - **Bradford Hill criteria:** A set of guidelines, articulated by epidemiologist Sir Austin Bradford Hill in 1965, for evaluating whether an observed association in observational data is likely to be causal.
 - **Concomitant variation:** The condition that cause and effect must co-vary: changes in the explanatory variable must be associated with changes in the outcome variable.
 - **Explanatory variable:** The variable hypothesized to influence or predict the response. Also called the independent variable or predictor.
 - **Randomization:** The process of using chance to assign participants to treatment and control groups, ensuring that pre-existing differences are approximately balanced.
 - **Response variable:** The variable being studied as the outcome. Also called the dependent variable.
 - **Retrospective study:** A study that looks backward in time, using existing records or recall to reconstruct what happened to participants.
-

2.11 Further Reading and References

The following works are cited in this chapter or provide valuable additional context for the ideas covered here.

On the Facebook emotional contagion study: Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24), 8788–8790.

On WEIRD sampling and its consequences: Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2–3), 61–83.

On the Literary Digest poll: Squire, P. (1988). Why the 1936 Literary Digest poll failed. *Public Opinion Quarterly*, 52(1), 125–133.

On response bias and eyewitness testimony: Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior*, 13(5), 585–589.

On research ethics and the Belmont Report: National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont Report*. U.S. Department of Health, Education, and Welfare.

On power, data, and algorithmic bias: D'Ignazio, C., & Klein, L. F. (2020). *Data feminism*. MIT Press. Available open access at data-feminism.mitpress.mit.edu.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

ProPublica. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. Angwin, J., Larson, J., Mattu, S., & Kirchner, L. [propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing](https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing).

On Henrietta Lacks: Skloot, R. (2010). *The immortal life of Henrietta Lacks*. Crown.

On facial recognition bias: Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, 81, 1–15.

On the Bradford Hill criteria: Hill, A. B. (1965). The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine*, 58(5), 295–300. (The lecture that introduced the nine criteria for evaluating causal claims from observational data, discussed in this chapter's section on when observational evidence can support causation.)

2.12 Exercises

2.12.1 Check Your Understanding

1. In the Facebook emotional contagion study, researchers manipulated users' News Feeds and measured changes in the emotional tone of their posts. Was this an observational study or an experiment? Explain your reasoning.
2. A school district wants to know whether a new math curriculum improves test scores. They implement the new curriculum at five schools and keep the old curriculum at five other schools, then compare end-of-year test scores. Is this an experiment or an observational study? If it is an experiment, is it a randomized experiment? What concerns might you have about this design?

3. Explain the difference between stratified sampling and cluster sampling. In stratified sampling, you sample *from* each subgroup. In cluster sampling, you sample *entire* subgroups. Why does this distinction affect the precision of your estimates?
4. A researcher surveys 1,000 adults by calling landline phones between 9 a.m. and 5 p.m. on weekdays. Identify at least two sources of bias in this sampling approach and explain the direction in which each bias would likely push the results.
5. A news article reports that children who eat dinner with their families at least five nights per week have higher grades than children who eat family dinners less often. A parent reads this and concludes that making their family eat dinner together will improve their child's grades. What is the problem with this reasoning? Identify at least one plausible confounding variable.
6. Explain the difference between random sampling and random assignment. Which one helps with generalizability (applying results to the broader population)? Which one helps with establishing causation?
7. A fitness app reports that its users lose an average of 12 pounds in the first three months. Why might this statistic be misleading? Identify the specific type of bias most likely at work.
8. In your own words, explain why increasing the sample size reduces random error but does not reduce bias. Use an analogy or example.
9. A pharmaceutical company tests a new drug by giving it to 500 volunteers who responded to an advertisement and comparing their outcomes to published data on patients with the same condition who did not receive the drug. Identify at least two problems with this design.
10. Describe a scenario in which a quasi-experiment might be used because a true randomized experiment would be impractical or unethical. Explain what makes the study quasi-experimental rather than a true experiment.

2.12.2 Apply It

1. You want to estimate the average commute time for employees at a company with 2,000 workers spread across four office locations (New York with 800 employees, Chicago with 500, Denver with 400, and Atlanta with 300). Design a stratified sampling plan with a total sample of 200 employees. How many employees would you sample from each location? Why is stratified sampling preferable to simple random sampling here?
2. A university wants to survey students about their experience with mental health services on campus. They have a list of all 12,000 enrolled students. Design a systematic sampling plan to select 400 students. What is the sampling interval? What would you do if the list were sorted by class year (all freshmen first, then sophomores, etc.), and how might that ordering affect your sample?

3. A hospital wants to study whether a new discharge procedure reduces patient readmission rates. They implement the new procedure on Floor 3 and keep the old procedure on Floor 5. Over six months, they compare readmission rates for patients from the two floors. Identify the explanatory variable, the response variable, and at least two potential confounding variables. Is this a true experiment? Why or why not?
4. A polling organization surveys 2,000 registered voters about their support for a proposed policy. The overall response rate is 15%. Among respondents, 62% support the policy. The polling organization reports that “62% of voters support the policy.” What type of bias is most concerning here? In which direction might the bias push the result, and why? What additional information would help you assess the severity of the bias?
5. Consider the following pairs of variables. For each pair, identify a plausible confounding variable that could explain the association, and explain how the confounder is related to both variables.
 - a. Number of hours spent studying and exam score
 - b. Neighborhood income level and life expectancy
 - c. Amount of organic food consumed and frequency of illness
 - d. Number of books in a home and children’s test scores
6. A tech company runs an A/B test on its website. Half of randomly selected visitors see a redesigned checkout page (Version B) and half see the original (Version A). After one week, Version B has a 4.2% conversion rate compared to Version A’s 3.8% conversion rate, based on 50,000 visitors per group. What makes this study an experiment? What advantage does this design have over simply redesigning the page and comparing sales before and after the change?
7. A researcher wants to study the effect of class size on elementary school student performance. Propose two study designs, one observational and one experimental. For each design, describe the sample, the explanatory variable, the response variable, and one key limitation.
8. A local government claims that its new public transit system has reduced traffic congestion by 20%, based on a comparison of average commute times before and after the system opened. What information would you need before accepting this claim? Identify at least three factors that could confound the comparison.
9. You are asked to evaluate a survey sent to all customers who made a purchase in the past year. The survey asks about satisfaction on a 1-to-7 scale and includes open-ended questions about complaints. The response rate is 8%. Write a brief memo (3 to 5 sentences) to a manager explaining why the results should be interpreted with caution. Be specific about the type of bias and its likely direction.
10. Using publicly available data from a source such as the U.S. Census Bureau (data.census.gov) or the World Bank (data.worldbank.org), find two variables for a set of geographic units (states,

countries, or counties) that are positively correlated. Identify a plausible confounding variable for the association. Explain why the correlation between the two variables does not necessarily mean that one causes the other.

2.12.3 Think Deeper

1. The Facebook emotional contagion study was reviewed by Cornell University's IRB, which determined that its review was not required because Facebook, not Cornell, had collected the data. Evaluate this reasoning. Should the ethical standards for research conducted by companies be the same as those for research conducted by universities? If not, should they be higher, lower, or different in some other way? What are the implications for the growing volume of research conducted by technology companies using data from their platforms?
2. *Data Feminism* by D'Ignazio and Klein argues that "what gets counted counts," meaning that the decision about what to measure and what to ignore shapes what society pays attention to. Identify a social issue where you believe important data is not being collected. What would you want to measure, and why? Who has the power to decide whether this data gets collected, and what barriers might prevent it from being collected?
3. Consider the tension between two goals in research design. On one hand, we want research to produce generalizable knowledge that helps many people. On the other hand, we want to protect individuals from being harmed by the research process. Using the Henrietta Lacks case and the Facebook study as examples, discuss how these goals can come into conflict. How should researchers and institutions balance them?
4. Survivorship bias affects many areas of everyday reasoning, including career advice, business strategy, and personal finance. Choose one domain outside of academic research and describe a specific example of survivorship bias that could lead to bad decisions. What data would you need to correct for the bias, and how easy or difficult would that data be to obtain?
5. Algorithmic decision-making tools like COMPAS are increasingly used in criminal justice, hiring, lending, and other high-stakes domains. ProPublica's analysis showed that COMPAS satisfied one mathematical definition of fairness while violating another. Research has since shown that certain definitions of fairness are mathematically incompatible, meaning no algorithm can satisfy all of them simultaneously. Given this impossibility, who should decide which definition of fairness to use in a given context? What role should statistical analysis play in this decision, and what are its limits? In thinking about this, you may find it useful to consult Cathy O'Neil's *Weapons of Math Destruction* (2016), which examines how the appearance of mathematical objectivity can obscure embedded bias in high-stakes algorithmic systems.

Chapter 3

Summarizing Data with Numbers

3.1 Two Economies, One Country

In January 2024, the U.S. economy found itself in a strange position. Depending on which cable news channel you turned to, it was either thriving or broken.

On one channel, a commentator pointed out that the American economy had grown at an impressive rate, that unemployment was near historic lows, and that the country's gross domestic product per capita was the envy of the developed world. The economy, this person said, was strong. On another channel, a different commentator pointed out that the average American was struggling to afford groceries, that housing costs had spiraled beyond the reach of young families, and that real wages for the bottom half of earners had barely budged in decades. The economy, this person said, was failing ordinary people.

Both were citing real data. Neither was lying. So how could they look at the same economy and reach opposite conclusions?

The answer lies in a choice that most viewers never notice, the choice of which number to report. When you summarize an entire distribution of data with a single value, you are making a decision about what matters. And that decision has consequences.

Consider household income in the United States. According to the U.S. Census Bureau, the median household income in 2022 was approximately \$74,580. The mean household income, calculated by the same bureau using the same data, was considerably higher, around \$105,000. That is a gap of roughly \$30,000 between two numbers that both claim to represent the “typical” American household.

The gap exists because income in the United States is heavily right-skewed. Most households earn somewhere between \$30,000 and \$100,000. But a relatively small number of households earn far more — hundreds of thousands, millions, or even billions per year. Those extreme values pull the mean upward while barely affecting the median. A politician who wants to paint a rosy picture reports the mean. A politician who wants to highlight the struggles of working families reports the median. Both are “correct.” Neither is complete.

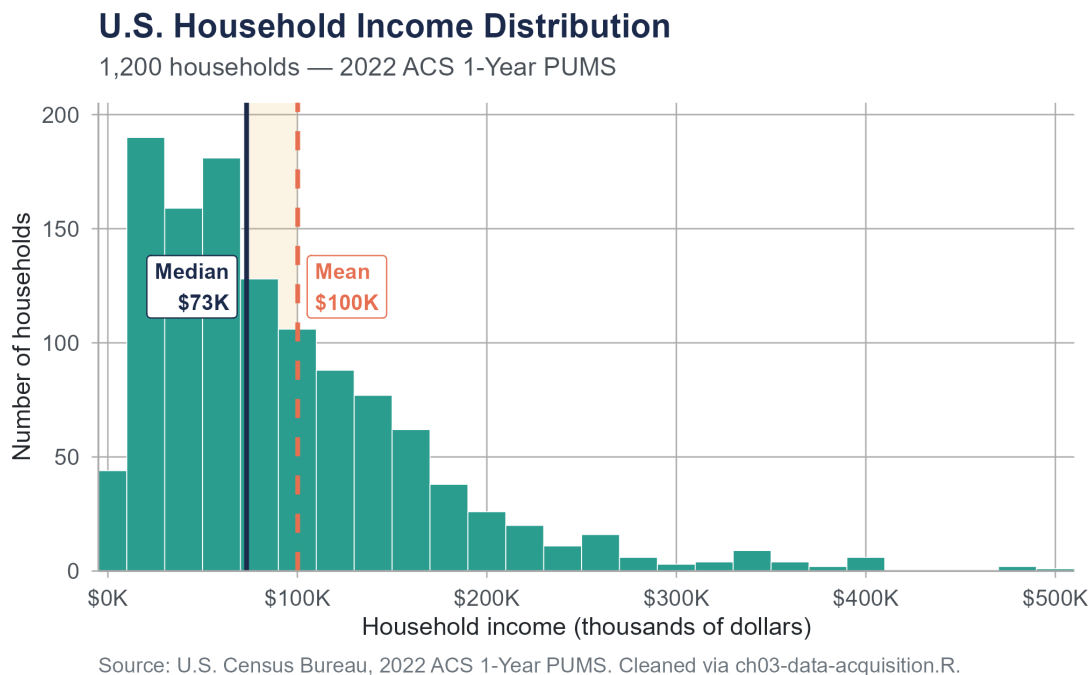


Figure 3.1: U.S. household income distribution with the mean and median marked. The right-skewed shape pulls the mean well above the median, illustrating how the choice of summary statistic can tell very different stories about the same data. Source: U.S. Census Bureau, 2022 ACS 1-Year PUMS.

This chapter is about the numbers we use to summarize data, what they tell us, what they hide, and how to choose among them wisely. By the end of it, you will understand why that \$30,000 gap exists, why it matters, and how to avoid being fooled by someone else’s choice of summary statistic.

3.2 Measures of Center

When someone asks, “What is the typical value in this dataset?”, they are asking for a measure of center. There are three common answers to that question, and they do not always agree.

3.2.1 The Mean

The **mean** is the arithmetic average. Add up all the values, divide by how many there are. If you have five test scores of 72, 85, 90, 68, and 95, the mean is

$$\bar{x} = \frac{72 + 85 + 90 + 68 + 95}{5} = \frac{410}{5} = 82$$

More formally, for a dataset with n observations $x_1, x_2, \dots, x_n$, the sample mean is

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

The mean is the most widely used measure of center. It has nice mathematical properties, it uses every value in the dataset, and it forms the basis of many statistical procedures we will encounter later in this book.

But the mean has a weakness. It is sensitive to extreme values. Suppose one of those five students missed most of the semester and scored a 15 instead of a 95. The mean drops from 82 to 66, even though four of the five scores are unchanged. A single extreme value pulled the center of the data in its direction.

This sensitivity is exactly what creates the gap between mean and median household income. Bill Gates walks into a diner with nine other people. Before he arrived, the mean net worth of the people in the room was maybe \$50,000. Now it is over ten billion dollars. Nothing about the financial reality of those nine people changed, but the mean tells a very different story.

3.2.2 The Median

The **median** is the middle value when the data is arranged in order. Half the observations fall below it, half fall above it.

To find the median, sort the data from smallest to largest. If the number of observations is odd, the median is the middle value. If the number of observations is even, the median is the average of the two middle values.

For the original five test scores, sorted: 68, 72, 85, 90, 95. The median is 85, the value in position 3.

Now replace 95 with 15: 15, 68, 72, 85, 90. The median is 72. It changed, but not nearly as dramatically as the mean did. The median is **resistant** to extreme values, meaning that outliers do not pull it around the way they pull the mean.

This resistance is why economists and government agencies typically report median household income rather than mean household income when they want to describe what life is like for a typical family. The median says, “Half of households earn less than this, and half earn more.” It is not distorted by the fortunes of billionaires.

3.2.3 The Mode

The **mode** is the most frequently occurring value in a dataset. In the set 3, 5, 5, 7, 8, the mode is 5. In the set 2, 2, 4, 4, 6, there are two modes (2 and 4), making the distribution **bimodal**. Some datasets have no repeated values at all, in which case the mode is not particularly useful.

The mode is the only measure of center that works for categorical data. If you survey 200 people about their favorite season and 78 say autumn, 52 say summer, 41 say spring, and 29 say winter, the mode is autumn. You cannot calculate a mean or median for seasons.

For numerical data, the mode is less commonly reported than the mean or median. But it becomes useful when you want to describe the shape of a distribution, particularly when that distribution has more than one peak.

3.2.4 When to Use Which

The choice between mean, median, and mode is not arbitrary. It depends on the data.

Use the mean when the data is roughly symmetric and has no extreme outliers. The mean uses all the information in the dataset and is the foundation of many inferential methods. For symmetric distributions, the mean and median will be close to each other anyway.

Use the median when the data is skewed or contains outliers. Income, home prices, hospital bills, commute times, and many other real-world variables are right-skewed, with a long tail of high values. For these, the median better represents what a typical observation looks like.

Use the mode for categorical data or when you specifically want to identify the most common category or value. The mode can also be informative for numerical data when the distribution is bimodal, revealing that the data may contain two distinct groups.

Here is a practical guideline. If you compute both the mean and the median and they are close together, the data is probably roughly symmetric and either measure works well. If they are far apart, the data is skewed, and you should think carefully about which one answers the question you are actually asking.

The choice matters in everyday life more than most people realize. Consider a housing market scenario that plays out in cities across the country. Imagine a real estate website reporting that the “average home price” in a particular suburb is \$620,000. Prospective buyers visit the neighborhood expecting to find homes near that price point, but most of the houses are priced between \$280,000 and \$400,000. The average is inflated by a handful of newly built luxury properties on large lots, including one listed at \$4.2 million. The median home price, around \$345,000, would have painted a far more accurate picture of what a typical buyer could expect to find. This pattern, a small number of high-end properties pulling the mean well above what most buyers will encounter, is characteristic of any market with substantial price variation at the top.

This is not an unusual scenario. It plays out in every city where a few high-end properties coexist with a larger stock of modest homes. Real estate agents know this, and the savvier ones choose their statistics deliberately. An agent trying to attract wealthy buyers to a neighborhood will quote the mean. An agent trying to reassure a young couple that a neighborhood is affordable will quote the median. Both are telling the truth. Both are also being strategic about which truth to tell.

The same dynamic appears in salary negotiations. A company might report that the “average salary” for its employees is \$95,000. That sounds impressive until you learn that the median is \$62,000, and the average is pulled up by a few executives earning seven figures. A job candidate who walks into an interview expecting \$95,000 based on the average may be disappointed. The median would have set more realistic expectations.

i Note

The relationship between mean and median can tell you the direction of **skewness**. In a right-skewed distribution (long tail to the right), the mean is greater than the median, because the

extreme high values pull the mean upward. In a left-skewed distribution (long tail to the left), the mean is less than the median. This is not a perfect diagnostic, but it is a useful rule of thumb. The income data is a textbook case. Mean income exceeds median income in virtually every country on Earth, because income distributions are almost universally right-skewed.

3.3 Why Measures of Center Are Not Enough

Imagine two classes of 20 students each take the same exam. Both classes have a mean score of 75. Are these two classes performing similarly?

Not necessarily. In Class A, every student scored between 70 and 80. In Class B, half the students scored around 95 and the other half scored around 55. The averages are identical, but the student experiences are completely different. Class A is consistent. Class B is split.

The mean told you where the center is. It told you nothing about how spread out the data is around that center. For that, you need measures of spread, also called measures of variability or dispersion.

This is more than an academic point. It has life-or-death consequences in fields like medicine. Consider two hospitals that both report an average surgical wait time of 14 days. At Hospital A, every patient waits between 12 and 16 days. The process is predictable, and patients can plan around it. At Hospital B, some patients are seen within 2 days while others wait 45 days. The average is the same, but the patient experience could hardly be more different. A patient at Hospital B faces real uncertainty about when they will receive care, and those stuck in the long tail of the wait time distribution may suffer worsening conditions. Reporting only the average conceals a serious problem in the way care is distributed.



Try It Online

The **Datasaurus Explorer** shows a set of datasets that all share the same mean, standard deviation, and correlation, yet look completely different when plotted. Open it now, before reading further. The visual makes the chapter's argument more forcefully than any example in text can.

The same principle applies in manufacturing. A pharmaceutical company producing pills with a target dosage of 200 milligrams needs to care about more than the average dosage across a batch. If the average is 200 mg but individual pills range from 150 mg to 250 mg, patients are receiving

inconsistent and potentially dangerous doses. The spread of the data, beyond its center, determines whether the product is safe.

Measures of spread answer the question, “How much do the values in this dataset differ from one another?” A small spread means the values are clustered tightly around the center. A large spread means they are scattered widely.

3.4 Measures of Spread

3.4.1 The Range

The simplest measure of spread is the **range**, which is the difference between the maximum and minimum values.

$$\text{Range} = \text{Maximum} - \text{Minimum}$$

For the test scores 68, 72, 85, 90, 95, the range is $95 - 68 = 27$.

The range is easy to compute and easy to understand. But it has a serious limitation: it depends on only two values, the largest and the smallest. One extreme observation can dramatically inflate the range, and the range tells you nothing about how the rest of the data is distributed between those two extremes.

Consider two datasets. Dataset A contains the values 50, 51, 52, 53, 54, 55, 100. Dataset B contains 50, 60, 70, 80, 90, 95, 100. Both have a range of 50, but they look very different. In Dataset A, most values are clustered around 50, with a single outlier at 100. In Dataset B, the values are spread evenly across the range. The range cannot distinguish between these situations.

3.4.2 The Interquartile Range (IQR)

A more resistant measure of spread is the **interquartile range**, or IQR. Rather than looking at the full range from minimum to maximum, the IQR measures the range of the middle 50% of the data.

$$\text{IQR} = Q_3 - Q_1$$

where Q_1 is the first quartile (25th percentile) and Q_3 is the third quartile (75th percentile). We will define quartiles more precisely in a moment, but the idea is straightforward. Q_1 is the value

below which 25% of the data falls, and Q_3 is the value below which 75% of the data falls. The IQR tells you how wide the middle half of the distribution is.

Because the IQR ignores the most extreme values at both ends, it is resistant to outliers, just as the median is resistant to outliers as a measure of center. The IQR is the natural companion to the median, and the two are often reported together.

3.4.3 Variance and Standard Deviation

The range and IQR tell you about the spread of data using just a few selected values. The **variance** and **standard deviation** take a different approach. They measure how far, on average, each observation is from the mean. These are the most commonly used measures of spread in statistical practice, and they are worth understanding in detail.

Let us build the intuition step by step.

Step 1: Deviations from the mean. For each observation, calculate how far it is from the mean. If the mean is $\bar{x}$ and an observation is x_i , the deviation is $x_i - \bar{x}$. Some deviations will be positive (values above the mean) and some will be negative (values below the mean).

Consider a small dataset: 4, 7, 8, 11, 15. The mean is $\bar{x} = \frac{4+7+8+11+15}{5} = 9$. The deviations are:

Observation (x_i)	Deviation ($x_i - \bar{x}$)
4	$4 - 9 = -5$
7	$7 - 9 = -2$
8	$8 - 9 = -1$
11	$11 - 9 = 2$
15	$15 - 9 = 6$

Step 2: Why not just average the deviations? You might think we could measure spread by averaging these deviations. The problem is that the positive and negative deviations always cancel out. Try it: $(-5) + (-2) + (-1) + 2 + 6 = 0$. The sum of the deviations from the mean is always zero. This is a mathematical property of the mean, not a coincidence.

Step 3: Squaring the deviations. To prevent the cancellation, we square each deviation. Squaring makes every value positive, and it gives extra weight to large deviations, which is often desirable because large deviations represent observations that are far from the center.

Observation (x_i)	Deviation ($x_i - \bar{x}$)	Squared Deviation ($(x_i - \bar{x})^2$)
4	-5	25
7	-2	4
8	-1	1
11	2	4
15	6	36

Step 4: Computing the sample variance. Sum the squared deviations and divide, but not by n . For a sample, we divide by $n - 1$ rather than n , so the result is not a simple average. The reason involves a concept called **degrees of freedom**, which we will address more thoroughly in later chapters. The short version is that a sample mean was used to compute the deviations, and using the sample mean rather than the true population mean uses up one degree of freedom. Dividing by $n - 1$ corrects for the fact that a sample tends to underestimate the true variability of the population.

The **sample variance** is

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

For our example:

$$s^2 = \frac{25 + 4 + 1 + 4 + 36}{5 - 1} = \frac{70}{4} = 17.5$$

Step 5: Taking the square root. The variance is in squared units. If your data is measured in dollars, the variance is in “squared dollars,” which is not a meaningful quantity. Taking the square root of the variance returns us to the original units and gives us the **sample standard deviation**.

$$s = \sqrt{s^2} = \sqrt{17.5} \approx 4.18$$

The standard deviation tells you, roughly, how far a typical observation is from the mean. For this dataset, observations are, on average, about 4.18 units away from the mean of 9.

i Note

You might wonder why we bother with variance at all, given that the standard deviation is more interpretable. The reason is that variance has cleaner mathematical properties. When you combine independent random variables, their variances add. Standard deviations do not. This property becomes essential in probability theory and inferential statistics, which is why variance appears in nearly every formula from Chapter 5 onward. Think of variance as the quantity that makes the math work and standard deviation as the quantity that makes the interpretation work.

3.4.4 Population vs. Sample Notation

Throughout this book, we distinguish between a population parameter and a sample statistic. The population mean is denoted μ (the Greek letter mu), and the population standard deviation is denoted σ (sigma). When we calculate these from sample data, we use $\bar{x}$ for the mean and s for the standard deviation.

The population variance formula divides by N (the population size):

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

The sample variance formula divides by $n - 1$:

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

In practice, you will almost always be working with samples, so the $n - 1$ version is the one to use unless you are told explicitly that your data represents the entire population.

3.4.5 Interpreting the Standard Deviation

The standard deviation is a central summary number in statistics, and developing an intuition for it will serve you well. Here are some guidelines for interpretation.

A standard deviation of zero means there is no spread at all. Every observation is the same.

A small standard deviation relative to the mean indicates that observations are tightly clustered around the center. If the mean exam score is 80 and the standard deviation is 3, most students scored close to 80.

A large standard deviation relative to the mean indicates wide dispersion. If the mean exam score is 80 and the standard deviation is 20, students' scores are scattered from very low to very high.

Here is a way to build practical intuition for the standard deviation. Think of it as a “typical distance from the center.” When a weather forecaster says that the average high temperature in July in Phoenix is 106 degrees Fahrenheit with a standard deviation of about 4 degrees, you know that most July days will be within a few degrees of 106. A day of 114 degrees would be two standard deviations above the mean, which is unusual but not unprecedented. A day of 90 degrees would be four standard deviations below the mean, which would be so unusual it would make the news. The standard deviation gives you a ruler for measuring how surprising any particular value is. Once you develop this habit of thinking, “How many standard deviations away is that?”, you will find yourself applying it everywhere, from stock market returns to your morning commute to the number of points your favorite basketball team scores per game.

The coefficient of variation provides a way to compare spread across datasets with different scales. It is the standard deviation divided by the mean, often expressed as a percentage:

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

If one factory produces bolts with a mean length of 10 cm and a standard deviation of 0.5 cm, and another produces beams with a mean length of 500 cm and a standard deviation of 2 cm, which process is more variable? The bolt factory has a CV of 5%, while the beam factory has a CV of 0.4%. Relative to what it is producing, the bolt factory has more variability.

3.5 Percentiles, Quartiles, and the Five-Number Summary

3.5.1 Percentiles

A **percentile** tells you the value below which a given percentage of the data falls. If your score on a standardized test is at the 85th percentile, that means 85% of test-takers scored below you. The percentile is not a percentage score on the test itself. You could score 72% of the questions correctly and still be at the 85th percentile if most people scored below 72%.

More formally, the p th percentile is the value such that $p\%$ of the observations fall at or below it.

Percentiles are used constantly in practical settings. Pediatricians track children's height and weight by percentile. Standardized tests report percentile ranks. Economic researchers use percentiles to describe income distributions with more nuance than a single number allows.

3.5.2 Quartiles

Quartiles divide the data into four equal parts using three specific percentiles.

- Q_1 (first quartile) = 25th percentile. One quarter of the data falls below this value.
- Q_2 (second quartile) = 50th percentile = the median. Half the data falls below this value.
- Q_3 (third quartile) = 75th percentile. Three quarters of the data falls below this value.

To compute quartiles by hand, first sort the data and find the median. Then find the median of the lower half of the data (that gives you Q_1) and the median of the upper half (that gives you Q_3).

Consider the sorted dataset: 12, 15, 18, 22, 25, 28, 30, 33, 40.

There are 9 values. The median (Q_2) is the 5th value, which is 25.

The lower half is 12, 15, 18, 22. The median of this group is $\frac{15+18}{2} = 16.5$, so $Q_1 = 16.5$.

The upper half is 28, 30, 33, 40. The median of this group is $\frac{30+33}{2} = 31.5$, so $Q_3 = 31.5$.

The IQR is $Q_3 - Q_1 = 31.5 - 16.5 = 15$.

Note that different software packages use slightly different algorithms to compute quartiles, so you may occasionally see small differences depending on which tool you use. Even within R, the `quantile()` function offers nine different definitions, and this can cause confusion when students compare their homework answers. The differences are typically small, often just a fraction of a unit, and they vanish entirely for large datasets. The conceptual interpretation is the same regardless of the method, so do not lose sleep over which algorithm your software uses. The important thing is understanding what the quartiles represent, not matching numbers to the last decimal place.

3.5.3 The Five-Number Summary

The **five-number summary** consists of five values that provide a compact description of a dataset's distribution.

1. **Minimum** (the smallest value)

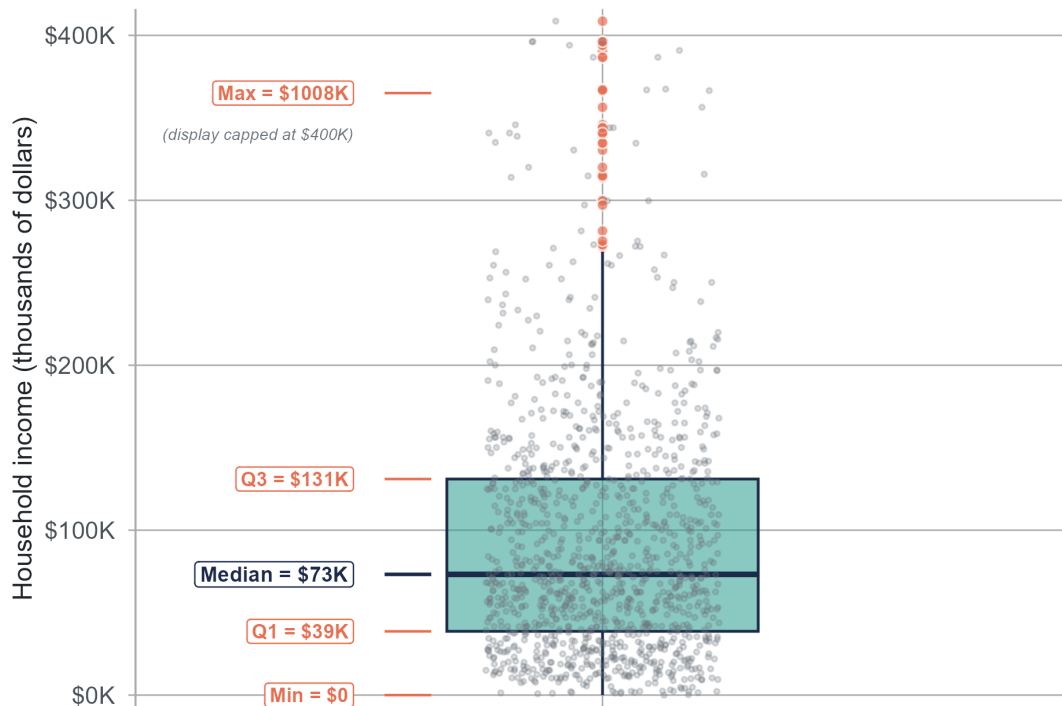
2. Q_1 (the first quartile)
3. **Median** (the second quartile)
4. Q_3 (the third quartile)
5. **Maximum** (the largest value)

For the dataset above, the five-number summary is: 12, 16.5, 25, 31.5, 40.

These five numbers tell you a lot. The minimum and maximum give you the range. Q_1 and Q_3 give you the IQR. The median gives you the center. The spacing between consecutive values tells you about the shape. If the distance from Q_1 to the median is much smaller than the distance from the median to Q_3 , the upper half of the distribution is more spread out than the lower half, suggesting right skewness.

The Five-Number Summary: Household Income

Each annotated value matches what you will compute in Exercise 3



Source: U.S. Census Bureau, 2022 ACS 1-Year PUMS. Coral points are IQR outliers.

Figure 3.2: A box plot of U.S. household income data with all five summary statistics annotated, alongside jittered data points showing the underlying distribution. The box plot compresses the entire distribution into a compact display built from the minimum, Q1, median, Q3, and maximum. Source: U.S. Census Bureau, 2022 ACS 1-Year PUMS.

The five-number summary is the foundation of the **box plot**, a visual display we will explore in Chapter 4. For now, just know that box plots are graphical representations of the five-number summary, with extensions for identifying potential outliers.

3.5.4 Describing the Income Distribution with Percentiles

Returning to U.S. household income, percentiles give us a much richer picture than any single number can. Data from the Economic Policy Institute and the U.S. Census Bureau allow us to see the full distribution.

In 2022, the approximate percentile breakdown for annual household income looked something like this:

Percentile	Approximate Income
10th	\$16,800
25th (Q_1)	\$33,000
50th (Median)	\$74,580
75th (Q_3)	\$133,000
90th	\$213,000
95th	\$295,000
99th	\$600,000+

Source: U.S. Census Bureau, Current Population Survey Annual Social and Economic Supplement, 2022; Economic Policy Institute State of Working America Data Library. These figures represent the national household income distribution and are not drawn from the acs-household-income.csv exercise dataset, which contains approximately 1,200 sampled households.

Look at the spacing. The distance from the 10th percentile to the median is about \$57,780. The distance from the median to the 90th percentile is about \$138,420. The upper half of the income distribution is far more spread out than the lower half. And the gap from the 90th to the 99th percentile, roughly \$387,000, is wider than the entire range from the 10th to the 90th. This is what extreme right skewness looks like in a real dataset. It is also why the gap between mean and median income is so large.

The IQR for household income is approximately $\$133,000 - \$33,000 = \$100,000$. This tells you that the middle 50% of American households earn between roughly \$33,000 and \$133,000 per year. That is itself a wide range, reflecting substantial economic diversity even within the middle half of the distribution.

3.6 Outliers

3.6.1 What Is an Outlier?

An **outlier** is an observation that is unusually far from the rest of the data. There is no universal agreement on exactly how far “unusually far” means, but there are standard methods for detecting them.

The word “outlier” sounds technical, but the concept is intuitive. If you are looking at the salaries of employees in a small company and most earn between \$40,000 and \$90,000, a salary of \$2,500,000 would be an outlier. It does not fit the general pattern.

Outliers can arise for several reasons, and understanding why an outlier exists is more important than deciding what to do with it mechanically.

Data entry errors. Someone typed 5000 instead of 50. This is a mistake, and fixing or removing it is straightforward and appropriate.

Measurement errors. A scale was miscalibrated. A sensor malfunctioned. Again, these errors should be corrected if possible.

Extreme but legitimate observations. Some values are far from the rest and belong there. Jeff Bezos’s income is an outlier in any sample of American incomes, but it is not an error. It is a real value that reflects something meaningful about the economy.

Different populations. An outlier may indicate that the observation comes from a different group than the rest of the data. If you are measuring the weights of house cats and one “cat” weighs 400 pounds, you may have accidentally included a lion.

The history of science is full of cases where outliers turned out to be the most interesting part of the data. In the 1990s, astronomers studying distant supernovae noticed that some of their brightness measurements did not fit the expected pattern. Rather than discarding these outliers, they investigated further and discovered that the expansion of the universe was accelerating, a finding that upended cosmology and earned the 2011 Nobel Prize in Physics. On the other hand, history is full of cases where outliers were simply measurement errors. In September 2011, physicists at the OPERA experiment reported that neutrinos traveling from CERN in Switzerland to a detector in Gran Sasso, Italy appeared to arrive about 60 nanoseconds faster than light, a result that would have overturned a century of physics. Rather than immediately announcing a discovery, the team spent months checking their equipment. In February 2012, they identified two sources of systematic error: a loose fiber optic cable connector and an oscillator running at the wrong frequency.

Both issues had introduced a consistent bias into the timing measurements. The anomalous results were not a breakthrough. They were a hardware problem masquerading as a signal. The lesson the OPERA team drew, and that the scientific community drew from the episode, is exactly the right one: when a measurement is far from what theory and prior evidence predict, the first hypothesis is usually instrument or process error, not a new law of nature.

The lesson is that outliers demand attention, not automatic deletion. They are signals that something unusual is happening, and the analyst's job is to figure out whether that something is a discovery, an error, or simply a reflection of natural variation in a messy world.

3.6.2 Detecting Outliers with the IQR Method

The most common method for identifying potential outliers uses the IQR. It defines boundaries called **fences** beyond which observations are flagged as potential outliers.

$$\text{Lower fence} = Q_1 - 1.5 \times \text{IQR}$$

$$\text{Upper fence} = Q_3 + 1.5 \times \text{IQR}$$

Any observation below the lower fence or above the upper fence is considered a potential outlier.

Let us work through an example. Suppose you have the following dataset of monthly electric bills (in dollars) for 15 households:

45, 52, 58, 62, 65, 68, 70, 72, 75, 80, 85, 92, 98, 110, 310

First, find the quartiles. There are 15 observations.

The median is the 8th value: 72.

The lower half is 45, 52, 58, 62, 65, 68, 70. The median of this group (the 4th value) is 62. So $Q_1 = 62$.

The upper half is 75, 80, 85, 92, 98, 110, 310. The median of this group is 92. So $Q_3 = 92$.

$$\text{IQR} = 92 - 62 = 30$$

$$\text{Lower fence: } 62 - 1.5(30) = 62 - 45 = 17$$

$$\text{Upper fence: } 92 + 1.5(30) = 92 + 45 = 137$$

No values fall below 17, so there are no low outliers. The value 310 exceeds 137, so it is flagged as a potential outlier.

The word “potential” matters. The IQR method identifies observations that are unusually far from the bulk of the data. It does not tell you whether those observations are errors, interesting phenomena, or something else entirely. That determination requires judgment and context.

3.6.3 The Ethics of Outlier Removal

Deciding what to do with outliers is one of the most consequential choices a data analyst makes, and one of the most ethically fraught.

Ethics Moment

Removing outliers changes your results. Sometimes dramatically. An analyst studying CEO compensation who removes the top 1% of earners will report a very different picture than one who keeps them in. A pharmaceutical researcher who drops patients with extreme side effects will present a rosier picture of drug safety than one who includes them. A school district that excludes the lowest-performing students from its average test scores will look more successful than it actually is.

None of these removals are inherently wrong. Outlier removal can be justified when values are clearly erroneous or when they come from a different population than the one you intend to study. But removal must be transparent. If you remove outliers, you should report that you did so, explain why, and show what the results look like both with and without the removed observations.

The danger is not in removing outliers. The danger is in removing them selectively, keeping the ones that support your conclusion and discarding the ones that do not, without telling anyone. This practice, whether intentional or unconscious, is a form of data manipulation. It can make a failing drug look effective, a bad policy look successful, or an unequal economy look fair. The rule is simple: if you would be embarrassed to explain your outlier removal decision to a skeptical colleague, you probably should not be making it.

AI Reality Check

When asked to “clean” a dataset, AI tools often flag and remove outliers mechanically, applying rules like “drop anything beyond 1.5 times the IQR” without asking whether those values are errors or legitimate observations. A large language model asked to prepare income data for analysis might automatically remove a household earning \$310,000 because

it exceeds the upper fence, destroying the very signal you are trying to study. The tool treats outlier detection as a cleaning step rather than an investigation step.

AI tools also tend to default to the mean when asked for “the average,” even when the data is heavily skewed and the median would be far more appropriate. Ask an AI to summarize a dataset of home prices in a neighborhood with one mansion, and it will likely report a mean that no typical home comes close to. It will do this confidently, without flagging that the distribution is skewed or that the median tells a more honest story.

The chapter’s lesson applies directly here: outliers demand investigation, not automatic deletion, and the choice between mean and median requires understanding the shape of your data. These are judgment calls that no algorithm should make for you.

3.6.4 How Outliers Affect Summary Statistics

The practical impact of outliers on summary statistics is worth seeing with concrete numbers.

Consider the following dataset of 10 home prices (in thousands of dollars) in a neighborhood:

185, 192, 205, 210, 218, 225, 230, 240, 255, 1850

Without the outlier (1850), the mean is about \$217,800 and the median is \$218,000. With the outlier included, the mean jumps to about \$381,000 while the median changes only to \$221,500. The standard deviation goes from about \$22,500 to about \$516,600.

A real estate agent who reports the mean home price as \$381,000 is technically correct but seriously misleading. Nine of the ten homes sold for under \$260,000. The median tells the true story of what a typical buyer should expect to pay.

This example illustrates why **resistant statistics** (summary statistics that are not strongly affected by outliers or extreme values) are often preferable when outliers are present or suspected. The median and IQR are the two most common resistant statistics. They are not immune to outliers, but they are far less affected by them than the mean and standard deviation.

3.7 Putting It All Together

Let us apply everything from this chapter to a single dataset to see how these measures work in concert.

The Economic Policy Institute publishes data on average hourly wages by income percentile in the United States. Suppose we have the following simplified dataset of annual incomes (in thousands of dollars) for 20 workers sampled from a mid-sized American city:

22, 28, 31, 34, 36, 38, 40, 42, 44, 47, 50, 53, 56, 60, 65, 72, 85, 105, 140, 310

In the exercises that follow, you will apply these same methods to a much larger dataset of 1,200 households drawn from the 2022 American Community Survey. The concepts are identical; only the scale changes.

Mean:

$$\bar{x} = \frac{22 + 28 + 31 + \cdots + 140 + 310}{20} = \frac{1,358}{20} = 67.9$$

The mean income in this sample is \$67,900.

Median:

With 20 values, the median is the average of the 10th and 11th values. The sorted data is already in order, so the median is $\frac{47+50}{2} = 48.5$, or \$48,500.

The mean is about \$19,400 higher than the median. This tells us the data is right-skewed, pulled upward by the high earners.

Quartiles and IQR:

Q_1 is the median of the lower 10 values (22, 28, 31, 34, 36, 38, 40, 42, 44, 47). With 10 values, $Q_1 = \frac{36+38}{2} = 37$, or \$37,000.

Q_3 is the median of the upper 10 values (50, 53, 56, 60, 65, 72, 85, 105, 140, 310). $Q_3 = \frac{65+72}{2} = 68.5$, or \$68,500.

$IQR = 68.5 - 37 = 31.5$, or \$31,500.

Five-Number Summary:

Minimum = 22, $Q_1 = 37$, Median = 48.5, $Q_3 = 68.5$, Maximum = 310.

(All in thousands of dollars.)

Outlier Detection:

Lower fence: $37 - 1.5(31.5) = 37 - 47.25 = -10.25$

Upper fence: $68.5 + 1.5(31.5) = 68.5 + 47.25 = 115.75$

No values fall below the lower fence. Two values, 140 and 310, exceed the upper fence of 115.75 and are flagged as potential outliers.

Standard Deviation:

Computing this by hand for 20 values is tedious, but working through the formula:

$$s = \sqrt{\frac{1}{19} \sum_{i=1}^{20} (x_i - 67.9)^2} \approx 56.9$$

The standard deviation is about \$56,900. This is large relative to the mean, confirming that the data has substantial spread. The coefficient of variation is $\frac{56.9}{67.9} \times 100\% \approx 83.8\%$, which is high, indicating substantial variability in incomes.

What the numbers tell us together. The mean (\$67,900) substantially exceeds the median (\$48,500), confirming right skewness. The IQR (\$31,500) is much smaller than the range (\$288,000), indicating that the range is heavily influenced by extreme values. The two flagged outliers at \$140,000 and \$310,000 are driving much of the discrepancy between the mean and the median. If a reporter wanted to describe the typical income in this sample, the median would be the better choice. If a researcher wanted to describe the total economic output of the group, the mean (or the sum) would be more relevant.

This is the key lesson. No single summary statistic is always right or always wrong. The right statistic depends on the question you are asking.

It is worth noting that this kind of analysis is far from an academic exercise. Journalists, policy analysts, and researchers make these choices every day. When the Bureau of Labor Statistics releases its monthly jobs report, the numbers that make headlines are themselves summary statistics chosen from a much larger set of available data. The unemployment rate is a kind of proportion. The number of jobs added is a count. The average hourly wage is a mean. Each tells a different story about the labor market, and each can be used selectively to support a particular narrative. A reader who understands the material in this chapter is better equipped to ask the right follow-up questions. When someone tells you “the average is X,” you should wonder about the median. When someone tells you a group is performing well “on average,” you should ask about the spread. These are not pedantic questions. They are the difference between understanding the data and being misled by it.

3.8 A Note on Income Data

The income data used throughout this chapter comes primarily from two sources. The **U.S. Census Bureau’s Current Population Survey** and **American Community Survey** are among the primary official sources for household income data in the United States. These surveys sample tens of thousands of households annually and produce the official income statistics cited by government agencies, researchers, and the media.

The **Economic Policy Institute (EPI)** publishes detailed analyses of income and wage data, often broken down by percentile, race, gender, education level, and other categories. Their State of Working America Data Library is freely available online and provides the kind of granular data that reveals patterns invisible in national averages.

The dataset used in this chapter’s exercises (`acs-household-income.csv`) draws directly from the 2022 ACS 1-Year Public Use Microdata Sample. It contains approximately 1,200 households sampled to represent variation across all 50 states, five regions, and five education levels. The data acquisition and sampling script (`ch03-data-acquisition.R`) is available on the companion website. See Appendix B for a full description of the dataset’s variables, sampling strategy, and education level classifications.

When you encounter income data in the news or in research, two questions worth asking are, “Is this mean or median?” and “What population does this represent?” National averages can mask enormous regional and demographic variation. The median household income in San Francisco is very different from the median household income in rural Mississippi, and neither is well-represented by the national figure.

3.9 Choosing the Right Summary Statistics

The following table provides a quick reference for when to use each measure.

Situation	Measure of Center	Measure of Spread
Symmetric data, no outliers	Mean	Standard deviation
Skewed data or outliers present	Median	IQR
Categorical data	Mode	Not applicable (use frequency tables)

Situation	Measure of Center	Measure of Spread
Comparing variability across different scales	Mean with CV	Coefficient of variation
Quick overview of distribution	Median	Five-number summary

When in doubt, report both the mean and the median. If they are close, the distribution is roughly symmetric and either works well. If they differ substantially, the discrepancy itself is informative, and you should investigate why.

3.10 Looking Ahead

Chapter 3 has given you a numerical vocabulary for describing data: where the center is, how spread out the values are, and how to identify observations that fall unusually far from the rest. Chapter 4 takes these same ideas and turns them into pictures. The five-number summary you computed here is the direct foundation of the box plot. The distinction between symmetric and skewed distributions, which we described in words here, becomes immediately visible in a histogram or density plot. The outlier-detection logic from this chapter reappears in visualizations that mark flagged observations as individual points. Numerical summaries and visual displays are not alternatives. They answer different questions about the same distribution, and the most complete analysis uses both.

3.11 Key Terms

- **Mean:** The arithmetic average of a dataset. Calculated by summing all values and dividing by the number of observations. Notation: $\bar{x}$ for a sample mean, μ for a population mean.
- **Median:** The middle value of a sorted dataset. Half the observations fall below it, half above.
- **Mode:** The most frequently occurring value in a dataset. The only measure of center applicable to categorical data.
- **Range:** The difference between the maximum and minimum values in a dataset.
- **Interquartile Range (IQR):** The difference between the third quartile and the first quartile ($Q_3 - Q_1$). Measures the spread of the middle 50% of the data. Resistant to outliers.

- **Variance:** The average of the squared deviations from the mean. Notation: s^2 for a sample, σ^2 for a population. Measured in squared units.
- **Standard Deviation:** The square root of the variance. Measures how far, on average, observations are from the mean. Notation: s for a sample, σ for a population. Measured in the same units as the data.
- **Coefficient of Variation (CV):** The standard deviation divided by the mean, expressed as a percentage. Useful for comparing variability across datasets with different units or scales.
- **Percentile:** The value below which a given percentage of observations fall. The 85th percentile is the value below which 85% of observations fall.
- **Quartiles:** The 25th (Q_1), 50th (Q_2 , also the median), and 75th (Q_3) percentiles. They divide the sorted data into four equal parts.
- **Five-Number Summary:** A set of five descriptive statistics consisting of the minimum, Q_1 , median, Q_3 , and maximum. Provides a compact overview of the distribution.
- **Outlier:** An observation that is unusually far from the rest of the data. Often identified using the IQR method (values beyond $Q_1 - 1.5 \times \text{IQR}$ or $Q_3 + 1.5 \times \text{IQR}$).
- **Fences:** The boundaries used in the IQR method for identifying potential outliers. The lower fence is $Q_1 - 1.5 \times \text{IQR}$ and the upper fence is $Q_3 + 1.5 \times \text{IQR}$. Observations beyond the fences are flagged as potential outliers.
- **Resistant Statistic:** A summary statistic that is not strongly affected by outliers or extreme values. The median and IQR are resistant. The mean and standard deviation are not.
- **Skewness:** A measure of the asymmetry of a distribution. Right-skewed distributions have a long tail to the right (mean > median). Left-skewed distributions have a long tail to the left (mean < median).
- **Bimodal:** A distribution with two distinct peaks, suggesting the data may contain two subgroups.
- **Degrees of Freedom:** The number of independent pieces of information in the data. In the context of sample variance, this is $n - 1$, reflecting the constraint that the deviations from the sample mean must sum to zero. The concept reappears in later chapters with the t, chi-square, and F distributions.

3.12 Further Reading and References

The following works are cited in this chapter or provide valuable additional context.

On U.S. household income data: U.S. Census Bureau. (2023). *Income in the United States: 2022* (Current Population Reports, P60-279). U.S. Government Publishing Office. The primary source for official household income statistics, including the \$74,580 median figure cited throughout this chapter.

Economic Policy Institute. *State of Working America Data Library*. Available at epi.org/data. Provides detailed percentile-level income breakdowns used in the chapter's income distribution table.

On the chapter dataset: The income distribution discussed in this chapter is drawn directly from the 2022 ACS 1-Year Public Use Microdata Sample. The data acquisition script (`ch03-data-acquisition.R`) is available on the companion website.

On the OPERA neutrino episode: The OPERA Collaboration. (2012). Measurement of the neutrino velocity with the OPERA detector in the CNGS beam. *Journal of High Energy Physics*, 2012(10), 93. The published version contains the corrected result, consistent with the speed of light. The original anomalous claim appeared in a September 2011 preprint (arXiv:1109.4897v1). The error, traced to a loose fiber optic cable, was identified and announced by the collaboration in February 2012.

On outlier judgment in scientific practice: Baggerly, K.A., & Coombes, K.R. (2009). Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *Annals of Applied Statistics*, 3(4), 1309–1334. A forensic investigation into published cancer genomics results that shows what happens when outliers and data errors go uninvestigated.

For further reading on summary statistics and economic inequality: Piketty, T. (2014). *Capital in the twenty-first century*. Harvard University Press. (Uses income and wealth percentile data extensively; the chapter's discussion of the gap between mean and median income connects directly to its empirical framework.)

3.13 Exercises

3.13.1 Check Your Understanding

1. A news headline reads, “Average American household earns over \$105,000 per year.” Another headline reads, “Typical American household earns about \$75,000 per year.” Both are based

- on the same government data. Explain how both can be correct. Which number is the mean and which is the median? Which better represents the experience of a typical household, and why?
- For the dataset 3, 5, 7, 7, 8, 10, 12, calculate the mean, median, and mode. Are any of these values the same? What does this suggest about the shape of the distribution?
 - Explain in your own words why the sum of the deviations from the mean always equals zero. Why does this mathematical property create a problem if we want to measure how spread out the data is?
 - A dataset has a mean of 50 and a standard deviation of 0. What can you conclude about the values in this dataset? Explain.
 - Two datasets each have a mean of 100. Dataset A has a standard deviation of 5 and Dataset B has a standard deviation of 25. Describe in words what the distributions of these two datasets would look like compared to each other.
 - What is the difference between the range and the interquartile range? Under what circumstances would the range be a misleading measure of spread? Give a specific example.
 - Explain why the median is described as “resistant” to outliers while the mean is not. Illustrate your answer with a simple numerical example.
 - A dataset has $Q_1 = 20$, median = 35, and $Q_3 = 42$. Without seeing the full dataset, what can you say about the shape of the distribution? Is it symmetric, left-skewed, or right-skewed? Explain your reasoning.
 - The formula for sample variance divides by $n - 1$ rather than n . What is this adjustment called, and why is it used? You do not need to prove it mathematically, just explain the intuition.
 - A student calculates the mean of a dataset and gets 60. They then calculate the median and get 82. What does this tell them about the shape of the distribution? Would they expect the data to be right-skewed or left-skewed? Explain.

3.13.2 Apply It

(See Appendix B for complete variable descriptions for all datasets used in these exercises.)

For the following problems, use the **acs-household-income.csv** dataset available on the companion website. This dataset is drawn from the 2022 American Community Survey 1-Year Public Use Microdata Sample, produced by the U.S. Census Bureau. It contains approximately 1,200 household records representing all 50 states, sampled to reflect the national distribution of income and educational attainment. The variables are `household_id` (a unique identifier), `state` (the household’s state of residence), `region` (geographic region: Northeast, Southeast, Midwest, Southwest,

or West), `household_income` (annual household income in 2022 dollars), `education_level` (the highest education level attained by the head of household: High School or Less, Some College or Associate, Bachelor, Master or Professional, or Doctorate), and `household_size` (number of people in the household).

1. Load the dataset and examine its structure. How many observations are there? How many variables? Identify each variable's type (categorical or numerical). For the numerical variables, briefly describe what each one measures.
2. Compute the mean and median of `household_income` across all households. Are these two values close to each other or far apart? What does the relationship between them tell you about the shape of the income distribution in this sample? How do your sample estimates compare to the 2022 national figures cited in this chapter (\$74,580 median, \$105,000 mean)?
3. Compute the five-number summary for `household_income`. Report the minimum, Q_1 , median, Q_3 , and maximum. Calculate the IQR. Interpret these values in the context of household income inequality.
4. Group the data by `state`. For each state, compute the mean and median `household_income`. Identify the five states with the highest mean household income. For each of these states, compute the difference between the mean and the median. Which state has the largest gap? What does a larger gap suggest about the income distribution within that state?
5. Using the IQR method on the full `household_income` variable, check whether any individual households are outliers. Show your fence calculations. How many households are flagged as outliers? Are the outliers concentrated in any particular states or regions?
6. Group the data by `education_level`. For each group, compute the mean, standard deviation, and coefficient of variation of `household_income`. Which education level has the most variability in absolute terms (largest SD)? Which has the most variability in relative terms (largest CV)? Does the Some College or Associate group behave as you would expect given its position between High School or Less and Bachelor? What does this tell you about income variation across education levels?
7. Choose any single state that has at least 10 households in the dataset. For that state, compute the 10th, 25th, 50th, 75th, and 90th percentiles of `household_income`. Compute the approximate range ($P_{90} - P_{10}$) and the IQR ($P_{75} - P_{25}$). What fraction of the approximate range is captured by the IQR? How does this compare to what you would expect for a symmetric distribution?
8. Group the data by `region`. For each region, compute the mean, median, and standard deviation of `household_income`. Which region has the highest mean? Which has the largest standard deviation? Do the mean and median differ substantially within any region, suggesting skewness?

9. Create a new variable called the “income gap ratio” for each state, defined as the ratio of the 90th percentile of `household_income` to the 10th percentile of `household_income` within that state (use only states with at least 10 observations). Compute this ratio for all qualifying states. What is the mean and median of this ratio across states? Which state has the highest ratio, and which has the lowest? What does this ratio capture about within-state inequality?
10. Pick three states: one with a high median income, one near the overall sample median, and one with a low median income. For each, compute the five-number summary of `household_income`. Compare the three distributions. Where do you see the largest differences, at the bottom, the middle, or the top of the income distribution?
11. Pick any 10 households from the dataset. Compute their sample variance and sample standard deviation by hand using the formulas $s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$ and $s = \sqrt{s^2}$. Show each step: the mean, each $(x_i - \bar{x})$, each squared deviation, the sum, the division by $n - 1$, and the square root. Then verify your answer with R’s `var()` and `sd()`. Why does the formula divide by $n - 1$ rather than n ? Tie your answer back to the conceptual question in *Check Your Understanding* problem 9.

3.13.3 Think Deeper

1. Politicians sometimes argue about whether the economy is “working” by citing different summary statistics. One camp cites GDP growth and mean incomes, pointing to expansion. The other cites median wages and cost-of-living data, pointing to stagnation for ordinary families. Is one side “right” and the other “wrong”? Or are they answering different questions? What set of summary statistics would you recommend for an honest, comprehensive assessment of economic well-being? Justify your choices.
2. A hospital administrator reports that the average emergency room wait time is 45 minutes. Patient advocacy groups counter that most patients wait over an hour. Explain how both claims could be true simultaneously. What additional information would you want before deciding which statistic better represents the patient experience? Discuss the ethical responsibility of the hospital when choosing which number to report publicly.
3. An analyst studying housing affordability removes all homes priced above \$2 million from the dataset before computing average home prices, arguing that luxury homes are “not relevant” to the typical buyer. Is this a defensible decision? Under what circumstances would it be appropriate, and under what circumstances would it be misleading? How should the analyst communicate this decision in a published report?
4. Consider two school districts. District A has a mean math score of 78 with a standard deviation of 5. District B has a mean math score of 78 with a standard deviation of 18. The districts serve

similar demographics. Which district is performing better, and why? A parent asks, “Is my child more likely to get a good education in District A or District B?” How would you answer this question using the concepts from this chapter? What additional information would you want?

5. The Gini coefficient condenses an entire income distribution into a single number between 0 and 1. What information is lost in this compression? Two countries could have the same Gini coefficient but very different income distributions. Construct a hypothetical example of two such countries and explain how summarizing complex data with a single number can obscure important differences. When, if ever, is a single summary statistic sufficient for making a policy decision?

Chapter 4

Summarizing Data with Pictures

4.1 The Maps That Showed What Everyone Already Knew

In the 1930s, the federal government's Home Owners' Loan Corporation sent agents into 239 American cities to evaluate neighborhoods for mortgage lending risk. The agents drew maps. They shaded neighborhoods in four colors. Green meant "best," areas deemed safe for investment. Blue meant "still desirable." Yellow was "definitely declining." Red was "hazardous."

The red zones, overwhelmingly, were where Black Americans and other minority populations lived.

These maps became the architecture of a system known as redlining. Banks used them, or maps like them, to deny mortgages to residents of red-shaded areas. The Federal Housing Administration used similar logic to guarantee loans in white suburbs while refusing them in integrated or minority neighborhoods. Over decades, this system funneled wealth into some communities and starved others. The effects compounded. Neighborhoods denied investment deteriorated. Deteriorating neighborhoods were cited as evidence that the original ratings were correct. The cycle fed itself for generations.

For years, researchers documented redlining's legacy using tables of numbers. They calculated disparities in homeownership rates, differences in median household wealth, gaps in school funding. The numbers were damning. They were also, for many audiences, abstract. A table showing that the median Black household holds one-tenth the wealth of the median white household is a fact. It is an important fact. But it can sit on a page without making the pattern immediately legible or hard to dismiss.

Then came the maps.

In 2016, a collaborative team of researchers at the University of Richmond’s Digital Scholarship Lab, Virginia Tech, and the University of Maryland launched *Mapping Inequality*, a project that digitized and overlaid those original 1930s HOLC maps onto modern cities. Suddenly, anyone with a web browser could see the red-shaded zones from the 1930s and compare them to today’s patterns of poverty, health outcomes, environmental hazards, and racial demographics. The overlap was striking. Not surprising to anyone who had lived through that history, but striking to see in color on a screen.

Researchers at the National Community Reinvestment Coalition used these visualizations alongside quantitative analysis to demonstrate that neighborhoods redlined decades ago remain disproportionately low-income and minority today. A 2018 study by the NCRC found that 74% of neighborhoods graded “hazardous” eight decades ago are still low-to-moderate income, and 64% are still majority-minority communities. The numbers told the story. The maps made you *see* it.

The figure below makes that 64% claim visible in data drawn directly from the Mapping Inequality project. It is a type of chart called a **box plot**, which we define formally in Section 4.4. For now, read each box as a compact summary of the distribution of minority population shares within neighborhoods of that historical HOLC grade: the line inside each box is the median, the box spans the middle half of the data, and the points are neighborhoods far from the typical pattern. The gradient from Grade A to Grade D is unmistakable.

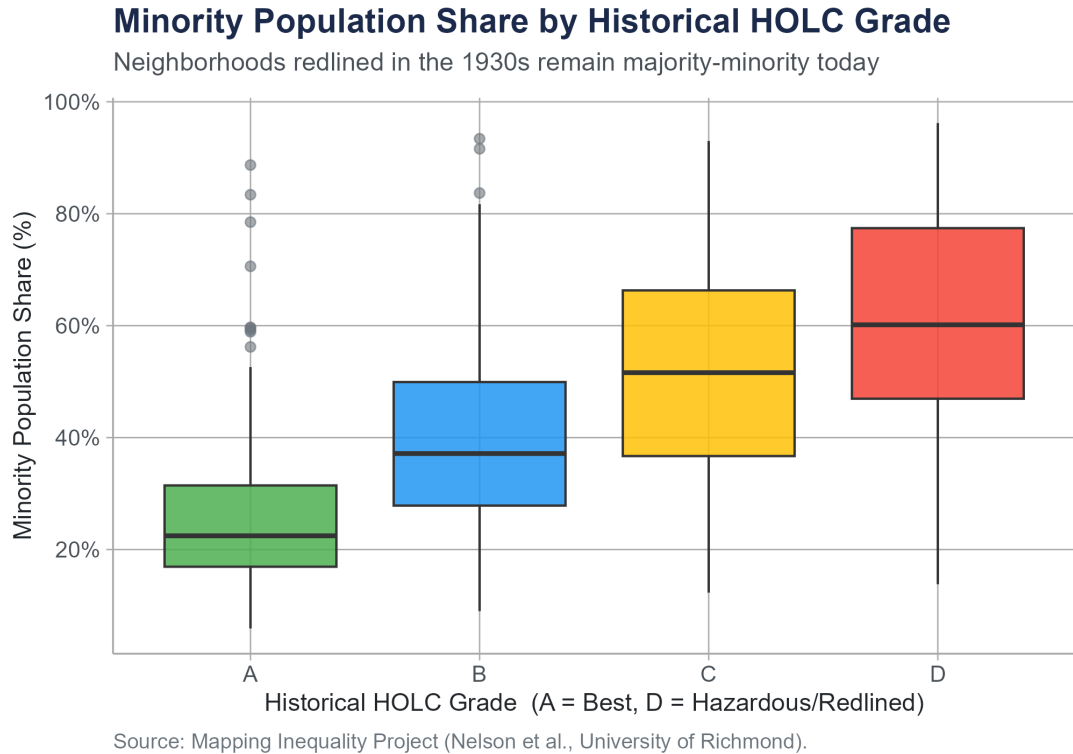


Figure 4.1: Minority population share (%) by historical HOLC grade, from A (best) to D (hazardous/redlined). Neighborhoods originally graded “D” have substantially higher minority population shares today, showing the persistent demographic legacy of 1930s redlining. Source: FiveThirtyEight analysis of the Mapping Inequality Project (Nelson et al., 2016, University of Richmond); 2020 U.S. Census.

i Try It Online

For an interactive StoryMap exploring the geographic footprint of redlining and its present-day consequences, visit: storymaps.arcgis.com/stories/e32a080e412446ddb073f1cf8d4e3fa9

The original *Mapping Inequality* interactive project, which lets you browse digitized HOLC maps for any major American city, is at: dsl.richmond.edu/panorama/redlining

This is what data visualization does at its best. It takes patterns that exist in tables and formulas and makes them visible, immediate, and hard to ignore. A well-made chart does more than present

information. It creates understanding. And sometimes, as in the case of redlining, it creates the kind of understanding that leads to action.

This chapter is about learning to create that kind of understanding. We will cover the major types of statistical visualizations, when to use each one, how to read them, and how to recognize the many ways they can deceive. Along the way, we will encounter a concept that ties visualization to analysis in a fundamental way, the idea of correlation, and we will spend serious time on why correlation is one of the most abused concepts in public discourse.

4.2 Principles of Effective Visualization

Before we get into specific chart types, let us establish some ground rules. Not every picture is worth a thousand words. Some are worth about twelve, and several of those words are unprintable.

Edward Tufte, a statistician and artist whose books on data visualization have become foundational texts in the field, proposed a concept he called the **data-ink ratio**. The idea is simple. In any visualization, some of the ink on the page (or pixels on the screen) represents actual data. The rest is decoration: gridlines, background shading, 3D effects, clip-art icons, that sort of thing. Tufte argued that we should maximize the proportion of ink devoted to data and minimize everything else. Every drop of ink that does not convey information is visual noise that makes the data harder to read.

This does not mean charts should be ugly or spartan. It means every element should earn its place. A gridline that helps the reader estimate a value is useful. A gradient background that exists because it looks fancy is not. A legend that identifies groups is necessary. A decorative border is clutter.

Here are some practical principles that will serve you well throughout this course and in any work involving data.

Show the data. The purpose of a visualization is to reveal patterns in data. If the chart obscures the data, something has gone wrong. Three-dimensional bar charts look impressive and make it nearly impossible to accurately compare bar heights because of the perspective distortion. A flat bar chart is always better.

Choose the right chart for your question. Different chart types are designed for different purposes. A histogram shows the distribution of a single numerical variable. A bar chart compares counts or proportions across categories. A scatter plot shows the relationship between two numerical variables. Using the wrong chart type is like using a hammer to drive screws.

Label clearly. Every axis needs a label with units. Every chart needs a title or caption that tells the reader what they are looking at. If colors represent groups, include a legend. If the data comes from a particular source, say so. A chart without labels is a puzzle, and your audience did not sign up for puzzle hour.

Do not distort the data. This is both a design principle and an ethical one. Truncating axes, manipulating aspect ratios, using misleading scales, or cherry-picking time ranges can make data appear to say something it does not. The section on data distortion later in this chapter treats these violations as the ethical failures they are, not merely as technical mistakes.

Respect your audience. A chart intended for an executive summary needs to be simpler than one in a technical report. A chart in a newspaper needs to be interpretable by a general audience. Know who you are communicating with and design accordingly.

i Note

Tufte's *The Visual Display of Quantitative Information* (1983) remains one of the most influential books on data visualization ever written. If you find yourself drawn to this topic, it is worth reading in full. His other books, *Envisioning Information*, *Visual Explanations*, and *Beautiful Evidence*, are also excellent. Fair warning: after reading Tufte, you will never look at a PowerPoint chart the same way again. This may or may not improve your life.

4.3 Histograms and Density Plots

Let us start with the most fundamental question you can ask about a numerical variable: how are its values distributed?

Suppose you have salary data for 500 employees at a company. You could report the mean and the standard deviation, and those numbers would tell you something. But they would leave a lot out. Are salaries clustered tightly around the average, or are they spread out? Is there one peak or two? Are there extreme values pulling the average in one direction? A single number cannot answer these questions. A picture can.

4.3.1 Histograms

A **histogram** divides the range of a numerical variable into equal-width intervals called **bins** and counts how many observations fall into each bin. The result is a bar chart where the horizontal axis

represents the variable's values, the vertical axis represents frequency (count) or relative frequency (proportion), and the height of each bar shows how many observations land in that range.

Reading a histogram is straightforward. Where are the bars tallest? That is where the data is concentrated. How wide is the spread of bars? That tells you about variability. Is the distribution symmetric, or does it trail off more to one side?

Several features of a distribution become visible in a histogram.

Shape. Is the distribution roughly symmetric (a mirror image around the center), skewed right (a long tail extending toward higher values), or skewed left (a long tail toward lower values)? Income distributions are almost always skewed right. Most people earn modest amounts, and a smaller number earn very large amounts, pulling the tail rightward.

Center. Where does the distribution seem to be centered? In a symmetric distribution, this is near the peak. In a skewed distribution, the peak and the mean may be in noticeably different places.

Spread. How much do the values vary? A histogram where all the bars are packed tightly together indicates low variability. One where bars stretch across a wide range indicates high variability.

Outliers. Are there isolated bars far from the main body of the distribution? These represent unusual values that may deserve investigation.

Modality. Does the distribution have one peak (unimodal), two peaks (bimodal), or more? A bimodal distribution often suggests that two different groups have been mixed together. If you plotted the heights of all adults in a room without separating by sex, you might see two peaks: one near the average height for women and one near the average height for men. In general, whenever you see an unexpected second peak, ask whether your data might contain two distinct subgroups.

One important choice when creating a histogram is the number of bins. Too few bins and the histogram is too coarse to reveal the shape of the distribution. Too many bins and the histogram becomes noisy, with every small random fluctuation creating its own bar. There is no single correct number of bins. Most software chooses a reasonable default, but you should always try a few different options and see which tells the clearest story.

4.3.2 Density Plots

A **density plot** (also called a kernel density estimate) is a smoothed version of a histogram. Instead of counting observations in bins, it estimates a smooth curve that represents the distribution. You can think of it as a histogram with infinitely narrow bins that have been smoothed out.

The vertical axis of a density plot represents **probability density**, not frequency. The total area under the curve equals 1. This means the height of the curve at any point does not tell you the exact number or proportion of observations at that value. Instead, the area under the curve between any two values tells you the proportion of observations in that range.

Density plots have some advantages over histograms. They are not sensitive to bin width choices. They make it easy to overlay multiple distributions on the same plot for comparison. And they give a cleaner, more immediately readable sense of the distribution's shape.

The trade-off is that density plots can smooth away features that a histogram would reveal. A sharp cutoff in the data (say, no one in your dataset earns below minimum wage) might appear as a gentle taper in a density plot. And because the curve extends beyond the actual data range, a density plot might suggest values that are impossible (negative incomes, for instance). These are cosmetic issues, not serious analytical errors, but they are worth knowing about.

In practice, histograms and density plots are complementary. Many analysts create both, or overlay a density curve on top of a histogram to get the best of both representations.

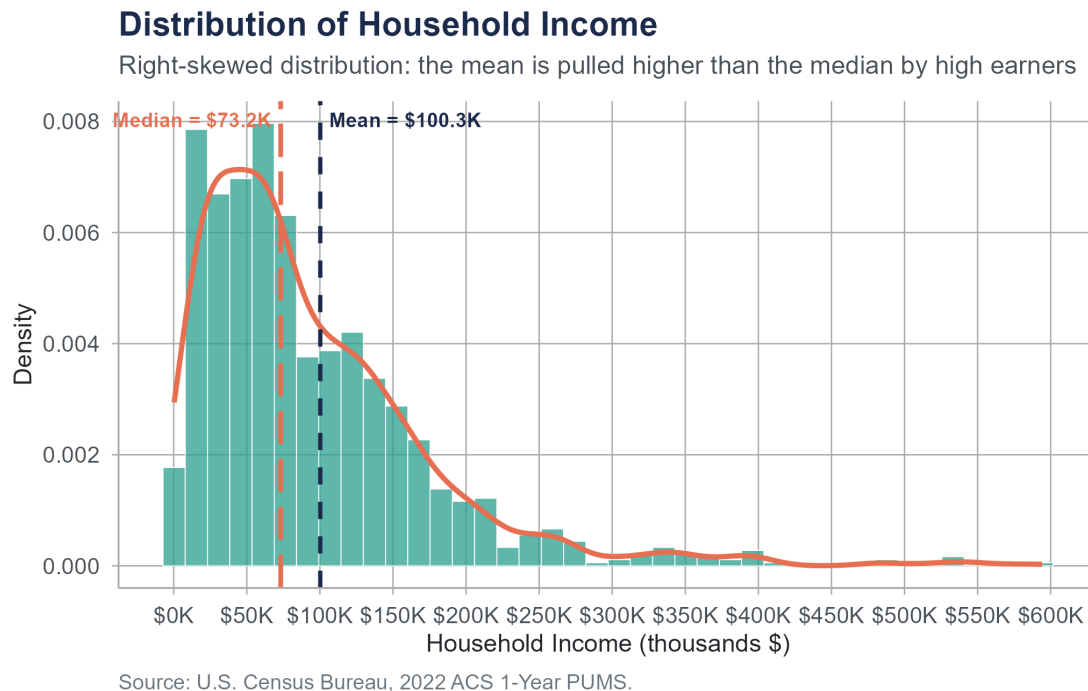


Figure 4.2: A histogram with density overlay showing the right-skewed distribution of U.S. household income. The mean is pulled higher than the median by high earners, a pattern typical of income data. Source: U.S. Census Bureau, 2022 ACS 1-Year PUMS.

4.4 Box Plots

Histograms and density plots are excellent for exploring a single distribution, but what if you want to compare distributions across groups? If you have salary data for five departments, overlaying five histograms on one chart becomes cluttered fast. This is where box plots come in.

In Chapter 3, we introduced the **five-number summary**: the minimum, first quartile (Q_1), median, third quartile (Q_3), and maximum of a dataset. A **box plot** (sometimes called a box-and-whisker plot) is a visual representation of exactly those five numbers, extended to flag potential outliers. Reading a box plot fluently requires knowing its parts.

The **box** spans from Q_1 (the 25th percentile) to Q_3 (the 75th percentile). This box contains the middle 50% of the data, a range called the **interquartile range** (IQR). A line inside the box marks the **median** (the 50th percentile).

The **whiskers** extend from the box to the most extreme data points that fall within 1.5 times the IQR from the edges of the box. Specifically, the lower whisker extends to the smallest value at or above $Q_1 - 1.5 \times \text{IQR}$, and the upper whisker extends to the largest value at or below $Q_3 + 1.5 \times \text{IQR}$.

Any observations beyond the whiskers are plotted individually as points and are considered **potential outliers**. These are values that are unusually far from the center of the distribution. As emphasized in Chapter 3, potential outliers are not automatically errors. They deserve investigation, not automatic deletion.

Box plots are compact. You can line up ten or twenty of them side by side and compare distributions at a glance. Where is the median higher? Which group has more variability? Which distributions are symmetric, and which are skewed? Are there outliers, and in which groups?

One important limitation: box plots do not show the shape of the distribution within the box. A symmetric distribution and a bimodal distribution could produce identical box plots if they have the same five-number summary. For this reason, some analysts prefer **violin plots**, which combine the compactness of box plots with the shape information of density plots by showing a mirrored density curve on each side of the box. The figure below shows violin plots of the same redlining data as Figure 4.1, making visible the distributional shape within each HOLC grade that the box plot alone cannot show.

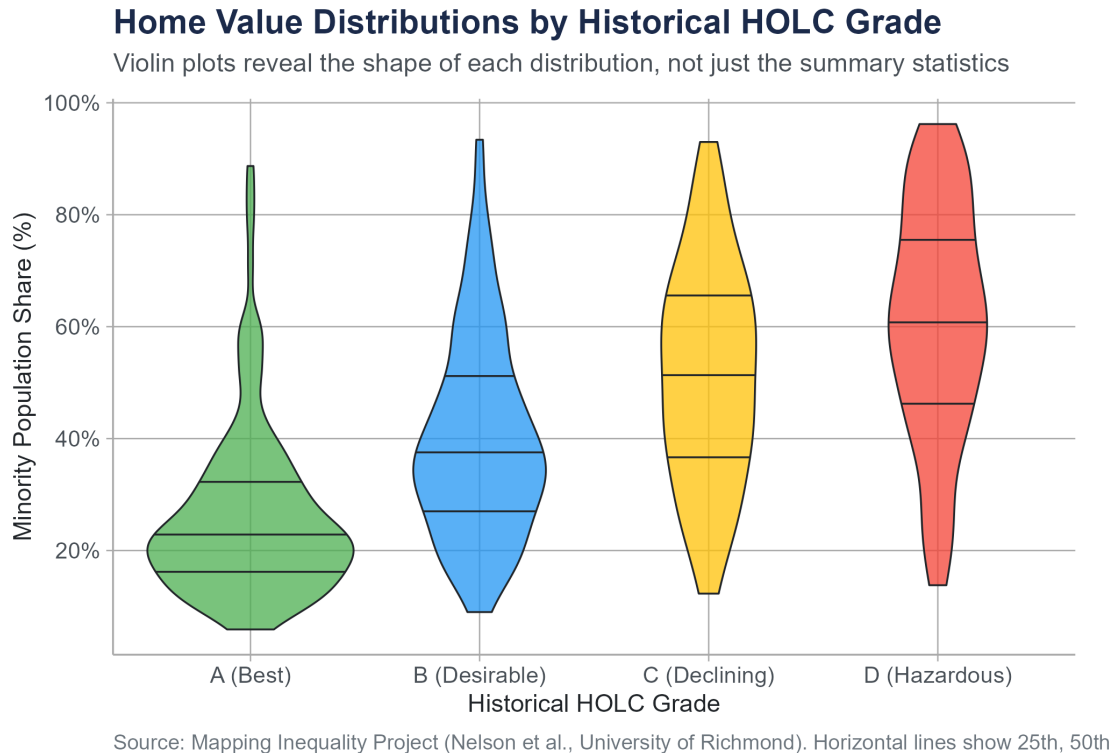


Figure 4.3: Minority population share (%) by historical HOLC grade, shown as violin plots. Unlike box plots, violin plots reveal the full shape of each distribution, including skewness and the concentration of values. Horizontal lines show the 25th, 50th, and 75th percentiles. Source: Mapping Inequality Project (Nelson et al., 2016, University of Richmond).

Box plots are also less intuitive for audiences unfamiliar with them. If you are presenting to a general audience, explain what the parts represent before expecting interpretation. For technical and academic audiences, box plots are a standard and efficient tool.

4.5 Bar Charts

Bar charts are the workhorses of categorical data visualization. They represent counts, proportions, or some summary statistic for each category as the height (or length) of a bar.

This sounds simple, and it is. But simplicity does not mean bar charts are foolproof. They are, in fact, one of the most frequently botched chart types in business presentations and news media.

4.5.1 When to Use Bar Charts

Use a bar chart when you want to compare a quantity across categories. How many customers are in each region? What proportion of survey respondents chose each option? What is the average satisfaction score for each product?

Bar charts work for nominal categories (regions, product names, blood types) and ordinal categories (education levels, satisfaction ratings). For ordinal categories, order the bars in their natural sequence. For nominal categories, order them by frequency (tallest to shortest) to make comparisons easier, unless there is a conventional ordering your audience expects.

4.5.2 Dos and Don'ts

Do start the vertical axis at zero. This is the most important rule for bar charts. The visual comparison of bar heights works because the height of each bar is proportional to the value it represents. If the axis starts at some number other than zero, the visual proportions are distorted and the chart becomes misleading. A bar that appears twice as tall as another should represent a value that is actually twice as large.

Do keep it simple. One set of bars, clearly labeled, with a clear title. If you need to compare two sets of categories, a grouped bar chart (bars side by side within each category) or a stacked bar chart can work, but these get complicated fast. If you find yourself creating a chart with twelve colors and a legend that takes up half the page, reconsider your approach.

Don't use 3D effects. Three-dimensional bar charts add visual noise and make it harder to compare bar heights accurately. There is no situation in which a 3D bar chart communicates information better than a 2D one.

Don't use bar charts for continuous data. If your horizontal axis represents a continuous variable like time or temperature, a line chart or histogram is usually more appropriate. A bar chart for stock prices over 60 consecutive trading days is technically possible but makes the chart harder to read than a simple line.

Don't overload. A bar chart with 40 categories is not a bar chart. It is a wall of color. If you have many categories, consider filtering to the most important ones, grouping smaller categories into an “other” category, or using a different visualization entirely.

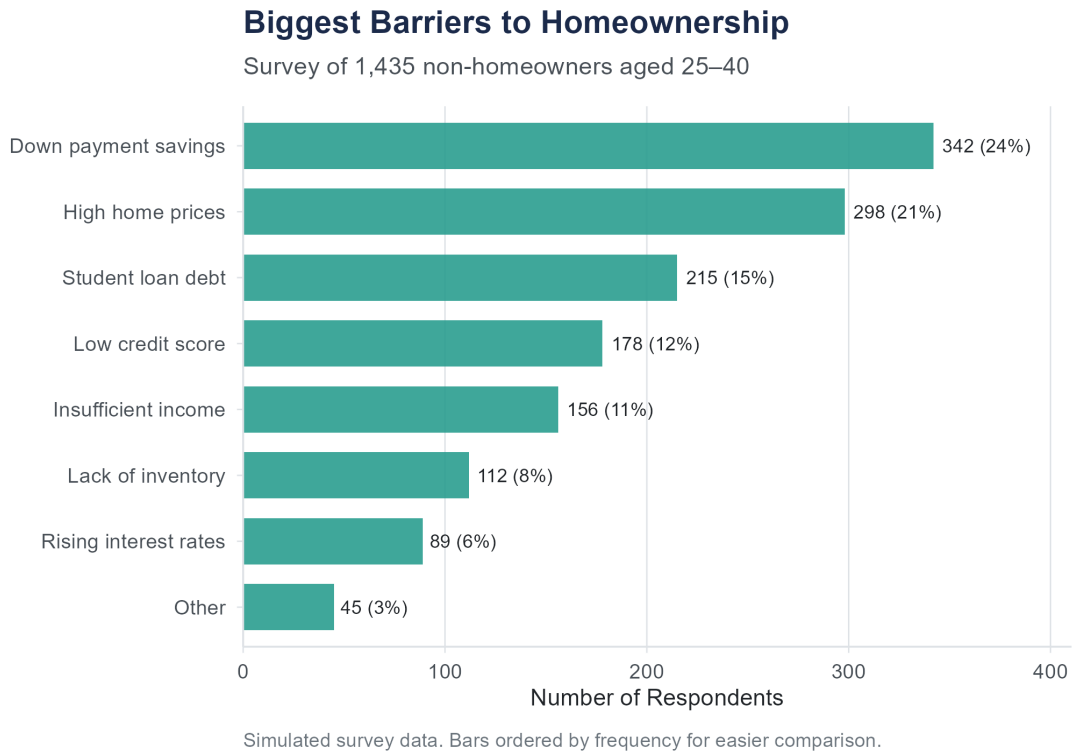


Figure 4.4: A clean horizontal bar chart showing survey responses about barriers to homeownership. Bars are ordered by frequency, making comparisons easy. The axis starts at zero, and each bar is clearly labeled. Source: Simulated survey data.

4.5.3 A Special Case: Pie Charts

We should talk about pie charts, because you are going to encounter them everywhere, and because many data visualization experts would prefer you did not use them.

A pie chart represents proportions as slices of a circle. The idea is intuitive. The whole circle is 100%, and each slice represents a category's share.

The problem is that humans are not very good at comparing angles and areas. We are much better at comparing lengths. Studies in visual perception research have consistently shown that people estimate proportions more accurately from bar charts than from pie charts. When there are more than a few categories, or when the proportions are similar in size, pie charts become very difficult to read accurately.

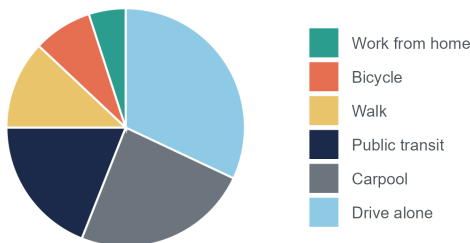
Pie charts can work when you have only two or three categories with very different proportions and the primary message is something like “Category A dominates.” In almost every other situation, a bar chart is more effective.

Use pie charts sparingly, if at all. If you do use one, never use a 3D pie chart. The perspective distortion makes slices in the front appear larger than equally-sized slices in the back.

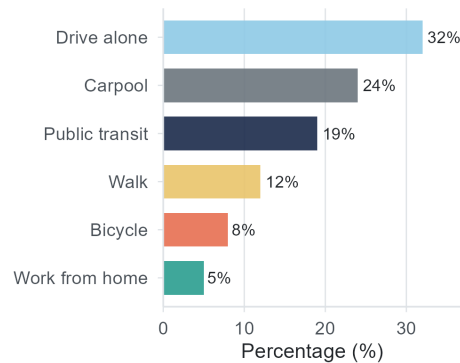
Same Data, Two Chart Types: Which Is Easier to Read?

Comparing proportions is much easier with bar lengths than with pie slice angles

Pie Chart



Bar Chart



Simulated commute mode share data.

Can you tell the exact difference between Carpool and Public Transit in the pie chart?

Figure 4.5: The same transportation data shown as a pie chart (left) and a bar chart (right). Try to estimate the exact difference between “Carpool” and “Public Transit” in the pie chart – it is much easier in the bar chart. Source: Simulated commute mode share data.

4.6 Scatter Plots

Everything we have discussed so far involves visualizing one variable at a time, or one variable broken down by categories of another. But some of the most interesting questions in statistics are about relationships between two numerical variables. Does more education lead to higher income? Does advertising spending relate to sales? Does temperature affect the number of bike-share riders on a given day?

A **scatter plot** answers these questions visually. It places one variable on the horizontal axis (often called x) and another on the vertical axis (often called y). Each observation is a point on the plot, positioned according to its values on both variables.

When you look at a scatter plot, you are looking for patterns.

Direction. Do the points tend to go upward from left to right (a positive association, meaning higher values of x tend to go with higher values of y)? Or downward (a negative association, meaning higher x goes with lower y)? Or is there no apparent trend?

Form. Is the pattern roughly linear (the points scatter around a straight line), or is it curved? A linear pattern is simpler to describe and model. A curved pattern requires more sophisticated tools.

Strength. How tightly do the points cluster around the pattern? If the points hug a line closely, the relationship is strong. If the points form a loose cloud, the relationship is weak.

Unusual observations. Are there points far from the overall pattern? These could be data errors, or they could be unusual cases worth investigating.

Scatter plots are powerful because they show you the raw data. Unlike summary statistics that condense information into a few numbers, a scatter plot preserves every observation. You can see clusters, gaps, outliers, and nonlinear patterns that no single number could capture.

Before the next figure, a brief preview of two terms we will define more fully later. The **Pearson correlation coefficient** r is a number between -1 and $+1$ that measures the strength and direction of a linear relationship between two variables; it is formally introduced in Section 4.7. The **confidence band** around a trend line (a shaded region showing uncertainty in the estimated relationship across all values of x , distinct from the confidence intervals for parameters defined formally in Chapter 7) can be read as saying: “the true line probably falls somewhere within this region.”

Study Hours and Exam Performance

Each dot represents one student; the trend line shows the positive linear relationship



Simulated data ($n = 120$). Shaded band shows 95% confidence interval for the regression line.

Figure 4.6: A scatter plot showing the positive relationship between weekly study hours and exam scores ($r = 0.72$). Each dot is one student. The trend line and shaded confidence band summarize the linear pattern. Source: Simulated data ($n = 120$).

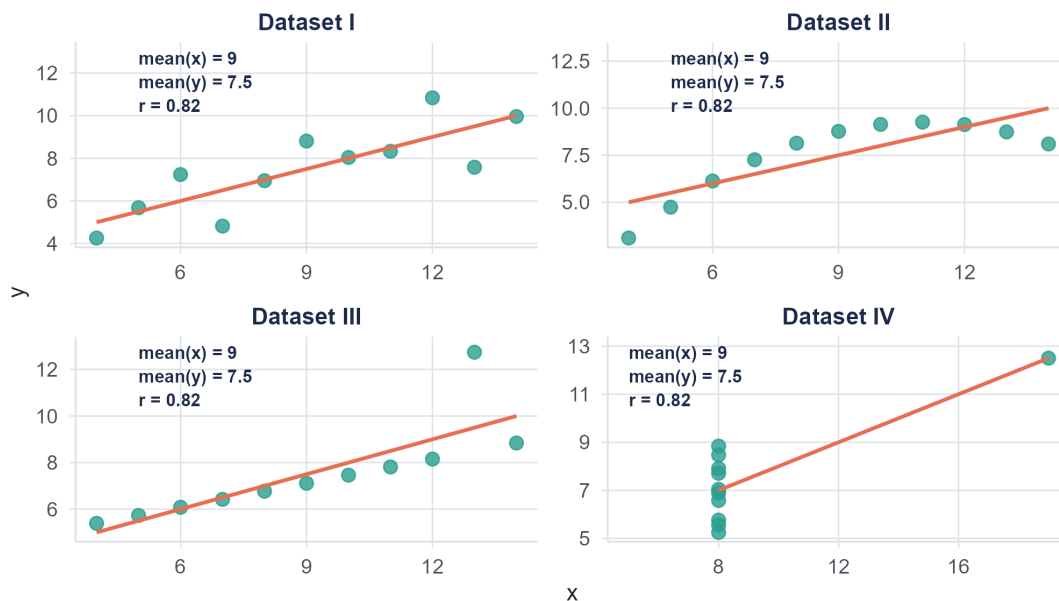
i Note

In 1973, Francis Anscombe, then at Yale University, published “Graphs in Statistical Analysis” (*The American Statistician*, 27(1), 17–21), in which he constructed four datasets that have approximately the same means, variances, correlation, and regression line, yet look completely different when plotted. Anscombe’s Quartet remains one of the most famous demonstrations of why you should always visualize your data before relying on summary statistics alone. More recently, Alberto Cairo created the original “Datasaurus” dataset; Justin Matejka and George Fitzmaurice of Autodesk Research extended it into the “Datasaurus Dozen” (CHI 2017): thirteen datasets, including one shaped like a dinosaur, that all share the same summary statistics. The lesson is clear. Numbers alone can hide what a picture reveals.

Open the **Datasaurus Explorer** to interact with these datasets directly and see the point illustrated.

Anscombe's Quartet: Same Statistics, Different Stories

All four datasets share nearly identical means, correlations, and regression lines



Anscombe (1973). Each dataset has $\text{mean}(x) = 9$, $\text{mean}(y) \approx 7.5$, $r \approx 0.82$, and the same regression line.

Figure 4.7: Anscombe's Quartet: four datasets with nearly identical means, correlations ($r \approx 0.82$), and regression lines, yet strikingly different visual patterns. Always visualize your data. Source: Anscombe (1973), *The American Statistician*, 27(1), 17–21.

4.7 Correlation

Looking at a scatter plot and describing what you see in words like “positive,” “strong,” or “linear” is useful but subjective. Two people looking at the same scatter plot might disagree about whether the relationship is “moderately strong” or just “moderate.” We need a number.

4.7.1 Pearson's Correlation Coefficient

The most common measure of the linear relationship between two numerical variables is **Pearson's correlation coefficient**, denoted r . It ranges from -1 to $+1$. This is the same r that appeared in the scatter plot figures above, and now we can define it precisely.

Here is what the values mean.

- $r = +1$ means a perfect positive linear relationship. Every point falls exactly on a line that slopes upward.
- $r = -1$ means a perfect negative linear relationship. Every point falls exactly on a line that slopes downward.
- $r = 0$ means no linear relationship. Knowing the value of x tells you nothing about the value of y , at least not through a linear pattern.
- Values between 0 and ± 1 indicate varying degrees of linear association. The closer $|r|$ is to 1, the tighter the points cluster around a straight line.

The formula for Pearson's r for a sample of n paired observations $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ is

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

where $\bar{x}$ is the sample mean of the x values and $\bar{y}$ is the sample mean of the y values.

This formula looks intimidating, but the logic is elegant. The numerator multiplies each x value's deviation from its mean by the corresponding y value's deviation from its mean, then sums these products. If x and y tend to be above their means together and below their means together (positive association), the products are mostly positive, making the sum positive. If high x tends to go with low y (negative association), the products are mostly negative. The denominator standardizes the result to fall between -1 and $+1$.

An equivalent formula that some textbooks prefer uses z-scores:

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

where s_x and s_y are the sample standard deviations of x and y respectively. This version makes it clear that r is the average product of the standardized values. When both variables are above their means (both z-scores positive) or both below (both negative), the product is positive. When one is above and the other below, the product is negative.

4.7.2 Interpreting Correlation

How strong is “strong”? There is no universal standard, but here are some commonly used rough guidelines for interpreting the magnitude of r .

$ r $	Interpretation
0.00 to 0.19	Very weak or negligible
0.20 to 0.39	Weak
0.40 to 0.59	Moderate
0.60 to 0.79	Strong
0.80 to 1.00	Very strong

These are rough categories, not rigid thresholds. Context matters substantially. In physics, a correlation of 0.6 between two quantities that theory says should be perfectly related would be considered poor, since measurement and experimental error should be much smaller than that. In psychology or education research, a correlation of 0.4 between a personality measure and a behavioral outcome might be considered quite strong, given the enormous complexity of human behavior and the imperfect quality of most measurements. Always interpret r in the context of your field and your question.

There are several important limitations of Pearson’s r that you need to understand.

Correlation measures linear relationships only. If x and y have a perfect curved relationship (say, $y = x^2$), the correlation can be zero or near zero even though the variables are perfectly related. Always look at the scatter plot.

Correlation is sensitive to outliers. A single extreme observation can dramatically change the value of r . One data point far from the cloud of other points can inflate or deflate the correlation, depending on where it falls. Again, look at the scatter plot.

Correlation does not imply causation. This is so important it gets its own section.

Correlation is unitless. It does not matter whether you measure height in inches or centimeters, or weight in pounds or kilograms. The correlation between height and weight will be the same regardless of the units, because the formula standardizes both variables.

Correlation treats both variables symmetrically. The correlation of x with y is the same as the correlation of y with x . This is different from regression, where the roles of x and y are distinct. We will encounter regression in later chapters.

4.7.3 Building Your Correlation Intuition

One of the best ways to understand correlation is to practice estimating it. Look at a scatter plot and guess the correlation before calculating it. Over time, your visual intuition will calibrate to the numbers.

Here are some benchmarks to anchor your intuition.

When r is near 0, the scatter plot looks like a shapeless cloud. There is no discernible upward or downward trend. If you squint, you might think you see a pattern, but that is your brain finding patterns where none exist, a tendency we will discuss more in the chapter on probability.

When $|r|$ is around 0.3, there is a slight trend. You can sort of see it, but the cloud of points is wide relative to the trend. Think of the relationship between shoe size and GPA in a sample that mixes adults and children of different ages: very slight association driven by age as a confounding variable.

When $|r|$ is around 0.7, the trend is clear. The points form a band rather than a cloud, and the direction is obvious. The relationship between height and weight in a sample of adults is often in this range.

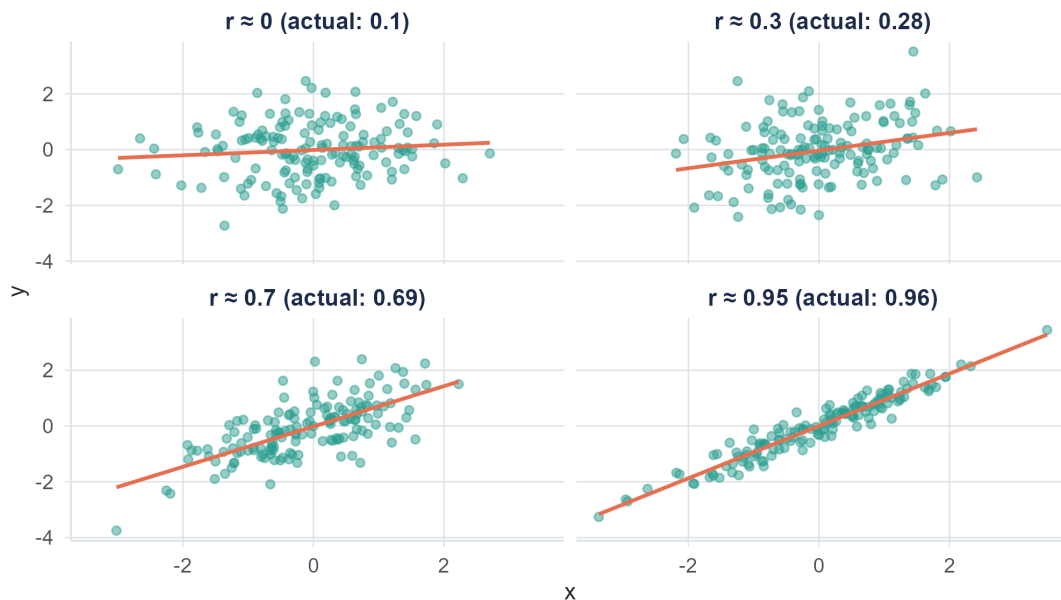
When $|r|$ is above 0.9, the points nearly form a line. There is very little scatter. Physical measurements of the same object taken with slightly different instruments often correlate this strongly.

Try It Online

Open the **Correlation Game** on the companion website. It shows scatter plots and asks you to estimate r before revealing the answer. Play it enough times and you will develop a reliable visual sense for what different correlation strengths look like, a skill that transfers directly to reading scatter plots in research and practice.

What Different Correlations Look Like

As r increases, points cluster more tightly around the trend line



Simulated data ($n = 150$ per panel). Red line shows the linear trend.

Figure 4.8: Four scatter plots showing what different correlation strengths look like: r near 0 (shapeless cloud), r near 0.3 (slight trend), r near 0.7 (clear band), and r near 0.95 (nearly a line). Source: Simulated data ($n = 150$ per panel).

4.8 The Correlation-Causation Trap

If there is one lesson from this chapter that you carry with you for the rest of your life, let it be this.

Correlation does not imply causation.

You have probably heard this before. You will hear it again. You will hear it so many times that you may start to think it is a cliché that does not need repeating. And then you will read a headline that says “Study Finds Coffee Drinkers Live Longer” and think “I should drink more coffee,” and at that moment you will be falling into the correlation-causation trap just like everyone else.

Let us be precise about what the problem is.

When two variables are correlated, there are several possible explanations.

X causes Y. This is the explanation that news headlines jump to. More education causes higher income. Smoking causes lung cancer. Sometimes it is true. But correlation alone cannot establish it.

Y causes X. The direction of causation can be the reverse of what you assume. Maybe higher income allows people to pursue more education, rather than education causing higher income. In reality, it is probably a bit of both.

Z causes both X and Y. This is the **confounding variable** (or lurking variable) problem, and it is the most common trap. Ice cream sales and drowning deaths are positively correlated. Ice cream does not cause drowning. Hot weather causes both more ice cream consumption and more swimming, and more swimming means more drownings. The confounding variable is temperature. In Chapter 2, we examined this problem in detail and showed how directed acyclic graphs (DAGs) can make confounding relationships visible. The three conditions for establishing causality discussed in Chapter 2 (concomitant variation, time order, and elimination of alternatives) are exactly what prevents us from jumping from correlation to causation.

The correlation is coincidental. With enough variables, some will correlate by pure chance. This is where things get both funny and alarming.

4.8.1 Spurious Correlations

Tyler Vigen, then a student at Harvard Law School, created a website and later a book called *Spurious Correlations* (2015; tylervigen.com/spurious-correlations) that pairs variables with high correlations and no plausible causal connection. The results are memorable.

The number of films Nicolas Cage appeared in between 2000 and 2009 correlates with the number of people who drowned by falling into a swimming pool during those same years ($r = 0.664$). Per capita cheese consumption in the United States correlates almost perfectly with the number of people who died by becoming tangled in their bedsheets ($r = 0.947$). The divorce rate in Maine correlates with per capita consumption of margarine ($r = 0.993$).

All three correlations were computed by Vigen from real government data: the CDC National Center for Health Statistics, USDA food supply data, and state-level vital statistics. The correlations are real. The patterns do appear in the data over these periods. That is exactly what makes them useful as teaching examples. These are memorable because the causal explanations are absurd. Nicolas Cage movies do not cause drowning. Cheese does not cause bedsheet entanglement. But the correlations are real, in the sense that the numbers actually do move together over those time periods.

The point is not that correlation is useless. Correlation is a valuable tool. The point is that correlation tells you two things move together. It does not tell you *why*. Figuring out why requires the kind of study design discussed in Chapter 2, ideally a controlled experiment with random assignment, combined with domain knowledge and critical thinking.

4.8.2 Why We Keep Falling for It

The correlation-causation trap is persistent because our brains are wired to seek causal explanations. This is a feature, not a bug. Early humans who assumed the rustling in the grass was caused by a predator survived more often than those who assumed it was just the wind. But in a world of complex data and multiple interacting variables, our instinct to see causation in every pattern leads us astray.

Media coverage makes it worse. “Correlation found between X and Y” is a boring headline. “X causes Y” is a compelling one. Science journalists, pressed for clicks and constrained by space, often frame correlational findings as causal ones. Even when the article itself includes appropriate caveats, the headline does not.

What can you do? Three things.

First, whenever you see a claimed association, ask yourself what confounding variables might explain it. If a study finds that people who own dogs have lower blood pressure, could it be that people who are healthier and more active are more likely to own dogs? Could it be that wealthier people are more likely to both own dogs and have access to better healthcare?

Second, ask whether the study design supports causal claims. Observational studies, no matter how large or sophisticated, cannot establish causation on their own. Only experiments with random assignment can do that (Chapter 2). There are advanced statistical methods that try to approximate experimental conditions using observational data, but even these come with assumptions and limitations.

Third, be especially skeptical of causal claims about topics where confounding is likely rampant: topics like nutrition (“eating X causes Y”), lifestyle (“doing X makes you live longer”), and anything involving self-selected groups (“people who choose X are more successful”).

⚠ AI Reality Check

AI tools can compute correlations instantly, but they struggle with the context that separates useful analysis from misleading analysis. Ask an AI to analyze the relationship between two variables and it will report the correlation coefficient, often slipping into causal language. “Higher values of X lead to higher values of Y” when the data shows only association. The word “lead” implies a direction of influence that correlation cannot establish, but AI tools reach for causal phrasing because it sounds more informative.

AI chart-making tools introduce a different set of problems. Default visualization choices in many platforms favor pie charts over bar charts, add unnecessary 3D effects, truncate axes to make small differences look dramatic, and choose color schemes that obscure rather than reveal patterns. An AI asked to “make a chart of this data” will produce something that looks polished and professional while violating several of the principles discussed in this chapter. The human analyst’s job is twofold: choose the visualization that tells the truth about the data, and describe the patterns without overstating what the data can support. Correlation is a tool for measuring association. Turning it into a causal claim requires the kind of study design discussed in Chapter 2, not a confident-sounding sentence from a language model.

4.9 Distortion as an Ethical Failure

In Section 4.2, we listed “Do not distort the data” as one of the five core principles of effective visualization. It is worth pausing to make something explicit: the violations described in this section are not merely technical mistakes. When a chart manipulates scales, cherry-picks time windows, or removes axes to create a false impression, it misrepresents reality to an audience. That is an ethical failure, not just a design flaw. The question to ask about any visualization is not only “does it look right?” but “does it accurately represent what the data say?”

4.9.1 Truncated Axes

The most common way to mislead with a visualization is to truncate the vertical axis, starting it at some value other than zero. For bar charts, this is almost always deceptive, because it exaggerates differences between bars.

Imagine a bar chart comparing approval ratings for two candidates. Candidate A has 51% approval and Candidate B has 49%. If the axis runs from 0% to 100%, the two bars look nearly identical,

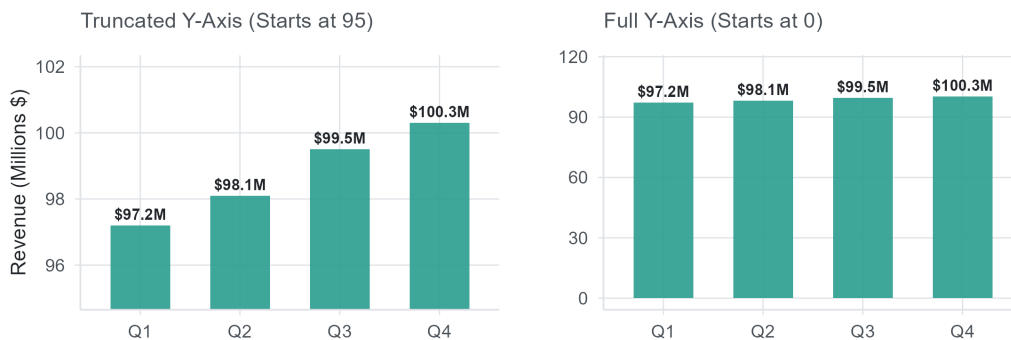
which accurately reflects a very small difference. If the axis runs from 47% to 53%, the bars look dramatically different, with Candidate A's bar appearing to tower over Candidate B's. The underlying numbers are the same. The visual impression is completely different.

For line charts, axis truncation is more of a judgment call. A line chart showing a company's revenue over time might reasonably start the axis at a value near the lowest data point, because the purpose is to show the trend, not the absolute magnitude. But even with line charts, a truncated axis can exaggerate small fluctuations into what looks like dramatic volatility.

The key question is whether the visual impression matches the substantive reality. If a small difference is substantively unimportant, the chart should not make it look dramatic. If it is substantively important, the chart should help the reader see it, but with clear labeling that indicates the scale.

The Same Data, Two Stories: Y-Axis Manipulation

Truncating the y-axis exaggerates small differences



Same simulated data in both panels. Only the y-axis range differs.

Figure 4.9: The same quarterly revenue data shown with two different y-axis ranges. The truncated axis starting at 95 makes small differences look dramatic, while the full axis starting at 0 reveals that all four quarters are nearly identical. Source: Simulated data.

4.9.2 Cherry-Picked Time Ranges

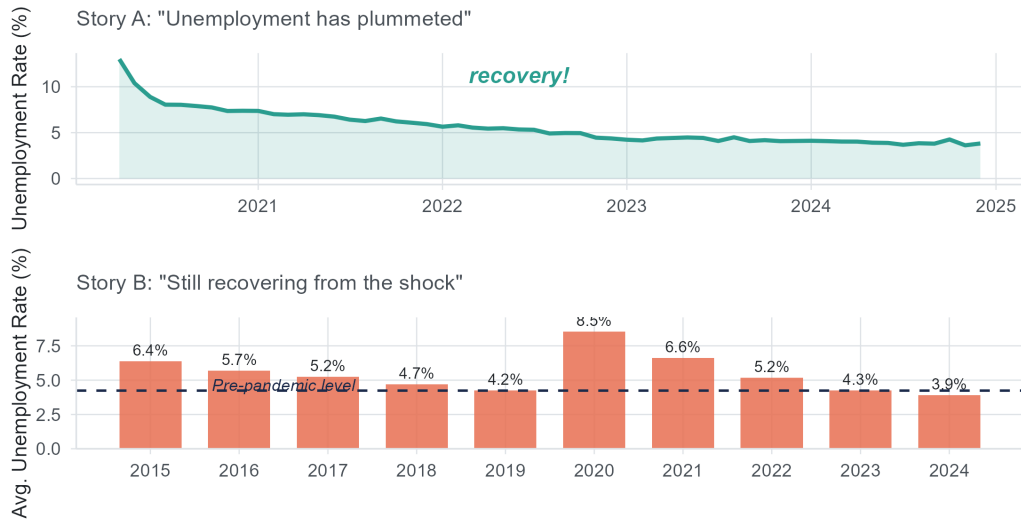
Choosing where to start and end a time series can completely change the story. A stock that has risen 30% over five years looks great. That same stock, if it dropped 50% in year one and has been recovering since, tells a different story. A politician can claim economic growth by choosing

a starting point in a recession and an ending point during recovery. A critic can claim economic decline by choosing the reverse.

Always look at the full axis range and ask yourself whether a different starting point would change the impression.

Same Data, Two Stories: How Framing Changes the Message

Cherry-picking time ranges and aggregation can tell opposite stories



Simulated monthly unemployment data, 2015–2024. Both panels use the same underlying dataset.

Figure 4.10: The same unemployment data framed two ways. Story A cherry-picks the post-pandemic recovery period to show a dramatic decline. Story B shows annual averages over the full decade, revealing that rates have not yet returned to pre-pandemic levels. Source: Simulated data based on published BLS unemployment patterns.

4.9.3 Manipulated Aspect Ratios

Stretching a chart vertically exaggerates changes. Squashing it horizontally compresses them. The same data can look like a dramatic trend or a flat line depending on the aspect ratio. There is no single “correct” aspect ratio, but extreme ratios should make you suspicious.

4.9.4 Misleading Area and Volume Comparisons

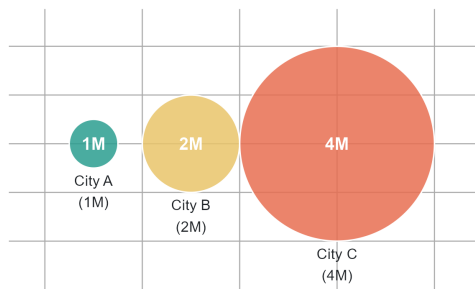
When visual elements are scaled by area or volume rather than by length, differences get distorted in ways that are easy to overlook. Suppose a chart represents city populations as circles scaled by radius. City A has twice the population of City B, so the designer makes City A's radius twice as large. But area grows as the square of radius: City A's circle has *four* times the area of City B's. It looks four times as large rather than twice as large. Three-dimensional representations are worse: doubling each linear dimension produces an eightfold increase in volume.

The rule is simple: if you are comparing quantities, scale visual elements by length (bar charts, dot plots), not by area or volume. When you see bubble charts or pictograms with scaled icons, check how the scaling was done. If the caption says “area is proportional to value,” verify it against the numbers shown. If the caption says nothing about scaling, assume the designer may not have checked.

Area Distortion: Why Circles Mislead

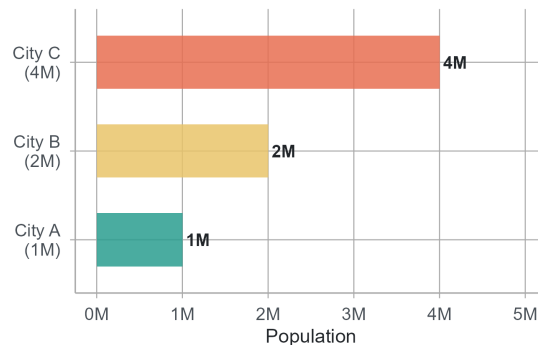
Doubling the radius makes a circle 4× larger in area, not 2×

Panel A: Circles scaled by radius (misleading)



City C appears 16× larger than City A, but is only 4× larger.

Panel B: Bar chart scaled by length (correct)



City C's bar is exactly 4× longer than City A's.

Figure 4.11: Area distortion in practice. Panel A scales three city populations by circle radius: City C appears 16 times larger than City A, but is only 4 times larger. Panel B uses a bar chart where length is proportional to value — City C's bar is exactly 4 times longer. Always compare quantities by length, not area or volume.

4.9.5 The Dual-Axis Trap

Charts with two different vertical axes, one on the left and one on the right, allow two variables to be plotted on the same chart. This sounds helpful, but it creates a serious problem. By choosing

the scales for the two axes, you can make any two variables appear to track each other closely, even if the relationship is meaningless.

To see why, consider a chart showing Nicolas Cage film releases on the left axis and U.S. pool drownings on the right. If you scale the two axes so that both series occupy roughly the same vertical space on the chart, the lines appear to move in tandem, because the designer forced them to through scale choice. This is, in fact, how many spurious correlations are visualized. If you see a dual-axis chart, be skeptical. Ask whether the apparent relationship would survive a different choice of scales.

The Dual-Axis Trap: You Can Make Any Two Variables Track Together

Pool drownings and Nicolas Cage films (1999–2009, $r = 0.664$). Panel A hides the fact that one series ranges from 85–123 deaths and the other from 1–4 films.

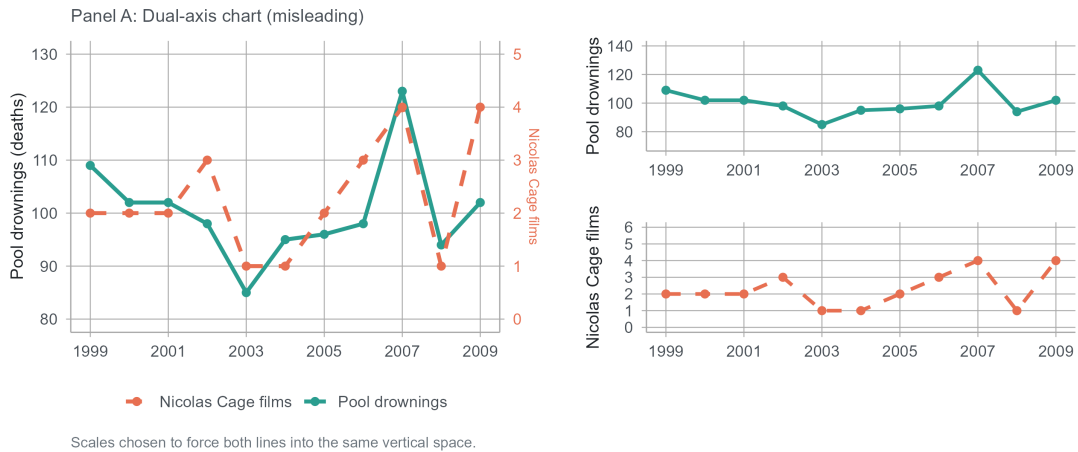


Figure 4.12: The dual-axis trap illustrated with Vigen’s Nicolas Cage / pool drowning data ($r = 0.664$). Panel A uses a dual axis scaled so both lines fill the same vertical space — they appear to track each other closely. Panel B shows each variable at its true scale: one series ranges from 85 to 123 deaths, the other from 1 to 4 films. The apparent relationship in Panel A is entirely an artifact of scale choice. Source: Tyler Vigen, *Spurious Correlations* (2015); data from CDC and IMDb.

! Ethics Moment

In September 2015, Representative Jason Chaffetz displayed a chart at a House Oversight Committee hearing that appeared to show Planned Parenthood’s cancer screening services plummeting while its abortion services skyrocketed. The chart was created by Americans United for Life and used no vertical axis at all. The two trend lines were placed on the same

graphic with no common scale, making it impossible to judge actual magnitudes. When replotted with a proper axis, the picture was far less dramatic. Cancer screenings had indeed declined (from about 2 million to about 935,000 annually between 2006 and 2013), but they still substantially exceeded abortion services (roughly 300,000 annually). The original chart was not technically lying about the direction of the trends, but by removing the scale, it created a deeply misleading visual impression.

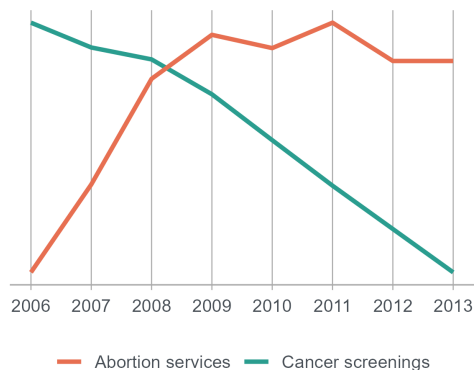
This is not a partisan point. Misleading charts appear across the political spectrum and in corporate communications, news media, and advocacy from every direction. When someone removes the axis, ask yourself why. When someone uses visual tricks to exaggerate a difference, ask who benefits from that exaggeration. Your job as a statistically literate person is not to take any chart at face value, but to read the data behind the picture.

For a comprehensive treatment of statistical deception, see *Calling Bullshit: The Art of Skepticism in a Data-Driven World* (2020) by Carl Bergstrom and Jevin West of the University of Washington, or their freely available course materials at callingbullshit.org.

The Same Data, Two Very Different Impressions

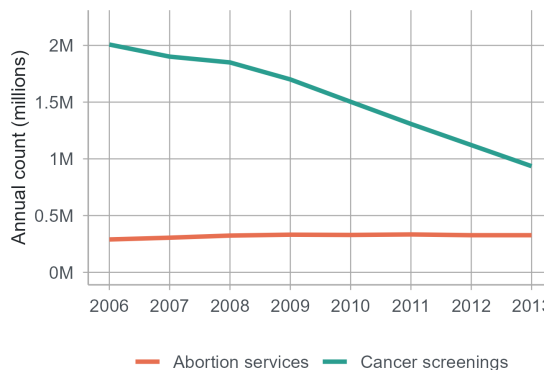
Removing the y-axis makes the lines appear to cross — but they never do

Panel A: Original chart (no common axis)



Each line scaled independently to fill the panel — no shared y-axis.

Panel B: Corrected chart (shared axis, actual counts)



Source: Planned Parenthood annual reports, 2006-2013. 2009-2010 estimates.

Figure 4.13: The same Planned Parenthood data shown two ways. Panel A removes the y-axis so both trend lines fill the same vertical space regardless of their actual magnitudes — the impression is that the lines are crossing dramatically. Panel B uses a shared axis: cancer screenings (teal) still substantially exceed abortion services (coral) throughout. Source: Planned Parenthood annual reports (approximate values).

4.10 Putting It All Together

Choosing the right visualization is not a matter of personal preference. It is driven by the type of data you have and the question you are trying to answer. Here is a quick reference.

Your question	Variable types	Best chart type
What does the distribution of one numerical variable look like?	One numerical	Histogram or density plot
How do distributions compare across groups?	One numerical, one categorical	Side-by-side box plots (or violin plots)
How many observations are in each category?	One categorical	Bar chart
What proportions do categories represent?	One categorical	Bar chart (or pie chart, reluctantly)
What is the relationship between two numerical variables?	Two numerical	Scatter plot
How does a numerical variable change over time?	One numerical, one time variable	Line chart
How do multiple numerical variables relate to each other?	Multiple numerical	Correlation matrix or scatter plot matrix

This table is a starting point, not a constraint. As you gain experience, you will develop an intuition for when standard chart types work and when a creative variation might be more effective. But when in doubt, stick with the basics. A clear, well-labeled standard chart beats a clever but confusing novel visualization every time.

Here is a principle that ties together everything in this chapter. The purpose of data visualization is not to impress. It is not to decorate. It is to make the truth easier to see. The redlining maps at the start of this chapter did exactly that. They took patterns that existed in tables of numbers and made them visible in a way that was immediate and undeniable. That is the standard to aim for.

4.11 Looking Ahead

Chapter 4 completes the descriptive foundations of the book. You now have both the numerical vocabulary from Chapter 3 (mean, median, IQR, standard deviation) and the visual vocabulary from this chapter (histograms, box plots, scatter plots, and the ethical dimension of visualization). Chapter 5 marks a shift. Instead of describing data we have already collected, we will begin asking probabilistic questions: how likely is a given outcome, how do we reason about uncertainty before observing data, and how does the math of probability connect to everything the rest of the book will build on? The transition from descriptive statistics to probability and inference is the most important conceptual step in the course, and Chapter 5 begins it by grounding the ideas in a scenario most readers encountered directly: diagnostic testing during a pandemic.

4.12 Key Terms

- **Bar chart:** A visualization that uses the height or length of bars to represent counts, proportions, or summary statistics for categories of a categorical variable.
- **Bin:** An interval used to group values when constructing a histogram. The number and width of bins affect the histogram's appearance and the features it reveals.
- **Bimodal:** A distribution with two distinct peaks, often suggesting that the data contains two different subgroups.
- **Box plot (box-and-whisker plot):** A visual representation of the five-number summary. The box spans Q_1 to Q_3 , the line inside marks the median, whiskers extend to the most extreme non-outlier observations, and individual points beyond the whiskers are flagged as potential outliers.
- **Confounding variable (lurking variable):** A variable that influences both the explanatory variable and the response variable, creating a misleading association between them. See Chapter 2 for a full treatment.
- **Correlation:** A measure of the strength and direction of the linear relationship between two numerical variables.
- **Data-ink ratio:** A concept introduced by Edward Tufte referring to the proportion of ink in a visualization that represents actual data rather than decoration or non-data elements.
- **Density plot (kernel density estimate):** A smoothed curve that estimates the distribution of a numerical variable. The total area under the curve equals 1.

- **Five-number summary:** The minimum, Q_1 , median, Q_3 , and maximum of a dataset. Introduced in Chapter 3 and visualized by the box plot.
- **Histogram:** A visualization that divides the range of a numerical variable into bins and uses bar heights to represent the frequency or relative frequency of observations in each bin.
- **Interquartile range (IQR):** The difference $Q_3 - Q_1$. Represents the spread of the middle 50% of the data. Resistant to outliers.
- **Outlier:** An observation that falls unusually far from the rest of the data. In a box plot, observations beyond $1.5 \times \text{IQR}$ from the nearest quartile are flagged as potential outliers.
- **Pearson's correlation coefficient (r):** A sample statistic measuring the strength and direction of the linear relationship between two numerical variables, ranging from -1 (perfect negative linear relationship) to $+1$ (perfect positive linear relationship). The corresponding population parameter is ρ (rho).
- **Scatter plot:** A visualization that plots two numerical variables against each other, with each observation represented as a point.
- **Skewed distribution:** A distribution that is not symmetric. A right-skewed distribution has a long tail extending toward higher values (mean $>$ median). A left-skewed distribution has a long tail extending toward lower values (mean $<$ median).
- **Spurious correlation:** A correlation between two variables that exists statistically but has no causal explanation, often arising from coincidence or a shared common cause.
- **Truncated axis:** An axis on a chart that does not start at zero, which can exaggerate visual differences between values.
- **Violin plot:** A visualization that combines the compactness of a box plot with the shape information of a density plot, showing a mirrored density curve on each side of a central summary. Useful when comparing full distributional shapes across groups.

4.13 Further Reading and References

The following works are cited in this chapter or provide valuable additional context.

On redlining and the Mapping Inequality project: Nelson, R.K., Winling, L., Marciano, R., Connolly, N.D.B., et al. (2016). Mapping inequality: Redlining in New Deal America. In R.K. Nelson & E.L. Ayers (Eds.), *American Panorama*. University of Richmond Digital Scholarship Lab. dsl.richmond.edu/panorama/redlining.

Mitchell, B., & Franco, J. (2018). *HOLC “redlining” maps: The persistent structure of segregation and economic inequality*. National Community Reinvestment Coalition. The source for the 74% and 64% persistence figures cited in the opening section.

On data visualization principles: Tufte, E.R. (1983). *The visual display of quantitative information*. Graphics Press. The source of the data-ink ratio concept.

Cairo, A. (2016). *The truthful art: Data, charts, and maps for communication*. New Riders. A thorough, practical guide to visualization design and the ethics of representation.

Bergstrom, C.T., & West, J.D. (2020). *Calling bullshit: The art of skepticism in a data-driven world*. Random House. Covers misleading charts and statistical deception in accessible, example-driven form. Course materials available at callingbullshit.org.

On Anscombe’s Quartet: Anscombe, F.J. (1973). Graphs in statistical analysis. *The American Statistician*, 27(1), 17–21. The original paper constructing the four datasets that share identical summary statistics yet look completely different when plotted.

Matejka, J., & Fitzmaurice, G. (2017). Same stats, different graphs: Generating datasets with varied appearance and identical statistics through simulated annealing. *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, 1290–1294. The Datasaurus Dozen, extending Anscombe’s insight to thirteen datasets including one shaped like a dinosaur.

On spurious correlations: Vigen, T. (2015). *Spurious correlations*. Hachette Books. Website at tylervigen.com/spurious-correlations. All correlations cited in Section 4.8.1 are from Vigen’s calculations using public government data sources (CDC National Center for Health Statistics, USDA food supply data, and state vital statistics). The underlying data is real; the correlations are real; the causal connections are not.

On the Chaffetz chart and misleading visualization in advocacy: The original Planned Parenthood chart used in the 2015 House Oversight Committee hearing was created by Americans United for Life. Reanalysis and visualization of the corrected chart appeared widely in fact-checking journalism, including PolitiFact and The Washington Post.

4.14 Exercises

4.14.1 Check Your Understanding

1. What is the data-ink ratio, and why did Edward Tufte argue it should be maximized? Give an example of a chart element that adds ink but not information.
2. You are given a histogram of household incomes in a large city. The histogram has a single peak near \$40,000 and a long right tail extending past \$500,000. Describe the shape of this distribution using the terms introduced in this chapter. Would you expect the mean or the median to be higher? Explain why, using what you learned in Chapter 3.
3. Explain the difference between a histogram and a bar chart. Give one situation where you would use each.
4. A box plot shows $Q_1 = 20$, median = 35, $Q_3 = 45$. The lower whisker extends to 5 and the upper whisker extends to 68. Three points are plotted individually above 68, at values 85, 95, and 110.
 - a. What is the IQR?
 - b. Calculate the upper and lower fences using the $1.5 \times \text{IQR}$ rule. Show your work.
 - c. Confirm that all three individual points (85, 95, 110) exceed the upper fence. Are any of them within the fence?
5. Two scatter plots are shown. In Plot A, the points form a tight band sloping upward from left to right. In Plot B, the points form a loose cloud with no discernible trend. Estimate a plausible value of r for each plot and explain your reasoning.
6. Suppose $r = 0$ for two variables x and y . Does this mean that x and y are unrelated? Explain, and give an example where $r = 0$ but a clear relationship exists.
7. A news headline reads “People Who Eat Organic Food Have 25% Lower Cancer Rates.” List at least three possible explanations for this correlation that do not involve organic food directly preventing cancer. For each explanation, name the type (reverse causation, confounding, or coincidence).
8. Why is starting the vertical axis of a bar chart at a value other than zero misleading? Is the same rule always true for line charts? Explain the difference.
9. Explain why pie charts are generally less effective than bar charts for comparing proportions across categories. Under what limited circumstances might a pie chart be acceptable?
10. The Pearson correlation coefficient between two variables is $r = -0.85$. Describe what a scatter plot of these two variables would look like, including the direction, form, and strength of the pattern.

4.14.2 Apply It

- The following table shows data for ten cities, listing average daily temperature (in degrees Fahrenheit) and the number of bike-share rides on a randomly selected day.

City	Temp (°F)	Rides
A	32	120
B	45	310
C	55	485
D	62	580
E	70	710
F	75	750
G	80	820
H	85	680
I	90	590
J	95	430

- Create a scatter plot with temperature on the horizontal axis and rides on the vertical axis.
 - Describe the pattern. Is it linear? What happens at very high temperatures?
 - Would Pearson's r be a good summary of this relationship? Why or why not?
- A researcher collects data on hours of sleep and self-reported mood scores (1 to 10) for eight participants.

Participant	Hours of Sleep	Mood Score
1	4	3
2	5	4
3	6	5
4	6	6
5	7	7
6	7	6
7	8	8
8	9	9

- Compute the means $\bar{x}$ and $\bar{y}$.

- b. Using the formula $r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \cdot \sqrt{\sum (y_i - \bar{y})^2}}$, calculate r by hand. Show your work.
- c. Interpret the value of r in context.
- d. Can you conclude that more sleep causes better mood? Explain using the Chapter 2 framework for causality (concomitant variation, time order, elimination of alternatives).
3. You have data on starting salaries of recent graduates from four departments at a university: Engineering, Business, Education, and Liberal Arts. You want to compare the distribution of starting salaries across departments. Which visualization would you choose and why? Describe what you would look for in the resulting chart, using specific features of the distributions (center, spread, shape, outliers).
4. Consider two variables measured on 200 college students: hours spent on social media per week and GPA. Suppose the correlation is $r = -0.32$.
- Describe the direction and strength of this correlation.
 - Does this correlation tell you that increasing social media use will lower a student's GPA? Explain.
 - Name two confounding variables that might help explain this correlation.
5. A marketing team creates a bar chart comparing quarterly sales for two products. Product A sold 1,010 units and Product B sold 990 units. The team starts the vertical axis at 980.
- Sketch or describe the chart with the axis starting at 0 instead.
 - How does the visual impression change?
 - In this case, does the small difference matter for the decision being made? How should that influence the choice of axis range?
6. Find a data visualization in a news article, social media post, or advertisement. Identify at least two design choices that either help or hinder accurate interpretation. If the visualization is misleading, describe specifically how and suggest an improvement.
7. The following table shows exam scores for two sections of a statistics course.

Section A	Section B
62	71
68	73
72	75
75	78
78	79
80	80

Section A	Section B
82	82
85	84
88	86
95	88

- a. Compute the five-number summary for each section. (Recall the method from Chapter 3.)
 - b. Draw side-by-side box plots (by hand or using software).
 - c. Compare the two distributions. Which section has a higher median? Which has more variability? Are there any potential outliers? (Use the $1.5 \times \text{IQR}$ rule.)
8. A dataset contains information about 1,000 houses sold in a metropolitan area, including sale price, square footage, number of bedrooms, lot size, year built, and neighborhood. For each of the following questions, state which type of visualization you would use and which variables would be on each axis.
- a. What is the distribution of sale prices?
 - b. How do sale prices differ across neighborhoods?
 - c. What is the relationship between square footage and sale price?
 - d. How many houses were sold in each neighborhood?
9. Explain why a correlation of $r = 0.993$ between the divorce rate in Maine and per capita margarine consumption does not mean that margarine causes divorce. Apply the four-part framework from this chapter (X causes Y, Y causes X, Z causes both, coincidence). What would a study designed to actually test a causal connection need to look like?
10. Using the z-score formula $r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$, explain in plain language:
- a. Why the correlation is positive when high values of x tend to go with high values of y .
 - b. Why it is negative when high values of x tend to go with low values of y .
 - c. What happens to each product when both variables are above their means? When one is above and the other is below?

The following three exercises use the **housing-redlining.csv** dataset available on the companion website. This dataset is drawn from the Mapping Inequality project (Nelson et al., 2016, University of Richmond, Virginia Tech, and University of Maryland), which digitized the original HOLC neighborhood security maps and linked them to current U.S. Census demographic data. It contains 551 neighborhoods across 138 metropolitan areas, with variables including neighborhood_id, metro_area, holc_grade (historical HOLC grade: A, B, C, or D), total_population, pct_white, pct_black, pct_hispanic, pct_asian, and pct_minority.

11. Using the `housing-redlining.csv` dataset, create side-by-side box plots of `pct_minority` grouped by `holc_grade` (A through D). Describe the pattern you observe. Which grade has the highest median percentage of minority residents? Which has the most variability? Are there potential outliers? Write two to three sentences interpreting what these box plots reveal about the persistent demographic legacy of redlining. (For reference, the mean `pct_minority` values by grade are approximately 26% for Grade A, 40% for Grade B, 51% for Grade C, and 61% for Grade D.)
12. Create a scatter plot of `pct_white` (horizontal axis) versus `pct_minority` (vertical axis), using color or shape to distinguish the four HOLC grades. Compute the Pearson correlation between `pct_white` and `pct_minority` across all 551 neighborhoods. Describe the direction, strength, and form of the relationship. Does the pattern appear consistent across HOLC grades, or do certain grades cluster in particular regions of the plot? What does the clustering reveal about the relationship between historical redlining grades and present-day racial demographics?
13. Choose any two of the 138 metropolitan areas in the dataset. For each metro area, create a bar chart showing the mean `pct_minority` for each HOLC grade. Place the two bar charts side by side. Do both metro areas show the same pattern across grades, or does one show a larger gap between Grade A and Grade D neighborhoods? What does this comparison suggest about how redlining's demographic legacy may vary across metropolitan areas?

4.14.3 Think Deeper

1. The *Mapping Inequality* project made redlining data visible in a way that tables of numbers could not. But visualization can also be used to mislead. How should we think about the power of visualization when it can be used both to reveal truth and to distort it? What responsibilities do data analysts have when choosing how to visualize data on sensitive topics like housing discrimination or public health?
2. A pharmaceutical company funds a study comparing its new drug to a competitor. The study finds the new drug is slightly more effective ($r = 0.12$ between drug type and patient improvement). The company creates a visualization for a marketing brochure that uses truncated axes, strategic color choices (the competitor's bar in gray, their drug's bar in bright green), and a carefully chosen time window. No numbers on the chart are wrong. Every data point is real. Is this ethical? Should there be standards for how statistical results are visualized in advertising? Who should set and enforce those standards?
3. Pearson's r only measures linear relationships. Many real-world relationships are nonlinear. A medication might help at low doses and harm at high doses (an inverted-U relationship). Economic growth might accelerate and then plateau. Student performance might improve with

study time up to a point, then decline as fatigue sets in. Given that Pearson's r could be near zero for all of these relationships, what does this suggest about the risk of using a single summary statistic to describe a relationship? How should analysts guard against this risk?

4. Dual-axis charts are popular in business presentations and news media. Critics argue they are inherently misleading because the creator can choose scales that make any two variables appear related. Defenders argue that with careful labeling and honest scale choices, dual-axis charts can be informative. Where do you stand, and why? Can you think of a specific case where a dual-axis chart would be useful, and another where it would be deceptive?
5. In the age of social media, data visualizations can go viral, reaching millions of people who have no training in statistical literacy. A misleading chart shared widely enough can shape public opinion before any correction reaches even a fraction of the original audience. What obligations, if any, do social media platforms have to flag or contextualize potentially misleading data visualizations? How would such a system work in practice, and what are the risks of getting it wrong?

Part II

Probability and Inference

Chapter 5

Probability: Thinking About Chance

5.1 The Test Said Yes

In the spring of 2020, as COVID-19 spread across the world, a question consumed millions of people. If my test comes back positive, does that mean I have the virus?

The answer, it turned out, was “not necessarily.” And the reason is a foundational idea in statistics.

Here is a simplified but realistic scenario from the early months of the pandemic. Suppose you lived in a community where about 1% of the population was actually infected at a given time. Suppose you took a rapid antigen test with the following performance characteristics, which were roughly in line with many tests on the market. If you truly had the virus, the test correctly came back positive about 90% of the time. If you did not have the virus, the test correctly came back negative about 95% of the time.

Now suppose you tested positive. How worried should you have been?

Most people, including many physicians, would have said something like “the test is 90% accurate, so there is about a 90% chance you are infected.” That answer is wrong. It is not a little wrong. It is sharply wrong.

Let’s work through the math with a concrete group. Imagine 10,000 people from this community get tested. Since 1% are truly infected, that gives us 100 people with the virus and 9,900 without it. Of the 100 infected people, 90% test positive, so 90 get a correct positive result. Of the 9,900 uninfected people, 5% get a false positive, so 495 people who do not have the virus are told they do.

Add up all the positive tests. That's $90 + 495 = 585$ positive results. Of those 585 people staring at a positive test, only 90 actually have the virus. That means the probability that you truly have COVID given a positive test is $90/585 \approx 15.4\%$.

Read that again. A positive result on this test, in this community, meant about a 15% chance you were actually infected. More than 8 out of 10 people with positive results were false alarms.

This is not a failure of the test. The test performed exactly as advertised. This is a consequence of a low **base rate**, the proportion of people in the population who actually had the condition. When the base rate is low, even a good test produces more false positives than true positives, simply because the “healthy” group is so much larger than the “sick” group.

Understanding this requires understanding probability, and specifically a result called Bayes' theorem. That is the destination of this chapter. Along the way, we will build the fundamental rules of probability from the ground up, because Bayes' theorem does not come out of nowhere. It is a logical consequence of how probabilities combine.

During the pandemic, this misunderstanding had real consequences. People who tested positive entered mandatory quarantine, cancelled surgeries, pulled children from school, and isolated from dying relatives, sometimes on the basis of a result that was more likely to be wrong than right. Meanwhile, people who tested negative felt invincible and went about their lives, even though some of them were actually infected. The errors ran in both directions, and the human cost was substantial.

None of this means testing was useless. It means that interpreting a test result requires more information than the result itself. You need to know the base rate, the test's **sensitivity**, and the test's **specificity**. And you need to know how to combine them. That is what this chapter teaches.

Try It Online

Before you read another word, open the **Base Rate Lab** on the companion website and go to the **Natural Frequencies** tab. It shows the 10,000-person tree from the COVID story, live. Drag the prevalence slider from 1% down toward 0.1% and watch the “true positive” and “false positive” piles at the bottom of the tree. The positive predictive value — the chance you're really infected given a positive test — collapses toward zero even though the test itself hasn't changed at all. That collapse is the whole lesson of this chapter, and it is much easier to feel than to memorize.

5.2 What Probability Means

Before we calculate anything, we need to agree on what probability is. This turns out to be less obvious than it sounds, and professional statisticians have argued about it for more than a century.

There are three main interpretations.

The **classical** interpretation defines probability as the number of favorable outcomes divided by the total number of equally likely outcomes. The probability of rolling a 3 on a fair die is $1/6$ because there are six equally likely faces and one of them is a 3. This works well for dice, cards, and coins, but it struggles when outcomes are not equally likely, or when you cannot enumerate them.

The **frequentist** interpretation defines probability as the long-run relative frequency of an event. If you flip a coin thousands of times, the proportion of heads will settle near 0.5. The probability of heads *is* that long-run proportion. This interpretation is practical and intuitive for repeatable events, but it gets awkward when you want to talk about the probability of a one-time event, like the probability that a specific candidate wins an election.

The **Bayesian** (or subjectivist) interpretation defines probability as a *degree of belief* in a proposition, conditional on the evidence available. A doctor estimating a 10% chance that a patient's tumor is malignant before any biopsy, or an election analyst pegging a 65% probability that a particular candidate wins on Tuesday, is expressing a degree of belief about a single, non-repeatable situation. This interpretation handles one-time events naturally, but it requires a framework for updating beliefs as new evidence arrives. That framework, as you might guess, is Bayes' theorem.

For this book, we will use whichever interpretation makes sense for the problem at hand. In practice, most working statisticians move between interpretations depending on context, even if they have strong philosophical preferences. What matters for now is the mathematical machinery, which is the same regardless of interpretation.

5.2.1 The Language of Probability

An **experiment** (or **random process**) is any procedure that produces an uncertain outcome. Rolling a die, drawing a card, testing a patient, surveying a voter.

The **sample space**, denoted S , is the set of all possible outcomes. For a single coin flip, $S = \{H, T\}$. For rolling a die, $S = \{1, 2, 3, 4, 5, 6\}$.

An **event** is any subset of the sample space. “Rolling an even number” is the event $\{2, 4, 6\}$. “Getting heads” is the event $\{H\}$.

The **probability** of an event A , written $P(A)$, is a number between 0 and 1 (inclusive) that measures how likely the event is to occur. A probability of 0 means the event cannot happen. A probability of 1 means the event is certain. Everything else falls in between.

Two basic rules anchor everything that follows.

Rule 1. For any event A , $0 \leq P(A) \leq 1$.

Rule 2. The probabilities of all outcomes in the sample space must add up to 1. Something has to happen.

5.3 Basic Probability Rules

5.3.1 The Complement Rule

The **complement** of an event A , written A^c (read “A complement”), is the event that A does not occur. If A is “it rains today,” then A^c is “it does not rain today.”

Since either A happens or it doesn’t, and those are the only two options,

$$P(A) + P(A^c) = 1$$

which means

$$P(A^c) = 1 - P(A)$$

This is often the easiest way to compute a probability. If the probability of at least one defective item in a shipment is hard to calculate directly, it may be much simpler to calculate the probability that *no* items are defective and subtract from 1.

Example. A weather forecast says there is a 35% chance of snow tomorrow. What is the probability it does not snow?

$$P(\text{no snow}) = 1 - P(\text{snow}) = 1 - 0.35 = 0.65$$

Simple, but the complement rule becomes especially powerful when combined with other rules. If you want to know the probability that at least one person in a room of 30 shares your birthday, it is much easier to compute $1 - P(\text{nobody shares your birthday})$ than to enumerate all the ways one or more people could share it. We will see more examples like this as the chapter progresses.

5.3.2 The Addition Rule

Suppose you want to know the probability that event A or event B occurs (or both). The answer depends on whether A and B can happen at the same time.

Mutually exclusive events cannot occur simultaneously. If you roll a single die, the events “roll a 2” and “roll a 5” are mutually exclusive. You cannot get both on one roll.

For mutually exclusive events, the probability of A or B is simply the sum of their individual probabilities.

$$P(A \text{ or } B) = P(A) + P(B) \quad (\text{if } A \text{ and } B \text{ are mutually exclusive})$$

The probability of rolling a 2 or a 5 on a fair die is $1/6 + 1/6 = 2/6 = 1/3$.

But most events in the real world are **not** mutually exclusive. If you draw a card from a standard deck, the events “draw a heart” and “draw a queen” can both happen, because the queen of hearts exists. If you just add $P(\text{heart}) + P(\text{queen}) = 13/52 + 4/52 = 17/52$, you have counted the queen of hearts twice.

The general addition rule corrects for this double-counting.

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

For our card example, $P(\text{heart or queen}) = 13/52 + 4/52 - 1/52 = 16/52 = 4/13$.

Notice that when events are mutually exclusive, $P(A \text{ and } B) = 0$, so the general rule simplifies to the simpler rule. The general rule always works.

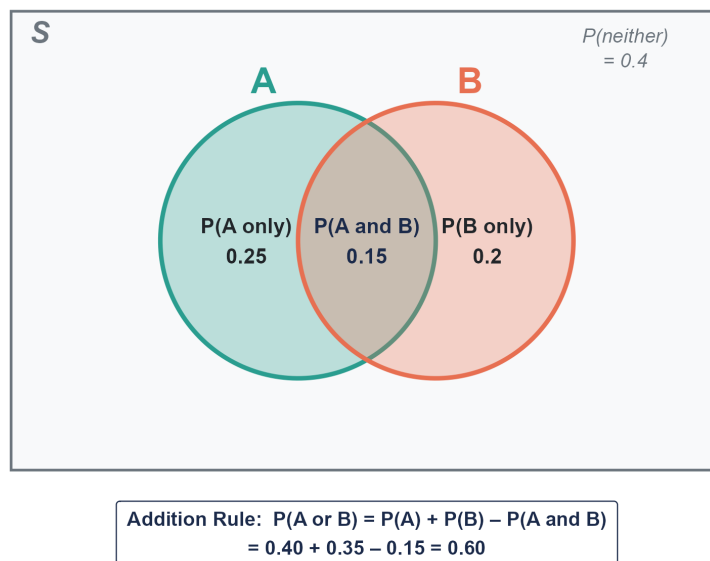
i Note

The addition rule has a visual analog. If you draw a Venn diagram with two overlapping circles, the total area covered by at least one circle equals the area of circle A plus the area of circle B minus the overlap region, which you would otherwise count twice. Venn diagrams

are more than decorations from middle school math. They are a useful way to keep track of which events overlap and which do not.

Probability Rules: The Addition Rule

$P(A \text{ or } B)$ accounts for the overlap to avoid double-counting



The overlapping region $P(A \text{ and } B)$ must be subtracted once

Figure 5.1: Venn diagram illustrating the addition rule with two overlapping events A and B. The overlap region $P(A \text{ and } B)$ must be subtracted once to avoid double-counting when computing $P(A \text{ or } B)$.

5.3.3 The Multiplication Rule and Independence

Now suppose you want to know the probability that event A and event B both occur. This is where the concept of **independence** becomes critical.

Two events are **independent** if the occurrence of one does not change the probability of the other. Flipping a coin and rolling a die are independent. Getting heads on the coin does not make any particular die outcome more or less likely.

For independent events, the probability that both occur is the product of their individual probabilities.

$$P(A \text{ and } B) = P(A) \times P(B) \quad (\text{if } A \text{ and } B \text{ are independent})$$

The probability of flipping heads and rolling a 6 is $1/2 \times 1/6 = 1/12$.

If events are **not** independent, we need the general multiplication rule, which uses conditional probability.

$$P(A \text{ and } B) = P(A) \times P(B | A)$$

We will return to this shortly, once we have defined conditional probability properly.

A word of warning. Independence is often assumed when it should not be. “What is the probability that both of my flights are delayed?” is only $P(\text{flight 1 delayed}) \times P(\text{flight 2 delayed})$ if the two delays are independent. But if your second flight is a connection from the first, or if both fly out of the same airport during the same weather system, they are not independent, and the simple multiplication gives you the wrong answer.

Example. Suppose a manufacturing process has three independent quality checks. Each check catches a defective product with probability 0.80. What is the probability that a defective product passes all three checks undetected?

Each check independently fails to catch the defect with probability $1 - 0.80 = 0.20$. Since the checks are independent, the probability of slipping past all three is

$$P(\text{all three miss}) = 0.20 \times 0.20 \times 0.20 = 0.008$$

Less than 1%. This is the logic behind redundant systems, from quality control to airplane safety. Each individual check might miss something, but stacking independent checks on top of each other drives the probability of total failure down rapidly.

Try It Online

Open the **Probability Rules Sandbox** on the companion website. It lets you set $P(A)$, $P(B)$, and $P(A \text{ and } B)$ with sliders and watch every rule in this section update at once on a Venn diagram, a proportional unit-square view, and a tree diagram. Toggle **Force indepen-**

dent to see the multiplication rule snap into place; toggle **Force mutually exclusive** to see the addition rule simplify. A small set of preset scenarios (cards, dice, weather, handedness) is included so you can check your intuition against worked examples.

5.4 Conditional Probability and Two-Way Tables

5.4.1 Conditional Probability

The **conditional probability** of A given B , written $P(A \mid B)$, is the probability that A occurs given that we already know B has occurred. The vertical bar $\mid$ is read as “given.”

The formula is

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$$

provided $P(B) > 0$. In words, to find the probability of A given B , take the probability that both happen and divide by the probability of B . You are essentially restricting your attention to the world where B has happened and asking how often A also happens in that restricted world.

Think of it this way. If I tell you that a randomly selected person from a room is wearing glasses, and I ask you the probability that person is over 60, you are no longer thinking about the entire room. You are mentally filtering down to only the glasses-wearers and asking what fraction of that subgroup is over 60. That mental filter is what conditional probability formalizes.

Let’s make this concrete with a two-way table.

5.4.2 Working with Two-Way Tables

Consider a hypothetical survey of 800 recent graduates from a mid-sized university about their employment status and their major category. Here are the results.

	Employed Full-Time	Employed Part-Time	Seeking Employment	Total
STEM	180	40	30	250
Business	150	45	25	220

	Employed Full-Time	Employed Part-Time	Seeking Employment	Total
Humanities	100	60	50	210
Social Sciences	80	25	15	120
Total	510	170	120	800

From this table, we can answer many probability questions by treating the 800 graduates as our population and computing proportions.

What is the probability that a randomly selected graduate is employed full-time?

$$P(\text{full-time}) = \frac{510}{800} = 0.6375$$

What is the probability that a randomly selected graduate is a humanities major?

$$P(\text{humanities}) = \frac{210}{800} = 0.2625$$

What is the probability that a graduate is both a STEM major and employed full-time?

$$P(\text{STEM and full-time}) = \frac{180}{800} = 0.225$$

Now for conditional probability.

Given that a graduate is a STEM major, what is the probability they are employed full-time?

We restrict our attention to the 250 STEM majors and ask how many of them are employed full-time.

$$P(\text{full-time} \mid \text{STEM}) = \frac{180}{250} = 0.72$$

Given that a graduate is employed full-time, what is the probability they are a STEM major?

Now we restrict our attention to the 510 full-time employed graduates.

$$P(\text{STEM} \mid \text{full-time}) = \frac{180}{510} \approx 0.353$$

Notice something important. $P(\text{full-time} \mid \text{STEM}) = 0.72$ is not the same as $P(\text{STEM} \mid \text{full-time}) \approx 0.353$. These are different questions with different answers. The first asks, “Among STEM majors, what fraction work full-time?” The second asks, “Among full-time workers, what fraction were STEM majors?” Confusing the two is one of the most common errors in probabilistic reasoning, and it will come back with full force when we discuss Bayes’ theorem.

To drive the point home with a more everyday example, consider this. The probability of being wet given that it is raining is very high, close to 1 if you are outside. But the probability of it raining given that you are wet is much lower. You could have just stepped out of a pool, or walked through a sprinkler, or spilled a drink. $P(\text{wet} \mid \text{rain}) \neq P(\text{rain} \mid \text{wet})$. The direction of the conditional matters.

5.4.3 Testing for Independence with Two-Way Tables

Two events are independent if knowing one occurred does not change the probability of the other. In terms of conditional probability, A and B are independent if and only if

$$P(A \mid B) = P(A)$$

Let’s test whether major category and employment status are independent in our graduate data. Is being a STEM major independent of being employed full-time?

$$P(\text{full-time}) = 510/800 = 0.6375$$

$$P(\text{full-time} \mid \text{STEM}) = 180/250 = 0.72$$

Since $0.72 \neq 0.6375$, knowing that someone is a STEM major does change the probability they are employed full-time. So these events are **not** independent.

Now check humanities majors. $P(\text{full-time} \mid \text{humanities}) = 100/210 \approx 0.476$. This is lower than the overall rate, so employment status and being a humanities major are also not independent. Major and employment status appear to be associated in this dataset.

Whether this association reflects a causal relationship (your major *causes* different employment outcomes) or is confounded by other variables (perhaps students who choose different majors also differ in other ways) is a question we cannot answer from this table alone. But probability gives us the tools to precisely describe the pattern.

i Note

When people say “correlation is not causation,” they are making exactly this point. Statistical association between two variables, whether measured by conditional probabilities, correlations, or any other metric, tells you the variables move together. It does not tell you why. Maybe A causes B . Maybe B causes A . Maybe some third factor C causes both. Maybe it's a coincidence in this particular dataset. Moving from “associated” to “causes” requires much more than a probability calculation. It requires careful study design, which we explored in Chapter 2.

5.5 Bayes' Theorem

5.5.1 The Problem of Inverse Probability

Let's return to the COVID testing scenario from the opening. The fundamental issue was that people confused two different conditional probabilities.

Sensitivity is $P(\text{positive test} \mid \text{infected})$, the probability of a positive result given that the person actually has the disease. In our example, this was 0.90.

What people wanted to know was $P(\text{infected} \mid \text{positive test})$, the probability of actually having the disease given a positive result. This is called the **positive predictive value**.

These are not the same number. Confusing them is sometimes called the **prosecutor's fallacy** or the **confusion of the inverse**, and it shows up in medicine, criminal justice, and everyday life with alarming regularity.

Bayes' theorem provides the bridge between these two conditional probabilities. The formula is

$$P(A \mid B) = \frac{P(B \mid A) \times P(A)}{P(B)}$$

In the context of diagnostic testing, this becomes

$$P(\text{infected} \mid \text{positive}) = \frac{P(\text{positive} \mid \text{infected}) \times P(\text{infected})}{P(\text{positive})}$$

The denominator, $P(\text{positive})$, is the total probability of testing positive, which includes both true positives and false positives. We can expand it using the **law of total probability**.

$$P(\text{positive}) = P(\text{positive} \mid \text{infected}) \times P(\text{infected}) + P(\text{positive} \mid \text{not infected}) \times P(\text{not infected})$$

Plugging in our COVID numbers,

$$P(\text{positive}) = 0.90 \times 0.01 + 0.05 \times 0.99 = 0.009 + 0.0495 = 0.0585$$

$$P(\text{infected} \mid \text{positive}) = \frac{0.90 \times 0.01}{0.0585} = \frac{0.009}{0.0585} \approx 0.154$$

About 15.4%, exactly what we found with our frequency approach earlier.

5.5.2 Why Natural Frequencies Work Better

If the formula above felt like a lot of abstract symbol manipulation, you are not alone. Research by Gerd Gigerenzer and his colleagues has shown that people, including physicians, are far better at reasoning about probabilities when information is presented as **natural frequencies** rather than as percentages or probabilities.

Gigerenzer et al. (2007), in their paper “Helping Doctors and Patients Make Sense of Health Statistics,” demonstrated that when doctors were given disease prevalence and test accuracy as conditional probabilities, a large fraction, sometimes a majority, got the answer wrong. But when the same information was presented as natural frequencies (“out of 10,000 people...”), accuracy improved dramatically.

We already used this approach in the opening. Let’s formalize it with a frequency table.

Step 1. Start with a convenient population size. We will use 10,000.

Step 2. Apply the base rate. 1% of 10,000 = 100 infected, 9,900 not infected.

Step 3. Apply the test performance to each group.

	Test Positive	Test Negative	Total
Infected	90	10	100
Not Infected	495	9,405	9,900
Total	585	9,415	10,000

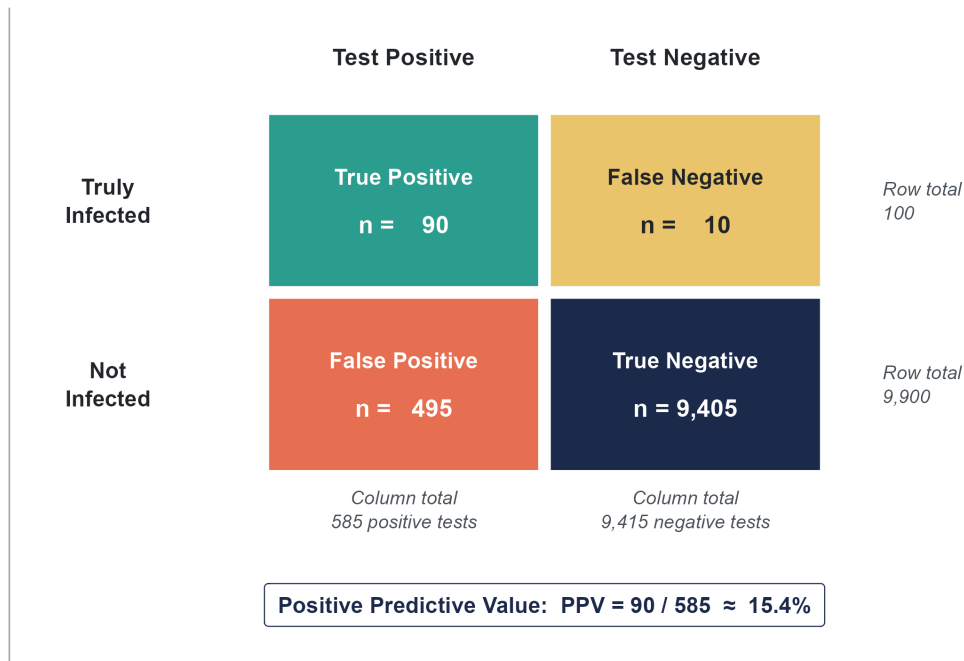
Step 4. Read the answer. Of the 585 who tested positive, 90 are truly infected.

$$P(\text{infected} \mid \text{positive}) = \frac{90}{585} \approx 0.154$$

No formulas needed. Just counting.

Why a Positive Test Doesn't Mean You're Positive

Population: 10,000 | Prevalence: 1% | Sensitivity: 90% | Specificity: 95%



When prevalence is low, the much larger 'Not Infected' group produces more false positives (495) than the 'Truly Infected' group produces true positives (90) — even though the test is 90% sensitive.

Figure 5.2: Confusion matrix for a population of 10,000 tested with 1% disease prevalence, 90% sensitivity, and 95% specificity. Of the 585 positive tests (column total below the grid), only 90 are true positives, giving a positive predictive value of 15.4%. Source: Hypothetical population computed from the stated test characteristics.

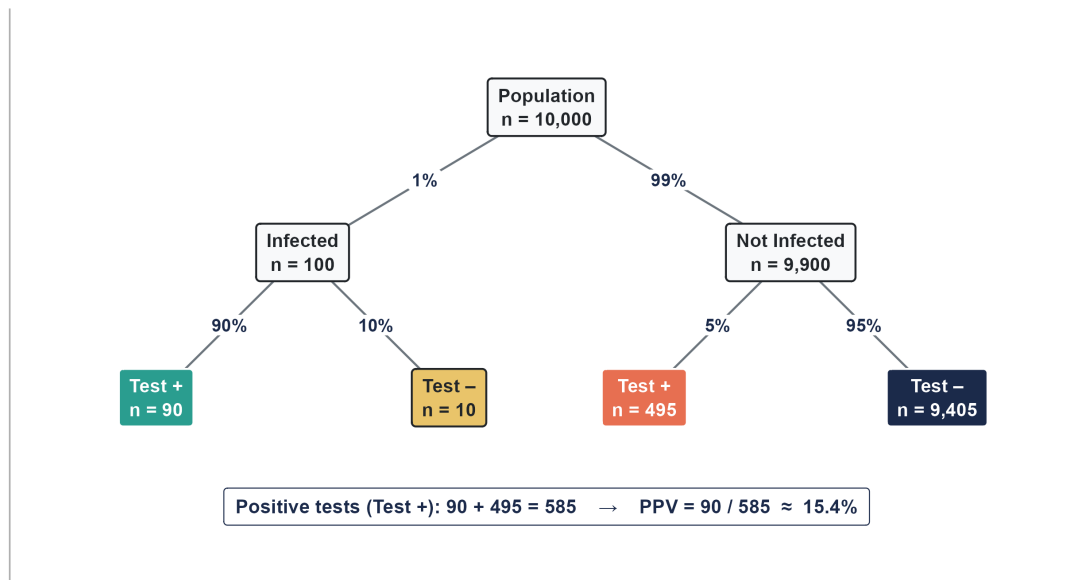
This is not a trick or a shortcut. It is mathematically identical to using Bayes' formula. The frequency table simply makes the same calculation transparent by converting abstract proportions into concrete counts that our brains handle more naturally. Whenever you face a Bayesian reasoning problem, the frequency table approach is a reliable path to the correct answer.

5.5.3 The Tree Diagram Approach

Another visual approach is the probability tree, which maps out the sequence of events and their probabilities. The first split applies the base rate (1% infected, 99% not infected), and the second split applies the test's sensitivity and specificity to each branch.

Probability Tree for COVID Testing

Population: 10,000 | Prevalence: 1% | Sensitivity: 90% | Specificity: 95%



Top branch shows the base rate; second-level branches apply the test's sensitivity and specificity to each group.

Figure 5.3: Probability tree for the COVID testing scenario. The top branch splits the population by base rate; the bottom branches apply test sensitivity (90%) and specificity (95%). The four leaves are the same four counts shown in Figure 5.2.

To find $P(\text{infected} \mid \text{positive})$, look at all the “Test +” leaves. There are $90 + 495 = 585$ positive tests. Of those, 90 come from the “Infected” branch.

$$P(\text{infected} \mid \text{positive}) = \frac{90}{90 + 495} = \frac{90}{585} \approx 0.154$$

The tree makes the structure visible. You can see that the false positive branch (495) dwarfs the true positive branch (90) because the “Not Infected” group is so much larger to begin with.

5.5.4 What Happens When the Base Rate Changes

The base rate is the engine that drives the result. Watch what happens when the prevalence changes but the test stays the same.

Scenario 1. Prevalence = 1% (our original scenario).

$$P(\text{infected} \mid \text{positive}) = 90/585 \approx 15.4\%$$

Scenario 2. Prevalence = 10%. Now 1,000 are infected and 9,000 are not.

True positives: $1,000 \times 0.90 = 900$. False positives: $9,000 \times 0.05 = 450$. Total positives: 1,350.

$$P(\text{infected} \mid \text{positive}) = 900/1,350 \approx 66.7\%$$

Scenario 3. Prevalence = 50%. Now 5,000 are infected and 5,000 are not.

True positives: $5,000 \times 0.90 = 4,500$. False positives: $5,000 \times 0.05 = 250$. Total positives: 4,750.

$$P(\text{infected} \mid \text{positive}) = 4,500/4,750 \approx 94.7\%$$

The same test. The same sensitivity. The same specificity. But the meaning of a positive result changes substantially depending on how common the disease is in the population being tested. This is why mass screening programs for rare diseases must be designed with extreme care, and why public health officials recommended against certain widespread testing strategies during COVID-19 when community prevalence was very low.

There is a practical lesson here for how testing was deployed during the pandemic. If you were a healthcare worker in a COVID ward, your exposure risk was high, the effective “base rate” for you was much higher than 1%. A positive test in that context was much more likely to be correct. If you were working from home in a rural area with very few cases, the same positive test was much more likely to be a false alarm. The test itself did not change. Your situation did. Probability is always conditional on context.

This is also why **confirmatory testing** is standard practice in medicine. A positive result on a screening test is usually followed up with a second, often more specific, test. The logic is Bayesian. Once someone tests positive on the first test, their updated probability of having the disease (the posterior probability) becomes the new base rate for interpreting the second test. If the first test raised the probability from 1% to 15%, and the second test is also positive, the probability climbs much higher. Two independent tests agreeing is far more informative than either test alone.

 Try It Online

Return to the **Base Rate Lab** with everything you now know about Bayes' theorem. Three tabs reinforce what you just read.

- **Three Views** puts the classical outcome-counting, frequentist long-run simulation, and Bayesian prior-to-posterior pictures side by side for the same problem. All three give the same PPV — they just get there by different mental routes.
- **Scenarios** lets you cycle through preset real-world tests (COVID rapid, screening mammography, HIV ELISA, workplace drug testing, airport behavioral screening, home pregnancy). Guess the PPV before you look, then compare. Low base rates are where intuition breaks most dramatically.
- **Sequential Testing** implements exactly the confirmatory-testing logic above: the posterior after test 1 becomes the prior for test 2. Watch how two independent positives can push the COVID probability from 1% to 15% and then to 77%.

 Ethics Moment: Who Pays for the Errors?

Every diagnostic test has two types of errors. A **false positive** tells a healthy person they are sick. A **false negative** tells a sick person they are healthy. No test eliminates both.

Here is the ethical tension. Different stakeholders experience these errors differently.

For the **individual** tested during COVID-19, a false positive might have meant a two-week quarantine away from family, lost wages, missed school, and considerable anxiety, all for a disease they did not have. For someone who relied on hourly wages and had no paid sick leave, the cost was beyond psychological; it was financial.

For the **community**, a false negative was arguably worse. A person told they were not infected might have continued going to work, visiting elderly relatives, and spreading a virus they did not know they carried. One false negative could seed an outbreak.

For the **employer**, the calculus depended on the industry. A hospital administrator might have preferred more false positives (better to over-quarantine staff than risk an outbreak on the ward). A warehouse manager facing a labor shortage might have felt differently.

These tradeoffs are more than technical. They are moral. When public health agencies set the threshold for what counts as a “positive” result, they are implicitly making a value judgment about which type of error is more acceptable. Making that judgment transparent, rather than hiding it behind technical language, is an ethical obligation. Probability gives us the tools to

quantify the tradeoffs. It does not tell us which tradeoff to choose. That is a human decision, and it should be made openly.

⚠️ AI Reality Check: When the Machine Ignores the Base Rate

Large language models and AI systems can be very good at many tasks, but they sometimes stumble on exactly the kind of probabilistic reasoning we have been discussing.

When researchers have posed classic Bayesian reasoning problems to large language models, the results have been sobering. Binz and Schulz (2023), in a study published in the *Proceedings of the National Academy of Sciences*, ran a battery of cognitive psychology tasks against GPT-3 and found the same base-rate neglect that has been documented in humans for half a century. Macmillan-Scott and Musolesi (2024) extended this to seven different LLMs, including GPT-4 and Claude, and reported that none of them reasoned consistently well on standard probability tasks; their failures were not even consistent in the same direction across runs. A typical failure mode looks like this. When asked, “A disease affects 1 in 1,000 people. A test has a 5% false positive rate and a 99% true positive rate. If someone tests positive, what is the probability they have the disease?”, some models confidently answered “99%” or “95%” — committing the same confusion of the inverse that trips up humans. The correct answer, by Bayes’ theorem, is closer to 2%.

The issue is subtle. These models have seen Bayes’ theorem in their training data. They can recite the formula. They can sometimes solve textbook problems that are formatted in familiar ways. But when the problem is rephrased, or when the base rate is not highlighted, or when the question is embedded in a realistic scenario, they can fall back on the same heuristic that humans use. They match the answer to whichever number “feels” like it should be the answer, usually the sensitivity or the specificity.

This matters because AI tools are increasingly used in medical, legal, and financial contexts where base rates carry real weight. If you ask an AI assistant, “My patient tested positive on a screening test with 95% sensitivity and 90% specificity. The disease prevalence is 2%. Should I be concerned?”, you need the system to walk through the Bayesian calculation, not just report the sensitivity. As of this writing, some systems do this well and some do not, and the user has no easy way to tell the difference without understanding the math themselves. The lesson is not “do not use AI for probability.” The lesson is that your understanding of base rates and conditional probability is your defense against both your own intuitive errors and the errors of any tool you rely on.

5.6 Random Variables and Expected Value

5.6.1 Random Variables

We have been talking about events, things like “rolling a 6” or “testing positive.” A **random variable** takes this a step further by assigning a numerical value to each outcome of a random process.

A random variable is typically denoted by a capital letter like X . The specific values it can take are denoted by lowercase letters.

Example. Let X be the number of heads in three coin flips. The possible values of X are 0, 1, 2, and 3.

Example. Let Y be the number of points scored by a basketball team in a game. Y could be any non-negative integer, though in practice it falls within a certain range.

Example. Let Z be the time (in minutes) a customer waits on hold before reaching a service representative. Z can take any positive real value.

A **discrete** random variable takes on a countable number of values (like the number of heads, or the number of defective items in a batch). A **continuous** random variable can take any value within an interval (like waiting time, or temperature). This chapter focuses on discrete random variables. We will tackle continuous ones in the next chapter.

Why bother with the concept of a random variable when we were doing fine with events? Because random variables let us bring the tools of algebra and calculus to bear on probability. Once we assign numbers to outcomes, we can compute averages, measure spread, plot distributions, and build mathematical models. The language of random variables is the bridge between the probability rules we have learned and the statistical methods we will develop in the coming chapters.

5.6.2 Probability Distributions

A **probability distribution** for a discrete random variable specifies the probability of each possible value. It must satisfy two conditions: every probability is between 0 and 1, and all probabilities sum to 1.

For X = number of heads in three flips of a fair coin, the distribution is

x	0	1	2	3
$P(X = x)$	1/8	3/8	3/8	1/8

Where do these come from? With three flips, there are $2^3 = 8$ equally likely outcomes (HHH, HHT, HTH, THH, HTT, THT, TTH, TTT). One outcome has 0 heads, three have exactly 1 head, three have exactly 2 heads, and one has 3 heads.

5.6.3 Expected Value

The **expected value** (or **mean**) of a discrete random variable X is the weighted average of its possible values, with the probabilities serving as weights.

$$E(X) = \sum_i x_i \cdot P(X = x_i)$$

For our coin-flip example,

$$E(X) = 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} = 0 + \frac{3}{8} + \frac{6}{8} + \frac{3}{8} = \frac{12}{8} = 1.5$$

You cannot flip 1.5 heads, of course. The expected value is not a prediction of any single outcome. It is the long-run average. If you repeated the three-flip experiment thousands of times and recorded the number of heads each time, the average of all those numbers would approach 1.5.

Example. A small insurance company offers a one-year policy that pays \$50,000 if the policyholder is hospitalized during the year. Based on actuarial data, the probability of hospitalization is 0.02. The company charges \$1,200 per policy. What is the expected profit per policy?

Let X be the company's profit on a single policy.

- If no hospitalization (probability 0.98), the company collects \$1,200 and pays nothing. Profit = \$1,200.
- If hospitalization (probability 0.02), the company collects \$1,200 and pays \$50,000. Profit = \$1,200 - \$50,000 = -\$48,800.

$$E(X) = 1,200 \times 0.98 + (-48,800) \times 0.02 = 1,176 - 976 = \$200$$

On average, the company expects to make \$200 per policy. This does not mean every policy is profitable. Some will result in large payouts. But across a large number of policies, the average profit per policy should be near \$200.

5.6.4 Variance and Standard Deviation of a Random Variable

The expected value tells us the center of a distribution, but not how spread out the values are. The **variance** of a random variable measures that spread.

$$\text{Var}(X) = E[(X - \mu)^2] = \sum_i (x_i - \mu)^2 \cdot P(X = x_i)$$

where $\mu = E(X)$.

The standard deviation is the square root of the variance, $\text{SD}(X) = \sqrt{\text{Var}(X)}$, and it has the same units as the original variable.

For the insurance example with $E(X) = 200$,

$$\begin{aligned} \text{Var}(X) &= (1,200 - 200)^2 \times 0.98 + (-48,800 - 200)^2 \times 0.02 \\ &= (1,000)^2 \times 0.98 + (-49,000)^2 \times 0.02 \\ &= 1,000,000 \times 0.98 + 2,401,000,000 \times 0.02 \\ &= 980,000 + 48,020,000 = 49,000,000 \end{aligned}$$

$$\text{SD}(X) = \sqrt{49,000,000} = \$7,000$$

The expected profit is \$200, but the standard deviation is \$7,000. This tells us that the outcome on any individual policy is highly variable, even though the long-run average is positive. This is exactly why insurance companies need a large pool of policyholders. The individual outcomes are volatile, but the *average* across many policies stabilizes around the expected value, a phenomenon we will formalize later as the law of large numbers.

5.7 The Binomial Distribution

5.7.1 When Binomial Applies

Many real-world situations share a common structure. You repeat the same basic trial multiple times, each trial has only two possible outcomes (success or failure), the probability of success stays the same from trial to trial, and the trials are independent of each other.

Some examples that fit this structure (at least approximately):

- A free-throw shooter attempts 20 shots during a game. Each shot either goes in (success) or does not (failure).
- A quality inspector examines 50 items from a production line. Each item is either defective (success, if you are counting defects) or acceptable.
- A pollster contacts 500 registered voters. Each voter either supports a particular candidate (success) or does not.
- A pharmaceutical company enrolls 200 patients in a trial. Each patient either experiences a side effect or does not.

When these four conditions are met (fixed number of trials, two outcomes per trial, constant probability of success, independence), the number of successes follows a **binomial distribution**.

A brief note on the word “success.” In the context of the binomial distribution, “success” does not mean something good happened. It simply means the outcome you are counting occurred. If you are counting defective items, a defective item is a “success.” If you are counting patients who experience an adverse reaction, an adverse reaction is a “success.” The terminology is unfortunate, but it is standard, and you will encounter it everywhere.

5.7.2 The Binomial Formula

If X is the number of successes in n independent trials, each with probability of success p , then

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

for $k = 0, 1, 2, \dots, n$.

Let’s unpack each piece.

$\binom{n}{k}$ is the **binomial coefficient**, read “n choose k.” It counts the number of ways to choose which k of the n trials will be the successes. The formula is

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

where $n! = n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1$ is the factorial of n , with $0! = 1$ by convention.

p^k is the probability of getting success on each of the k successful trials.

$(1-p)^{n-k}$ is the probability of getting failure on each of the remaining $n-k$ trials.

The logic here is straightforward once you see it. Any particular sequence of k successes and $n-k$ failures has probability $p^k(1-p)^{n-k}$, because the trials are independent and you multiply probabilities. But there are $\binom{n}{k}$ different orderings that produce exactly k successes. Since these orderings are mutually exclusive (only one can happen), we add them up, which is the same as multiplying $\binom{n}{k}$ by the probability of any single arrangement.

5.7.3 A Worked Example

Suppose a basketball player has a 75% free-throw percentage, and they attempt 8 free throws in a game. What is the probability they make exactly 6?

Here $n = 8$, $k = 6$, $p = 0.75$.

$$\binom{8}{6} = \frac{8!}{6! \cdot 2!} = \frac{8 \times 7}{2 \times 1} = 28$$

$$P(X = 6) = 28 \times (0.75)^6 \times (0.25)^2$$

$$= 28 \times 0.17798 \times 0.0625$$

$$= 28 \times 0.01112$$

$$\approx 0.3115$$

There is about a 31% chance they make exactly 6 of 8 free throws.

What about the probability they make 6 or more? That means $P(X \geq 6) = P(X = 6) + P(X = 7) + P(X = 8)$.

$$P(X = 7) = \binom{8}{7}(0.75)^7(0.25)^1 = 8 \times 0.13348 \times 0.25 \approx 0.2670$$

$$P(X = 8) = \binom{8}{8}(0.75)^8(0.25)^0 = 1 \times 0.10011 \times 1 \approx 0.1001$$

$$P(X \geq 6) \approx 0.3115 + 0.2670 + 0.1001 = 0.6786$$

There is about a 68% chance they make 6 or more of their 8 free throws.

5.7.4 Mean and Standard Deviation of the Binomial

The expected value and standard deviation of a binomial random variable have particularly simple formulas.

$$E(X) = np$$

$$SD(X) = \sqrt{np(1-p)}$$

For our free-throw shooter, $E(X) = 8 \times 0.75 = 6$ and $SD(X) = \sqrt{8 \times 0.75 \times 0.25} = \sqrt{1.5} \approx 1.22$.

On average, they would make 6 out of 8, with a standard deviation of about 1.22. So making 5 or 7 (within about one standard deviation of the mean) would be completely unsurprising, while making 2 (more than three standard deviations below the mean) would be very unusual.

5.7.5 Building Intuition

Let's use the binomial distribution to think about something more consequential. Suppose a new drug has a 40% response rate. A doctor prescribes it to 10 patients. What is the probability that none of them respond?

$$P(X = 0) = \binom{10}{0}(0.40)^0(0.60)^{10} = 1 \times 1 \times 0.006047 \approx 0.006$$

About 0.6%. Very unlikely, which is reassuring if you are the drug manufacturer. But let's turn this around. What is the probability that fewer than 3 patients respond?

$$P(X < 3) = P(X = 0) + P(X = 1) + P(X = 2)$$

$$P(X = 0) \approx 0.006$$

$$P(X = 1) = \binom{10}{1}(0.40)^1(0.60)^9 = 10 \times 0.40 \times 0.01008 \approx 0.040$$

$$P(X = 2) = \binom{10}{2}(0.40)^2(0.60)^8 = 45 \times 0.16 \times 0.01680 \approx 0.121$$

$$P(X < 3) \approx 0.006 + 0.040 + 0.121 = 0.167$$

About a 17% chance. So even with a drug that works 40% of the time, a doctor treating 10 patients has a non-trivial chance of seeing fewer than 3 responses. This is important context for clinical decision-making. A doctor who sees only 2 of 10 patients respond might lose confidence in the drug, but the binomial distribution tells us this outcome is not that unusual. Small samples produce noisy results.

This connects to one of the deepest themes in statistics. Variability in small samples does not mean something is broken. It means you need more data before drawing conclusions.

Try It Online

Open the **Binomial Explorer** on the companion website. Set n and p with the sliders and watch the PMF update live. Three things to do before you move on.

- **Distribution tab.** Reproduce the free-throw example: set $n = 8$, $p = 0.75$, and choose the highlight $P(X = k)$ with $k = 6$. The bar chart should show the same 0.3115 we computed by hand. Then switch the highlight to $P(X \geq k)$ to see the 0.6786 from the next paragraph.
- **Shape Lab tab.** Hold p fixed at 0.30 and step n through 10, 30, and 100. Watch the bars become smoother and more symmetric — a preview of the Central Limit Theorem in Chapter 6.
- **Calculator tab.** Pick any (n, p) and find the smallest k with $P(X \leq k) \geq 0.95$. That's the 95th percentile of the distribution, the same idea that will show up in confidence intervals in Chapter 7.

5.7.6 When the Binomial Does Not Apply

It is worth being clear about when the binomial model breaks down, because forcing data into the wrong model produces misleading results.

Trials are not independent. If you are drawing cards from a deck without replacement, each draw changes the composition of the remaining deck. The probability of drawing a heart on the second card depends on what happened on the first draw. For small samples from large populations, the dependence is negligible and the binomial is a reasonable approximation. But for larger samples, or for populations that are not much larger than the sample, you need a different distribution (the hypergeometric, which we will not cover in this book but which you may encounter in more advanced work).

The probability of success changes. If a salesperson's probability of closing a deal improves as they gain experience throughout the day, the constant-probability assumption is violated. If the probability of a manufacturing defect increases as a machine wears out over a shift, same problem.

More than two outcomes. If each trial can result in three or more categories (a customer chooses brand A, brand B, or brand C), the binomial does not apply directly. You would need its generalization, the multinomial distribution.

Recognizing when a model fits and when it does not is just as important as knowing how to use the model when it does fit.

5.8 Tying It Together

Let's return once more to where we started. The COVID testing scenario is ultimately a story about conditional probability, Bayes' theorem, and the power of base rates. But it also involves every concept in this chapter.

The test result is a random variable (positive or negative). The probability rules let us combine the base rate with the test characteristics. The complement rule reminds us that $P(\text{not infected}) = 1 - P(\text{infected})$. The multiplication rule connects the joint probability of being infected and testing positive to the conditional probability of testing positive given infection. And Bayes' theorem ties the whole calculation together, allowing us to flip the conditional probability from what the test tells us to what we actually want to know.

If you take away one idea from this chapter, let it be this. Probabilities are not isolated numbers. They exist in context. A 90% accuracy rate means different things depending on what you are

testing for, who you are testing, and how common the condition is. Understanding probability is not about memorizing formulas, though the formulas are useful. It is about developing the habit of asking, “What is the full picture? What other information do I need before I can interpret this number?”

That habit of mind, thinking carefully about context and conditions, is the foundation of every statistical method we will build in the rest of this book.

And it is a habit that transfers far beyond statistics. In business, in medicine, in policy, in your personal decisions, the question “what is the base rate?” is almost always the right starting point. Before you get excited about a positive result, ask how common the thing you are looking for is in the first place. Before you panic about a risk, ask how frequently that risk materializes in the relevant population. Before you trust a prediction, ask what the prior probability was before the evidence arrived.

Probability is more than a mathematical tool. It is a way of thinking carefully about uncertainty. And in a world full of uncertainty, thinking carefully is never wasted effort.

5.9 Looking Ahead

Everything in this chapter has dealt with discrete outcomes: events you can list and count, like coin flips, test results, or the number of defective items in a batch. But many real-world quantities, such as height, weight, income, or time, are continuous. For a continuous random variable, the probability of observing any single exact value is zero; instead, we calculate probabilities over ranges. That shift from counting outcomes to measuring areas under a curve is the subject of the next chapter, where we meet the most important continuous distribution in statistics: the normal distribution. Along the way, we will encounter the Empirical Rule, z-scores, and ultimately the Central Limit Theorem, the result that makes nearly all of statistical inference possible.

5.10 Further Reading and References

The following works are cited in this chapter or provide valuable additional context for the ideas covered here.

On Bayes’ theorem and natural frequencies: Gigerenzer, G. (2002). *Calculated risks: How to know when numbers deceive you*. Simon & Schuster. (An accessible treatment of how natural

frequency representations, like the 10,000-person table used in this chapter, help people reason about conditional probability far more accurately than traditional probability notation.)

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, 102(4), 684–704. (The research paper demonstrating that physicians and laypeople alike perform dramatically better on Bayesian reasoning tasks when information is presented as natural frequencies rather than conditional probabilities.)

On the base rate fallacy: Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80(4), 237–251. (The foundational paper documenting how people systematically ignore base rates when making probabilistic judgments.)

On COVID-19 testing and probabilistic reasoning: Watson, J., Whiting, P. F., & Brush, J. E. (2020). Interpreting a COVID-19 test result. *BMJ*, 369, m1808. (A clear explanation of positive predictive value, sensitivity, specificity, and prevalence applied to COVID-19 testing, closely paralleling the scenario that opens this chapter.)

For further reading on probability: Blitzstein, J. K., & Hwang, J. (2019). *Introduction to probability* (2nd ed.). CRC Press. (A thorough, example-driven probability textbook that develops the concepts introduced in this chapter at much greater depth, with applications across disciplines.)

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, 8(2), 53–96. (A comprehensive review of how statistical illiteracy among both physicians and patients leads to poor health decisions, extending the natural-frequencies approach introduced in this chapter.)

On large language models and probabilistic reasoning: Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6), e2218523120. (Documents that GPT-3 reproduces several classic human reasoning biases, including base-rate neglect, on standard cognitive psychology tasks.)

Macmillan-Scott, O., & Musolesi, M. (2024). (Ir)rationality and cognitive biases in large language models. *Royal Society Open Science*, 11(6), 240255. (Tests seven large language models on classic probability and reasoning problems and finds that none reason consistently well, with failures that vary even across repeated runs of the same prompt.)

5.11 Key Terms

- **Sample space** (S): The set of all possible outcomes of a random process.
- **Event**: A subset of the sample space, representing one or more outcomes of interest.
- **Probability**: A number between 0 and 1 that measures how likely an event is to occur.
- **Complement**: The complement of event A , written A^c , is the event that A does not occur. $P(A^c) = 1 - P(A)$.
- **Mutually exclusive events**: Events that cannot occur at the same time. $P(A \text{ and } B) = 0$.
- **Addition rule**: $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$.
- **Independent events**: Events where the occurrence of one does not affect the probability of the other. $P(A \text{ and } B) = P(A) \times P(B)$.
- **Conditional probability**: The probability of A given that B has occurred, $P(A \mid B) = P(A \text{ and } B)/P(B)$.
- **Bayes' theorem**: $P(A \mid B) = P(B \mid A) \times P(A)/P(B)$. Allows us to update probabilities in light of new evidence.
- **Base rate**: The overall probability of a condition in the population, before any test or evidence is considered.
- **Sensitivity**: $P(\text{positive test} \mid \text{disease present})$. The probability that a test correctly identifies someone who has the condition.
- **Specificity**: $P(\text{negative test} \mid \text{disease absent})$. The probability that a test correctly identifies someone who does not have the condition.
- **Positive predictive value**: $P(\text{disease present} \mid \text{positive test})$. The probability that someone with a positive test actually has the condition.
- **False positive**: A test result that says the condition is present when it is not.
- **False negative**: A test result that says the condition is absent when it is present.
- **Natural frequencies**: Presenting probability information as counts in a hypothetical population (e.g., “out of 10,000 people”) rather than as percentages. Tends to improve human reasoning.
- **Random variable**: A numerical quantity whose value is determined by a random process.
- **Probability distribution**: A specification of all possible values of a random variable and their associated probabilities.
- **Expected value**: The long-run average of a random variable, calculated as $E(X) = \sum x_i \cdot P(X = x_i)$.
- **Variance**: A measure of how spread out the values of a random variable are around the expected value.

- **Binomial distribution:** The distribution of the number of successes in n independent trials, each with success probability p . $P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$.
- **Binomial coefficient:** $\binom{n}{k} = n! / [k!(n - k)!]$. The number of ways to choose k items from n .
- **Law of total probability:** $P(B) = P(B | A) \times P(A) + P(B | A^c) \times P(A^c)$. Allows us to compute the overall probability of B by considering all the ways B can occur.
- **Confusion of the inverse:** The error of treating $P(A | B)$ as if it were $P(B | A)$.
- **Continuous random variable:** A random variable that can take any value within an interval, such as height or temperature. Covered in detail in Chapter 6.
- **Discrete random variable:** A random variable that takes on a countable number of values, such as the number of heads in ten coin flips.
- **Prosecutor's fallacy:** The error of treating $P(A | B)$ as if it were $P(B | A)$, particularly dangerous in legal and medical contexts.

5.12 Exercises

5.12.1 Check Your Understanding

These questions test whether you have grasped the core concepts. You should be able to answer them without extensive calculation.

1. A weather app says there is a 30% chance of rain tomorrow. What is the probability it does not rain?
2. Are the events “a student is left-handed” and “a student is right-handed” mutually exclusive? Are they exhaustive (do they cover all possibilities)? Explain.
3. A bag contains 5 red marbles and 3 blue marbles. You draw one marble. Are the events “draw a red marble” and “draw a blue marble” independent? Are they mutually exclusive? Explain the difference between these two concepts.
4. In your own words, explain why $P(\text{disease} | \text{positive test})$ is not the same as $P(\text{positive test} | \text{disease})$. Give a non-medical example that illustrates the same idea.
5. A test has 99% sensitivity and 95% specificity. The disease prevalence is 0.1%. Without doing a formal calculation, explain whether you would expect the positive predictive value to be high or low, and why.
6. Suppose $P(A) = 0.4$, $P(B) = 0.5$, and $P(A \text{ and } B) = 0.2$. Are A and B independent? How do you know?

7. What does it mean to say that two events are independent? Give an example of two events that are independent and two events that are not.
8. Why does the expected value of a random variable not have to be a value the variable can actually take? Illustrate with an example.
9. List the four conditions that must be satisfied for a situation to follow a binomial distribution. For each condition, give an example of a situation where that particular condition would be violated.
10. A friend says, “My coin came up heads five times in a row, so it is due for tails.” What concept from this chapter does your friend misunderstand? Explain.

5.12.2 Apply It

These problems require calculations. Show your work and interpret your results in context.

1. A company’s records show that 60% of its employees have a college degree, 45% have more than five years of experience, and 30% have both. What is the probability that a randomly selected employee has a college degree or more than five years of experience (or both)?
2. At a large hospital, 8% of patients are admitted to the ICU. Among ICU patients, 20% require mechanical ventilation. Among non-ICU patients, 1% require mechanical ventilation. What is the probability that a randomly selected patient requires mechanical ventilation?
3. Using the hospital data from problem 2, if a patient is on mechanical ventilation, what is the probability they are in the ICU? Use Bayes’ theorem or a frequency table approach.
4. A rapid strep test has a sensitivity of 86% and a specificity of 95%. In a pediatric clinic during winter, the prevalence of strep throat among children presenting with sore throat is about 30%. A child tests positive. What is the probability they actually have strep throat? Construct a frequency table using a hypothetical population of 1,000 children to find your answer.
5. A quality control process involves inspecting 15 items from each production batch. Historical data shows a 3% defect rate. Let X be the number of defective items in a sample of 15.
 - (a) What is the probability that the sample contains no defective items?
 - (b) What is the probability that the sample contains exactly one defective item?
 - (c) What is the probability that the sample contains two or more defective items?
 - (d) What are the expected value and standard deviation of X ?
6. Two independent fire alarms are installed in a building. During a fire, alarm A activates with probability 0.95 and alarm B activates with probability 0.90.

- (a) What is the probability both alarms activate during a fire?
- (b) What is the probability neither alarm activates? (This is the scenario the building manager worries about most.)
- (c) What is the probability at least one alarm activates?
- 7.** A genetic screening test for a rare condition has the following characteristics. Sensitivity is 99.5% and specificity is 99%. The condition affects 1 in 5,000 newborns. In a city with 20,000 births per year, how many positive test results would you expect? How many of those would be true positives? Calculate the positive predictive value.
- 8.** A small business owner sells handmade candles at a farmer's market. On any given Saturday, the probability of selling 0 candles is 0.05, selling 1-5 candles is 0.20, selling 6-10 candles is 0.40, selling 11-15 candles is 0.25, and selling more than 15 candles is 0.10. The owner makes a profit of \$5 per candle but pays \$30 to rent the booth. Using the midpoints of each range (0, 3, 8, 13, 18), find the expected number of candles sold and the expected profit.
- 9.** A recent survey found that 72% of adults in a city support expanding public transit. You randomly sample 12 adults from this city.
- (a) What is the probability that exactly 9 support the expansion?
- (b) What is the probability that 10 or more support the expansion?
- (c) What is the expected number of supporters in your sample?
- (d) Would it be unusual to find that only 5 of the 12 support the expansion? Justify your answer using the mean and standard deviation.
- 10.** Two events A and B satisfy $P(A) = 0.3$, $P(B | A) = 0.6$, and $P(B | A^c) = 0.2$.
- (a) Find $P(A \text{ and } B)$.
- (b) Find $P(B)$ using the law of total probability.
- (c) Find $P(A | B)$ using Bayes' theorem.
- (d) Are A and B independent? Explain.

The following three exercises use the covid-testing.csv dataset available on the companion website. This is a simulated dataset mathematically modeled to reflect real-world patterns from published sensitivity and specificity estimates for diagnostic tests during the COVID-19 pandemic. No real patients are represented. The dataset contains 2,000 observations with variables including patient_id, truly_infected (Yes/No), test_result (Positive/Negative), age_group (18–29, 30–44, 45–59, 60–74, 75+), and risk_category (Low/Medium/High).

- 11.** Using the covid-testing.csv dataset, construct a 2×2 frequency table of truly_infected (rows) versus test_result (columns). From this table, calculate the sensitivity (true positive rate)

and specificity (true negative rate) of the test. Then calculate the positive predictive value: among all patients who tested positive, what proportion were truly infected? Compare this value to the sensitivity and explain why they differ.

12. Compute the prevalence of infection (proportion with `truly_infected = Yes`) separately for each `age_group`. Display these five proportions in a table. Which age group has the highest prevalence? Now compute the positive predictive value for each age group separately (among those who tested positive within each age group, what fraction were truly infected?). How does PPV change across age groups, and why? Connect your answer to Bayes' theorem.

13. Restrict your analysis to the `High risk` category only. Within this subgroup, construct a new 2×2 frequency table of infection status versus test result. Calculate the prevalence, sensitivity, specificity, and positive predictive value for this subgroup. Compare these values to the overall values you computed in Problem 11. Does the test perform differently in a high-risk population? Explain why or why not, using the concepts of base rate and Bayes' theorem.

5.12.3 Think Deeper

These questions ask you to reflect on the concepts in broader context. There is not always a single correct answer.

1. During the early months of COVID-19, many workplaces required employees to test negative before returning to work. Using what you have learned about base rates and predictive values, discuss the strengths and limitations of this policy. Under what conditions (in terms of prevalence, test sensitivity, and specificity) would such a policy be most effective? Under what conditions might it cause more harm than good?

2. In criminal trials, forensic evidence is sometimes presented with statements like “the probability of finding this DNA match if the defendant were innocent is 1 in 10 million.” Explain why this statement does not mean “the probability that the defendant is innocent is 1 in 10 million.” What additional information would a juror need to properly evaluate this evidence? How does this connect to Bayes' theorem?

3. Insurance companies use probability calculations to set premiums, as illustrated in the expected value example in this chapter. Some critics argue that using demographic data (such as age, gender, or zip code) to set different premiums for different groups amounts to discrimination, even if the probabilities are accurate. Others argue that charging everyone the same rate unfairly forces low-risk groups to subsidize high-risk groups. Where do you stand on this issue? What role should

probability and statistical risk assessment play in decisions that affect individuals? Are there cases where actuarially accurate pricing should be overridden by other considerations?

4. Consider two screening tests for the same disease. Test A has 99% sensitivity and 90% specificity. Test B has 90% sensitivity and 99% specificity. The disease prevalence is 2%.

- (a) Calculate the positive predictive value and negative predictive value for each test.
- (b) If this disease is fatal when undetected, which test would you prefer for screening? Why?
- (c) If the treatment for this disease has severe side effects, which test would you prefer? Why?
- (d) What does this exercise reveal about the claim that one test is “better” than another?

5. Gigerenzer and colleagues have shown that presenting probability information as natural frequencies substantially improves people’s ability to reason about Bayesian problems. Yet most medical test results are still communicated to patients as percentages (“this test has a 95% accuracy rate”). Why do you think the medical establishment has been slow to adopt frequency-based communication? What are the potential consequences of this communication gap? If you were designing a patient-facing health app that reported test results, how would you present the information?

Chapter 6

The Normal Distribution and the Central Limit Theorem

6.1 Every Baby Gets Weighed

Within minutes of being born, before most newborns have even opened their eyes, someone places them on a scale. The number that appears, measured in grams, gets recorded in a chart, compared against a threshold, and sometimes used to make fast decisions about that infant's care.

A full-term baby in the United States weighs, on average, about 3,400 grams, roughly 7.5 pounds. Most babies cluster near that number. Some weigh a bit more, some a bit less, and as you move further from the average in either direction, you find fewer and fewer babies. A 2,200-gram baby is uncommon. A 4,800-gram baby is uncommon. A baby under 1,500 grams is rare and almost certainly in serious medical trouble.

If you were to plot the birth weights of thousands of full-term newborns on a histogram, the shape would be unmistakable. It would rise smoothly to a peak near the center, then fall off symmetrically on both sides, forming the shape that statisticians call the **normal distribution** and everyone else calls the bell curve.

Hospitals do not use this shape merely as a curiosity. They use it to set clinical thresholds. A birth weight below 2,500 grams is classified as “low birth weight.” Below 1,500 grams is “very low birth weight.” These cutoffs are not arbitrary round numbers picked for convenience. They correspond roughly to positions on the normal distribution where outcomes begin to worsen in measurable,

predictable ways. Babies in the tails of the distribution, those unusually small or unusually large, face higher risks of complications, and the further into the tail they fall, the higher the risk.

The pediatrician examining a newborn is, in a very real sense, asking a statistical question. Given what we know about the distribution of birth weights, how unusual is this baby's weight? And if it is unusual, how concerned should we be?

That question, “how unusual is this value given the distribution?”, is the question this entire chapter is built to answer. The normal distribution gives us a framework for quantifying “unusual.” The Central Limit Theorem, which we will reach by the end of the chapter, explains why this framework works far more broadly than you might expect.

Birth weights are also a good reminder that statistics is never just about numbers on a page. Behind every data point is a baby, a family, a set of circumstances. The statistical tools we develop here are means to an end, and that end is better decisions, better care, and better understanding of the world.

6.2 The Normal Distribution

6.2.1 Shape, Center, and Spread

The normal distribution is a continuous probability distribution defined by two parameters. The mean, μ (the Greek letter mu), determines where the center of the bell sits on the number line. The standard deviation, σ (the Greek letter sigma), determines how wide or narrow the bell is.

A normal distribution with $\mu = 100$ and $\sigma = 15$ is centered at 100 and has most of its values falling within about 15 units of the center on either side. A normal distribution with $\mu = 100$ and $\sigma = 5$ is centered at the same place but is much more tightly concentrated. A normal distribution with $\mu = 0$ and $\sigma = 1$ is centered at zero and is the version we will use most often as a reference. That particular version has a name, the **standard normal distribution**, and we will return to it shortly.

Every normal distribution, regardless of its mean or standard deviation, shares the same fundamental shape. It is

- **Symmetric** around the mean. The left half is a mirror image of the right half.
- **Unimodal**, meaning it has a single peak, which occurs at the mean.
- **Asymptotic**, meaning the tails extend infinitely in both directions, getting ever closer to the horizontal axis but never quite touching it.

That last property sometimes bothers students. If the tails go on forever, doesn't that mean a normal distribution assigns some probability to absurd values, like a negative birth weight? Technically, yes. But the probability in the far tails is so vanishingly small that it is practically zero. A normal model with $\mu = 3400$ grams and $\sigma = 500$ grams assigns a probability of less than one in a billion to a negative birth weight. The model is not perfect, but it is useful, and "useful but imperfect" is the description of every statistical model ever built.

The statistician George Box captured this idea in a now-famous line. "Essentially, all models are wrong, but some are useful," he wrote in 1987 (a shorter version of the same idea had appeared in his 1976 paper *Science and Statistics*). The normal distribution is wrong as a literal description of birth weights, test scores, or anything else. But as an approximation, it is useful enough to support good decisions and precise enough to ground serious scientific research.

The normal distribution turns up in a wide range of places. Heights of adult humans in a given demographic group follow it closely. So do repeated measurements of the same physical quantity, like the diameter of a machined part or the temperature reading from a well-calibrated thermometer. IQ scores are designed to be normally distributed, with a mean of 100 and a standard deviation of 15, because the test is constructed to produce that shape. Blood pressure readings, SAT scores, the weights of apples from a single orchard, and the amount of cereal a machine dispenses into a box all approximate the bell curve with varying degrees of fidelity.

Why does this one shape keep appearing? Part of the answer is mathematical, and we will encounter it formally when we reach the Central Limit Theorem later in this chapter. The informal version is that when a quantity is influenced by many small, independent factors, none of which dominates the others, the resulting distribution tends to be approximately normal. An adult's height is the product of dozens of genetic factors, nutritional influences, and environmental conditions, each contributing a little, none contributing overwhelmingly. The amount of cereal dispensed into a box depends on the speed of the conveyor, the density of the cereal, minor vibrations of the machine, and air currents in the factory. Add up enough small, independent contributions and the bell curve emerges, almost as if by invitation.

6.2.2 The Density Curve

The normal distribution is described by a **probability density function** (PDF). For a normal distribution with mean μ and standard deviation σ , the density function is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

You do not need to memorize this formula. You certainly do not need to compute it by hand. But it is worth pausing to notice a few things about it.

First, the formula depends on only two quantities, μ and σ . Once you know the mean and the standard deviation, you know everything there is to know about a normal distribution. That is an elegant feature. Many real-world phenomena are approximately characterized by just those two numbers.

Second, the function produces the *height* of the curve at any given point x . The height itself is not a probability. For continuous distributions, probabilities correspond to *areas* under the curve. The probability that a randomly selected value falls between two points a and b is the area under the density curve from a to b . The total area under the entire curve equals 1, which makes sense because the probability that a value falls *somewhere* is 100%.

Third, the expression $(x - \mu)/\sigma$ appears inside the exponent. This quantity, the distance of x from the mean measured in standard deviation units, will turn out to be important enough to earn its own name.

6.3 Z-Scores

6.3.1 Standardizing Values

Suppose a baby is born weighing 2,900 grams. Is that weight concerning? The answer depends on context, specifically on what the typical weight is and how much variation there is. If the average birth weight is 3,400 grams with a standard deviation of 500 grams, then 2,900 grams is one standard deviation below the mean. Not typical, but not alarming.

The **z-score** (also called the standard score) converts any value from a normal distribution into the number of standard deviations it falls above or below the mean.

$$z = \frac{x - \mu}{\sigma}$$

For our 2,900-gram baby, with $\mu = 3400$ and $\sigma = 500$:

$$z = \frac{2900 - 3400}{500} = \frac{-500}{500} = -1.0$$

A z-score of -1.0 means the value is one standard deviation below the mean. A z-score of $+2.0$ means two standard deviations above. A z-score of 0 means the value is exactly at the mean.

6.3.2 Why Z-Scores Matter

Z-scores let you compare values across different scales. Imagine you want to compare two students' performances. Akira scored 680 on the SAT Math section, where the mean is 528 and the standard deviation is 120. Sasha scored 29 on the ACT Math section, where the mean is 20.7 and the standard deviation is 5.4.

Who performed better relative to their peers?

For Akira:

$$z = \frac{680 - 528}{120} = \frac{152}{120} \approx 1.27$$

For Sasha:

$$z = \frac{29 - 20.7}{5.4} = \frac{8.3}{5.4} \approx 1.54$$

Sasha's z-score is the higher of the two, meaning that score sits further above the mean of its test's distribution, measured in standard deviations. By this comparison, Sasha's relative performance was stronger.

Z-scores make the incomparable comparable. Grams, test scores, temperatures, blood pressure readings, it does not matter what the original units are. Once you convert to z-scores, everything is on the same scale, measured in standard deviations from the mean.

This is why z-scores show up everywhere in applied work. Researchers compare treatment effects across different outcome measures by converting to z-scores. Coaches evaluate athletes across different events using z-scores. Economists compare economic indicators with different units and scales by standardizing them. Whenever someone says "after adjusting for the scale of measurement," they are often talking about something equivalent to a z-score transformation.

Z-scores also provide the foundation for many everyday decisions that people encounter without thinking about the underlying statistics. When a pediatrician plots a toddler's weight on a growth chart and says "your child is at the 30th percentile," that percentile came from a z-score calculation. The child's weight was compared to a reference population, standardized, and then converted to a

percentile. When a credit bureau assigns you a credit score, the underlying model uses standardized variables, each individual factor in your credit history is compared to the distribution of that factor across millions of consumers. Even the draft picks in professional sports rely on z-score logic. A scout comparing a quarterback's arm strength to a wide receiver's speed needs a common scale, and standardization provides one.

Note

Z-scores also help you spot data entry errors. If you are looking at a dataset of adult heights and you see a z-score of 7.2, something is almost certainly wrong. A value more than four or five standard deviations from the mean is so improbable under a normal distribution that a typo or measurement error is a far more likely explanation than an actual extreme observation.

Ethics Moment

Clinical thresholds derived from the normal distribution shape real lives, and the reference population behind the distribution matters. Pediatric growth charts, for example, are built from data on specific populations. For decades, the primary U.S. growth charts were based largely on formula-fed, predominantly white infants. A breastfed baby from a different background whose growth pattern was perfectly healthy could be flagged as “below the 5th percentile,” triggering unnecessary interventions, parental anxiety, and supplemental feeding that undermined breastfeeding.

The World Health Organization addressed this in 2006 by releasing growth standards based on breastfed children from six countries across five continents. The new charts changed which children were classified as underweight or overweight, with real consequences for clinical decisions and public health policy.

The broader lesson applies wherever normal distribution thresholds are used. Standardized test score cutoffs, insurance risk categories, diagnostic thresholds, all of them depend on a reference population. When that reference population does not represent the people being evaluated, the thresholds can systematically misclassify entire groups. A z-score is only as fair as the distribution it was computed from. Asking “whose data defined this distribution?” is not a technical question. It is an ethical one.

6.3.3 From Z-Scores to the Standard Normal

When you compute a z-score, you are transforming a value from an $N(\mu, \sigma)$ distribution into the corresponding value on the $N(0, 1)$ distribution, the standard normal. Every normal distribution, after z-score transformation, becomes the standard normal. This is powerful because it means we only need one table, one set of calculations, and one reference distribution to handle all normal distributions.

6.4 The Empirical Rule

Before we get into precise probability calculations, there is a useful approximation that you should commit to memory. For any normal distribution, approximately

- **68%** of values fall within 1 standard deviation of the mean (between $\mu - \sigma$ and $\mu + \sigma$)
- **95%** of values fall within 2 standard deviations of the mean (between $\mu - 2\sigma$ and $\mu + 2\sigma$)
- **99.7%** of values fall within 3 standard deviations of the mean (between $\mu - 3\sigma$ and $\mu + 3\sigma$)

This is called the **Empirical Rule**, sometimes the 68-95-99.7 Rule.

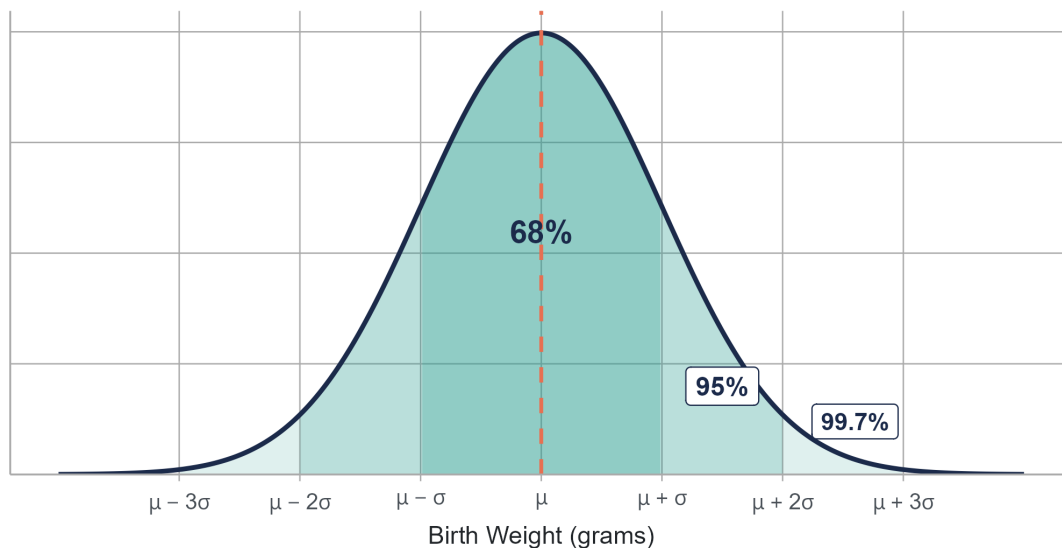
For birth weights with $\mu = 3400$ grams and $\sigma = 500$ grams, the Empirical Rule tells us

- About 68% of babies weigh between 2,900 and 3,900 grams
- About 95% weigh between 2,400 and 4,400 grams
- About 99.7% weigh between 1,900 and 4,900 grams

That last number means roughly 3 in every 1,000 full-term babies fall outside the range of 1,900 to 4,900 grams. These are the babies in the extreme tails of the distribution, the ones who are most likely to need specialized care.

The Normal Distribution and the Empirical Rule

Birth weights: mean = 3,400g, standard deviation = 500g



Approximately 68% of values fall within 1 SD, 95% within 2 SD, 99.7% within 3 SD

Figure 6.1: The normal distribution for birth weights (mean = 3,400g, SD = 500g) with shaded regions showing the Empirical Rule. Approximately 68% of values fall within one standard deviation, 95% within two, and 99.7% within three standard deviations of the mean.

The Empirical Rule is not a precise calculation tool. It is a quick mental check, a way to develop intuition about what values are common, uncommon, and rare under a normal distribution. When someone quotes a number to you and you want to know whether it is reasonable, the Empirical Rule gives you an instant frame of reference.

The more precise number for “within 2 standard deviations” is actually 95.44%, and the shortcut uses 2 rather than the exact value of 1.96. For quick estimates, those differences do not matter. For precise calculations, we turn to z-tables or software.

The Empirical Rule also serves as a useful warning system for when the normal model might not be appropriate. If you have a dataset and you find that 50% of values fall within one standard deviation of the mean instead of the expected 68%, or that 10% of values fall beyond two standard deviations instead of the expected 5%, those discrepancies are a hint that your data may not be well-described by a normal distribution. Financial return data, for instance, routinely violate the Empirical Rule in the tails. Stock market returns experience extreme moves (three, four, or five standard deviations

from the mean) far more often than a normal distribution would predict. This mismatch between the normal model and reality contributed to the 2008 financial crisis, when risk models built on the assumption of normality dramatically underestimated the probability of catastrophic losses. The models said certain events were virtually impossible. The market disagreed.

6.5 Normal Probability Calculations

6.5.1 Using Z-Tables

A **z-table** (also called the standard normal table) gives you the cumulative probability to the left of any z-score. That is, for a given z-value, the table tells you $P(Z \leq z)$, the probability that a standard normal variable takes a value less than or equal to z .

For example, looking up $z = -1.00$ in a z-table gives approximately 0.1587. This means about 15.87% of values in a standard normal distribution fall below $z = -1.00$.

Let us return to birth weights. What proportion of full-term newborns weigh less than 2,500 grams (the clinical threshold for low birth weight)?

Step 1. Convert to a z-score.

$$z = \frac{2500 - 3400}{500} = \frac{-900}{500} = -1.80$$

Step 2. Look up $z = -1.80$ in the z-table.

$$P(Z \leq -1.80) \approx 0.0359$$

Step 3. Interpret.

About 3.59% of full-term newborns are expected to have a birth weight below 2,500 grams. Out of every 1,000 full-term births, roughly 36 would be classified as low birth weight based on this model.

6.5.2 Common Probability Questions

Most normal probability questions fall into a few patterns. Here is how to handle each.

“Less than” questions. What is $P(X < a)$? Convert a to a z-score, look it up in the table. The table gives you the answer directly.

“Greater than” questions. What is $P(X > a)$? Convert a to a z-score, look it up in the table, then subtract from 1. Because the total area is 1, $P(X > a) = 1 - P(X \leq a)$.

“Between” questions. What is $P(a < X < b)$? Convert both a and b to z-scores, look up both, and subtract. $P(a < X < b) = P(Z < z_b) - P(Z < z_a)$.

Example. What proportion of full-term newborns weigh between 3,000 and 4,000 grams?

$$z_1 = \frac{3000 - 3400}{500} = -0.80 \quad \Rightarrow \quad P(Z < -0.80) \approx 0.2119$$

$$z_2 = \frac{4000 - 3400}{500} = 1.20 \quad \Rightarrow \quad P(Z < 1.20) \approx 0.8849$$

$$P(3000 < X < 4000) = 0.8849 - 0.2119 = 0.6730$$

About 67.3% of full-term newborns weigh between 3,000 and 4,000 grams.

6.5.3 Finding Values from Probabilities (Working Backward)

Sometimes the question runs in reverse. Instead of “what proportion falls below this value?”, the question is “what value separates the bottom 10%?”

Example. Below what weight do the lightest 10% of full-term newborns fall?

Step 1. Find the z-score that has 10% of the area to its left. Looking in the z-table for the value closest to 0.1000, we find $z \approx -1.28$.

Step 2. Convert back to the original scale.

$$x = \mu + z \cdot \sigma = 3400 + (-1.28)(500) = 3400 - 640 = 2760 \text{ grams}$$

About 10% of full-term newborns weigh less than 2,760 grams.

This type of calculation is called finding a **percentile** or a **quantile**. The value 2,760 grams is the 10th percentile of the birth weight distribution under this model.

Percentile calculations show up constantly in health care, education, and human resources. When a pediatrician tells a parent that their child is “in the 40th percentile for height,” they are performing exactly this kind of calculation. The child’s height corresponds to a z-score, and that z-score maps to a cumulative probability. The 40th percentile means 40% of children the same age are shorter and 60% are taller. It does not mean the child is short in any concerning way. It means the child is modestly below average, which is perfectly typical.

Try It Online

Open the **Normal Curve Explorer** on the companion website. Set μ and σ with the sliders, pick a region (less than a , greater than a , between a and b , outside $[a, b]$, or within $\pm k$ standard deviations of μ), and watch the shaded area and the probability update live.

Three things to try before you move on.

- **Probability tab.** Reproduce the low-birth-weight calculation: load the *Newborn birth weights* preset, set the region to “ $X < a$ ” with $a = 2,500$. The shaded area should give the same 3.6% we found by hand.
- **Standardize tab.** With the same setup, look at the standard normal panel on the right. The cutoff $a = 2,500$ in the original-units plot maps to $z = -1.80$ in the standardized plot, and the shaded area is identical. That is the whole reason the z-table works.
- **Empirical Rule tab.** Switch to the *Standard normal* preset. The exact areas — 68.27%, 95.45%, 99.73% — are the precise numbers behind the 68–95–99.7 shorthand.

6.6 Assessing Normality

The normal distribution is a model, and like all models, it fits some data well and other data poorly. Before using normal distribution calculations, it is worth checking whether your data are at least approximately normal. Two common approaches are the QQ plot and the Shapiro-Wilk test.

6.6.1 QQ Plots

A **quantile-quantile plot** (QQ plot) graphs the observed data values against the values you would expect if the data came from a normal distribution. If the data are normally distributed, the points fall approximately along a straight diagonal line. Departures from that line reveal *how* the data deviate from normality, and the shape of the departure tells you what kind of non-normality you are looking at.

The five patterns in Figure 6.2 cover most of what you will encounter in practice.

- **Points close to the line.** The data are approximately normal. Proceed with normal-based calculations.
- **A “smile” (both ends curving above the line).** The data are right-skewed. The right tail extends further than a normal distribution would predict, and the left tail is shorter — both effects pull points above the diagonal.
- **A “frown” (both ends curving below the line).** The data are left-skewed. Mirror image of the previous case.
- **An S-shape (lower-left below the line, upper-right above the line).** The data are **heavy-tailed**. Extreme values are more common than a normal distribution would predict in *both* directions. Financial returns and reaction times often look like this.
- **A reverse S-shape (lower-left above the line, upper-right below the line).** The data are **light-tailed**. Extreme values are *rarer* than a normal distribution would predict. Bounded measurements like exam scores or proportions often look like this.
- **A few points deviating only at the extremes.** Common and usually not a serious problem, especially with larger samples. A few outliers do not invalidate the normal model for the bulk of the data.

Five QQ-plot patterns to recognize

Sample quantiles (y-axis) plotted against theoretical normal quantiles (x-axis). The orange line

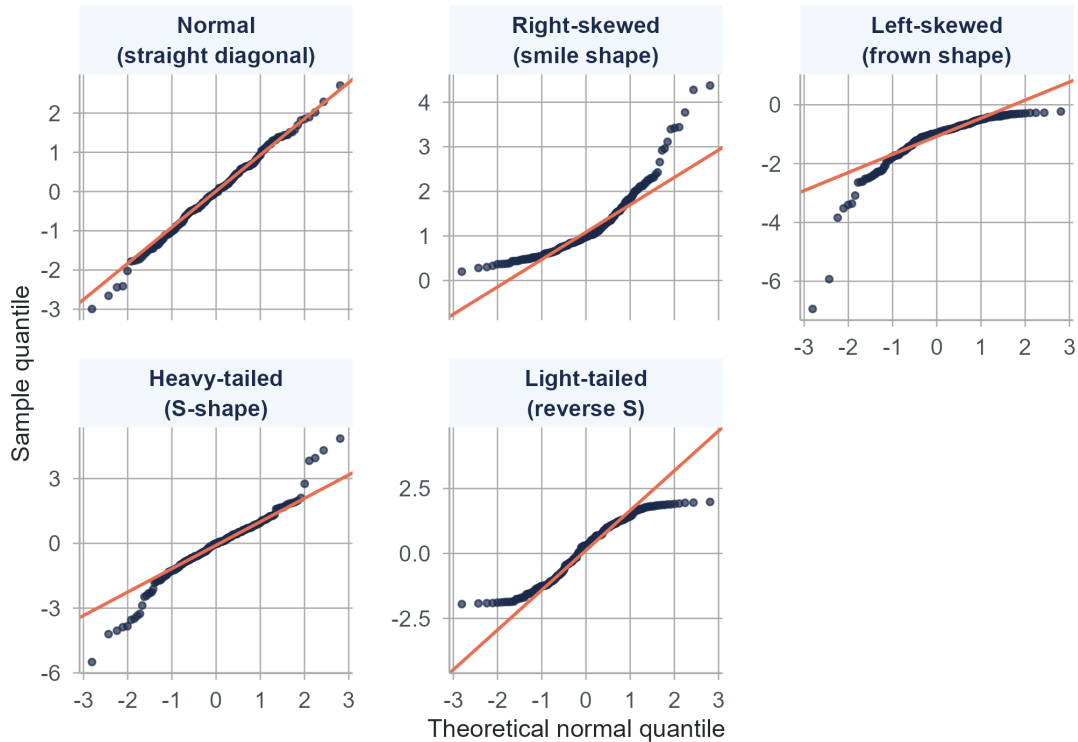


Figure 6.2: Five canonical QQ-plot patterns. Sample quantiles on the y-axis are plotted against theoretical normal quantiles on the x-axis; the orange line is what perfectly normal data would look like. Read the shape, not the individual points.

QQ plots are more useful than histograms for assessing normality because they make deviations from the expected pattern visually obvious, especially in the tails — which is where normal-based calculations tend to mislead you the most.

6.6.2 The Shapiro-Wilk Test

The **Shapiro-Wilk test** is a formal hypothesis test for normality. The null hypothesis is that the data come from a normal distribution. A small p -value is interpreted as evidence against that hypothesis, suggesting the data are not normal.

A word of caution before we go further. You will see references throughout this book and in the wider literature to a p -value being “below 0.05” as a kind of automatic rejection trigger. That convention has a long history, but treating any single threshold as a hard line between “real” and “not real” is one of the most over-criticized habits in modern statistics. Chapter 8 takes this on directly, including the practice known as p -hacking and the broader replication crisis it helped expose. For now, just note that a Shapiro-Wilk p -value of 0.04 is not categorically different from one of 0.06, and either way you should look at a QQ plot before drawing a conclusion.

A few additional cautions about the test itself.

First, with large samples, the Shapiro-Wilk test will reject normality even when the departure is trivially small and practically irrelevant. A dataset of 10,000 values might fail the test while being close enough to normal that all normal-based calculations work perfectly well.

Second, with small samples, the test has low power, meaning it may fail to detect non-normality even when the data are clearly not normal.

For these reasons, the QQ plot is often more informative than the formal test. Use both together when in doubt, but let the QQ plot guide your judgment about whether the non-normality is large enough to matter for your analysis.

A practical approach is this. If the QQ plot looks reasonably straight and the Shapiro-Wilk p -value is not unusually small, you can proceed with normal-based methods. If the QQ plot shows clear curvature in a small sample, or if the test produces a very small p -value alongside a visibly bent QQ plot, take that seriously and consider alternatives like nonparametric methods or data transformations. And if you have a very large sample and the test rejects normality but the QQ plot looks fine, the departure is probably too small to worry about.

AI Reality Check

Ask an AI tool to analyze a dataset and it will almost certainly start computing means, standard deviations, and z -scores. It may even generate confidence intervals and p -values without checking whether the underlying data is approximately normal. Most AI tools treat normality as the default assumption rather than something to be verified.

This matters because many real-world distributions are not normal, and some are far from normal. Income data is right-skewed. Medical cost data has extreme outliers. Customer lifetime value data often follows a power law. For these distributions, the mean and standard deviation can be misleading summaries, and z -scores computed from them can mischaracterize how unusual a value really is. A z -score of 3.0 under a normal distribution corresponds

to the 99.87th percentile. Under a heavy-tailed distribution, a value three standard deviations above the mean might be far more common than that.

Before trusting any AI-generated analysis that relies on normal distribution calculations, check the shape of your data. A histogram or QQ plot takes seconds to produce and can save you from conclusions that are precisely computed and fundamentally wrong.

6.7 From Individual Values to Sample Means

Everything we have discussed so far concerns individual observations. The z-score for a single baby's weight. The probability that a single randomly selected baby falls below a threshold. But in practice, researchers and decision-makers usually work not with single values but with *averages* computed from samples.

A hospital administrator does not ask, "What is the probability that one baby weighs less than 2,500 grams?" The question becomes, "If we look at the average birth weight across the 200 deliveries in our unit last month, what can we infer about the population of mothers we serve?"

A quality control engineer does not ask, "What is the probability that one widget is too heavy?" They ask, "If the average weight across a sample of 50 widgets exceeds a threshold, should we stop the production line?"

This shift, from reasoning about individuals to reasoning about sample averages, is where the most important ideas in this chapter live. And it is a shift that mirrors how decisions are made in the real world. A school board evaluating a new reading curriculum does not look at one student's test score. They look at the average across hundreds of students. A public health official monitoring air quality does not rely on one sensor reading from one afternoon. They look at the average across many readings over time. The entire apparatus of modern evidence-based decision-making, from clinical trials to A/B testing on websites, rests on the behavior of sample averages rather than individual observations.

6.8 Sampling Distributions

6.8.1 The Concept

Imagine the following thought experiment. You have a large population, say, all full-term newborns in the United States in a given year, roughly 3.6 million babies. The population has a mean birth weight of $\mu = 3400$ grams and a standard deviation of $\sigma = 500$ grams.

Now imagine you randomly select 25 babies and compute the mean weight of your sample. You might get $\bar{x} = 3,452$ grams. Put that sample back (or just imagine the population is so large it does not matter). Draw another sample of 25 and compute the mean. You might get $\bar{x} = 3,378$ grams. Do it again: $\bar{x} = 3,421$ grams. Again: $\bar{x} = 3,339$ grams.

If you did this thousands of times, drawing a fresh sample of 25 each time and recording the sample mean, you would build up a collection of sample means. That collection has its own distribution, its own shape, center, and spread. This distribution of sample means is called the **sampling distribution of the sample mean**.

The sampling distribution is not the distribution of the data. It is the distribution of the *statistic* (in this case, the mean) computed across many hypothetical samples. This distinction is absolutely critical to understanding inference.

6.8.2 Properties of the Sampling Distribution

The sampling distribution of the sample mean has three key properties. All of them follow from mathematical probability theory, and all of them can be verified through simulation.

1. Center. The mean of the sampling distribution equals the population mean.

$$\mu_{\bar{x}} = \mu$$

If the population mean birth weight is 3,400 grams, then the average of all possible sample means (across all possible samples of size n) is also 3,400 grams. The sample mean is an **unbiased estimator** of the population mean. It does not systematically overshoot or undershoot.

2. Spread. The standard deviation of the sampling distribution is smaller than the population standard deviation, by a factor of $\sqrt{n}$.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

This quantity, the standard deviation of the sampling distribution, has a special name. It is called the **standard error** of the mean.

For our birth weight example with $\sigma = 500$ and $n = 25$:

$$\sigma_{\bar{x}} = \frac{500}{\sqrt{25}} = \frac{500}{5} = 100 \text{ grams}$$

Individual birth weights vary with a standard deviation of 500 grams, but sample means of 25 babies vary with a standard deviation of only 100 grams. Averages are less variable than individuals. This should match your intuition. It would be surprising to find one baby weighing 4,400 grams (2 standard deviations above the mean), but it would be very surprising to find that the average weight of 25 randomly selected babies was 4,400 grams. That would be 10 standard errors above the population mean, an event so improbable it essentially cannot happen by chance.

3. Shape. This is where the Central Limit Theorem enters the picture, and it deserves its own section.

6.9 Standard Error vs. Standard Deviation

This is the concept that trips up more introductory statistics students than perhaps any other. Let us be as clear as we can.

The **standard deviation** (σ or s) measures how much *individual values* spread out around the mean of the distribution. A standard deviation of 500 grams for birth weights means individual babies vary by about 500 grams from the average.

The **standard error** ($\sigma_{\bar{x}}$ or SE) measures how much *sample means* spread out around the population mean if you were to take many samples of the same size. A standard error of 100 grams (when $n = 25$) means that sample means of 25 babies vary by about 100 grams from the true population average.

Here is another way to think about it. The standard deviation is a property of the *population* (or data). It stays the same no matter how many observations you collect. If birth weights have a standard deviation of 500 grams, that is true whether you sample 10 babies or 10,000 babies. The underlying variability in the population does not change because you measured more of it.

The standard error is a property of the *sampling process*. It depends on both the population standard deviation and the sample size. And it *shrinks* as the sample size grows. With $n = 25$, the standard error is 100 grams. With $n = 100$, it drops to 50 grams. With $n = 400$, it is 25 grams.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

This formula encodes a foundational idea in statistics. Larger samples give you more precise estimates. But the improvement follows a square root pattern, not a linear one. To cut the standard error in half, you need to quadruple the sample size. To cut it to a third, you need nine times as many observations. There are diminishing returns to collecting more data, which is why sample size planning matters so much in research design.

i Note

When reading research papers, pay attention to whether authors report “standard deviation” or “standard error” on their graphs and tables. They are different quantities that serve different purposes, and confusing them is one of the most common errors in published research. A graph showing error bars with standard errors looks more precise than the same graph with standard deviation bars, which can mislead readers about how variable the underlying data are.

Let us explore a concrete comparison to drive this home.

Suppose you survey commuters about their one-way commute time. The population has $\mu = 27$ minutes and $\sigma = 12$ minutes.

- **Individual commute times** range widely. About 68% of people commute between 15 and 39 minutes (the mean plus or minus one standard deviation).
- **Sample means with $n = 36$** have a standard error of $12/\sqrt{36} = 2$ minutes. About 68% of sample means of 36 commuters would fall between 25 and 29 minutes, a much narrower range.
- **Sample means with $n = 144$** have a standard error of $12/\sqrt{144} = 1$ minute. About 68% of these sample means would fall between 26 and 28 minutes.

The individual data are noisy. The sample means are much less noisy. And as the sample gets bigger, the sample means cluster ever more tightly around the true population mean.

6.10 The Central Limit Theorem

We now arrive at what is, without exaggeration, the most important result in introductory statistics. It is the reason that so many of the methods in the rest of this book work. It is the reason we can build confidence intervals and perform hypothesis tests. It is the bridge between probability theory and statistical practice.

6.10.1 The Statement

The **Central Limit Theorem** (CLT) states the following.

If you draw random samples of size n from *any* population with mean μ and finite standard deviation σ , the sampling distribution of the sample mean $\bar{x}$ will be approximately normal with mean μ and standard deviation $\sigma/\sqrt{n}$, provided n is sufficiently large.

In notation:

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \quad \text{approximately, for large } n$$

Or equivalently, the standardized version:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

follows approximately a standard normal $N(0, 1)$ distribution.

6.10.2 Why This Result Matters

The Central Limit Theorem has a long history. Versions of it were proved by Abraham de Moivre in 1733 (for the binomial case) and generalized by Pierre-Simon Laplace in the early 1800s, but it took over a century of further mathematical work before the theorem was stated and established in full generality. The name “Central Limit Theorem” itself dates to George Pólya in 1920, who called it “central” not because it occupies a central position in a distribution, but because it is of central importance to the field of probability. Without this theorem, the framework of confidence intervals, hypothesis tests, and regression analysis that fills the rest of this book would not hold together.

Read the statement again, especially the part that says “from *any* population.”

The population could be normally distributed. It could be skewed. It could be uniform. It could be bimodal. It could have a bizarre, irregular shape unlike anything in a textbook. It does not matter. As long as the population has a finite mean and a finite standard deviation, the sampling distribution of the mean will approach normality as the sample size grows.

This is what makes the result so useful in practice: we do not need to know the shape of the population distribution to make inferences about the mean. We only need a large enough sample. The normal distribution is more than one model among many. It is the universal destination of the sampling distribution of the mean, regardless of where the journey starts.

The Central Limit Theorem in Action

As sample size grows, the distribution of sample means becomes normal

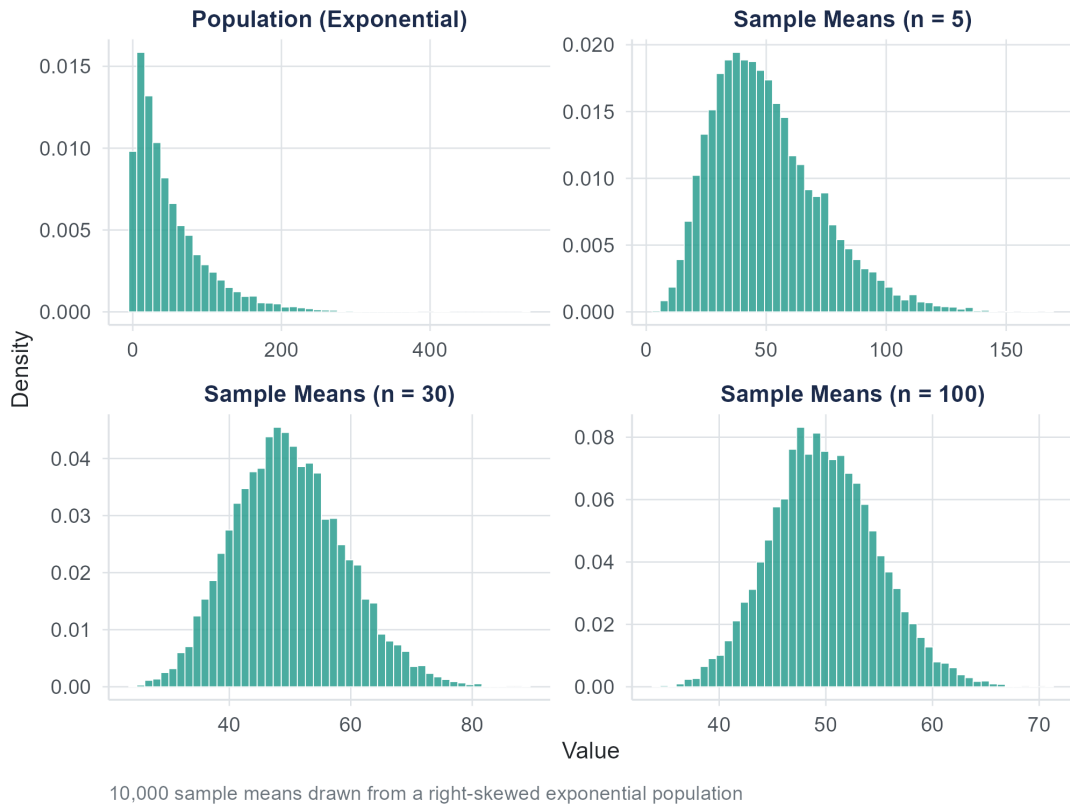


Figure 6.3: The Central Limit Theorem demonstrated with a right-skewed exponential population. As sample size increases from 5 to 30 to 100, the distribution of sample means becomes progressively more symmetric and bell-shaped, regardless of the skewed population from which the samples were drawn.

6.10.3 How Large Is “Large Enough”?

The CLT says the approximation improves as n increases, but how large does n need to be? The answer depends on how far the population distribution is from normal.

- **If the population is already normal**, the sampling distribution of the mean is exactly normal for any sample size, even $n = 1$. The CLT’s approximation is perfect from the start.

- **If the population is roughly symmetric** (like a uniform distribution), $n = 15$ or $n = 20$ is often enough for the sampling distribution to look very close to normal.
- **If the population is moderately skewed**, $n = 30$ is the traditional guideline, though this is a rule of thumb, not a law of nature.
- **If the population is extremely skewed or has very heavy tails**, you might need $n = 50$, $n = 100$, or even more.

The traditional advice of “ $n \geq 30$ and you’re fine” is a simplification. It works in many common situations but not all. The best practice is to think about the shape of your population and adjust your expectations accordingly. Simulation can help. If you generate random samples from a distribution that resembles your population and plot the resulting sample means, you can see directly whether the sampling distribution looks approximately normal at your sample size.

Try It Online

Return to the **Distribution Explorer** and switch to the Sampling Distribution view. Choose a skewed population shape, set the sample size to 5, and draw repeated samples. Then increase n to 30, then 100. Watch the sampling distribution of the mean transform from skewed to approximately normal as n grows. There is no better way to internalize the Central Limit Theorem than to watch it happen in real time.

6.10.4 A Visual Walkthrough

Here is what you would see if you ran a simulation.

Start with a right-skewed population. Think of income data or response time data, a large cluster of low-to-moderate values with a long tail stretching to the right.

Draw samples of size $n = 5$ and plot the means. The distribution of sample means is still somewhat right-skewed, though less so than the population. The center is at the population mean, but the shape is not yet normal.

Increase to $n = 15$. The distribution of sample means is now closer to symmetric. The right-skew is fading, though you might still detect a slight asymmetry.

Increase to $n = 30$. The distribution looks quite normal. A QQ plot would show points falling close to a straight line.

Increase to $n = 100$. The distribution is, for all practical purposes, perfectly normal. The spread has also narrowed considerably compared to $n = 5$.

At every stage, the center stays the same, at the population mean. At every stage, the spread decreases. And the shape steadily approaches the bell curve.

6.10.5 The CLT in Action

Let us work through a complete example that puts all the ideas together.

A city's department of transportation wants to study commute times. They know from census data that the average commute time in their city is $\mu = 32$ minutes with a standard deviation of $\sigma = 14$ minutes. The distribution of individual commute times is right-skewed (many short commutes, fewer very long ones).

A researcher draws a random sample of $n = 49$ commuters. What is the probability that the sample mean commute time exceeds 35 minutes?

Step 1. Identify the sampling distribution.

By the CLT, the sampling distribution of $\bar{X}$ is approximately normal with

$$\mu_{\bar{x}} = 32 \quad \text{and} \quad \sigma_{\bar{x}} = \frac{14}{\sqrt{49}} = \frac{14}{7} = 2 \text{ minutes}$$

Step 2. Standardize.

$$z = \frac{35 - 32}{2} = \frac{3}{2} = 1.50$$

Step 3. Find the probability.

$$P(\bar{X} > 35) = P(Z > 1.50) = 1 - P(Z \leq 1.50) = 1 - 0.9332 = 0.0668$$

There is about a 6.7% chance that a random sample of 49 commuters will have a mean commute time exceeding 35 minutes.

Notice that we used the normal distribution even though the population of commute times is right-skewed. The CLT is what licenses that move. We are not saying individual commute times are normal. We are saying that the *average* of 49 randomly selected commute times is approximately normal.

6.10.6 What the CLT Does Not Say

A few common misconceptions are worth addressing.

The CLT does not say the data become normal. It says the *sampling distribution of the mean* becomes normal. If you take a sample of 100 incomes from a right-skewed population, the 100 income values in your sample will still look right-skewed. But the distribution of all possible sample means from samples of 100 will be approximately normal. The distinction is subtle and important.

The CLT does not apply to all statistics. The theorem is about the mean. There are similar results for some other statistics (proportions, for example, which we will see in later chapters), but not all statistics have sampling distributions that approach normality. The median, the range, and the maximum all have sampling distributions with different properties.

The CLT does not eliminate the need for random sampling. The theorem assumes your samples are drawn randomly from the population. If your sample is biased, if you only survey people who volunteer, or only measure widgets from one machine, the CLT cannot rescue you. A beautiful normal sampling distribution is useless if it is centered at the wrong value because the sampling process was flawed.

The CLT does not guarantee useful results with small samples from extreme distributions. If your population distribution is wildly skewed or has very heavy tails (meaning extreme values are common), the CLT's approximation may require a much larger sample than the traditional rule of thumb suggests. Consider the distribution of wealth, as opposed to income, in the United States. Wealth is so extremely concentrated that sample means can behave erratically even with samples of several hundred observations. A single billionaire in a sample of 500 people will have a dramatic effect on the sample mean, and it takes a very large sample before the CLT's reassurance of approximate normality becomes practically reliable. Knowing that the CLT exists does not excuse you from thinking carefully about whether your particular data and sample size meet its assumptions in a meaningful way.

6.11 Putting It All Together

Let us trace the arc of this chapter's logic, because the pieces build on each other.

Individual values from a normal distribution can be standardized using z-scores, which tell you how many standard deviations a value is from the mean. Using the standard normal distribution, you can calculate the probability of observing values in any range.

When you shift from individual values to sample means, a new distribution emerges, the sampling distribution. This distribution is centered at the population mean, has a standard deviation called the standard error (which equals $\sigma/\sqrt{n}$), and, by the Central Limit Theorem, is approximately normal regardless of the shape of the underlying population, provided the sample is large enough.

This chain of ideas, from individual variation to sampling variation to the guaranteed normality of the sampling distribution, is what makes statistical inference possible. In the next two chapters, we will use these ideas to build confidence intervals and to test hypotheses. Every confidence interval you construct, every p -value you calculate, relies on the sampling distribution being approximately normal. The Central Limit Theorem is why that reliance is justified.

Let us close by returning to where we started. When a hospital weighs a newborn and compares that weight to a distribution, it is applying the ideas from the first half of this chapter, z-scores, areas under the curve, and the Empirical Rule. When a researcher collects data on hundreds of births and uses the average to draw conclusions about a population, they are relying on the ideas from the second half, sampling distributions, standard error, and the Central Limit Theorem. The individual measurement and the sample average are both governed by the mathematics of the normal distribution, but in different ways and for different reasons. Understanding that distinction is the conceptual core of statistical inference, and it is what the next several chapters will build upon.

6.12 Looking Ahead

With the normal distribution and the Central Limit Theorem in hand, we now have the machinery to move from description to inference. We know that sample means follow an approximately normal distribution centered at the population mean, with a spread that shrinks predictably as the sample grows. The next chapter puts this machinery to work. Given a single sample mean, how do we construct a range of plausible values for the population mean? How precise is our estimate, and what determines that precision? Those are the questions of confidence intervals, and they are the first formal tool of statistical inference. If the Central Limit Theorem is the engine, confidence intervals are the first thing we build with it.

6.13 Further Reading and References

The following works are cited in this chapter or provide valuable additional context for the ideas covered here.

On the normal distribution and its history: Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press. (A scholarly but readable history covering the development of the normal distribution by de Moivre, Laplace, and Gauss, and the emergence of the Central Limit Theorem.)

On birth weight data and clinical thresholds: Martin, J. A., Hamilton, B. E., Osterman, M. J. K., & Driscoll, A. K. (2019). Births: Final data for 2018. *National Vital Statistics Reports*, 68(13). National Center for Health Statistics. (The CDC's authoritative source for U.S. birth weight distributions, the data underlying this chapter's examples.)

On the Central Limit Theorem: Fischer, H. (2011). *A history of the Central Limit Theorem: From classical to modern probability theory*. Springer. (A comprehensive account of how the CLT evolved from de Moivre's early work through Lindeberg, Lévy, and Feller, for readers interested in the mathematical backstory.)

On the misuse of normality assumptions in finance: Taleb, N. N. (2007). *The black swan: The impact of the highly improbable*. Random House. (A forceful, accessible argument about how assuming normality in financial markets led to catastrophic underestimation of tail risks, contributing to the 2008 financial crisis discussed in this chapter.)

For further reading on distributions and sampling: Diez, D. M., Çetinkaya-Rundel, M., & Barr, C. D. (2019). *OpenIntro Statistics* (4th ed.). OpenIntro. (A free, open-source statistics textbook with clear coverage of the normal distribution, sampling distributions, and the Central Limit Theorem, with worked examples and exercises.)

On pediatric growth standards and reference populations: WHO Multicentre Growth Reference Study Group. (2006). *WHO child growth standards: Length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: Methods and development*. World Health Organization. (The source for the growth standards discussed in this chapter's Ethics Moment, based on breastfed children from six countries across five continents.)

6.14 Key Terms

- **Normal distribution:** A continuous probability distribution defined by its mean (μ) and standard deviation (σ), with a symmetric, bell-shaped density curve.
- **Standard normal distribution:** The normal distribution with $\mu = 0$ and $\sigma = 1$. The reference distribution used in z-tables.
- **Probability density function (PDF):** A function that describes the relative likelihood of a continuous random variable taking any given value. Areas under the curve represent probabilities.
- **Z-score:** The number of standard deviations a value is above or below the mean. Calculated as $z = (x - \mu)/\sigma$.
- **Empirical Rule (68-95-99.7 Rule):** For a normal distribution, approximately 68% of values fall within 1 standard deviation of the mean, 95% within 2, and 99.7% within 3.
- **Z-table (Standard normal table):** A table giving the cumulative probability $P(Z \leq z)$ for values of z in a standard normal distribution.
- **Percentile (Quantile):** The value below which a specified percentage of observations fall. The 10th percentile is the value below which 10% of the data lie.
- **QQ plot (Quantile-quantile plot):** A graphical tool for assessing whether data follow a theoretical distribution, typically the normal. Points falling near a straight line suggest normality.
- **Shapiro-Wilk test:** A formal hypothesis test for normality. A small p -value suggests the data are not normally distributed.
- **Sampling distribution:** The probability distribution of a statistic (such as the sample mean) computed from all possible samples of a given size from a population.
- **Standard error:** The standard deviation of the sampling distribution of a statistic. For the sample mean, $SE = \sigma/\sqrt{n}$ when σ is known, or $SE = s/\sqrt{n}$ when it is estimated from the sample.
- **Central Limit Theorem (CLT):** The theorem stating that the sampling distribution of the sample mean approaches a normal distribution as the sample size increases, regardless of the population's shape, provided the population has a finite mean and finite standard deviation.
- **Unbiased estimator:** A statistic whose sampling distribution is centered at the true population parameter. The sample mean is an unbiased estimator of the population mean.

6.15 Exercises

6.15.1 Check Your Understanding

These questions test whether you grasp the core concepts. You should be able to answer them without performing calculations.

1. What two parameters completely determine a normal distribution? What does each one control about the shape of the curve?
2. A z-score of -2.5 indicates that a value is _____ standard deviations _____ the mean. In terms of the Empirical Rule, would you consider this value common or unusual?
3. Explain in your own words why the total area under a normal density curve equals 1.
4. The Empirical Rule says about 95% of values fall within 2 standard deviations of the mean. If you are told that a dataset has a mean of 50 and a standard deviation of 8, what range captures about 95% of the values (assuming the data are approximately normal)?
5. You look at a QQ plot and see points that follow the diagonal line closely in the middle but curve sharply upward at the right end. What does this suggest about the data's distribution?
6. A classmate says, "The Central Limit Theorem tells us that if we collect a large enough sample, our data will be normally distributed." What is wrong with this statement?
7. Why is the standard error always smaller than the standard deviation (assuming $n > 1$)? Explain the logic, not just the formula.
8. Two populations have the same mean. Population A has $\sigma = 10$ and population B has $\sigma = 30$. You draw samples of $n = 100$ from each. Which population's sample means will have a larger standard error? By what factor?
9. Suppose a population is heavily right-skewed. A researcher draws random samples of size $n = 5$ and plots the sampling distribution of the mean. Another researcher does the same with $n = 200$. Whose sampling distribution will look more like a normal distribution? Why?
10. Explain why the Central Limit Theorem is necessary for confidence intervals and hypothesis tests to work, even when the population is not normally distributed.

6.15.2 Apply It

These problems require calculations. Show your work.

1. Adult male heights in a certain country are approximately normally distributed with $\mu = 175$ cm and $\sigma = 7$ cm.
 - (a) What is the z-score for a man who is 168 cm tall?
 - (b) What proportion of men are taller than 185 cm?
 - (c) What proportion of men have heights between 170 cm and 180 cm?
2. A standardized reading test has scores that are normally distributed with $\mu = 500$ and $\sigma = 80$.
 - (a) What score separates the top 5% of students from the rest?
 - (b) What score marks the 25th percentile?
 - (c) A student scores 620. What is this student's z-score, and what percentage of students scored lower?
3. The time it takes a barista at a busy coffee shop to prepare an espresso drink is normally distributed with $\mu = 4.2$ minutes and $\sigma = 0.8$ minutes.
 - (a) What proportion of drinks take more than 5.5 minutes?
 - (b) What proportion of drinks are prepared in under 3 minutes?
 - (c) The shop guarantees drinks within 6 minutes or they are free. What percentage of drinks will be free?
4. A factory produces ball bearings with a target diameter of 10.00 mm. The actual diameters are normally distributed with $\mu = 10.00$ mm and $\sigma = 0.05$ mm. Ball bearings are rejected if their diameter is less than 9.90 mm or greater than 10.10 mm.
 - (a) What z-scores correspond to the rejection thresholds?
 - (b) What proportion of ball bearings are rejected?
 - (c) If the factory produces 20,000 ball bearings per day, how many are expected to be rejected daily?
5. A population of monthly cellphone bills has $\mu = \$85$ and $\sigma = \$32$. A researcher takes a random sample of $n = 64$ customers.
 - (a) What are the mean and standard error of the sampling distribution of $\bar{X}$?
 - (b) What is the probability that the sample mean exceeds \$90?
 - (c) What is the probability that the sample mean is between \$80 and \$90?
6. The amount of soda dispensed by a vending machine is normally distributed with $\mu = 12.0$ oz and $\sigma = 0.3$ oz.
 - (a) What is the probability that a single cup contains less than 11.5 oz?

- (b) If a quality inspector measures 9 cups, what is the probability that the sample mean is less than 11.5 oz?
- (c) Compare your answers to (a) and (b) and explain why they differ.
- 7.** Household electricity consumption in a region has $\mu = 950$ kWh per month and $\sigma = 200$ kWh. The distribution is right-skewed.
- (a) Can you use the normal distribution to find the probability that a single randomly selected household uses more than 1,200 kWh? Why or why not?
- (b) Can you use the normal distribution to find the probability that the mean consumption of a sample of $n = 50$ households exceeds 1,000 kWh? Justify your answer.
- (c) If the answer to (b) is yes, compute the probability.
- 8.** Resting heart rates for healthy adults are approximately normally distributed with $\mu = 72$ bpm and $\sigma = 8$ bpm.
- (a) What percentage of adults have resting heart rates below 60 bpm?
- (b) Above what heart rate do the top 10% of adults fall?
- (c) A doctor's office measures the average resting heart rate of 16 patients during a wellness day. What is the probability that this average exceeds 76 bpm?
- 9.** An airline knows that the weight of checked baggage per passenger is normally distributed with $\mu = 42$ pounds and $\sigma = 10$ pounds. A small regional flight carries 36 passengers.
- (a) What are the mean and standard error of the sampling distribution of the mean baggage weight per passenger?
- (b) The cargo hold can safely handle a mean baggage weight of up to 45 pounds per passenger. What is the probability that the mean weight exceeds this limit?
- (c) If the airline wanted the probability in (b) to be less than 1%, how many passengers would they need to sample or account for? (Hint: find the n that makes $P(\bar{X} > 45) < 0.01$.)
- 10.** A professor's exam scores are approximately normally distributed with $\mu = 74$ and $\sigma = 11$.
- (a) What proportion of students score an A (90 or above)?
- (b) The professor curves the exam so that the top 15% get A's. What is the minimum score for an A under this curve?
- (c) If a class has 40 students, what is the probability that the class average is below 70?

The following three exercises use the birth-weights.csv dataset available on the companion website. This dataset contains 944 observations drawn from the OpenIntro births14 dataset, a real sample of 2014 CDC natality data. Variables include baby_id, birth_weight_g (birth weight in

grams), gestational_weeks, mother_age, prenatal_visits, and smoking_status (Yes/No). Because the dataset includes preterm births, the distribution of birth weights is slightly left-skewed (skewness = -0.985). The full-term subset (gestational weeks ≥ 37) contains 829 observations and is more symmetric.

11. Using the birth-weights.csv dataset, compute the sample mean and standard deviation of birth_weight_g for all 944 infants. Create a histogram of birth weights and overlay (or sketch) a normal curve using your computed mean and standard deviation. Does the distribution appear approximately normal? Create a QQ plot of birth weights to further assess normality. You should find a sample mean of approximately 3,266 grams and a standard deviation of approximately 593 grams. Note any left skew caused by the inclusion of preterm births. Then repeat the histogram and QQ plot for only the full-term infants (gestational weeks ≥ 37). Comment on how restricting to full-term births affects the shape of the distribution.

12. Using the sample mean and standard deviation you computed in Problem 11 (approximately 3,266 g and 593 g), calculate the z-score for a birth weight of 2,500 grams (the clinical threshold for low birth weight). You should find $z \approx -1.29$, giving a normal-model prediction of about 9.9% below 2,500 g. Now count the actual number of infants in the dataset with birth_weight_g below 2,500 and compute the observed proportion (you should find it is close to 9.8%). How does the observed proportion compare to the normal model's prediction? Also compute the z-score for 4,000 grams ($z \approx 1.24$) and compare the predicted and observed proportions above that threshold (observed: about 10.8%). What do these comparisons tell you about how well the normal model fits these data, particularly in the tails?

13. Separate the data by smoking_status (846 non-smokers and 80 smokers). For smokers and non-smokers separately, compute the mean and standard deviation of birth_weight_g. You should find that the mean birth weight for non-smokers is approximately 3,296 grams and for smokers approximately 2,948 grams, a difference of about 348 grams. Now suppose you draw random samples of $n = 25$ non-smoking mothers from this dataset. Using the Central Limit Theorem, compute the standard error of the sample mean. What is the probability that a random sample of 25 non-smoking mothers would produce a sample mean birth weight below 3,200 grams?

6.15.3 Think Deeper

These questions require more interpretation and critical thinking. There may not be a single correct answer.

1. Birth weights are approximately, but not perfectly, normally distributed. The normal model works well for full-term singleton births but less well when you include premature births, multiple births (twins, triplets), or do not account for differences across demographic groups. Discuss the tradeoffs of using a single normal model for birth weights versus using separate models for different subpopulations. When might a single model be adequate, and when could it lead to poor clinical decisions?
2. The Empirical Rule tells us that about 99.7% of values in a normal distribution fall within 3 standard deviations of the mean. In the financial world, events beyond 3 standard deviations are sometimes called “tail events” or “black swans.” During the 2008 financial crisis, several market moves were described as “25-sigma events,” meaning they were 25 standard deviations from the expected value. If the normal model were correct, such events would be so improbable that they should essentially never happen in the lifetime of the universe. What does this tell you about whether financial returns follow a normal distribution? What are the practical consequences of assuming normality when the true distribution has heavier tails?
3. A pharmaceutical company runs a clinical trial with $n = 400$ patients and finds that the sample mean improvement in a symptom score is $\bar{x} = 2.1$ points, with a standard error of 0.5 points. Another company runs a smaller trial with $n = 25$ patients studying a different treatment and finds $\bar{x} = 4.8$ points, with a standard error of 3.2 points. Both improvements are positive. Which result gives you more confidence that the treatment actually works? Explain your reasoning in terms of the sampling distribution and the relationship between sample size, standard error, and the reliability of the sample mean as an estimate.
4. A news article reports that “the average American checks their phone 144 times per day, based on a survey of 1,000 adults.” The standard deviation of phone-checking frequency is likely very large, perhaps 80 or more, given how widely this behavior varies. Despite this large individual variation, the article treats the average of 144 as a precise number. Using your understanding of the standard error, calculate the approximate standard error for this estimate. Is the article’s implied precision justified? What additional information would help you evaluate the claim?
5. The Central Limit Theorem is sometimes described as “the reason statistics works.” But the theorem requires random sampling, which is rarely achieved perfectly in practice. Most surveys have some nonresponse bias. Most experiments lose participants to dropout. Most observational datasets come from convenience samples. If the CLT’s assumption of random sampling is so routinely violated, why do statistical methods still produce useful results in practice? Under what circumstances would violations of random sampling be severe enough to make CLT-based inference misleading?

Chapter 7

Confidence Intervals

7.1 The Night the Polls “Failed”

On November 8, 2016, most major polling organizations had Hillary Clinton leading Donald Trump in the presidential race. The New York Times needle pointed overwhelmingly toward Clinton. FiveThirtyEight gave her about a 71% chance of winning. Other forecasters were even more confident. When the results came in and Trump won, the reaction was immediate and volcanic. The polls were wrong. The pollsters had failed. Data was dead.

Except the polls were not wrong. Not in the way people meant it.

The final national polling average had Clinton ahead by about 3.2 percentage points. She won the national popular vote by 2.1 points. That is a miss of roughly one point, well within the stated margin of error of most national polls, which typically runs plus or minus 3 to 4 percentage points. By the standards of polling accuracy, the 2016 national polls were fairly ordinary.

The problem was not accuracy. The problem was interpretation. When people saw “Clinton +3,” they read it as “Clinton wins.” They mentally deleted the margin of error. They ignored the part where the poll was saying, in effect, “our best estimate is Clinton +3, but it could easily be Clinton +6 or Trump +1, and we would not be surprised by either.” A race that is within the margin of error is a race that could go either way. The data said exactly that. People heard something else.

Fast forward to 2020. The polls overestimated Joe Biden’s margin in several key states by larger amounts, roughly 4 to 5 points in some battleground states. This time the errors were bigger, and the postmortems more painful. But even in 2020, most final results fell within or just outside the reported confidence intervals. The story was not that polling was broken. The story was that

the public, the media, and even some analysts consistently misunderstood what a margin of error means.

This chapter is about that misunderstanding and how to fix it. The margin of error is, after all, what gave this book its title. It is one of the most useful ideas in all of statistics, and one of the most widely misinterpreted. A confidence interval is how we express what we know and what we do not know from a sample of data. Learning to read them correctly, construct them, and think critically about what they can and cannot tell you is one of the most practical skills you will take from this book.

The confusion is understandable. We live in a culture that rewards certainty and punishes hedging. A politician who says “I am confident this policy will work” sounds strong. A politician who says “our best estimate suggests this policy will work, but there is a plausible range of outcomes that includes modest failure” sounds weak, even though the second statement is almost always more honest. Confidence intervals are the statistical equivalent of that second statement. They are an exercise in disciplined honesty about what we know, what we suspect, and what could still surprise us.

7.2 A Single Number Is Never Enough

Let us start with a simple question. Suppose you want to know the average amount of student loan debt carried by graduates of four-year universities in the United States. You cannot survey every graduate, so you take a random sample of 400 recent graduates and compute the sample mean. You get \$28,950.

Is that the answer? Is the average debt for all graduates exactly \$28,950?

Of course not. If you took a different random sample of 400 graduates, you would get a different number. Maybe \$29,300. Maybe \$28,410. Every sample gives a slightly different result because every sample contains a slightly different set of people. The sample mean of \$28,950 is a **point estimate**, a single number that serves as our best guess of the population parameter we care about.

Point estimates are useful. They are also incomplete. A point estimate tells you where the center of your best guess is, but it tells you nothing about how much uncertainty surrounds that guess. A point estimate based on a sample of 10 people carries much more uncertainty than one based on a sample of 10,000 people, but if all you report is the number, they look identical.

This is where **interval estimates** come in. Instead of reporting a single number, we report a range of plausible values for the population parameter. The sample mean is \$28,950, and based on the

variability in our data and the size of our sample, we are confident that the true population mean is somewhere between \$27,600 and \$30,300. That range is a confidence interval. It communicates both our estimate and our uncertainty.

i Note

The tension between point estimates and interval estimates shows up everywhere, well beyond statistics. When a contractor gives you a quote for a kitchen renovation, the smart ones give a range. “Somewhere between \$35,000 and \$45,000, depending on what we find when we open the walls.” The ones who give you a single number either know something the others do not or are about to surprise you with change orders. A range is more honest. It admits uncertainty. Statistics formalizes this honesty.

7.3 The Logic of Confidence Intervals

A confidence interval has a simple structure. At its core, every confidence interval follows the same recipe.

$$\text{point estimate} \pm \text{margin of error}$$

The point estimate is the statistic you computed from your sample, the sample mean, the sample proportion, whatever quantity you are using to estimate the population parameter. The margin of error captures how far off that estimate might plausibly be. A bigger margin of error means more uncertainty. A smaller one means less.

The margin of error itself has two ingredients. One is a measure of how much variability exists in your data, usually captured by the standard error of your statistic. The other is a multiplier that depends on how confident you want to be, which we will call the critical value. So the full recipe looks like this.

$$\text{point estimate} \pm (\text{critical value}) \times (\text{standard error})$$

This structure is the same whether you are estimating a mean, a proportion, or a difference between groups. The specific formulas change, but the logic does not. You start with your best guess. You figure out how much that guess could reasonably wiggle if you took a different sample. You stretch a band around your estimate that covers the most plausible values for the true parameter.

Think of it like throwing darts. Your point estimate is where the dart lands. The margin of error is the circle you draw around the dart to account for the fact that you are not a perfect dart thrower. A better dart thrower, analogous to having a larger sample or less variable data, gets a smaller circle. Someone who demands a higher probability that the bullseye is inside the circle, analogous to wanting higher confidence, needs a bigger circle.

7.4 Confidence Intervals for Means

Let us start with the most common case. You have a sample of numerical data and you want to estimate the population mean, μ . You know your sample mean, $\bar{x}$, and your sample standard deviation, s .

To assemble a confidence interval, we need three pieces, in roughly this order.

1. The **standard error** — a measure of how much the sample mean would wobble from sample to sample. The Central Limit Theorem in Chapter 6 guarantees that this wobble has a predictable size when the sample is large enough.
2. A **critical value** — a multiplier that says how many standard errors wide the interval should be in order to achieve the confidence we want. Because we have to estimate the population standard deviation from our sample, the appropriate distribution to read this multiplier from is not the normal distribution but a closely related one called the ***t*-distribution**. We will introduce it shortly.
3. A small bookkeeping number called the **degrees of freedom** that tells us *which* *t*-distribution to use. For a single-sample mean it equals $n - 1$.

Combining these, the confidence interval for the mean is

$$\bar{x} \pm t^* \times \frac{s}{\sqrt{n}}$$

where n is the sample size and t^* is the critical value from the *t*-distribution with $n - 1$ degrees of freedom at your chosen confidence level.

The next three subsections unpack each piece in turn.

7.4.1 The Standard Error

The quantity $\frac{s}{\sqrt{n}}$ is the **standard error** of the sample mean. In Chapter 6, you learned that the sampling distribution of $\bar{x}$ has a standard deviation of $\frac{\sigma}{\sqrt{n}}$, where σ is the population standard deviation. But in practice, we almost never know σ . We have to estimate it using s , the sample standard deviation. When we plug s in for σ , the result is the standard error.

The standard error tells you how much the sample mean would bounce around from sample to sample. If the standard error is small, different samples would produce similar means, and your estimate is precise. If it is large, your estimate is less precise, and the confidence interval will be wider.

Notice that the standard error depends on two things. It gets smaller when s is smaller, meaning when the individual data points are less spread out. And it gets smaller when n is larger, because more data means more precision. That square root in the denominator is important. To cut the standard error in half, you need to quadruple your sample size. Precision does not come cheap.

7.4.2 Why t Instead of z

If you have already learned about the normal (z) distribution, you might wonder why we are using something called the t -distribution instead. The answer comes down to a practical problem.

When we use s to estimate σ , we introduce additional uncertainty. We are not just uncertain about the population mean, we are also uncertain about the population standard deviation. For large samples, this does not matter much, because s is a very good estimate of σ when you have a lot of data. But for smaller samples, s can be noticeably off from σ , and ignoring that extra wobble leads to confidence intervals that are too narrow.

The t -distribution, developed by William Sealy Gosset in 1908 while working at the Guinness brewery in Dublin, accounts for this extra uncertainty. It looks like the normal distribution but with heavier tails, meaning it assigns more probability to values far from the center. This makes the critical value t^* larger than the corresponding z^* from the normal distribution, which in turn makes the confidence interval wider. That extra width is the price you pay for not knowing σ .

Gosset's story is worth a brief detour. He was a chemist at Guinness, not a professor, and he was trying to figure out how to draw useful conclusions from very small samples of barley and hops. The brewery would not let him publish under his own name, worried that competitors might realize Guinness was using statistical methods, so he published under the pen name "Student." That is why the distribution is called "Student's t -distribution" in every textbook you will ever

open. One of the most important tools in modern statistics was invented by a brewer working on quality control for beer, and published under a pseudonym because his employer thought statistics was a trade secret. The history of science is full of moments like this, where practical problems in unglamorous settings produced ideas that changed entire fields.

As the sample size grows, the t -distribution gets closer and closer to the normal distribution. By the time you have several hundred observations, the difference is negligible. But for samples of 10, 20, or 50, the difference matters, and using the t -distribution gives you more honest intervals.

7.4.3 Degrees of Freedom

The t -distribution is not one distribution but a whole family of distributions, one for every possible number called the **degrees of freedom**. For a confidence interval about a single mean, the degrees of freedom equal $n - 1$.

Why $n - 1$? Think of it this way. To compute the sample standard deviation, you first need the sample mean. Once you know the mean, the last data point is determined by the others, because the values have to add up to produce that mean. So only $n - 1$ of the data values are free to vary. You have $n - 1$ independent pieces of information contributing to the estimate of variability.

When degrees of freedom are small, say 5 or 10, the t -distribution has noticeably heavier tails than the normal, and t^* is larger. When degrees of freedom are large, the t -distribution is nearly identical to the normal. For a 95% confidence interval, t^* with 10 degrees of freedom is about 2.228, while z^* is 1.96. With 100 degrees of freedom, t^* drops to about 1.984. The difference shrinks, but it never hurts to use the t -distribution when estimating a mean, regardless of sample size.

7.4.4 A Worked Example

A hospital administrator wants to estimate the average wait time in the emergency department. The administrator randomly selects 36 patient visits from the past month and records the wait time in minutes. The sample yields $\bar{x} = 47.2$ minutes and $s = 14.8$ minutes. The goal is a 95% confidence interval.

Step 1: Identify the pieces. $\bar{x} = 47.2$, $s = 14.8$, $n = 36$, confidence level = 95%.

Step 2: Degrees of freedom. $df = n - 1 = 35$.

Step 3: Find the critical value. For a 95% confidence interval with 35 degrees of freedom, $t^* = 2.030$ (from a t-table or software).

Step 4: Compute the standard error. $SE = \frac{s}{\sqrt{n}} = \frac{14.8}{\sqrt{36}} = \frac{14.8}{6} = 2.467$.

Step 5: Compute the margin of error. $ME = t^* \times SE = 2.030 \times 2.467 = 5.008$.

Step 6: Build the interval. 47.2 ± 5.008 , which gives $(42.19, 52.21)$.

The administrator can say with 95% confidence that the average ER wait time for patients at this hospital is between 42.2 and 52.2 minutes. We will discuss what “95% confident” actually means shortly, because it does not mean what you probably think it means.

7.4.5 Conditions for the t-Interval

The t -interval for a mean requires certain conditions to produce trustworthy results.

Independence. The observations in your sample must be independent of one another. One patient’s wait time should not affect another’s. Random sampling from the population generally ensures this. A useful rule of thumb is that the sample should be less than 10% of the population.

Normality (or a large enough sample). The t -interval assumes the population distribution is approximately normal, or the sample size is large enough for the Central Limit Theorem to kick in. For practical purposes, if $n \geq 30$, the CLT provides enough cover unless the data is extremely skewed. For smaller samples, you should check whether the data has a roughly symmetric shape without extreme outliers.

7.5 Confidence Intervals for Proportions

Now suppose you are not estimating a mean but a proportion. What fraction of voters support a particular candidate? What percentage of manufactured parts are defective? What share of customers will click on an advertisement?

When the variable is categorical with two outcomes (support/oppose, defective/not defective, click/no click), the parameter of interest is the population proportion p , and we estimate it with the **sample proportion**

$$\hat{p} = \frac{x}{n}$$

where x is the number of successes and n is the sample size.

7.5.1 Where the Formula Comes From

This is also the first time in the book that we are working directly with a proportion as our parameter of interest, so it is worth taking a moment to see how the standard-error formula here connects to ideas we already have in hand.

In Chapter 5 we learned that if each of n independent trials has success probability p , the count of successes X follows a binomial distribution with mean np and standard deviation $\sqrt{np(1-p)}$. The sample proportion $\hat{p} = X/n$ is just that count divided by a constant, so it has

$$E(\hat{p}) = p \quad \text{and} \quad \text{SD}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}.$$

So $\hat{p}$ is unbiased for p , and its variability shrinks like $1/\sqrt{n}$ — the same square-root pattern we have seen for sample means.

Chapter 6's Central Limit Theorem then tells us that for sufficiently large n this sampling distribution of $\hat{p}$ is approximately normal. That is what licenses the use of the standard-normal critical value z^* below. In practice we do not know p , so we plug in our sample estimate $\hat{p}$ and use $\sqrt{\hat{p}(1-\hat{p})/n}$ as the standard error.

7.5.2 The Formula

The confidence interval for a proportion is

$$\hat{p} \pm z^* \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Notice that this formula uses z^* rather than t^* . The standard error of a proportion is computed directly from $\hat{p}$ and n , so we do not have the extra layer of estimating a separate spread parameter the way we did for the mean. The z-distribution (standard normal) is appropriate here.

For a 95% confidence interval, $z^* = 1.96$. For a 90% interval, $z^* = 1.645$. For a 99% interval, $z^* = 2.576$.

7.5.3 A Worked Example

A polling firm surveys 1,200 likely voters and finds that 648 of them support a proposed infrastructure bill. What is a 95% confidence interval for the true proportion of likely voters who support the bill?

Step 1: Compute the sample proportion. $\hat{p} = \frac{648}{1200} = 0.54$.

Step 2: Compute the standard error. $SE = \sqrt{\frac{0.54 \times 0.46}{1200}} = \sqrt{\frac{0.2484}{1200}} = \sqrt{0.000207} = 0.01439$.

Step 3: Find the critical value. For 95% confidence, $z^* = 1.96$.

Step 4: Compute the margin of error. $ME = 1.96 \times 0.01439 = 0.0282$.

Step 5: Build the interval. 0.54 ± 0.0282 , which gives $(0.512, 0.568)$.

We are 95% confident that the true proportion of likely voters who support the bill is between 51.2% and 56.8%. Notice the margin of error is about ± 3 percentage points, which is typical for a national poll with a sample of about 1,200 people. Now you know where that “plus or minus 3 points” comes from whenever you hear it on the news.

7.5.4 Conditions for the Proportion Interval

Independence. Same as before. Observations must be independent, and the sample should be less than 10% of the population.

Success-failure condition. The approximation works well when the expected number of successes and failures are both at least 10. That means $n\hat{p} \geq 10$ and $n(1 - \hat{p}) \geq 10$. In the polling example, $1200 \times 0.54 = 648$ and $1200 \times 0.46 = 552$, both well above 10, so the condition is easily met.

When $\hat{p}$ is very close to 0 or 1, or the sample is very small, this condition may not be satisfied, and alternative methods are needed.

7.5.5 A Practical Illustration

To see how proportion intervals work in everyday life, consider a scenario outside politics. A large bakery chain wants to estimate what fraction of its customers would pay a dollar more for organic ingredients. They survey 500 customers at various locations and find that 210 say yes. The sample proportion is $\hat{p} = 210/500 = 0.42$.

The 95% confidence interval gives $0.42 \pm 1.96 \times \sqrt{\frac{0.42 \times 0.58}{500}} = 0.42 \pm 0.043$, or roughly (0.377, 0.463).

What does this tell the bakery's management? Somewhere between 38% and 46% of their customers would pay more for organic ingredients. That is a meaningful range. If the break-even point requires at least 45% of customers to pay the premium, the confidence interval tells management that the true proportion might be high enough, but it also might not. The interval includes values above and below 45%. Acting as if exactly 42% of customers will pay the premium, and building a business plan around that single number, ignores the uncertainty that the interval is designed to communicate. The smart move is to factor the entire range into the planning process, perhaps running a small pilot at a few locations before committing to the change chain-wide.

This kind of reasoning is available to anyone who takes confidence intervals seriously. The math is not complicated. The discipline lies in resisting the urge to treat the point estimate as settled fact.

Note

The “plus or minus 3 points” you hear in election coverage has become so standard that it borders on meaningless repetition. People hear it without absorbing it. A poll showing Candidate A at 49% and Candidate B at 47%, with a 3-point margin of error, is telling you that the race is basically a coin flip. But the headline will read “Candidate A Leads.” The margin of error is the most important number in any poll result, and it is the number most reliably ignored. Every time you see a poll, look at the margin of error first, the topline number second.

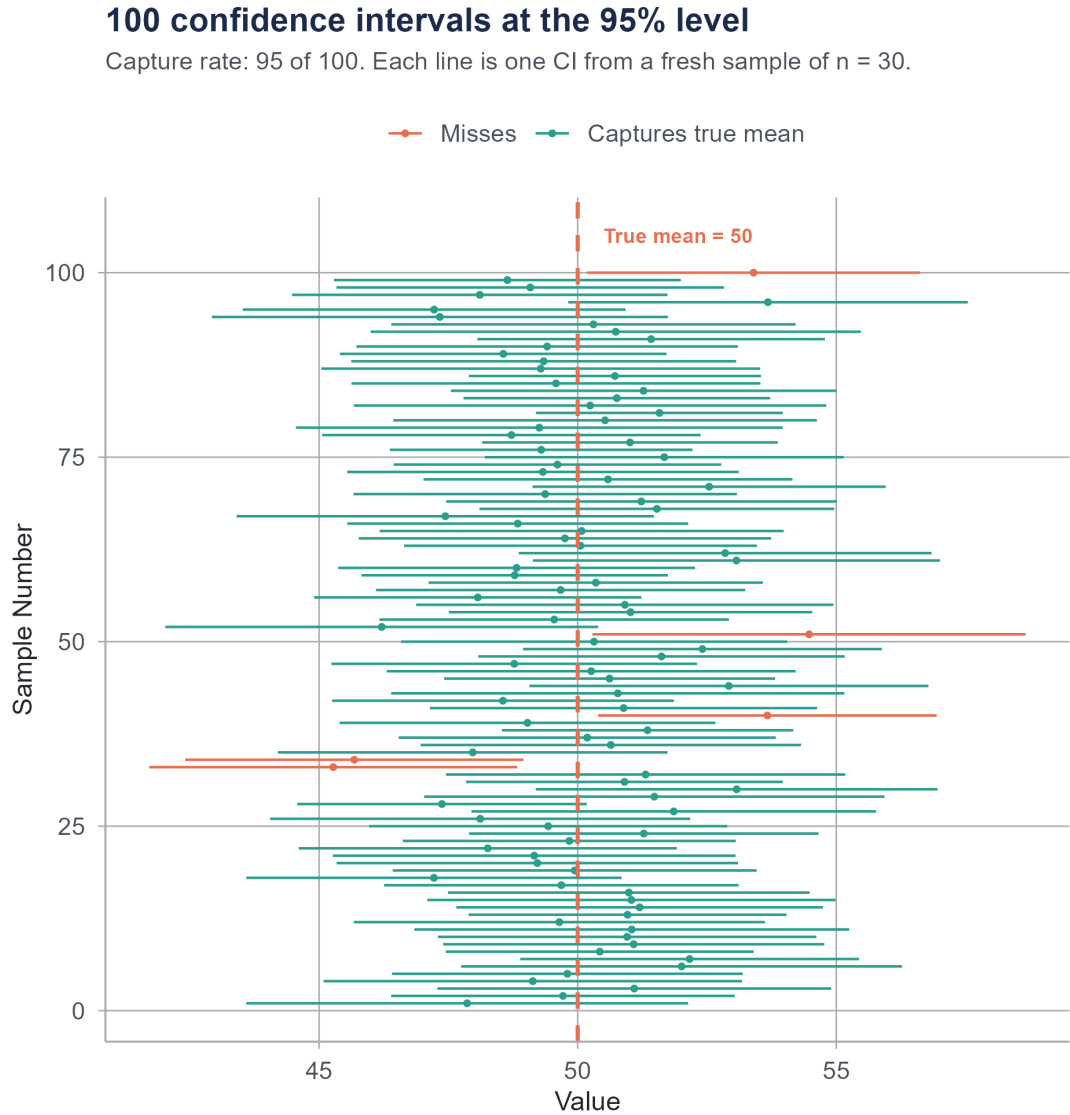
7.6 What “95% Confident” Actually Means

This is the section that matters most in this chapter. If you take only one thing from these pages, take this.

When we say we are “95% confident” that the population mean is between 42.2 and 52.2 minutes, we are **not** saying there is a 95% probability that the true mean falls in this interval.

Read that again. The most common interpretation of a confidence interval is wrong.

The true population mean, μ , is a fixed number. It is not random. It does not move around. Either μ is inside the interval (42.2, 52.2), or it is not. The probability is either 1 or 0, and we simply do not know which.



About 95 of every 100 such intervals contain the true mean.

Figure 7.1: One hundred 95% confidence intervals constructed from 100 independent random samples. Most intervals (shown in teal) capture the true population mean, while a few (shown in coral) miss it. This is what “95% confidence” means: about 95 out of every 100 intervals will contain the true parameter.

So what does “95% confidence” mean?

It means this: if we were to repeat the entire sampling process many, many times, each time drawing a new random sample of 36 patients and computing a new 95% confidence interval, then about 95% of those intervals would contain the true population mean. And about 5% would miss it.

The confidence level describes the long-run reliability of the method, not the probability of any single interval being correct. It is a statement about the procedure, not about the result.

Think of it like a factory that manufactures fishing nets. The factory claims that 95% of its nets are strong enough to hold a 20-pound fish. You buy one net. Does your specific net have a 95% chance of holding? No. Your net is either strong enough or it is not. The 95% describes the production process. Out of every 100 nets the factory makes, about 95 of them will hold and about 5 will tear. You just do not know which kind you got.

Similarly, your specific confidence interval either caught the true parameter or it did not. The “95%” tells you that the method you used to build it is one that catches the true parameter 95% of the time across many uses.

7.6.1 Why This Distinction Matters

You might think this is philosophical hairsplitting. It is not. The distinction has practical consequences.

If you believe “95% probability the parameter is in my interval,” you might treat the interval as near-certainty and make bold decisions based on it. If you understand that the interval was produced by a method with a 5% miss rate, you stay appropriately humble. One in twenty intervals produced this way misses the mark. And you cannot tell by looking at the interval whether this is one of the nineteen good ones or the one bad one.

This understanding also helps explain why different researchers studying the same question can get different confidence intervals, and why they sometimes disagree. Two polling firms survey the same electorate and get different results. Neither one is wrong. Both are working with different samples, and their intervals reflect the variability inherent in sampling. Sometimes one interval will contain the true value and the other will not, and that is exactly what a 5% miss rate predicts.

 Try It Online

Open the **Confidence Interval Simulator** on the companion website. Set the true population mean and generate 100 confidence intervals from random samples. Watch how most intervals capture the true mean, but a few miss. Change the confidence level from 95% to 90% to 99% and see how the miss rate changes. This is the single best way to internalize what confidence level means. Do not skip this one.

7.6.2 Common Misinterpretations (and Corrections)

Let us go through the most frequent ways people get confidence intervals wrong.

Wrong: “There is a 95% chance that the true mean is between 42.2 and 52.2.” **Correct:** “We are 95% confident that the true mean is between 42.2 and 52.2, meaning the method we used produces intervals that capture the true mean 95% of the time.”

Wrong: “95% of the data falls between 42.2 and 52.2.” **Correct:** The confidence interval is about the population mean, not about individual data points. Individual ER wait times could range from 5 minutes to 3 hours. The interval estimates where the average is, not where individual values are.

Wrong: “If we repeat the study, there is a 95% chance the new sample mean will fall in this interval.” **Correct:** A new sample mean will have its own sampling variability and could land anywhere. The interval estimates the population parameter, not future sample statistics.

Wrong: “This interval proves the population mean is not 40.” **Correct:** If 40 is outside the interval, the data provides evidence against 40 as the population mean. But a confidence interval does not “prove” anything. A different sample might yield a different interval that does include 40.

Wrong: “The confidence interval is so narrow that we basically know the exact value.” **Correct:** A narrow interval means the estimate is precise, but precision is not the same as accuracy. If the sampling method was biased (for instance, if the ER only tracked wait times for patients who checked in electronically, missing those who arrived by ambulance), the interval could be precisely wrong. It would precisely estimate the wrong quantity. Narrow intervals are nice, but they cannot fix problems that originate before the math begins.

These misinterpretations extend well beyond the classroom. They appear in published research, in news coverage, and in corporate boardrooms. A 2014 study by Hoekstra and colleagues surveyed researchers and students across multiple disciplines and found that a majority could not correctly

identify the meaning of a 95% confidence interval when given multiple-choice options. The people producing the intervals often misunderstand them just as badly as the people reading them. If that feels discouraging, consider the upside: understanding this one concept puts you ahead of a large portion of practicing professionals.

7.7 What Makes Confidence Intervals Wider or Narrower

Understanding what drives the width of a confidence interval gives you practical control over the precision of your estimates. Three factors determine the width, and each one works in an intuitive way.

7.7.1 Confidence Level

Wanting to be more confident means casting a wider net. A 99% confidence interval is wider than a 95% interval, which is wider than a 90% interval. The critical value (t^* or z^*) increases as the confidence level goes up.

For a z-based interval, z^* at 90% is 1.645, at 95% is 1.96, and at 99% is 2.576. Moving from 95% to 99% confidence increases the critical value by about 31%, which stretches the interval substantially. You are trading precision for confidence. Most of the time, 95% represents a reasonable compromise, which is why it has become the default in most fields.

7.7.2 Sample Size

Larger samples produce narrower intervals. The standard error has $\sqrt{n}$ in the denominator, so as n grows, the standard error shrinks, and the margin of error follows.

But notice the square root. Doubling your sample size does not cut the margin of error in half. It reduces it by a factor of $\sqrt{2} \approx 1.41$, which means about a 29% reduction. To halve the margin of error, you need four times the sample size. To cut it to one-third, you need nine times the data. This is the law of diminishing returns in sampling. The first hundred observations buy you a lot of precision. Going from 10,000 to 10,100 buys you almost nothing.

Bigger Samples, Smaller Margins (But with Diminishing Returns)

Margin of error = $1.96 \times \sigma / \sqrt{n}$ (95% confidence, $\sigma = 15$)

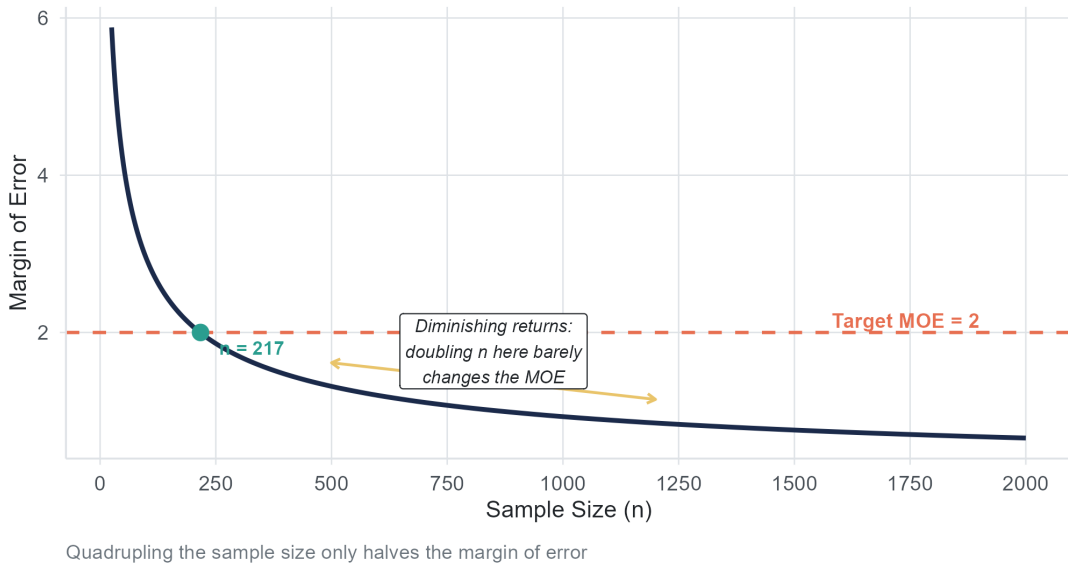


Figure 7.2: Margin of error plotted against sample size, showing the characteristic curve of diminishing returns. The margin of error drops steeply at first but flattens out for larger samples, illustrating why quadrupling the sample size only halves the margin of error.

7.7.3 Variability in the Data

More variability in the population, captured by s in the formula, leads to wider confidence intervals. This makes sense. If everyone in the population is similar, a small sample tells you a lot. If there is enormous variation, you need more data to pin down the center.

You have limited control over population variability. It is a property of the thing you are studying. Wait times in an emergency department are inherently more variable than, say, the weight of machine-produced bolts. But you can sometimes reduce variability by narrowing your population (studying wait times for a specific type of visit, for example) or by using more precise measurement instruments.

Here is a concrete way to feel the interplay of these three factors. Imagine you are trying to estimate the average household income in a medium-sized city. You survey 100 households and compute

a 95% confidence interval. It runs from \$52,000 to \$68,000, a range of \$16,000. You want the range to be narrower. You have three options, and they are not equally practical.

Option one: lower the confidence level from 95% to 90%. This narrows the interval, but now you are less confident it contains the true mean. You have traded safety for precision.

Option two: increase the sample size. If you survey 400 households instead of 100, you cut the standard error roughly in half, and the interval shrinks to maybe a \$8,000 range. This costs money and time, but it gives you both precision and the same level of confidence.

Option three: reduce variability. Instead of surveying all households, you could restrict your study to, say, households with two working adults. This group is likely more homogeneous in income, so s drops and the interval narrows. But now your interval estimates the mean income only for that subgroup, not for all households. You have gained precision by asking a narrower question.

Each option involves a tradeoff. Statistics is, at its heart, the art of managing these tradeoffs wisely.

7.7.4 A Summary Table

To make the interval...	You can...
Narrower (more precise)	Increase sample size
Narrower (more precise)	Accept a lower confidence level
Narrower (more precise)	Reduce variability (if possible)
Wider (more conservative)	Decrease sample size
Wider (more conservative)	Demand a higher confidence level
Wider (more conservative)	Study a more variable population

Try It Online

Return to the **Sampling Explorer** you first met in Chapter 2. Its confidence-interval and sample-size sections put a slider on each of the three knobs you have just read about — confidence level, sample size, and variability — and show the resulting margin of error live. Hold the population shape and the confidence level fixed, then drag the sample-size slider from 25 up to 400 and watch the margin of error fall, but at a slowing rate.

7.8 Choosing the Right Sample Size

In many practical situations, you know how precise you want your estimate to be before you collect the data. A pharmaceutical company needs to estimate a drug's effect within a certain margin. A polling firm promises a margin of error of plus or minus 3 percentage points. A quality engineer needs to estimate the average defect rate within 0.5 percentage points.

In all these cases, the question is the same. How large a sample do I need to achieve a desired margin of error?

7.8.1 Sample Size for a Mean

Starting from the confidence interval formula, the margin of error for a mean is

$$ME = t^* \times \frac{s}{\sqrt{n}}$$

If we solve for n , we get

$$n = \left(\frac{t^* \times s}{ME} \right)^2$$

In practice, at the planning stage we usually do not know s yet (we have not collected the data), and t^* depends on n through the degrees of freedom, creating a circular problem. The common approach is to use z^* instead of t^* for planning purposes and to estimate s from a pilot study, prior research, or an educated guess. The planning formula becomes

$$n = \left(\frac{z^* \times \sigma}{ME} \right)^2$$

where σ is your best prior estimate of the population standard deviation.

For example, suppose you want to estimate the average monthly grocery spending for families in a city within $\pm \$25$ at the 95% confidence level. From previous surveys, you believe the standard deviation is about \$180. Then

$$n = \left(\frac{1.96 \times 180}{25} \right)^2 = \left(\frac{352.8}{25} \right)^2 = (14.112)^2 = 199.1$$

Round up to $n = 200$. You need a sample of about 200 families. Always round up when computing required sample sizes, because rounding down gives you a margin of error slightly larger than what you wanted.

7.8.2 Sample Size for a Proportion

The margin of error for a proportion is

$$ME = z^* \times \sqrt{\frac{p(1-p)}{n}}$$

Solving for n gives

$$n = \frac{(z^*)^2 \times p(1-p)}{ME^2}$$

Here we need a prior estimate of p , the population proportion. If you have no idea what p is, the conservative approach is to use $p = 0.5$, because $p(1-p)$ is maximized at 0.5 and the resulting sample size will be the largest (most conservative) it could be. You will never underestimate the required sample size this way.

For instance, a polling organization wants a margin of error of ± 3 percentage points (0.03) at the 95% confidence level. Using $p = 0.5$,

$$n = \frac{(1.96)^2 \times 0.5 \times 0.5}{(0.03)^2} = \frac{3.8416 \times 0.25}{0.0009} = \frac{0.9604}{0.0009} = 1067.1$$

Round up to 1,068. This is why national polls typically survey around 1,000 to 1,200 people. That sample size gives a margin of error in the neighborhood of ± 3 percentage points at the 95% confidence level, regardless of the size of the population. A poll of 1,100 Americans has the same margin of error whether the country has 300 million people or 3 billion. The precision depends on the sample size, not on what fraction of the population you sample.

That last point surprises most people. It seems like a sample of 1,100 out of 330 million could not possibly tell you anything useful. But the mathematics says otherwise. The key is that each sampled person provides an independent piece of information about the population. As long as the sample is random, 1,100 independent data points constrain the population proportion to within about 3 percentage points with 95% confidence.

7.9 Connecting Back to the Polls

Now you have the tools to understand what happened in those election nights.

When a pre-election poll reports “Candidate A: 48%, Candidate B: 45%, margin of error ± 3.5 points,” here is what it is actually saying. The 95% confidence interval for Candidate A’s support is roughly 44.5% to 51.5%. The interval for Candidate B is roughly 41.5% to 48.5%. Those intervals overlap considerably. The poll is consistent with Candidate A winning comfortably, with Candidate B winning, or with a near tie.

But the headline says “Candidate A Leads.” And readers think the race is decided.

The 2016 and 2020 elections illustrated this gap between what the data said and what people heard. In 2016, many state-level polls were within their margins of error, but those margins of error were wide enough to include a Trump win. People who understood confidence intervals were less surprised. Nate Silver, who gave Trump a 29% chance, was widely mocked before the election for being too generous to Trump, and widely cited afterward for being the forecaster who got it “least wrong.” His model was not magic. It was just taking the uncertainty in the polls seriously.

In 2020, the polling errors were larger, pointing to real methodological challenges with reaching certain segments of the population. Even so, most results fell within the 95% confidence bands when you account for the full range of uncertainty including potential correlated errors across states. The lesson is not that polls are useless. The lesson is that the margin of error is not a decoration. It is the most informative part of the result.

There is a useful habit you can develop the next time you see a poll in the news. Before looking at who is “ahead,” look at the margin of error. Subtract the margin of error from the leading candidate’s number and add it to the trailing candidate’s number. If the result flips the lead, the poll is telling you the race is too close to call. If the lead survives even the worst-case interpretation, you have something more meaningful. This takes about five seconds and it transforms you from a passive consumer of poll headlines into someone who actually reads the data. Most people never bother. You should.

7.10 Ethics of Sample Size

We have talked about sample size as a matter of precision. But it is also a matter of ethics.

Consider a medical study designed to test whether a new drug reduces blood pressure. If the researchers enroll too few patients, the confidence interval for the drug's effect will be wide. So wide that even if the drug works, the study will not be able to detect the effect. The confidence interval will include both “the drug works” and “the drug does not work,” and the study will be inconclusive.

This is more than a statistical failure. It is an ethical one.

Ethics Moment

An **underpowered study**, one with a sample too small to detect the effect it is looking for, wastes resources and can cause real harm. Patients in the study were exposed to the risks of an experimental treatment without any chance of producing useful knowledge. Grant funding was consumed without returning answers. And the inconclusive results might discourage future research on a treatment that actually works, depriving future patients of a beneficial therapy.

Research ethics boards require investigators to justify their sample sizes before a study begins. A sample size calculation, based on the desired margin of error and the expected effect size, is more than a statistical formality. It is an ethical obligation. Running a study you know is too small to answer the question is a misuse of participants' trust, their time, and in many cases their bodies.

This principle extends beyond medicine. Business A/B tests with samples too small to detect meaningful differences waste company resources and delay decisions. Surveys with inadequate sample sizes in certain demographic groups can render those groups statistically invisible, leading to policies and products that do not account for their needs. Whenever you plan a study, asking “Is my sample large enough?” is more than a technical question. It is a question about whether you are being responsible with the resources and the people involved.

7.11 Putting It All Together

Let us revisit the structure of this chapter by walking through one more complete example that ties together all the pieces.

A chain of coffee shops wants to estimate the average amount customers spend per visit. They randomly sample 50 transactions from across their locations and record the amount spent. The sample mean is $\bar{x} = \$6.85$ and the sample standard deviation is $s = \$2.40$.

Constructing a 95% confidence interval for the mean.

1. **Point estimate.** $\bar{x} = 6.85$.
2. **Degrees of freedom.** $df = 50 - 1 = 49$.
3. **Critical value.** t^* for 95% confidence with 49 degrees of freedom is approximately 2.010.
4. **Standard error.** $SE = \frac{2.40}{\sqrt{50}} = \frac{2.40}{7.071} = 0.3394$.
5. **Margin of error.** $ME = 2.010 \times 0.3394 = 0.682$.
6. **Confidence interval.** $6.85 \pm 0.682 = (6.17, 7.53)$.

Interpretation. We are 95% confident that the average amount spent per visit across all customers at this coffee shop chain is between \$6.17 and \$7.53. This means that if we were to repeat this process many times, drawing new random samples of 50 transactions each time and computing a confidence interval, about 95% of those intervals would contain the true average spending amount.

What if the chain wanted a narrower interval? They could increase the sample size. If they sampled 200 transactions instead of 50, the standard error would drop to $\frac{2.40}{\sqrt{200}} = 0.1697$, and the margin of error would shrink to about 0.34, giving an interval roughly half as wide.

What if they wanted 99% confidence instead of 95%? The critical value would increase to about 2.680, and the margin of error would grow to $2.680 \times 0.3394 = 0.910$, giving a wider interval of (5.94, 7.76). More confidence, less precision.

Now suppose the company wants to plan a future, larger study. They want the margin of error to be no more than $\pm \$0.50$ at the 95% confidence level. Using the planning formula with $\sigma \approx 2.40$,

$$n = \left(\frac{1.96 \times 2.40}{0.50} \right)^2 = \left(\frac{4.704}{0.50} \right)^2 = (9.408)^2 = 88.5$$

They need at least 89 transactions. This is manageable. If they wanted $\pm \$0.25$, they would need

$$n = \left(\frac{1.96 \times 2.40}{0.25} \right)^2 = (18.816)^2 = 354.0$$

At least 354 transactions. Notice how halving the desired margin of error nearly quadrupled the required sample.

 Try It Online

Two apps work well together as you finish this chapter.

- The **Confidence Interval Simulator** is the place to feel what “95% confidence” actually means — generate hundreds of intervals and record what fraction capture the true parameter at the 90%, 95%, and 99% levels.
- The **Sampling Explorer**’s sample-size calculator is the place to plan a study before you collect data — pick the kind of estimate, the desired margin of error, and the confidence level, and the app tells you the required n that the formulas in this section produce.

7.12 A Few Final Cautions

Confidence intervals are only as good as your data. A confidence interval based on a biased sample is a precise estimate of the wrong thing. If the coffee shop only sampled transactions from its busiest location on Saturday mornings, the interval would be precise but misleading. Random sampling matters.

Confidence intervals assume the conditions are met. If the data is not independent, or if the sample is too small and the distribution is heavily skewed, the interval’s stated confidence level may not match its actual reliability. Always check conditions.

A narrow confidence interval does not mean the estimate is correct. It means the estimate is precise. Precision and accuracy are different things. You can be precisely wrong if your sampling method is biased or your measurement is flawed. A confidence interval quantifies sampling variability. It does not account for other sources of error, like measurement bias, nonresponse bias, or data processing mistakes.

Do not confuse statistical confidence with practical importance. A 95% confidence interval of (0.01, 0.03) for the difference in click-through rates between two website designs is very precise and estimates the true difference to fall between 1 and 3 percentage points. Whether that difference matters for your business depends on context, not on the confidence level.

AI Reality Check

Ask an AI tool to analyze a dataset and it will almost always give you a point estimate. “The average is 47.3.” Sometimes it will even give you a confidence interval. But watch what it does with that interval.

AI tools frequently report confidence intervals without checking whether the conditions for constructing them are met. They may compute a confidence interval for a mean from a

sample of 12 observations drawn from a heavily skewed distribution, where the t -interval is unreliable. They may construct a confidence interval for a proportion when the sample size is too small for the normal approximation (the success-failure condition fails). They may even generate a confidence interval from a convenience sample and present it as if it quantifies uncertainty about the population, when in fact the biggest source of uncertainty is the non-random sampling, which no formula can fix.

Perhaps more subtly, AI tools tend to present confidence intervals as final answers rather than as starting points for interpretation. A confidence interval is not a conclusion. It is a tool for thinking about uncertainty. Whether the interval is wide enough to change your decision, whether it includes practically meaningful values, whether the sample it is based on actually represents the population you care about, these are judgment calls that require context the AI does not have. Use AI to compute the interval. Use your own thinking to decide what it means.

7.13 Looking Ahead

This chapter introduced the first formal tool of statistical inference: the confidence interval. You now know how to move from a single sample statistic to a range of plausible values for a population parameter, and you understand what “95% confident” actually means (and what it does not mean). But confidence intervals answer only one kind of question: “What are the plausible values?” They do not answer “Is there evidence against a specific claim?” That is the subject of hypothesis testing. In the next chapter, we will learn to formalize the question “Could this result have happened by chance?” and develop a rigorous framework for deciding when the data provides evidence against a null hypothesis. The logic of hypothesis testing will draw directly on the sampling distributions and standard errors we have built up over the past two chapters.

7.14 Further Reading and References

The following works are cited in this chapter or provide valuable additional context for the ideas covered here.

On the interpretation of confidence intervals: Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bul-*

letin & Review, 21(5), 1157–1164. (A study documenting how frequently researchers misinterpret confidence intervals, even among those with statistical training.)

On polling methodology and election forecasting: Silver, N. (2012). *The signal and the noise: Why so many predictions fail — but some don't*. Penguin. (An accessible treatment of forecasting, uncertainty, and the role of probability in interpreting polls, written by the founder of FiveThirtyEight.)

Kennedy, C., Blumenthal, M., Clement, S., Clinton, J. D., Durand, C., Franklin, C., McGeeney, K., Miringoff, L., Olson, K., Rivers, D., Saad, L., Witt, G. E., & Wlezien, C. (2018). An evaluation of the 2016 election polls in the United States. *Public Opinion Quarterly*, 82(1), 1–33. (The American Association for Public Opinion Research's official post-mortem on what went wrong with the 2016 polls and what it means for margin-of-error reporting.)

On sample size planning and ethical dimensions: Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician*, 55(3), 187–193. (A concise discussion of the practical and ethical considerations involved in choosing a sample size, emphasizing that underpowered studies waste resources and participants' contributions.)

For further reading on confidence intervals: Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge. (A clear, visual approach to confidence intervals and effect sizes that argues for moving beyond p -values toward interval estimation as the primary tool of inference.)

7.15 Key Terms

- **Point estimate:** A single value computed from sample data that serves as the best guess for a population parameter. Examples include the sample mean $\bar{x}$ and the sample proportion $\hat{p}$.
- **Interval estimate:** A range of plausible values for a population parameter, constructed from sample data. A confidence interval is the most common type.
- **Confidence interval:** A range of plausible values for a population parameter, constructed from sample data as (point estimate $\pm$ margin of error). A 95% confidence interval means that 95% of intervals constructed this way would contain the true parameter.
- **Margin of error:** The amount added to and subtracted from the point estimate to form the confidence interval. Equals the critical value times the standard error.

- **Standard error:** The estimated standard deviation of a sampling distribution. Measures how much a sample statistic would vary across repeated samples. For a mean, $SE = \frac{s}{\sqrt{n}}$. For a proportion, $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.
 - **Critical value:** The multiplier (t^* or z^*) that determines how many standard errors wide the confidence interval extends. Depends on the desired confidence level and, for the t -distribution, the degrees of freedom.
 - **t -distribution:** A family of bell-shaped distributions, indexed by degrees of freedom, used for inference about means when the population standard deviation is unknown. Has heavier tails than the normal distribution, especially at low degrees of freedom.
 - **Degrees of freedom:** A parameter of the t -distribution equal to $n - 1$ for a single-sample mean. Reflects the amount of independent information available for estimating variability.
 - **Confidence level:** The long-run proportion of confidence intervals that would contain the true parameter if the sampling process were repeated many times. Common levels are 90%, 95%, and 99%.
 - **Sample size determination:** The process of calculating how large a sample is needed to achieve a desired margin of error at a given confidence level.
 - **Underpowered study:** A study with a sample too small to detect the effect of interest with reasonable confidence, raising both statistical and ethical concerns.
-

7.16 Exercises

7.16.1 Check Your Understanding

1. Explain in your own words why a point estimate alone is insufficient for drawing conclusions about a population parameter.
2. Write out the general form of a confidence interval. Identify each component and explain what it contributes.
3. A 95% confidence interval for the average commute time in a city is (22.4, 28.6) minutes. A student says, “There is a 95% probability that the true average commute time is between 22.4 and 28.6 minutes.” What is wrong with this interpretation? Provide a correct one.
4. Explain why we use the t -distribution rather than the z -distribution when constructing a confidence interval for a mean. Under what circumstances does the choice between t and z matter most?

5. What are degrees of freedom, and why does a confidence interval for a single mean use $n - 1$ degrees of freedom?
6. List three factors that affect the width of a confidence interval. For each one, state whether increasing it makes the interval wider or narrower, and explain why.
7. A researcher constructs a 90% confidence interval and a 99% confidence interval from the same data. Which interval is wider? Why?
8. A poll reports that 52% of voters support a ballot measure, with a margin of error of ± 4 percentage points. A news anchor says, “The measure has majority support.” Is this claim justified by the poll? Explain.
9. Suppose you double the sample size while everything else stays the same. By what factor does the margin of error change? Show your reasoning.
10. Explain what the success-failure condition is for a confidence interval for a proportion, and why it is necessary.

7.16.2 Apply It

1. A random sample of 25 college students reports the following weekly study hours. The sample mean is $\bar{x} = 14.3$ hours and the sample standard deviation is $s = 4.1$ hours. Construct a 95% confidence interval for the mean weekly study hours of all college students at this university. Show all steps and interpret the result.
2. In a sample of 800 adults, 328 report that they have visited a dentist in the past six months. Construct a 99% confidence interval for the true proportion of adults who have visited a dentist in the past six months. Verify that the conditions for the interval are met.
3. A food safety inspector takes a random sample of 40 chicken breast packages from a processing plant and measures the weight of each. The sample yields $\bar{x} = 1.02$ pounds and $s = 0.08$ pounds. The packages are labeled as containing 1.00 pound. Construct a 95% confidence interval for the mean weight. Based on your interval, does the evidence suggest the packages are accurately labeled?
4. A technology company surveys 1,500 of its users and finds that 63% are satisfied with the latest software update. Construct a 95% confidence interval for the true proportion of satisfied users. If the company’s goal was to achieve at least 60% satisfaction, what can you conclude?
5. A researcher wants to estimate the average daily screen time for teenagers in the United States within ± 15 minutes at the 95% confidence level. Previous studies suggest the standard deviation of daily screen time is about 90 minutes. How large a sample does the researcher need?

6. A political campaign wants to estimate the proportion of voters who recognize the candidate's name, with a margin of error of no more than ± 2 percentage points at the 95% confidence level. The campaign has no prior estimate of name recognition. What sample size is needed?
7. A hospital collects data on the recovery time (in days) for patients after a certain surgery. A sample of 18 patients has a mean recovery time of $\bar{x} = 5.8$ days and a standard deviation of $s = 1.9$ days. Construct a 90% confidence interval for the mean recovery time. What additional assumption should you check given the relatively small sample size?
8. Two polling organizations survey the same electorate. Poll A, based on $n = 1,000$, reports 48% support for a candidate with a margin of error of $\pm 3.1\%$. Poll B, based on $n = 2,500$, reports 51% support with a margin of error of $\pm 2.0\%$. The two confidence intervals overlap. Are the polls contradicting each other? Explain.
9. A quality control manager samples 60 light bulbs and measures their lifetimes. The sample mean is $\bar{x} = 1,020$ hours and $s = 85$ hours. Construct 90%, 95%, and 99% confidence intervals for the true mean lifetime. Present all three intervals and comment on how the width changes as the confidence level increases.
10. A nonprofit organization wants to estimate the proportion of residents in a neighborhood who would use a proposed community garden. They conduct a pilot survey of 50 residents and find that 34 would use the garden. (a) Construct a 95% confidence interval based on the pilot survey. (b) The organization considers this interval too wide and wants a margin of error of ± 5 percentage points. Using the pilot estimate as a guide, how many residents should they survey?

The following three exercises use the `polling-data.csv` dataset available on the companion website. This is real polling data drawn from FiveThirtyEight's pollster ratings dataset. The dataset contains 50 observations with variables including `poll_id`, `pollster` (polling organization), `date`, `sample_size`, `candidate_a_pct`, `candidate_b_pct`, `margin_of_error`, and `confidence_level`.

11. Using the `polling-data.csv` dataset, select any single poll and verify its reported margin of error. Using the reported `candidate_a_pct` as your sample proportion $\hat{p}$ and the poll's `sample_size` as n , compute the margin of error for a 95% confidence interval using the formula $ME = z^* \sqrt{\hat{p}(1 - \hat{p})/n}$. How close is your calculated margin of error to the value reported in the `margin_of_error` column? What might explain any difference?
12. Compute the mean and standard deviation of `candidate_a_pct` across all 50 polls. Construct a 95% confidence interval for the mean of `candidate_a_pct` across polls. Then identify how many of the 50 individual polls have a `candidate_a_pct` value outside this interval. Is the observed count consistent with what you would expect? Now create a scatter plot of `sample_size` (hori-

zontal axis) versus `margin_of_error` (vertical axis). Describe the pattern. Does the relationship match what the formula $ME \propto 1/\sqrt{n}$ predicts?

- Sort the polls by `sample_size`. Compare the five polls with the largest sample sizes to the five polls with the smallest sample sizes. For each group, compute the mean and standard deviation of `candidate_a_pct`. Which group shows more variability in its estimates? Construct a 95% confidence interval for the mean of `candidate_a_pct` within each group. Which group's interval is narrower? Explain why, using the concept of the sampling distribution.

7.16.3 Think Deeper

- A news article reports, "A new study finds that the average American spends 2.5 hours per day on social media, with a 95% confidence interval of (2.3, 2.7) hours." The study was conducted using an online survey promoted through social media platforms. Discuss at least two reasons why the confidence interval, despite being correctly calculated, might not give a trustworthy estimate of the true average for all Americans. What types of error does a confidence interval account for, and what types does it not?
- During the COVID-19 pandemic, early estimates of the virus's infection fatality rate varied widely across studies, partly because different studies had very different sample sizes and sampling methods. Some early estimates from small samples had very wide confidence intervals. How should policymakers use confidence intervals when making urgent decisions with limited data? Is it better to act on a point estimate from a small study or to wait for more precise data? What are the risks of each approach?
- Consider two studies investigating the same question: whether a job training program increases employment rates among participants. Study A enrolls 40 participants and finds an employment rate of 68%, with a 95% confidence interval of (53%, 83%). Study B enrolls 500 participants and finds an employment rate of 61%, with a 95% confidence interval of (57%, 65%). Study A's point estimate is higher, but Study B's interval is much narrower. Which study provides more useful evidence for deciding whether to expand the program? What does each study tell you that the other does not? How does sample size relate to the ethical obligation to produce actionable knowledge?
- A pharmaceutical company runs a clinical trial with 5,000 participants and finds that its new drug lowers cholesterol by an average of 3 mg/dL, with a 95% confidence interval of (2.1, 3.9) mg/dL. The interval does not include zero, so the effect is statistically distinguishable from no effect. However, physicians consider a reduction of less than 10 mg/dL to be clinically meaningless. Discuss the difference between statistical confidence and practical importance.

Should this drug be brought to market based on this evidence? What additional information would you want?

5. Return to the opening story about election polling. Some commentators have suggested that polls should stop reporting a single topline number and margin of error, and instead only report the full confidence interval or a probability distribution of possible outcomes. What would be gained and what would be lost by this approach? Consider the needs of different audiences, including voters, journalists, campaign strategists, and election administrators. Is there a tension between statistical accuracy and public communication?

Chapter 8

Hypothesis Testing

8.1 Did It Actually Work?

In 2007, a company called Opower began partnering with electric utilities across the United States to send home energy reports to residential customers. The reports were simple. Each household received a letter showing how much electricity it used compared to its neighbors, complete with smiley faces for efficient homes and gentle nudges for the less efficient ones. The idea drew on decades of behavioral science research suggesting that social comparisons can motivate people to change their behavior. If you knew that your neighbors were using less electricity than you, the theory went, you might turn off a few lights, adjust your thermostat, or replace that ancient refrigerator.

The program was a nonprofit's dream, a low-cost intervention that could scale to millions of households. But before anyone invested further, there was a question that needed answering.

Did it work?

That question is harder than it sounds.

Hunt Allcott, an economist then at NYU, set out to answer it rigorously. He studied Opower's program across multiple utility companies serving hundreds of thousands of households. The study design was careful. In each utility service area, households were randomly assigned to either receive the energy reports (the treatment group) or not receive them (the control group). Then Allcott compared electricity usage between the two groups over time.

The data showed that households receiving the reports used, on average, about 2% less electricity than households in the control group. Two percent. That was the number.

Now, two reactions are possible here. The first is to shrug. Two percent sounds small. Is it even worth the cost of printing and mailing all those reports? The second reaction is to ask a more fundamental question, one that comes before any discussion of whether the effect is big enough to care about.

Is the difference real?

Those households that got the reports used 2% less energy, on average, than those that did not. But averages vary from sample to sample. Even if the reports had absolutely zero effect on behavior, you would not expect the treatment and control groups to use exactly the same amount of electricity. Random variation alone would produce some difference. Maybe the treatment group just happened to include more naturally frugal people. Maybe the weather was different in the areas where reports were delivered. Maybe the 2% difference is nothing more than the kind of random fluctuation you would expect even if the reports did absolutely nothing.

This is the question that hypothesis testing was built to answer. Not “is the effect big enough to matter” (that comes later), but “is there an effect at all, or are we fooling ourselves?”

Allcott’s analysis concluded that the effect was real and very unlikely to have arisen by chance alone. His 2011 paper, published in the *Journal of Public Economics*, became widely cited in behavioral economics. The Opower program went on to reach more than 50 million households, and the company was acquired by Oracle in 2016. But to understand why Allcott’s conclusion was justified, and when similar conclusions are not, you need to understand the logic, machinery, and limitations of hypothesis testing.

That is what this chapter is about. It is also the densest chapter in this book, and the one whose ideas show up most often in subsequent chapters.

8.2 The Logic of Hypothesis Testing

8.2.1 Starting with a Question

Every hypothesis test begins with a question about a population. Does this drug lower blood pressure? Do customers who see the redesigned website buy more? Did the energy reports reduce electricity consumption?

We cannot observe the entire population, so we collect data from a sample and use that sample to make an inference about the population. We have been building toward this since Chapter 1.

Hypothesis testing formalizes the process by forcing us to state our question precisely, collect evidence, and evaluate that evidence against a specific standard.

8.2.2 Null and Alternative Hypotheses

The formal framework requires two competing statements.

The **null hypothesis**, written H_0 , is the default position. It states that nothing interesting is happening, that there is no effect, no difference, no relationship. In the Opower example, the null hypothesis would be that the energy reports have no effect on electricity usage, that the true mean difference in consumption between the treatment and control groups is zero.

$$H_0 : \mu_{\text{treatment}} - \mu_{\text{control}} = 0$$

The **alternative hypothesis**, written H_a (or sometimes H_1), is the claim that something is going on. It states that there is an effect, a difference, a relationship. For Opower, the alternative hypothesis is that the reports do change electricity consumption.

$$H_a : \mu_{\text{treatment}} - \mu_{\text{control}} \neq 0$$

Notice the structure. The null hypothesis is specific (the difference is exactly zero), while the alternative hypothesis is broad (the difference is anything other than zero). This asymmetry is deliberate and important. We do not try to prove the alternative hypothesis. Instead, we ask whether the data provides enough evidence to reject the null hypothesis. The burden of proof falls on the claim that something is happening.

8.2.3 The Courtroom Analogy

The most common way to explain this logic is through a courtroom analogy, and it is useful enough to walk through, even though it has limitations we should be honest about.

In a criminal trial in the United States, a defendant is presumed innocent until proven guilty. The prosecution must present evidence beyond a reasonable doubt to convince the jury to convict. If the evidence is insufficient, the jury returns a verdict of “not guilty.” Note that “not guilty” is not the same as “innocent.” It means the prosecution failed to meet its burden.

Hypothesis testing follows a similar structure.

- The null hypothesis (H_0) is like the presumption of innocence. It stands unless the evidence overturns it.
- The alternative hypothesis (H_a) is like the prosecution's claim that the defendant is guilty.
- The data is the evidence presented at trial.
- The p -value (which we will define shortly) measures the strength of the evidence against the null hypothesis.
- Rejecting H_0 is like a guilty verdict. We have enough evidence to act on the claim.
- Failing to reject H_0 is like a not guilty verdict. We do not have enough evidence to reject the default position.

This analogy captures the key asymmetry. We never “accept” the null hypothesis, just as a jury never declares a defendant innocent. We either reject it or we fail to reject it. The absence of evidence against the null is not the same as evidence for the null.

i Note

The phrasing “fail to reject H_0 ” sounds awkward, and students sometimes wonder why we do not just say “accept H_0 .” The reason matters. Failing to reject H_0 means we did not find enough evidence against it. Maybe the null is true. Or maybe there is an effect, but our sample was too small to detect it, or our measurement was too noisy. We simply cannot tell. Saying “accept H_0 ” implies a confidence we do not have.

8.2.4 Where the Analogy Breaks Down

The courtroom analogy is helpful but imperfect, and glossing over its limitations leads to misunderstanding.

First, criminal trials use “beyond a reasonable doubt” as the standard. Hypothesis testing uses a numerical significance threshold — almost always 5% — which is a much lower bar than what a criminal court would accept. In a criminal trial, we tolerate a high standard because convicting an innocent person is a serious error. In statistics, the threshold depends on context, and the 5% level is a convention, not a law of nature. We will define this threshold formally as α in just a moment.

Second, juries weigh qualitative evidence, witness credibility, narratives, and context. Hypothesis testing reduces the evidence to a single number (the p -value). Important context can get lost in that reduction.

Third, and most critically, the courtroom analogy encourages binary thinking, guilty or not guilty, reject or fail to reject. But the strength of evidence exists on a continuum. A p -value of 0.049 and a p -value of 0.051 are practically identical, yet one crosses a conventional threshold and the other does not. Treating these as fundamentally different is one of the most persistent problems in statistical practice.

We will return to this problem. For now, let us look at how the evidence is actually measured.

8.3 Test Statistics and P-Values

8.3.1 The Test Statistic

Suppose you want to test whether a coin is fair. You flip it 100 times and get 58 heads. Is the coin biased? Your sample proportion is 0.58, which is not 0.50, but samples vary. Even a perfectly fair coin would not give exactly 50 heads every time.

The question is whether 58 heads out of 100 is unusual enough to cast doubt on the claim that the coin is fair. To answer that, we need to quantify “unusual enough.”

A **test statistic** is a number that measures how far our sample result falls from what the null hypothesis predicts. The specific formula depends on the type of test, but the logic is always the same. The test statistic captures the discrepancy between what we observed and what we would expect if H_0 were true, scaled by how much variability we would expect in our samples.

In its most general form, a test statistic looks like this:

$$\text{Test Statistic} = \frac{\text{Observed value} - \text{Expected value under } H_0}{\text{Standard error}}$$

A large test statistic (far from zero) means the data is far from what the null hypothesis predicts. A small test statistic (close to zero) means the data is consistent with the null.



Which Do I Need: SD or SE?

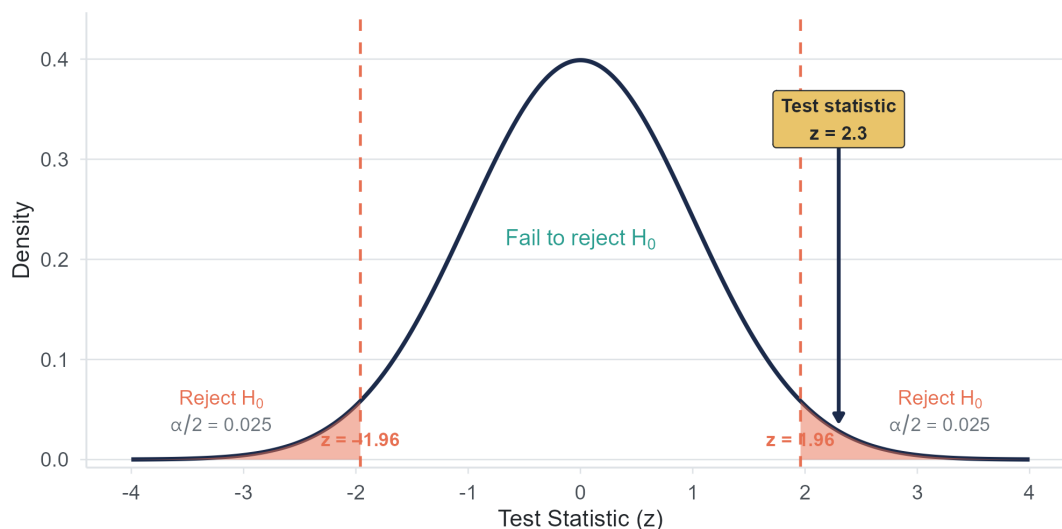
Standard deviation (SD) describes how much individual observations vary around the mean. It tells you about the spread in your data.

Standard error (SE) describes how much a sample statistic (like the mean) would vary if you repeated the study. It tells you about the precision of your estimate.

Use SD when describing your data. Use SE when making inferences about a population.

The Logic of Hypothesis Testing

Null distribution ($H_0: \mu = 0$) with two-tailed rejection regions ($\alpha = 0.05$)



The test statistic $z = 2.3$ falls in the rejection region, so we reject the null hypothesis

Figure 8.1: The null distribution (standard normal) with two-tailed rejection regions shaded in coral. A test statistic of $z = 2.3$ falls in the rejection region beyond the critical value of 1.96, leading to rejection of the null hypothesis at the 0.05 significance level.

8.3.2 What Is a P-Value?

The ***p*-value** is the probability of observing a test statistic as extreme as, or more extreme than, the one we actually calculated, assuming the null hypothesis is true.

Read that definition again. Slowly. It is a specific, technical definition, and every word matters. Many smart people get this wrong, and the consequences of getting it wrong are not academic.

The *p*-value answers this question: “If there were truly no effect (if H_0 were true), how likely would we be to see data at least this extreme?”

Think of it this way. Imagine you could rerun the study thousands of times in a universe where H_0 is true, where the treatment has no effect at all. In some of those repetitions, you would see small differences between groups. In a few, you would see larger differences, purely by luck. The *p*-value tells you what fraction of those repetitions would produce a result as extreme or more extreme than what you actually observed.

If the p -value is very small, it means that data like ours would be very unlikely under the null hypothesis. We take this as evidence against the null. If the p -value is not small, it means that data like ours would be quite plausible even if there were no effect. In that case, we have no compelling reason to reject the null.

For the coin example, the p -value would answer this: “If the coin were perfectly fair, what is the probability of getting 58 or more heads (or 42 or fewer) in 100 flips?” If that probability is very low, we have reason to suspect the coin is not fair.

8.3.3 The Significance Level

We need a threshold for “small enough.” The **significance level**, denoted α (alpha), is the probability threshold below which we reject H_0 . By convention, $\alpha = 0.05$ is the most common choice, meaning we reject the null hypothesis if the p -value is less than 0.05.

If that number looks familiar, it should. In Chapter 7, the same value appeared from the other side: a 95% confidence level corresponds to an α of $1 - 0.95 = 0.05$. Confidence intervals and hypothesis tests are using the same dial, just labeled differently. A two-sided test at level α rejects H_0 for exactly the values that fall outside a $(1 - \alpha)$ confidence interval.

The decision rule is straightforward:

- If $p \leq \alpha$, reject H_0 . The result is “statistically significant.”
- If $p > \alpha$, fail to reject H_0 . The result is “not statistically significant.”

Where did 0.05 come from? Ronald Fisher introduced it in his 1925 book *Statistical Methods for Research Workers*, where he proposed the 5% level as a convenient cutoff — “1 in 20 chance of being exceeded by chance.” Fisher framed it as a rule of thumb rather than a strict rule, and later in his career he was clearer that the threshold should depend on context. There is nothing magical about it. In some fields, a stricter threshold is standard (particle physics uses roughly $\alpha = 0.0000003$, or “five sigma,” before claiming a discovery). In others, a more relaxed threshold might be appropriate. The choice of α should be made before looking at the data and should reflect the consequences of making the wrong decision.

i Note

In 2018, a group of 72 prominent researchers published a paper in *Nature Human Behaviour* proposing that the default significance threshold be lowered from 0.05 to 0.005 for claims of new discoveries. Their argument was that the 0.05 threshold produces too many false

positives, contributing to the replication crisis. The proposal sparked intense debate and has not been universally adopted, but it highlights that the choice of α is a judgment call with real consequences for science.

8.3.4 What a P-Value Is NOT

The p -value is one of the most misunderstood concepts in all of statistics. Here are the most common misinterpretations, all of them wrong.

Wrong: “The p -value is the probability that the null hypothesis is true.” No. The p -value is calculated assuming the null hypothesis is true. It cannot tell you the probability that H_0 is true or false. That would require a fundamentally different framework (Bayesian statistics, which we touch on in Chapter 12).

Wrong: “A p -value of 0.03 means there is a 3% chance the results are due to chance.” This sounds close to the correct definition but subtly reverses the logic. The p -value is the probability of the data given the null, not the probability of the null given the data. These are not the same thing, just as the probability that a dog has four legs is not the same as the probability that a four-legged creature is a dog.

Wrong: “A smaller p -value means a bigger effect.” No. A tiny p -value can come from a huge effect in a small sample or from a tiny effect in a huge sample. The p -value mixes together the size of the effect and the size of the sample. To understand the size of the effect, you need a different measure (which we will cover shortly).

Wrong: “A p -value greater than 0.05 means there is no effect.” No. It means we did not find sufficient evidence against the null hypothesis. As Carl Sagan put it, the absence of evidence is not evidence of absence. The effect might exist but be too small for our sample to detect.

Wrong: “A p -value of 0.05 means the result has a 95% chance of being correct.” No. This confuses the p -value with the confidence level and misinterprets both.

If you take nothing else from this section, take this: a p -value is a measure of the compatibility between the data and a specific null hypothesis. It is not a measure of truth, importance, or certainty.

8.4 Type I and Type II Errors

When we make a decision about H_0 , we can be right or we can be wrong, and there are exactly two ways to be wrong.

8.4.1 Type I Error (False Positive)

A **Type I error** occurs when we reject H_0 even though it is actually true. We conclude there is an effect when there is not one. This is a false alarm.

The probability of a Type I error is α , the significance level we chose. When we set $\alpha = 0.05$, we are accepting a 5% chance that we will reject a true null hypothesis. This is not a bug in the system; it is a conscious tradeoff. We accept some risk of false positives in order to have a reasonable chance of detecting real effects.

8.4.2 Type II Error (False Negative)

A **Type II error** occurs when we fail to reject H_0 even though it is actually false. There is a real effect, but we missed it. This is a miss.

The probability of a Type II error is denoted β (beta). Unlike α , which we set in advance, β depends on several factors: how big the true effect is, how large our sample is, and how much variability there is in the data.

8.4.3 Power

Power is the probability of correctly rejecting a false null hypothesis. It is the complement of β :

$$\text{Power} = 1 - \beta$$

Power is the probability of detecting an effect that actually exists. A study with low power is like a metal detector with a weak signal: even if there is something buried there, you might walk right over it and hear nothing.

What determines power?

- **Effect size.** Bigger effects are easier to detect. A drug that lowers blood pressure by 20 points is easier to detect than one that lowers it by 2 points.
- **Sample size.** Larger samples give more precise estimates, making it easier to distinguish a real effect from noise.
- **Variability.** Less variability in the data (smaller standard deviations) makes effects easier to detect.
- **Significance level.** A more lenient α (say, 0.10 instead of 0.05) increases power but also increases the risk of Type I error. You cannot reduce both error types simultaneously without increasing sample size.

A common target for power is 0.80 (80%), meaning an 80% chance of detecting the effect if it exists. This is a convention, not a rule. For high-stakes studies, higher power might be warranted. For preliminary or exploratory work, lower power might be acceptable.

8.4.4 Why Power Matters in Practice

Underpowered studies are a recurring problem in published research, with consequences that reach beyond the academic literature. An underpowered study is one with too small a sample to have a reasonable chance of detecting the effect it is looking for. These studies are common. A review by Button and colleagues (2013) found that the median neuroscience study had statistical power of about 21% — only a 1 in 5 chance of detecting a real effect. Studies like these were unlikely to produce a significant result from the start, even when the underlying effect was real.

This creates two problems. First, the non-significant results are uninformative: we cannot tell whether the treatment did not work or whether the study was just too small to tell. Second, and more insidiously, the studies that do manage to find significant results in underpowered conditions are likely to be overestimates of the true effect size. This is because only the studies where random variation happened to inflate the effect will cross the significance threshold. It is a selection effect. The published literature ends up populated with inflated effects that fail to replicate, which is exactly what the replication crisis looks like from the inside.

The solution is simple in theory, if sometimes difficult in practice. Before you run a study, do a power analysis. Determine how large a sample you need to have a reasonable chance of detecting the smallest effect you would consider important. If you cannot afford a sample that large, be honest about the limitations of your study rather than running it anyway and treating a non-significant result as evidence that nothing is going on.

The following table summarizes the four possible outcomes:

	H_0 is True	H_0 is False
Reject H_0	Type I Error (α)	Correct Decision (Power = $1 - \beta$)
Fail to Reject H_0	Correct Decision ($1 - \alpha$)	Type II Error (β)

Think of it this way. Lowering α reduces your risk of false positives but increases your risk of false negatives. You become harder to fool, but you also become more likely to miss real effects. The only free lunch is a bigger sample, which reduces both error rates simultaneously.

Four Possible Outcomes in Hypothesis Testing

Every hypothesis test risks two types of errors

		REALITY	
		H0 is Actually False	H0 is Actually True
DECISION	Reject H0	Correct Decision (True Positive) Power = $1 - \beta$	Type I Error (False Positive) Probability = α
	Fail to reject H0	Type II Error (False Negative) Probability = β	Correct Decision (True Negative) Probability = $1 - \alpha$

Type I Error (α): Rejecting a true null hypothesis | Type II Error (β): Failing to reject a false null hypothesis

Figure 8.2: A 2x2 grid showing the four possible outcomes in hypothesis testing. Correct decisions appear on the diagonal (true positive with power = 1 minus beta, and true negative with probability 1 minus alpha), while errors appear off the diagonal (Type I error with probability alpha, and Type II error with probability beta).

Try It Online

Explore how sample size, effect size, and significance level affect power using the **Hypothesis Testing Playground**. Increase the sample size and watch the power curve shift. Then change the true effect size and see how the overlap between the null and alternative distributions changes.

8.5 Statistical Significance vs. Practical Significance

This is where many analysts, journalists, and even researchers go wrong.

A result is **statistically significant** if the p -value is below the chosen α threshold. That tells you the observed difference is unlikely to be entirely due to random chance. It does not tell you the difference is big, important, or worth acting on.

Return to the Opower study. The energy reports produced a statistically significant reduction in electricity consumption. The p -value was far below 0.05. But the effect size was about 2%. Two percent. Is that practically meaningful?

That depends on context. For an individual household, saving 2% on an electricity bill might be barely noticeable. But across millions of households, a 2% reduction aggregates to enormous energy savings, reduced carbon emissions, and measurable environmental impact. Practical significance depends on the scale and context of the application, beyond whether a p -value crossed a threshold.

8.5.1 The Large-Sample Problem

Here is an underappreciated fact. With a large enough sample, you can find statistically significant results for differences so tiny they are meaningless.

Suppose a company tests two versions of a website button, blue vs. green, and 500,000 people see each version. The blue button has a click rate of 3.01% and the green button has a click rate of 3.00%. With half a million people in each group, this 0.01 percentage point difference could easily produce a p -value below 0.05. Statistically significant. But is a company going to redesign its website over a difference of one click per ten thousand visitors? It should not.

Conversely, a small study might observe a large and practically important difference, say a 15-point reduction in blood pressure, but fail to reach statistical significance because the sample was too small and the uncertainty too wide.

Statistical significance and practical significance are different things. You need to assess both.

8.5.2 Effect Sizes

An **effect size** is a standardized measure of how large a difference or relationship is, independent of sample size.

The most common effect size for comparing two means is **Cohen's d**:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_p}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the sample means and s_p is the pooled standard deviation. Cohen's d tells you how many standard deviations apart the two group means are.

Jacob Cohen proposed rough benchmarks:

- $d = 0.2$: small effect
- $d = 0.5$: medium effect
- $d = 0.8$: large effect

These benchmarks are rough guidelines, not rigid rules. A “small” effect in Cohen's terms can still be important in certain contexts. Consider that a 0.2 standard deviation improvement in student test scores, applied across a national education program reaching millions of children, represents a sizable aggregate gain in learning. A drug that reduces the risk of heart attack by a seemingly small amount, applied to a population of millions of people at risk, can save thousands of lives. Conversely, a “large” effect in a trivial domain (people can distinguish Coke from Pepsi in a blind taste test) might not matter much at all.

For other types of comparisons, different effect size measures are used. For chi-square tests, **Cramer's V** measures the strength of association between two categorical variables. It is computed from the chi-square statistic itself,

$$V = \sqrt{\frac{\chi^2}{N \cdot \min(r - 1, c - 1)}},$$

where N is the total sample size and r and c are the numbers of rows and columns in the contingency table. The denominator scales away the influence of the table size, so V ranges from 0 (no association) to 1 (perfect association) regardless of how large the table or the sample is. As with Cohen's d , there are conventional benchmarks (Cohen, 1988): $V \approx 0.1$ is small, 0.3 is medium, and 0.5 is large for tables larger than 2×2 . For correlations, the correlation coefficient r itself serves as an effect size measure.

The key principle is this: always report effect sizes alongside p -values. A p -value tells you whether an effect is distinguishable from zero. An effect size tells you whether the effect is large enough to matter. A study that reports only the p -value is like a weather report that says “yes, it will rain” without telling you whether to expect a drizzle or a hurricane.

8.6 One-Sample t -Test

Let us now get specific about the mechanics. We will start with the simplest hypothesis test for a mean: the one-sample t -test.

8.6.1 When to Use It

Use a one-sample t -test when you want to test whether the mean of a single population differs from a specific hypothesized value. You have one sample of data and a claim about what the population mean should be.

8.6.2 Example

A coffee chain claims that its medium cups contain 16 ounces. A consumer group suspects the cups are being underfilled. They purchase 30 medium coffees from randomly selected locations and measure the contents. The sample mean is $\bar{x} = 15.7$ ounces with a sample standard deviation of $s = 0.8$ ounces.

Step 1: State the hypotheses.

The consumer group's research question is whether the cups are being underfilled, so the alternative hypothesis captures that suspicion:

$$H_a : \mu < 16$$

The null hypothesis is the logical complement: that the cups are *not* underfilled. Strictly speaking, that is

$$H_0 : \mu \geq 16,$$

but in practice the test focuses on the boundary case $\mu = 16$. The boundary is the version of H_0 that is hardest to distinguish from the alternative — any value of μ above 16 would produce data even less consistent with $H_a : \mu < 16$. So we compute everything assuming $\mu = 16$, and if the evidence is strong enough to reject that boundary case, it is automatically strong enough to reject every $\mu > 16$ above it. You will often see textbooks abbreviate the null as $H_0 : \mu = 16$ for this reason; it is the same test either way.

This is a one-tailed (left-tailed) test because the consumer group specifically suspects underfilling. We set the alternative first because it captures the research question. The null is then forced — it is whatever has to be true if H_a is false. Some textbooks reverse the order and set H_0 first; the procedure is the same in either case, but starting from H_a keeps the research question front and center.

For this example we will use $\alpha = 0.05$, the same conventional threshold introduced earlier. As that earlier discussion noted, this is a convention, not a hard line — the 0.04 / 0.05 / 0.06 region around the threshold is genuinely ambiguous, and a careful interpretation should look at effect sizes and confidence intervals alongside the p -value.

Step 2: Calculate the test statistic.

The one-sample t -statistic is:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Plugging in the values:

$$t = \frac{15.7 - 16}{0.8/\sqrt{30}} = \frac{-0.3}{0.146} = -2.05$$

Step 3: Find the p -value.

With $n - 1 = 29$ degrees of freedom, we look up the probability of getting a t -value of -2.05 or less. Using a t -table or software, the p -value is approximately 0.025.

Step 4: Make a decision.

At $\alpha = 0.05$, the p -value of 0.025 is less than 0.05, so we reject H_0 . There is sufficient evidence to conclude that the mean fill volume is less than 16 ounces.

Step 5: Interpret in context.

The data suggests the coffee chain is underfilling its medium cups. The sample mean was 15.7 ounces, which is 0.3 ounces below the advertised 16 ounces. Whether that shortfall is large enough to warrant action (a recall, a fine, a change in policy) is a practical question, not a statistical one. But the statistical test tells us the shortfall is unlikely to be explained by random variation alone.

8.6.3 Confidence Intervals and Hypothesis Tests: Two Sides of the Same Coin

You may have noticed a connection to Chapter 7. The 95% confidence interval for the mean fill volume in this example would be:

$$\bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}} = 15.7 \pm 2.045 \cdot \frac{0.8}{\sqrt{30}} = 15.7 \pm 0.299$$

This gives an interval of approximately (15.40, 16.00). Notice that the hypothesized value of 16 is right at the upper edge of this interval. If the interval had not included 16 at all, we would have rejected H_0 in a two-tailed test at $\alpha = 0.05$. This is not a coincidence. A two-sided hypothesis test at level α and a $(1 - \alpha)$ confidence interval always give the same answer: you reject H_0 exactly when the hypothesized value falls outside the confidence interval. The confidence interval has the advantage of also showing you a range of plausible values, beyond a yes-or-no verdict.

8.6.4 Assumptions

The one-sample *t*-test requires:

1. The data comes from a random sample (or at least an approximately representative one).
2. The observations are independent of each other.
3. The population distribution is approximately normal, or the sample size is large enough (roughly $n \geq 30$) for the Central Limit Theorem to apply.

If the sample is small and the data is heavily skewed, the *t*-test may not be appropriate, and a nonparametric alternative (such as the Wilcoxon signed-rank test) might be better. In practice, the *t*-test is fairly forgiving of mild departures from normality, especially with larger samples, thanks to the Central Limit Theorem. But with small samples, it pays to check whether the data is approximately symmetric before relying on the *t*-test.

8.7 Two-Sample *t*-Tests

More often, we want to compare two groups. Did the treatment group and control group differ? Do men and women have different average scores? Is the new process faster than the old one?

There are two flavors of two-sample *t*-tests, and choosing the right one depends on how the data was collected.

8.7.1 Independent Samples t-Test

Use this when the two groups are separate and unrelated. Different people, different items, different units. There is no pairing or matching between the groups.

8.7.1.1 Example

A nonprofit running a job training program wants to know whether participants earn more than non-participants one year after the program. They have salary data for 45 program participants (the treatment group) and 50 non-participants (the control group) who were similar in background at baseline.

- Treatment group: $\bar{x}_1 = 38,500$, $s_1 = 8,200$, $n_1 = 45$
- Control group: $\bar{x}_2 = 35,100$, $s_2 = 7,800$, $n_2 = 50$

Step 1: State the hypotheses.

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_a : \mu_1 - \mu_2 \neq 0$$

This is a two-tailed test because we are asking whether there is any difference, not specifying a direction in advance.

Step 2: Calculate the test statistic.

The independent samples t -statistic is:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$
$$t = \frac{38,500 - 35,100}{\sqrt{\frac{8,200^2}{45} + \frac{7,800^2}{50}}} = \frac{3,400}{\sqrt{\frac{67,240,000}{45} + \frac{60,840,000}{50}}}$$
$$t = \frac{3,400}{\sqrt{1,494,222 + 1,216,800}} = \frac{3,400}{\sqrt{2,711,022}} = \frac{3,400}{1,647} = 2.06$$

Step 3: Find the p -value.

For a one-sample *t*-test, the degrees of freedom were simply $n - 1$. Two-sample tests are more complicated, because we have two samples each contributing their own variability to the standard error. If you assume both populations have the same variance (the classical Student's *t*-test), the degrees of freedom are $n_1 + n_2 - 2$. If you do not assume equal variances — which is the safer default, as discussed in the next subsection — you use the **Welch–Satterthwaite approximation**, which produces a non-integer degrees-of-freedom value computed from s_1 , s_2 , n_1 , and n_2 . The result is bounded between $\min(n_1, n_2) - 1$ and $n_1 + n_2 - 2$, and you do not need to compute it by hand; any statistical package will produce it. Conceptually, the Welch–Satterthwaite df is a weighted compromise between what each sample alone could support, and it shrinks when the two samples disagree sharply about how variable the data is.

For this example, the Welch approximation gives approximately 91 degrees of freedom. With $t = 2.06$ and $df \approx 91$, the two-tailed *p*-value is approximately 0.042.

Step 4: Make a decision.

At $\alpha = 0.05$, the *p*-value of 0.042 is less than 0.05, so we reject H_0 . There is sufficient evidence of a difference in mean salary between the two groups.

Step 5: Interpret in context.

Program participants earned, on average, \$3,400 more per year than non-participants, and this difference is unlikely to be explained by chance alone. However, because this was not a randomized experiment (participants chose to enroll), we cannot be certain the difference was caused by the program. Other factors, such as motivation level, may differ between people who chose to participate and those who did not. This is a critical distinction. The hypothesis test tells us that a difference exists; it does not tell us why. Effect size: $d = 3,400/8,000 \approx 0.43$ (using a rough pooled standard deviation), which is a small-to-medium effect by Cohen's benchmarks.

8.7.1.2 A Note on Equal Variances

You may encounter two versions of the independent samples *t*-test. The classic version (sometimes called Student's *t*-test) assumes the two populations have equal variances. The version used above (Welch's *t*-test) does not make that assumption. Welch's version is almost always the safer choice, because the equal-variance assumption is often hard to justify and harder to verify. Most modern statistical software defaults to Welch's *t*-test, and that is the recommended default unless there is a specific subject-matter reason to assume equal variances.

8.7.2 Paired Samples t-Test

Use this when the two measurements come from the same individuals (or matched pairs). Before-and-after studies, where each person is measured twice, are the classic case.

8.7.2.1 Why Pairing Matters

When the same person is measured twice, the two measurements are not independent. A person who scores high on a pretest will likely score high on a posttest too, regardless of the treatment. The paired t -test accounts for this by analyzing the differences within each pair, effectively removing between-person variability.

8.7.2.2 Example

A school district implements a new reading intervention for 25 struggling readers. Each student takes a standardized reading assessment before the intervention and again after 12 weeks.

- Mean of the differences ($\bar{d}$, calculated as post minus pre for each student): $\bar{d} = 4.2$ points
- Standard deviation of the differences: $s_d = 6.1$ points
- Number of pairs: $n = 25$

Step 1: State the hypotheses.

$$H_0 : \mu_d = 0 \quad (\text{no change in reading scores})$$

$$H_a : \mu_d > 0 \quad (\text{scores improved})$$

This is a one-tailed test because the district specifically hypothesizes improvement.

Step 2: Calculate the test statistic.

$$t = \frac{\bar{d} - 0}{s_d / \sqrt{n}} = \frac{4.2}{6.1 / \sqrt{25}} = \frac{4.2}{1.22} = 3.44$$

Step 3: Find the p -value.

With $df = n - 1 = 24$ and $t = 3.44$ for a one-tailed test, the p -value is approximately 0.001.

Step 4: Make a decision.

At $\alpha = 0.05$, the p -value of 0.001 is far below 0.05, so we reject H_0 . There is strong evidence that reading scores improved.

Step 5: Interpret in context.

Students gained an average of 4.2 points on the reading assessment after the intervention. The improvement is statistically distinguishable from zero. The effect size is $d = 4.2/6.1 = 0.69$, which is a medium-to-large effect by Cohen's benchmarks. However, without a control group (students who did not receive the intervention), we cannot be certain the improvement was caused by the intervention rather than by natural growth, practice effects, or other factors. Students might have improved simply because they were 12 weeks older, or because taking the same test a second time is easier (a phenomenon called the testing effect). A stronger design would compare these students to a matched control group that did not receive the intervention, as we did in the independent samples example.

8.7.3 Independent vs. Paired: How to Choose

The distinction matters because using the wrong test throws away information or introduces error.

Ask yourself: "Is there a natural, meaningful pairing between observations in the two groups?"

- **Same people measured twice** (before/after, pretest/posttest): paired.
- **Matched pairs** (each treatment participant matched to a control participant on key characteristics): paired.
- **Two separate groups with no connection between individual observations**: independent.

If in doubt, think about what happens if you shuffle the data within one of the groups. In a paired design, shuffling destroys the pairing and ruins the analysis. In an independent design, the order within each group does not matter.

Here is a quick reference:

Scenario	Test
Same people, measured before and after a treatment	Paired
Spousal pairs compared on some variable	Paired
Left eye vs. right eye measurements on same patients	Paired
Students randomly assigned to two different classrooms	Independent
Men vs. women on a health outcome	Independent
Treatment plant A vs. treatment plant B	Independent

i Note

A common mistake in research is using an independent samples t -test when the data is actually paired. This typically results in a larger standard error and a less powerful test, making it harder to detect real effects. If you have paired data, use the paired test. You paid for that information when you designed the study, so use it.

8.8 When the Data Isn't Normal: Rank-Based Tests

The t -tests in the previous sections all assume either that the population is approximately normal or that the sample is large enough for the Central Limit Theorem to take over. They also implicitly assume the variable is measured on an interval or ratio scale — that the gap between two values carries quantitative meaning. What if neither holds?

Recall the variable types from Chapter 1. An **ordinal** variable, such as a customer satisfaction rating from 1 to 5 or a tumor stage from I to IV, carries an ordering but not necessarily a meaningful spacing: the gap between “1” and “2” is not guaranteed to equal the gap between “4” and “5”. Computing a sample mean of ordinal scores is mathematically possible but conceptually shaky. Or the data may be on an interval scale but heavily skewed, with a small enough sample that the t -test’s normality cover does not yet apply.

In these situations the standard alternative is to set aside the raw values and work with their **ranks**. Ranking throws away the magnitude information but preserves the ordering — which is all an ordinal scale carried in the first place. Tests that operate on ranks are called **nonparametric** because they make no assumption about a particular distributional shape.

Three rank-based tests cover most introductory situations.

- The **Wilcoxon signed-rank test** is the rank-based counterpart to the one-sample t -test (and, applied to within-pair differences, the paired t -test). It evaluates whether the median of a distribution differs from a hypothesized value.
- The **Wilcoxon rank-sum test**, also known as the **Mann–Whitney U test**, is the counterpart to the independent two-sample t -test. It evaluates whether two independent groups tend to produce systematically larger or smaller values.
- The **Kruskal–Wallis test** generalizes the rank-sum idea to three or more groups, and is the rank-based counterpart to the one-way ANOVA you will meet in Chapter 9.

Each of these tests produces a p -value with the same interpretation as the p -values in the t -test sections — the probability of seeing data this extreme if the null hypothesis were true — and the same decision rule applies: compare the p -value to a pre-specified α .

When should the rank-based version be the right choice? Three situations are common.

1. The variable is genuinely ordinal: a Likert-style satisfaction rating, a pain score, a class ranking. Median comparisons have a clear interpretation; mean comparisons do not.
2. The sample is small and visibly skewed or has clear outliers. The t -test's tolerance for non-normality weakens with small samples and heavy tails.
3. The research question is itself about a typical value (the median) rather than about the average.

The cost of using a rank-based test when the t -test's assumptions actually do hold is a modest loss of statistical power. The cost of using a t -test when the assumptions are clearly violated is a p -value that does not always mean what it appears to mean. When in doubt, run both, compare the conclusions, and report the rank-based version if the t -test's assumptions look shaky.

8.9 Chi-Square Tests

Not all data is numerical. When your variables are categorical, you need different tools. The chi-square (χ^2) family of tests handles questions about frequencies and proportions in categorical data.

8.9.1 Chi-Square Goodness of Fit Test

The goodness of fit test asks whether an observed distribution of frequencies matches an expected distribution.

8.9.1.1 Example

A city council member claims that complaints to the city are evenly distributed across the city's five districts. The city received 200 complaints last month, distributed as follows:

District	Observed	Expected (if equal)
North	52	40
South	38	40

District	Observed	Expected (if equal)
East	45	40
West	31	40
Central	34	40
Total	200	200

If complaints were truly evenly distributed, we would expect 40 per district ($200/5 = 40$).

Step 1: State the hypotheses.

H_0 : Complaints are evenly distributed across the five districts (each district has an equal proportion).

H_a : Complaints are not evenly distributed.

Step 2: Calculate the test statistic.

The chi-square statistic is:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed count and E_i is the expected count for each category. The intuition is straightforward: each term measures how far an observed count is from what the null predicts, squared so that overshoots and undershoots both count as deviations, and divided by the expected count so that the same absolute gap matters more in a small expected category than in a large one. A χ^2 value near zero means the observed counts look exactly like the null predicted; a large χ^2 value means at least one cell is far from its expected count, in either direction.

$$\chi^2 = \frac{(52 - 40)^2}{40} + \frac{(38 - 40)^2}{40} + \frac{(45 - 40)^2}{40} + \frac{(31 - 40)^2}{40} + \frac{(34 - 40)^2}{40}$$

$$\chi^2 = \frac{144}{40} + \frac{4}{40} + \frac{25}{40} + \frac{81}{40} + \frac{36}{40} = 3.6 + 0.1 + 0.625 + 2.025 + 0.9 = 7.25$$

Step 3: Find the p -value.

The degrees of freedom for a goodness of fit test are $k - 1$, where k is the number of categories. Here, $df = 5 - 1 = 4$.

Using a chi-square table or software, with $\chi^2 = 7.25$ and $df = 4$, the p -value is approximately 0.123.

Step 4: Make a decision.

At $\alpha = 0.05$, the p -value of 0.123 exceeds 0.05, so we fail to reject H_0 . We do not have sufficient evidence to conclude that complaints are unevenly distributed.

Step 5: Interpret in context.

Although the North district received more complaints than expected and the West district received fewer, these deviations from an even distribution are within the range of what we might see by chance. The data does not provide convincing evidence that some districts generate disproportionately more complaints. That said, the North district's count (52 vs. 40 expected) might warrant further investigation from a practical standpoint, even if the overall pattern is not statistically distinguishable from uniform.

8.9.2 Chi-Square Test of Independence

The test of independence asks whether two categorical variables are related to each other, or whether they are independent.

8.9.2.1 Example

A hospital wants to know whether patient satisfaction (satisfied vs. dissatisfied) is related to the time of admission (morning vs. afternoon vs. evening). They survey 300 patients and produce the following contingency table:

	Morning	Afternoon	Evening	Row Total
Satisfied	80	65	55	200
Dissatisfied	20	35	45	100
Column Total	100	100	100	300

Step 1: State the hypotheses.

H_0 : Patient satisfaction and time of admission are independent (not related).

H_a : Patient satisfaction and time of admission are not independent (there is a relationship).

Step 2: Calculate expected frequencies and test statistic.

Under independence, the expected frequency for each cell is:

$$E_{ij} = \frac{(\text{Row Total}_i)(\text{Column Total}_j)}{\text{Grand Total}}$$

For “Satisfied and Morning”:

$$E = \frac{200 \times 100}{300} = 66.67$$

Computing all expected frequencies:

	Morning	Afternoon	Evening
Satisfied	66.67	66.67	66.67
Dissatisfied	33.33	33.33	33.33

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$\chi^2 = \frac{(80 - 66.67)^2}{66.67} + \frac{(65 - 66.67)^2}{66.67} + \frac{(55 - 66.67)^2}{66.67} + \frac{(20 - 33.33)^2}{33.33} + \frac{(35 - 33.33)^2}{33.33} + \frac{(45 - 33.33)^2}{33.33}$$

$$\chi^2 = \frac{177.69}{66.67} + \frac{2.79}{66.67} + \frac{136.19}{66.67} + \frac{177.69}{33.33} + \frac{2.79}{33.33} + \frac{136.19}{33.33}$$

$$\chi^2 = 2.665 + 0.042 + 2.043 + 5.331 + 0.084 + 4.086 = 14.25$$

Step 3: Find the p -value.

Degrees of freedom for a test of independence: $df = (r - 1)(c - 1) = (2 - 1)(3 - 1) = 2$.

With $\chi^2 = 14.25$ and $df = 2$, the p -value is approximately 0.0008.

Step 4: Make a decision.

At $\alpha = 0.05$, the p -value of 0.0008 is far below 0.05. We reject H_0 . There is strong evidence that patient satisfaction is related to the time of admission.

Step 5: Interpret in context.

Patients admitted in the morning are considerably more likely to report satisfaction (80%) than patients admitted in the evening (55%). This relationship is very unlikely to be due to chance. The hospital might investigate what differs about the evening admission experience, whether it is staffing levels, wait times, the types of cases that present in the evening, or something else entirely. The chi-square test tells us the variables are related. It does not tell us why.

Notice a pattern in how we interpret chi-square results. The test gives a global answer: “yes, these variables are related” or “no, we cannot say they are.” It does not pinpoint where the relationship is strongest. To dig deeper, you can examine the individual contributions to the chi-square statistic (sometimes called standardized residuals) for each cell. In our example, the cells for “Dissatisfied and Evening” and “Satisfied and Morning” contributed the most to the overall chi-square, suggesting that the morning-evening contrast in satisfaction is the primary driver of the relationship.

8.9.3 Assumptions for Chi-Square Tests

1. The data consists of counts (frequencies) from a random sample.
2. The observations are independent.
3. Expected frequencies should be at least 5 in each cell. When this condition is not met, the chi-square approximation may be unreliable, and alternatives such as Fisher’s exact test may be more appropriate.

8.10 One-Tailed vs. Two-Tailed Tests

You may have noticed that some of our examples used one-tailed tests (the coffee underfilling, the reading intervention) and others used two-tailed tests (the salary comparison, the chi-square tests). The distinction is worth clarifying.

A **two-tailed test** evaluates evidence for a difference in either direction. The alternative hypothesis is $H_a : \mu \neq \mu_0$ (or $\mu_1 \neq \mu_2$). We reject H_0 if the test statistic is extreme in either tail of the distribution.

A **one-tailed test** evaluates evidence for a difference in a specific, pre-specified direction. The alternative hypothesis is either $H_a : \mu > \mu_0$ (right-tailed) or $H_a : \mu < \mu_0$ (left-tailed).

When should you use a one-tailed test? Only when you have a strong, pre-existing reason to predict the direction of the effect and when a difference in the other direction would be treated the same as no difference. The coffee consumer group had no interest in whether cups were overfilled; they specifically suspected underfilling. The school district hypothesized improvement, not decline.

When in doubt, use a two-tailed test. One-tailed tests are more powerful in the specified direction (the p -value is half of what the two-tailed p -value would be, but only when the observed effect is in the predicted direction; if the effect falls in the opposite direction, the one-tailed p -value exceeds 0.5 and the test cannot reject H_0 regardless of magnitude). This creates a tempting incentive for researchers to claim they predicted the direction all along. A researcher who observes a positive difference and then writes it up as a one-tailed test “because we predicted improvement” is engaging in a form of after-the-fact reasoning that inflates their chance of finding significance. This is a form of p -hacking, which we will discuss shortly. Some journals and reviewers have pushed back against one-tailed tests for exactly this reason, requiring researchers to justify the choice in advance.

8.11 The Hypothesis Testing Procedure, Summarized

Every hypothesis test, regardless of the specific test statistic, follows the same five-step procedure:

1. **State the hypotheses.** Write out H_0 and H_a in words and symbols.
2. **Choose the significance level.** Set α before looking at the data (0.05 is standard).
3. **Calculate the test statistic and p -value.** Use the appropriate formula for the type of data and question.
4. **Make a decision.** Compare the p -value to α .
5. **Interpret in context.** Translate the statistical conclusion back into the real-world question. Report effect sizes.

This structure is deliberately rigid. It forces clarity about what you are testing and what evidence would change your mind. Skipping steps, especially step 1 (stating hypotheses in advance) and step 5 (interpreting in context), is where most errors in practice occur.

! Ethics Moment: P-Hacking and the Replication Crisis

In 2011, psychologist Daryl Bem published a paper in the *Journal of Personality and Social Psychology* reporting nine experiments in which participants seemed to respond to stimuli before those stimuli appeared — that is, evidence consistent with precognition. The paper

followed the methodological conventions of the field at the time. The p -values were below 0.05.

Subsequent independent attempts to reproduce Bem's results have not consistently found the same effect. A registered, multi-lab replication of his procedures (Galak et al., 2012, with seven experiments and over 2,400 participants) reported effect sizes near zero. The disagreement between Bem's original results and the replications became a focal point in a larger conversation about how published research can show effects that fail to reappear when the same studies are run again.

The point of revisiting this episode is not to single out any one researcher. The methods Bem used were standard practice. The episode mattered because it exposed a set of practices, widespread across many fields, that collectively make it easier to find "significant" results that do not hold up under independent replication.

P-hacking refers to the practice of manipulating the data analysis process until a p -value below 0.05 emerges. This can include running multiple statistical tests and only reporting the one that "worked," trying different ways of excluding outliers to see which yields significance, measuring many outcome variables and highlighting the one that was significant, collecting data until the p -value dips below 0.05 and then stopping, and splitting data into subgroups until a significant result appears somewhere. None of these practices involve fabricating data. Each individual decision might seem reasonable in isolation. But together, they meaningfully inflate the chance of a false positive.

Researcher degrees of freedom is a related concept coined by Joseph Simmons, Leif Nelson, and Uri Simonsohn. In any study, researchers make dozens of choices about data collection, variable definition, inclusion criteria, control variables, and analysis methods. Each choice is a "degree of freedom" that affects the results. When researchers explore these choices until they find a combination that produces significance, they exploit these degrees of freedom, even if they do so unconsciously.

The cumulative effect of these practices on the published literature has been substantial.

In 2015, the **Reproducibility Project: Psychology** published the results of a coordinated effort to replicate 100 studies from leading psychology journals. About 36% of the replications produced statistically significant results in the same direction as the originals, and on average the replication effect sizes were less than half the size of the original effect sizes. This did not mean that 64% of the original findings were fabricated. It meant that the published record, shaped by selective reporting and many underpowered designs, gave a more confident picture than the underlying evidence could actually support.

The **Brian Wansink** case sits further along the same continuum. Wansink, a Cornell professor who led the Food and Brand Lab, was widely known for studies on eating behavior — “bottomless bowls” that purportedly caused people to eat more soup, the claim that serving vegetables first at buffets increased healthy eating. In 2017, after a blog post in which Wansink described encouraging a graduate student to slice a dataset in multiple ways until significant results emerged, other researchers reanalyzed his published work and identified widespread statistical inconsistencies. More than a dozen of his papers were retracted. Wansink resigned from Cornell in 2018.

Alex Reinhart’s book *Statistics Done Wrong* (freely available online) catalogs these and many other examples of how standard statistical practice goes astray. It is worth reading alongside this chapter.

What can you do about it?

- **Pre-register your hypotheses and analysis plan** before collecting data. If you decide what you are testing in advance, you cannot subconsciously adjust your analysis to get the result you want.
- **Report all tests you ran**, not just the significant ones. If you tested 20 variables, say so.
- **Adjust for multiple comparisons** when running multiple tests (methods like the Bonferroni correction exist for this purpose).
- **Focus on effect sizes and confidence intervals**, not just p -values.
- **Be skeptical of p -values just barely below 0.05**, especially in small studies with many analysis choices.
- **Value replication.** A single study with $p < 0.05$ is a preliminary finding, not a settled fact.

The replication crisis extends far beyond psychology or academia. Any field that uses statistical tests, medicine, business analytics, education research, public policy, is susceptible to the same pressures and the same errors.

AI Reality Check

AI tools can run hundreds of statistical tests in seconds. Hand an AI system a dataset and ask it to “find interesting patterns,” and it will oblige. It will test every variable against every other variable. It will try different transformations, subsets, and model specifications. And it will find “significant” results, lots of them, because with enough tests, false positives are guaranteed. Run 100 tests at $\alpha = 0.05$ on data that contains no real relationships, and you can expect about 5 to come back significant purely by chance. AI does not have

judgment about which tests make theoretical sense. It does not know which comparisons are scientifically meaningful and which are fishing expeditions. If you use AI to explore data, you must apply the same skepticism you would apply to any other analysis. A “significant” result generated by an AI running blind comparisons is worth exactly as much as a “significant” result generated by a researcher who tested 50 hypotheses and only reported the one that worked, which is to say, it is worth very little unless confirmed by an independent test on new data.

Try It Online

To see for yourself how easy it is to find significant results in random noise, try the **P-Hacking Simulator**. You will be given a dataset with no real relationships and a menu of analysis choices. See how quickly you can reach $p < 0.05$. Then consider what that means for published research that made the same choices without disclosing them.

8.12 Putting It All Together: The Opower Study Revisited

Let us return to where we started. Allcott’s study of Opower’s home energy reports was well-designed in ways that avoided many of the pitfalls we have discussed.

First, the study was a randomized experiment. Households were randomly assigned to receive reports or not, so the treatment and control groups were comparable on average. This rules out many confounding explanations.

Second, the sample sizes were enormous, hundreds of thousands of households across multiple utility companies. This gave the study high power to detect even small effects.

Third, the hypotheses were stated in advance. Allcott was testing whether the reports reduced energy consumption, not fishing through dozens of outcomes to see what changed.

The result: a statistically significant reduction of about 2% in electricity usage, with a p -value far below any conventional threshold. The effect size was small in absolute terms but potentially large in aggregate impact.

This is what a well-conducted hypothesis test looks like. Clear hypotheses stated in advance. Random assignment. Large samples. Transparent methods. And an honest assessment of what the results do and do not show. The 2% effect is real. Whether it justifies the cost of the program is a separate question, one that requires cost-benefit analysis, not just a p -value.

Allcott actually took this further. He estimated that the program cost about \$1 per household per year and saved about \$3 in electricity costs per household per year, plus additional social benefits from reduced carbon emissions. The effect was small for any individual household but positive in a cost-benefit sense and enormous in aggregate. This is what it looks like to move beyond “is it significant?” to “does it matter?”

Try It Online

Return to the **Hypothesis Testing Playground** to run simulated studies end to end. Set a significance level, run repeated experiments, and watch how often a true null hypothesis gets incorrectly rejected at that threshold.

8.13 Common Misconceptions and Pitfalls

Before we move to the exercises, let us collect several traps that consistently ensnare students and practitioners.

“We established the alternative hypothesis.” No. We rejected the null hypothesis. The difference is subtle but important. Rejecting H_0 means the data is inconsistent with the null. It does not mean we have demonstrated any particular alternative. Multiple alternative explanations could be consistent with the data.

“The p -value tells us the probability the null is true.” Already covered, but worth repeating because it is the single most common misinterpretation. The p -value is the probability of the data (or more extreme data) given the null, not the probability of the null given the data.

“We failed to reject H_0 , so the treatment does not work.” Absence of evidence is not evidence of absence. A non-significant result might reflect a true null, or it might reflect inadequate power, an insufficient sample to detect a real but small effect.

“Our p -value is 0.001, so the effect is huge.” The p -value reflects both the effect size and the sample size. A very small p -value could come from a very small effect measured in a very large sample.

“We should use a one-tailed test because we expect the effect to go in this direction.” Maybe. But if your “expectation” was formed after looking at the data, you are cheating. The direction must be specified before data collection.

Confusing independent and paired designs. If you measure the same people twice and analyze the data as if the groups were independent, you are ignoring the within-person correlation and losing power.

Ignoring assumptions. The t -test assumes approximate normality (or large samples) and independence. The chi-square test requires adequate expected frequencies. Violating these assumptions can invalidate your conclusions.

Treating the 0.05 threshold as a bright line. A result with $p = 0.048$ is not fundamentally different from one with $p = 0.053$. Yet the first gets published and the second gets filed away as a failure. This binary thinking has distorted entire fields of research. Report exact p -values and let readers evaluate the strength of the evidence themselves.

Running a test without thinking about power first. If you run a study with 15 participants and find $p = 0.12$, you have learned very little. The study was probably too small to detect anything but a very large effect. Before collecting data, ask yourself: “If there is an effect of the size I care about, is my sample large enough to have a reasonable chance of detecting it?” If the answer is no, either get a bigger sample or acknowledge the limitation upfront.

8.14 Looking Ahead

Hypothesis testing gives you a framework for asking, “Is the evidence strong enough to reject a specific claim?” But so far, we have tested claims about single means, single proportions, and differences between two groups. In many real-world situations, you need to compare three, four, or more groups simultaneously. Does teaching method A, B, C, or D produce the best outcomes? Do four different advertising strategies yield different click-through rates? Running multiple two-sample tests introduces a serious problem: the accumulation of false positive risk. The next chapter introduces ANOVA (Analysis of Variance), a method designed to compare multiple group means in a single test while controlling that risk. It extends the logic of hypothesis testing into settings with more complex group structures, and it connects directly to the regression models we will build in later chapters.

8.15 Further Reading and References

The following works are cited in this chapter or provide valuable additional context for the ideas covered here.

On the Opower study and behavioral nudges: Allcott, H. (2011). Social norms and energy conservation. *Journal of Public Economics*, 95(9-10), 1082–1095. (The study that provided the opening example for this chapter, demonstrating a statistically significant but small reduction in home energy use from behavioral nudges.)

On the ASA statement on p -values: Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p -values: Context, process, and purpose. *The American Statistician*, 70(2), 129–133. (The American Statistical Association’s statement clarifying what p -values are, what they are not, and common misinterpretations that have shaped research practice.)

On the replication crisis: Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716. (The coordinated study finding that 36% of 100 published psychology results replicated with statistical significance in the same direction, with average effect sizes less than half the original — a key reference in subsequent work on transparent and rigorous research practices.)

On statistical significance and the proposal to redefine it: Benjamin, D. J., Berger, J. O., Johannesson, M., et al. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2, 6–10. (The paper proposing a shift from $\alpha = 0.05$ to $\alpha = 0.005$ for claims of new discoveries, referenced in this chapter’s discussion of significance thresholds.)

For further reading on hypothesis testing: Ziliak, S. T., & McCloskey, D. N. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. University of Michigan Press. (A provocative critique of the overreliance on statistical significance at the expense of practical significance and effect sizes.)

On the replication crisis and p -hacking: Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100(3), 407–425. (The precognition study discussed in the Ethics Moment, whose publication in a major journal helped catalyze the broader conversation about replication.)

Galak, J., LeBoeuf, R. A., Nelson, L. D., & Simmons, J. P. (2012). Correcting the past: Failures to replicate ψ . *Journal of Personality and Social Psychology*, 103(6), 933–948. (A registered, multi-experiment attempt to reproduce Bem’s findings that found effect sizes near zero, frequently cited in subsequent discussions of the original paper.)

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366. (The paper that formalized the concept of “researcher degrees

of freedom” and demonstrated how ordinary analytical flexibility can produce false positives at alarming rates.)

Reinhart, A. (2015). *Statistics done wrong: The woefully complete guide*. No Starch Press. (A concise, accessible catalog of the most common statistical errors in published research, freely available online.)

On statistical power in neuroscience: Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5), 365–376. (The review documenting median statistical power of 21% in neuroscience, discussed in this chapter’s section on why power matters.)

8.16 Key Terms

- **Null hypothesis (H_0):** The default claim that there is no effect, no difference, or no relationship. The hypothesis that is tested and potentially rejected.
- **Alternative hypothesis (H_a):** The claim that there is an effect, a difference, or a relationship. The hypothesis the researcher hopes to find evidence for; supported when the null hypothesis is rejected.
- **Test statistic:** A numerical summary of the data that measures how far the sample result falls from what the null hypothesis predicts, relative to expected sampling variability.
- **P-value:** The probability of obtaining a test statistic as extreme as, or more extreme than, the one observed, assuming the null hypothesis is true. It is not the probability that the null hypothesis is true.
- **Significance level (α):** The pre-set threshold below which the p -value leads to rejection of H_0 . Commonly set at 0.05.
- **Type I error:** Rejecting H_0 when it is actually true (false positive). Probability equals α .
- **Type II error:** Failing to reject H_0 when it is actually false (false negative). Probability equals β .
- **Power:** The probability of correctly rejecting a false null hypothesis. Equals $1 - \beta$.
- **Statistical significance:** A result is statistically significant when the p -value is less than α .
- **Practical significance:** Whether the size of an effect is large enough to matter in context, regardless of statistical significance.

- **Effect size:** A quantitative measure of the magnitude of an effect, independent of sample size.
- **Cohen's d :** An effect size measure for comparing two means, calculated as the difference in means divided by the pooled standard deviation.
- **One-sample t -test:** A test of whether a population mean differs from a specified hypothesized value.
- **Independent samples t -test:** A test of whether the means of two unrelated groups differ.
- **Paired samples t -test:** A test of whether the mean difference between paired observations (e.g., before and after) differs from zero.
- **Chi-square goodness of fit test:** A test of whether an observed frequency distribution matches an expected distribution.
- **Chi-square test of independence:** A test of whether two categorical variables are associated, based on comparing observed cell counts in a contingency table to the counts expected under independence.
- **Degrees of freedom:** A parameter that determines the shape of the t or chi-square distribution, based on the sample size and number of groups or categories.
- **One-tailed test:** A hypothesis test where the alternative hypothesis specifies a direction (greater than or less than).
- **Two-tailed test:** A hypothesis test where the alternative hypothesis does not specify a direction (not equal to).
- **Welch's t -test:** A modification of the independent-samples t -test that does not assume equal variances in the two groups. Generally preferred as the default two-sample test.
- **Nonparametric test:** A hypothesis test that does not assume a specific distributional shape for the data. Rank-based tests are the most common nonparametric alternatives to the t -test.
- **Wilcoxon signed-rank test:** A rank-based alternative to the one-sample or paired t -test, evaluating whether the median (or median within-pair difference) differs from a hypothesized value.
- **Wilcoxon rank-sum test (Mann–Whitney U test):** A rank-based alternative to the independent two-sample t -test, evaluating whether two groups tend to produce systematically larger or smaller values.
- **Kruskal–Wallis test:** A rank-based generalization of the Wilcoxon rank-sum test to three or more groups; the rank-based counterpart to one-way ANOVA.
- **P-hacking:** The practice of manipulating data analysis (running multiple tests, trying different specifications) until a statistically significant result appears.
- **Researcher degrees of freedom:** The many choices available to a researcher during data analysis that can influence results, creating opportunities for p-hacking.
- **Replication crisis:** The finding, across multiple fields, that many published statistically significant results fail to replicate in subsequent studies.

8.17 Exercises

8.17.1 Check Your Understanding

1. In your own words, explain what a null hypothesis is and why hypothesis testing starts by assuming it is true.
2. A researcher tests whether a new fertilizer improves tomato yield and obtains a p -value of 0.03. Which of the following is the correct interpretation?
 - a. There is a 3% chance the fertilizer does not work.
 - b. There is a 97% chance the fertilizer works.
 - c. If the fertilizer had no effect, there is a 3% probability of observing results as extreme as (or more extreme than) what was actually observed.
 - d. The fertilizer increases yield by 3%.
3. Explain the difference between a Type I error and a Type II error. Give a real-world example of each in the context of medical testing (e.g., testing whether a new drug lowers cholesterol).
4. A study finds that a new teaching method produces a statistically significant improvement in test scores with $p = 0.04$ and Cohen's $d = 0.1$. What does this combination of results tell you? Would you recommend the school district adopt the new method? Why or why not?
5. Why is “fail to reject H_0 ” not the same as “accept H_0 ”? Explain using the courtroom analogy.
6. A marketing analyst runs an A/B test comparing two email subject lines and gets a p -value of 0.08. The analyst's manager says, “So the two subject lines perform the same.” Is the manager's conclusion correct? Why or why not?
7. Explain why increasing the sample size increases the power of a hypothesis test.
8. What is the difference between statistical significance and practical significance? Give an example where a result could be statistically significant but not practically significant.
9. A researcher plans to use a one-tailed test because “the treatment should help, not hurt.” The researcher has not yet collected data. Under what conditions is a one-tailed test justified? Under what conditions would it be inappropriate?
10. List three practices that constitute p-hacking. For each, explain why it inflates the probability of a Type I error.
11. A chi-square test of independence produces $\chi^2 = 3.84$ with $df = 1$ and a p -value of approximately 0.05. A different chi-square test produces $\chi^2 = 3.84$ with $df = 3$ and a p -value of

approximately 0.28. Explain why the same chi-square value leads to different p -values in these two cases.

12. In the context of hypothesis testing, what does “degrees of freedom” represent, and why does it matter?

8.17.2 Apply It

1. A factory claims its light bulbs last an average of 1,200 hours. A consumer agency tests 36 bulbs and finds a sample mean of 1,175 hours with a standard deviation of 80 hours. Perform a one-sample t -test at $\alpha = 0.05$ to determine whether the mean bulb life is less than 1,200 hours. State your hypotheses, calculate the test statistic, determine the p -value (approximately), and state your conclusion.
2. A fitness app claims its users walk an average of 10,000 steps per day. A researcher collects data from 40 app users and finds a mean of 9,450 steps with a standard deviation of 2,100 steps. Test whether the true mean differs from 10,000 at $\alpha = 0.05$. Calculate Cohen’s d and interpret it.
3. A pharmaceutical company tests a new headache medication. In a randomized trial, 60 patients receive the new medication and 60 receive a placebo. Time to headache relief is recorded.
 - New medication group: $\bar{x}_1 = 25$ minutes, $s_1 = 8$ minutes
 - Placebo group: $\bar{x}_2 = 30$ minutes, $s_2 = 10$ minutesPerform an independent samples t -test at $\alpha = 0.05$. Calculate Cohen’s d and interpret the result.
4. A psychologist measures anxiety scores for 20 patients before and after a mindfulness program. The mean difference (after minus before) is $\bar{d} = -5.3$ points (indicating a decrease in anxiety), with $s_d = 7.2$ points. Perform a paired t -test at $\alpha = 0.05$ to determine whether the program reduced anxiety. Calculate the effect size.
5. A political scientist claims that party affiliation is equally distributed among registered voters in a county: 35% Democrat, 35% Republican, 20% Independent, and 10% Other. A random sample of 500 voters yields: 195 Democrats, 160 Republicans, 105 Independents, and 40 Others. Perform a chi-square goodness of fit test at $\alpha = 0.05$.
6. An HR director wants to know whether employee retention (stayed vs. left within the first year) is related to the department of hire. Data from 400 recent hires:

	Engineering	Marketing	Sales	Operations	Total
Stayed	75	55	40	70	240

	Engineering	Marketing	Sales	Operations	Total
Left	25	45	60	30	160
Total	100	100	100	100	400

Perform a chi-square test of independence at $\alpha = 0.05$. Which department has the highest turnover rate? What might you investigate further?

- A researcher wants to detect a small effect ($d = 0.3$) using an independent samples t -test with $\alpha = 0.05$ and wants 80% power. Using the rough approximation that each group needs about $n = (16/d^2)$ participants, calculate the required sample size per group. What happens to the required sample size if the researcher wants to detect a medium effect ($d = 0.5$) instead?
- A brewery claims its “pint” glasses hold 16 ounces. You measure 25 glasses and find a mean of 15.6 ounces with a standard deviation of 1.0 ounce. Test whether the glasses hold less than 16 ounces at $\alpha = 0.01$ (note the stricter significance level). What changes compared to using $\alpha = 0.05$?
- Two sections of the same statistics course take the same final exam. Section A (taught with traditional lectures, $n = 35$) has a mean score of 78 with $s = 12$. Section B (taught with active learning methods, $n = 32$) has a mean score of 83 with $s = 11$. Test whether there is a difference in mean exam scores between the two teaching methods at $\alpha = 0.05$. Calculate the effect size and discuss both statistical and practical significance.
- A food delivery service surveys 600 customers about their satisfaction (satisfied or not satisfied) across three order types (restaurant meals, grocery delivery, and meal kits):

	Restaurant Meals	Grocery Delivery	Meal Kits	Total
Satisfied	160	130	110	400
Not Satisfied	40	70	90	200
Total	200	200	200	600

Test whether satisfaction level is independent of order type at $\alpha = 0.05$.

- A sleep researcher measures the number of hours of sleep for 15 participants on a night when they used a blue-light filter on their devices and a night when they did not (in random order). The mean difference (filter minus no filter) is $\bar{d} = 0.4$ hours, with $s_d = 0.9$ hours. Test whether the filter increases sleep at $\alpha = 0.05$. Is the result statistically significant? Regardless of your answer, discuss the practical significance of 0.4 hours (24 minutes) of additional sleep.

12. A city health department expects the following distribution of blood types in its population based on national data: Type O (44%), Type A (42%), Type B (10%), Type AB (4%). In a sample of 250 blood donors, they observe: 95 Type O, 120 Type A, 25 Type B, and 10 Type AB. Perform a goodness of fit test at $\alpha = 0.05$ and interpret the results.

The following three exercises use the energy-reports.csv dataset available on the companion website. This is a simulated dataset mathematically modeled to reflect real-world patterns from the Opower energy conservation study documented in Allcott (2011). No real households are represented. The dataset contains 200 observations (100 Treatment, 100 Control) with variables including household_id, group (Treatment/Control), energy_kwh_before (monthly energy usage before intervention), energy_kwh_after (monthly energy usage after intervention), home_size_sqft, num_occupants, and region.

13. Using the energy-reports.csv dataset, compute a new variable `energy_change = energy_kwh_after - energy_kwh_before` for each household. Compute the mean and standard deviation of `energy_change` separately for the Treatment and Control groups. Perform an independent samples *t*-test to determine whether the mean change in energy usage differs between Treatment and Control at $\alpha = 0.05$. State your hypotheses, report the test statistic and *p*-value, and interpret the result. Calculate Cohen's *d* and comment on the effect size.
14. Ignoring the Treatment/Control distinction for a moment, focus only on the Treatment group. Perform a one-sample *t*-test (or equivalently, a paired *t*-test comparing `energy_kwh_after` to `energy_kwh_before`) to test whether the Treatment group's energy usage decreased after the intervention. State the null and alternative hypotheses, report the test statistic and *p*-value, and state your conclusion at $\alpha = 0.05$. Then do the same for the Control group. Do the Control households show a statistically significant change? What does this comparison tell you about the effect of the energy reports?
15. The dataset includes `home_size_sqft` and `num_occupants` as additional variables. Group the Treatment households into two categories: those with `home_size_sqft` above the Treatment group median and those at or below it. Compute the mean `energy_change` for each subgroup. Is the energy reduction larger for bigger or smaller homes? Perform a two-sample *t*-test comparing the mean `energy_change` between these two subgroups at $\alpha = 0.05$. What does this suggest about whether the energy reports were equally effective for different types of households?
16. **Wilcoxon signed-rank test (paired).** Repeat the paired analysis from problem 14 using the rank-based **Wilcoxon signed-rank test** rather than the paired *t*-test. In R, `wilcox.test(after, before, paired = TRUE)` produces it. Compare the *p*-value from the rank-based test to the *p*-value from the paired *t*-test. Are the conclusions the same? Then create a histogram of the within-household differences (`energy_kwh_after - energy_kwh_before`); does the shape of that

distribution suggest the t -test's normality assumption is reasonable, or does the rank-based version seem like the safer choice?

17. **Wilcoxon rank-sum (Mann–Whitney U) test.** Repeat the two-group comparison from problem 13 using the **Wilcoxon rank-sum test** (also called the Mann–Whitney U test). In R, `wilcox.test(energy_change ~ group, data = ...)` produces it. The rank-sum test answers whether one group tends to produce systematically larger or smaller values than the other, without assuming normal distributions. Report the test statistic and p -value. Compare your conclusion to the t -test conclusion in problem 13.
18. **Kruskal–Wallis test.** Suppose the dataset had three groups instead of two — Treatment-A, Treatment-B, and Control. The rank-based generalization of the rank-sum test to three or more groups is the **Kruskal–Wallis test**, produced in R by `kruskal.test(energy_change ~ group, data = ...)`. Without running it on the actual dataset, write one paragraph explaining (a) when you would reach for Kruskal–Wallis instead of one-way ANOVA, and (b) what the relationship is between the rank-sum / Mann–Whitney test (two groups) and the Kruskal–Wallis test (three or more) — by analogy with the relationship between a two-sample t -test and one-way ANOVA.

8.17.3 Think Deeper

1. The 2017 proposal to lower the default significance threshold from 0.05 to 0.005 was controversial. What are the arguments in favor of a stricter threshold? What are the potential downsides? Consider the impact on different fields (medicine, social science, business analytics) and on researchers with limited resources for large studies.
2. Return to the Opower study discussed at the beginning of this chapter. Allcott found a statistically significant 2% reduction in energy usage. Suppose the study had been conducted with only 100 households (50 treatment, 50 control) instead of hundreds of thousands. The observed difference might have been similar, but the p -value almost certainly would have been above 0.05. Would you conclude that the energy reports do not work? What does this scenario reveal about the relationship between sample size and the conclusions of hypothesis testing?
3. The Brian Wansink episode is sometimes invoked as a stark example of what can go wrong when methodological practices are loose. But many of the practices that drew scrutiny — running multiple analyses, exploring data for interesting patterns, selectively reporting which results were written up — were common and even encouraged in some research environments at the time. To what extent is the replication crisis a problem of individual bad actors, and to what extent is it a systemic problem with incentives in academic publishing? What changes to the system would you recommend?

4. A pharmaceutical company tests 20 different compounds to see if any of them lower blood pressure. One compound produces a p -value of 0.03. The company announces that it has found a promising drug. What is wrong with this conclusion? What should the company do before making such a claim? Connect your answer to the concept of multiple comparisons.
5. Consider a criminal justice context. A new risk-assessment algorithm claims to predict which defendants will commit future crimes. The algorithm's developers report that it is "statistically significant" at predicting recidivism. Discuss the roles of Type I errors (false positives, flagging someone who would not reoffend) and Type II errors (false negatives, failing to flag someone who would reoffend) in this context. Which type of error is more harmful? Does the answer depend on whose perspective you take?
6. A colleague shows you a study with a p -value of 0.048 and says, "Just barely made it." Another study on the same topic reports a p -value of 0.052 and concludes "no effect found." What would you say to your colleague about the difference between these two results? How should these findings be interpreted together? What additional information would you want before forming a strong opinion?

Part III

Comparing and Modeling

Chapter 9

Comparing Groups

9.1 Are Emily and Greg More Employable Than Lakisha and Jamal?

In 2004, economists Marianne Bertrand and Sendhil Mullainathan published an experiment that was as simple in design as it was sobering in its findings. They created roughly 5,000 fictitious résumés and sent them in response to help-wanted ads in Boston and Chicago newspapers. The résumés were carefully constructed to be equivalent in quality, some with stronger credentials and some with weaker ones, but otherwise matched on experience, education, and formatting. The only thing the researchers varied was the name at the top.

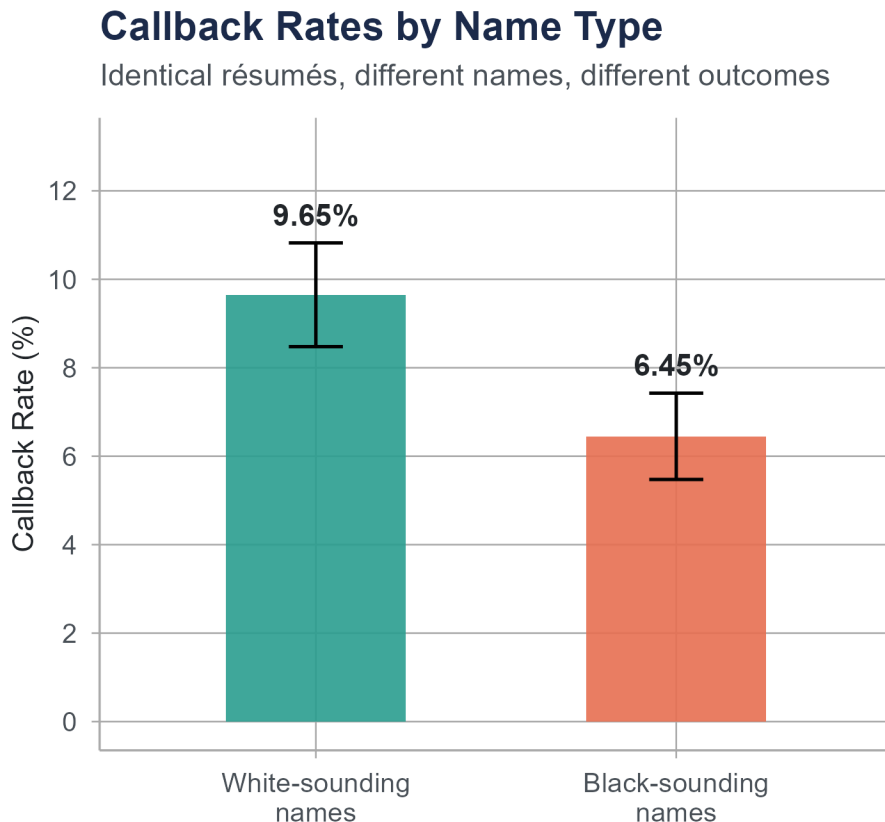
Half the résumés were assigned names that, based on prior surveys, were strongly associated with being White. Names like Emily Walsh, Greg Baker, Brendan Murray, Anne Ryan. The other half were assigned names strongly associated with being Black. Names like Lakisha Washington, Jamal Jones, Tamika Williams, Aisha Harris. (The last names here are illustrative; the original study used first names only as the experimental treatment.) Everything else about the paired résumés was identical or statistically equivalent.

Then the researchers waited for the phone to ring.

Résumés with White-sounding names received a callback about once in every 10 sent. Résumés with Black-sounding names received a callback about once in every 15. That is roughly a 50% gap in callback rates, driven entirely by the name at the top of the page, which served as a proxy for perceived race. The gap was consistent across occupations, industries, and employer sizes.

Here is what made the finding even more troubling. When the researchers improved the quality of the résumés by adding more experience, better credentials, and stronger formatting, the callback

rate for White-sounding names jumped noticeably. For Black-sounding names, higher quality résumés produced almost no additional benefit. The labor market was rewarding credentials for one group and largely ignoring them for another.



Source: Bertrand & Mullainathan (2004), Table 1.
95% confidence intervals computed from the published proportions and sample sizes.

Figure 9.1: Callback rates for résumés with White-sounding names versus Black-sounding names, with 95% confidence interval error bars. Identical résumés received different callback rates depending only on the name at the top, with White-sounding names receiving roughly 50% more callbacks.

The paper has been widely cited in labor economics. It is also, at its core, a study about comparing groups. Two groups of résumés, identical except for the name, sent into the same labor market, and the outcome measured across both groups. The statistical question is straightforward. Is the

difference in callback rates between these two groups large enough that we can rule out random chance as the explanation?

That is the kind of question this chapter teaches you to answer. But we are going to go beyond two groups. What happens when you have three groups, or five, or ten? What changes in your analysis when you move from comparing two means to comparing many? And what responsibilities come with analyzing group differences, especially when those differences involve race, gender, disability, income, or any other dimension along which people have been historically sorted and ranked?

This chapter marks the beginning of Part III of this book. In Parts I and II, you built the foundations: how to summarize data, how to think about probability, how to construct confidence intervals, and how to test hypotheses. Now we start applying those tools to increasingly realistic questions. Comparing groups is where statistics meets the messiness of the social world, where the numbers carry both uncertainty and meaning, and where the analyst's choices about framing and interpretation carry real consequences for real people.

9.2 The Problem with Running Multiple *t*-Tests

In Chapter 8, you learned to use a two-sample *t*-test to compare the means of two groups. If you want to know whether the average callback rate differs between résumés with White-sounding names and résumés with Black-sounding names, a two-sample *t*-test handles that cleanly. Two groups, one comparison, one test.

But suppose your research question is more complex. Imagine you are studying the effectiveness of four different teaching methods on exam scores. You have Group A (traditional lecture), Group B (flipped classroom), Group C (problem-based learning), and Group D (peer instruction). You want to know whether any of these methods produce different average exam scores.

Your first instinct might be to run all possible pairwise *t*-tests. Compare A to B. Compare A to C. Compare A to D. Compare B to C. Compare B to D. Compare C to D. That is six comparisons. With four groups, the number of pairwise comparisons is $\binom{4}{2} = 6$. With five groups it would be 10. With ten groups, 45.

This seems thorough. It is actually dangerous.

Here is the problem. Each *t*-test you run at the $\alpha = 0.05$ significance level has a 5% chance of producing a false positive, a Type I error, even when there is no real difference between the groups. If you run one test, your chance of at least one false positive is 5%. If you run six independent tests, each at 5%, the probability of getting at least one false positive is no longer 5%. It is

$$P(\text{at least one Type I error}) = 1 - (1 - \alpha)^k$$

where k is the number of comparisons. For six tests at $\alpha = 0.05$, this gives

$$1 - (1 - 0.05)^6 = 1 - (0.95)^6 = 1 - 0.7351 = 0.2649$$

That is roughly a 26% chance of at least one false positive. You started with a 5% threshold because you wanted to be reasonably sure your findings were not due to chance. By running six tests, you have inflated your actual error rate to more than one in four. Run 20 comparisons and your probability of at least one false positive exceeds 64%. You are now more likely than not to “find” a difference that does not exist.

This is the **multiple comparisons problem**, sometimes called the problem of multiplicity. It is not a minor technical nuisance. It is a common way that researchers fool themselves and their audiences. Every additional comparison you run is another lottery ticket for a false positive. Buy enough tickets and you will eventually “win.”

i Note

The multiple comparisons problem shows up far beyond formal hypothesis testing. Data journalists sometimes compare crime rates across dozens of cities and highlight whichever comparison looks most dramatic. Marketing analysts test ten different email subject lines and declare a winner after each line was tested on a small subsample. In each case, the more comparisons you make, the more likely you are to find something that looks like a pattern but is just noise. The first question to ask when someone presents a surprising group difference is always, “How many comparisons did they run to find this one?”

9.3 ANOVA: The Big Picture

Analysis of Variance, universally abbreviated as **ANOVA**, solves the multiple comparisons problem by testing all the groups at once. Instead of asking “Is Group A different from Group B?” and “Is Group A different from Group C?” and so on, ANOVA asks a single omnibus question.

“Is there any difference among the group means, or are they all essentially the same?”

More precisely, the null and alternative hypotheses for a one-way ANOVA are

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$$

$$H_A : \text{At least one } \mu_i \text{ differs from the others}$$

where $\mu_1, \mu_2, \dots, \mu_k$ are the population means of the k groups. The null hypothesis says all the groups have the same population mean. The alternative says at least one group's mean is different from at least one other group's mean. It does not say which groups differ or how many differ. Just that the means are not all equal.

This single test controls the Type I error rate at α across all groups simultaneously. You run one test, you get one p -value, and that p -value reflects the overall evidence against the null hypothesis that all group means are equal. No inflation. No lottery tickets.

The name “Analysis of Variance” is admittedly confusing when you first encounter it, because the method is used to compare means, not variances. But the name makes sense once you understand the logic, which is that ANOVA compares means by analyzing variances. Specifically, it compares two types of variability.

9.4 The F-Statistic: Between vs. Within

The core idea of ANOVA is surprisingly intuitive. Imagine you are looking at data from three groups. Each group has some spread in its values, and the group means are at different locations. The question ANOVA asks is this: are the group means more spread out than you would expect given how much variability exists within each group?

Think of it this way. Suppose three classes of students take the same exam. Class 1 averages 74, Class 2 averages 76, and Class 3 averages 75. The means are slightly different, but within each class the scores range from about 50 to 95. The means barely budge compared to the wide spread of individual scores within each class. That small wiggle among the means could easily be due to the random mix of students in each class. ANOVA would probably not reject the null hypothesis.

Now suppose the three class averages are 62, 78, and 91, while within each class the scores only range from about 5 points below the mean to 5 points above. The means are spread far apart relative to how tightly packed the individual scores are within each class. It would be very hard to explain that pattern by random variation alone. ANOVA would almost certainly reject the null hypothesis.

This is the logic of the ***F*-statistic**. It is a ratio.

$$F = \frac{\text{variability between groups}}{\text{variability within groups}}$$

When the group means are far apart relative to the spread within each group, the *F*-statistic is large. When the group means are close together relative to the within-group spread, the *F*-statistic is small, close to 1. An *F*-statistic of 1 means the between-group variability is about the same as the within-group variability, exactly what you would expect if the groups all came from the same population.

Here is another way to think about it. Imagine you are standing in a hallway and you can hear three different conversations happening in three rooms. If the rooms are full of people who are all talking at roughly the same volume (lots of within-group noise) and the three conversations are about equally loud on average (little between-group difference), you would have a hard time telling the rooms apart from the hallway. But if one room is whispering, one is speaking normally, and one is practically shouting, and within each room people are fairly consistent, you would immediately notice that the rooms are different. ANOVA formalizes that intuition. The question is always whether the signal between groups rises above the noise within groups.

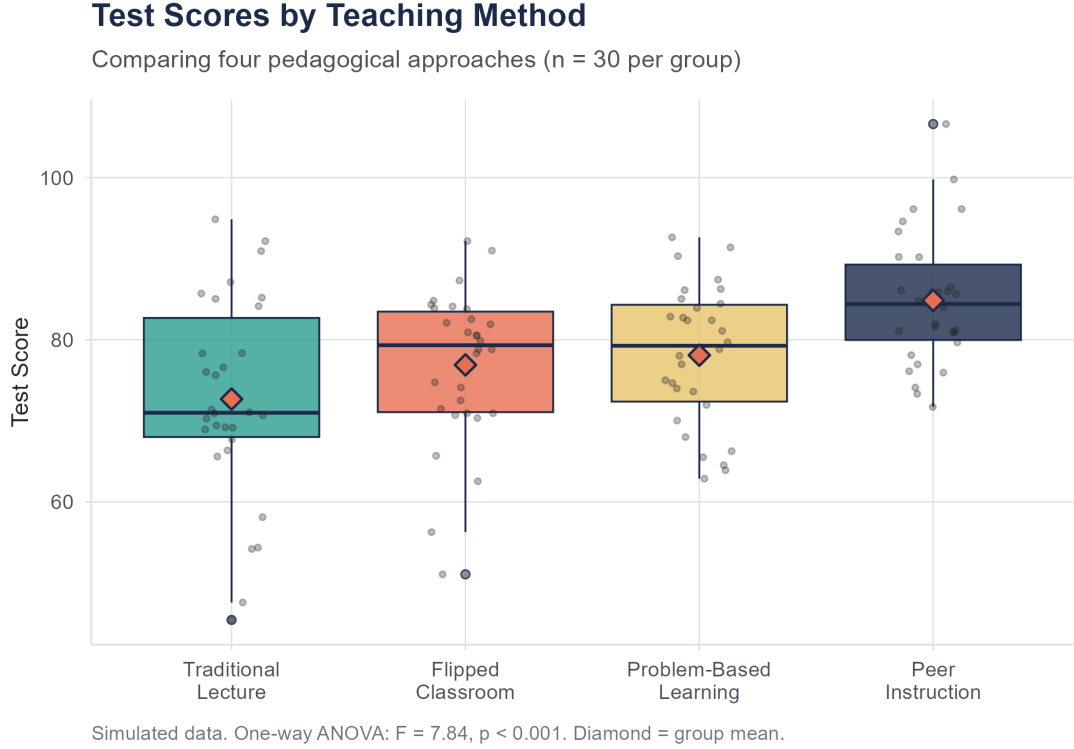


Figure 9.2: Boxplots comparing test scores across four teaching methods (traditional lecture, flipped classroom, problem-based learning, and peer instruction), with individual data points and group means marked. The variation between group means relative to the spread within each group is what the ANOVA F -statistic measures.

9.4.1 Making It Precise

Let us formalize this. Suppose you have k groups, with n_i observations in group i and a total of $N = n_1 + n_2 + \dots + n_k$ observations overall. Let $\bar{x}_i$ be the mean of group i and $\bar{x}$ be the overall mean of all observations combined, often called the **grand mean**.

The total variability in the data can be decomposed into two pieces.

Between-group variability measures how much the group means differ from the grand mean.

$$SS_{\text{between}} = \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2$$

Each group mean's distance from the grand mean is squared, weighted by the group size, and summed up. If the group means are all equal to the grand mean, this quantity is zero. The more the group means spread out, the larger it gets.

Within-group variability measures how much individual observations differ from their own group mean.

$$SS_{\text{within}} = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

This is the pooled variability inside all the groups. It captures the “noise” or natural spread that exists even within a single group. Think of it as the variability that has nothing to do with group membership.

The total sum of squares is

$$SS_{\text{total}} = SS_{\text{between}} + SS_{\text{within}}$$

This decomposition is fundamental. The total variability in your data can be split into the part explained by group membership and the part that remains unexplained, just the natural scatter within groups.

To compute the F -statistic, we need to convert these sums of squares into **mean squares** by dividing by their respective degrees of freedom.

$$MS_{\text{between}} = \frac{SS_{\text{between}}}{k - 1}$$

$$MS_{\text{within}} = \frac{SS_{\text{within}}}{N - k}$$

The between-group degrees of freedom are $k - 1$ because k group means have $k - 1$ independent pieces of information after accounting for the grand mean. The within-group degrees of freedom are $N - k$ because N total observations lose k degrees of freedom from estimating the k group means.

The F -statistic is then

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}}$$

This statistic follows an **F-distribution** with $k - 1$ and $N - k$ degrees of freedom, assuming the null hypothesis is true. The F -distribution is always positive and right-skewed. Under the null hypothesis, the expected value of F is approximately 1. Values much larger than 1 provide evidence against the null hypothesis.

9.4.2 The ANOVA Table

Results from an ANOVA are traditionally presented in a table that organizes all of these pieces.

Source	Sum of Squares	df	Mean Square	F	p -value
Between groups	SS_{between}	$k - 1$	MS_{between}	F	p
Within groups	SS_{within}	$N - k$	MS_{within}		
Total	SS_{total}	$N - 1$			

You will see this table in virtually every piece of statistical software. Learning to read it is essential.

9.4.3 A Worked Example

A researcher wants to know whether background music affects concentration. The researcher recruits 30 participants and randomly assigns 10 to each of three conditions: silence, classical music, and pop music with lyrics. Each participant completes a timed concentration task, and the researcher records the number of errors. The data produce the following group statistics.

Group	n	Mean errors	Standard deviation
Silence	10	4.2	1.8
Classical	10	5.1	2.0
Pop with lyrics	10	7.8	2.3

The grand mean is $\bar{x} = \frac{10(4.2) + 10(5.1) + 10(7.8)}{30} = \frac{171}{30} = 5.7$.

Between-group sum of squares.

$$SS_{\text{between}} = 10(4.2 - 5.7)^2 + 10(5.1 - 5.7)^2 + 10(7.8 - 5.7)^2$$

$$= 10(2.25) + 10(0.36) + 10(4.41) = 22.5 + 3.6 + 44.1 = 70.2$$

Within-group sum of squares. Using the relationship $SS_{\text{within}} = \sum_{i=1}^k (n_i - 1)s_i^2$, we get

$$SS_{\text{within}} = 9(1.8)^2 + 9(2.0)^2 + 9(2.3)^2 = 9(3.24) + 9(4.00) + 9(5.29) = 29.16 + 36.00 + 47.61 = 112.77$$

Mean squares.

$$MS_{\text{between}} = \frac{70.2}{3 - 1} = \frac{70.2}{2} = 35.1$$

$$MS_{\text{within}} = \frac{112.77}{30 - 3} = \frac{112.77}{27} = 4.177$$

F-statistic.

$$F = \frac{35.1}{4.177} = 8.40$$

With $df_1 = 2$ and $df_2 = 27$, an F -statistic of 8.40 yields a p -value of approximately 0.0015. At the $\alpha = 0.05$ level, we reject the null hypothesis and conclude that the mean number of errors is not the same across all three music conditions.

But here is a critical point. The F -test tells us that the three group means are not all equal. It does not tell us which groups differ from which. Is silence different from pop? Is classical different from pop? Is silence different from classical? The omnibus F -test cannot answer these questions. For that, we need post-hoc tests.

9.5 Assumptions of ANOVA

Before we move to post-hoc comparisons, let us be clear about what ANOVA requires to be trustworthy. There are three main assumptions, and they parallel the assumptions you encountered with t -tests.

9.5.1 Independence

The observations must be independent of one another. One participant's score should not influence another participant's score. Random assignment to groups (in an experiment) or random sampling from populations (in an observational study) generally ensures independence. Violations of independence are serious and can badly distort both the F -statistic and the p -value.

Common violations include repeated measures on the same participants (for which you need repeated-measures ANOVA, a topic for a more advanced course), clustering (students nested within classrooms, patients nested within hospitals), and time-series data where observations close in time are correlated.

9.5.2 Normality

ANOVA assumes that the observations within each group come from normally distributed populations. In practice, this assumption is less fragile than it sounds. The F -test is moderately resistant to violations of normality, especially when sample sizes are reasonably large and roughly equal across groups. A common rule of thumb is that with about 20 or more observations per group the Central Limit Theorem provides enough cover for the F -test even when individual observations are not normal — a heuristic, not a hard cutoff, and one you will see across most introductory and applied statistics texts.

With small samples and heavily skewed data, the normality assumption becomes more important. You can check it by examining histograms or normal probability plots of the data within each group. If the data look severely non-normal and your groups are small, you might consider a nonparametric alternative like the **Kruskal-Wallis test** (introduced in Chapter 8 alongside the Wilcoxon tests), which works on the ranks of the data and does not assume normality.

9.5.3 Equal Variances (Homogeneity of Variance)

ANOVA assumes that the population variances are the same across all groups. This is sometimes called the **homogeneity of variance** or **homoscedasticity** assumption. You can check it informally by comparing the standard deviations across groups. A common rule of thumb is that ANOVA is reasonably trustworthy as long as the largest group standard deviation is no more than about twice the smallest group standard deviation.

For a more formal check, **Levene's test** evaluates whether the group variances are statistically distinguishable. If you find evidence of unequal variances, you can use **Welch's ANOVA**, which

adjusts the degrees of freedom to account for the heterogeneity, much as Welch's t -test does for the two-group case.

In our music and concentration example, the standard deviations were 1.8, 2.0, and 2.3. The ratio of the largest to the smallest is $2.3/1.8 = 1.28$, well within the rule-of-thumb guideline. The equal variance assumption looks reasonable here.

9.6 Post-Hoc Tests: Finding Where the Differences Are

You have run your ANOVA and obtained a small p -value. You can conclude that the group means are not all equal. Now what?

You want to know which specific groups differ from which. This is the job of **post-hoc tests** (from the Latin for “after this”). These are pairwise comparisons that are run only after a statistically meaningful F -test, and they include adjustments to control the overall Type I error rate.

9.6.1 Tukey's Honestly Significant Difference (HSD)

The most widely used post-hoc test is **Tukey's HSD**, developed by the statistician John Tukey. It compares every possible pair of group means while controlling the family-wise error rate at α .

For each pair of groups i and j , the Tukey HSD computes

$$q = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{MS_{\text{within}}/n}}$$

where n is the number of observations per group (assuming equal group sizes) and q is compared to the studentized range distribution with parameters k (the number of groups) and $N - k$ (the within-group degrees of freedom).

In the case of unequal group sizes, a modified version called the Tukey-Kramer method is used, replacing n with a harmonic mean or adjusting the standard error for each pair. Most statistical software handles this automatically.

Tukey's HSD produces a confidence interval and a p -value for each pairwise comparison. A pair of groups is considered different if the confidence interval for their mean difference does not include zero, or equivalently, if the adjusted p -value is less than α .

Returning to our music and concentration example, Tukey's HSD would compare three pairs.

Comparison	Difference in means	Adjusted p -value	Conclusion
Silence vs. Classical	$5.1 - 4.2 = 0.9$	0.578	Not distinguishable
Silence vs. Pop	$7.8 - 4.2 = 3.6$	0.001	Different
Classical vs. Pop	$7.8 - 5.1 = 2.7$	0.014	Different

The pattern is clear. Pop music with lyrics leads to more errors than either silence or classical music. Silence and classical music do not produce distinguishably different error counts. The overall F -test told us something was different. Tukey's HSD told us what.

9.6.2 Other Post-Hoc Options

Tukey's HSD is not the only post-hoc procedure. Others you may encounter include

Bonferroni correction. Divide your significance level α by the number of comparisons. With three comparisons at $\alpha = 0.05$, each comparison is tested at $0.05/3 = 0.0167$. This is simple and conservative, meaning it is less likely to find differences that are real but also less likely to produce false positives. The Bonferroni correction works with any type of test, not just ANOVA follow-ups.

Scheffé's method. The most conservative of the common post-hoc tests. It controls the Type I error rate for all possible contrasts, not just pairwise comparisons. Because it is so conservative, it has less power to detect real differences.

Dunnett's test. Used when you want to compare each group to a single control group, rather than comparing all groups to each other. If you have one control condition and three treatment conditions, Dunnett's test makes only three comparisons instead of six.

For most introductory applications, Tukey's HSD is the standard choice. It balances Type I error control with statistical power and is available in every major statistics package.

In practice, you do not compute post-hoc tests by hand. R handles the bookkeeping: `TukeyHSD()` follows the base `aov()` function and produces a table of pairwise comparisons with adjusted p -values and confidence intervals. The companion R walkthrough on this book's website shows the full sequence end to end — fit the ANOVA, check assumptions, apply Tukey's HSD, and report effect sizes — for the music-and-concentration example used in this section. The point of doing

the calculation by hand on a small example, as we did above, is to make the logic concrete, not to suggest you would ever need to do it that way for real data.

One important detail about post-hoc tests. They should only be conducted after a statistically meaningful omnibus F -test. If the overall ANOVA fails to reject the null hypothesis, there is no justification for hunting through pairwise comparisons looking for differences. The F -test is the gatekeeper. If it says “no convincing evidence that the means differ,” you should not go looking for differences anyway, because any you find are likely to be Type I errors. Some researchers ignore this convention and run pairwise comparisons regardless, which brings you right back to the multiple comparisons problem that ANOVA was designed to avoid.

There is an exception worth mentioning. If you had planned comparisons (also called a priori contrasts) specified before collecting data, you can test those specific comparisons without first running an omnibus F -test. Planned comparisons are hypothesis-driven, not data-driven, and they carry less risk of capitalizing on chance. For example, if a researcher studying four teaching methods had a specific theory that problem-based learning would outperform traditional lecture, they could test that one comparison directly. The key is that the comparison was specified before looking at the data, not selected after seeing which means happened to be the farthest apart.

Note

John Tukey, who developed the HSD test, was a prolific statistician of the twentieth century. He also invented the box plot, is credited with early published uses of the terms “bit” (for binary digit) and “software,” and developed the Fast Fourier Transform algorithm with James Cooley. His 1977 book *Exploratory Data Analysis* helped shift statistics from a purely confirmatory discipline to one that also valued looking at data with fresh eyes before testing hypotheses. If you find yourself appreciating a box plot, you have Tukey to thank.

9.7 Effect Sizes for Group Comparisons

A small p -value tells you that the group means are probably not all equal. It does not tell you how much they differ, or whether the difference matters in practical terms. For that, you need an effect size.

9.7.1 Eta-Squared (η^2)

The most common effect size for ANOVA is **eta-squared**, defined as

$$\eta^2 = \frac{SS_{\text{between}}}{SS_{\text{total}}}$$

Eta-squared represents the proportion of total variability in the outcome that is accounted for by group membership. It ranges from 0 to 1. An η^2 of 0.10 means that 10% of the variability in the outcome is associated with group differences. The remaining 90% is within-group variability.

In our music and concentration example

$$\eta^2 = \frac{70.2}{70.2 + 112.77} = \frac{70.2}{182.97} = 0.384$$

About 38% of the variability in concentration errors is accounted for by the music condition. That is a large effect by conventional standards.

Cohen's (1988) guidelines for interpreting η^2 are

η^2	Interpretation
0.01	Small effect
0.06	Medium effect
0.14	Large effect

These are rough benchmarks, not rigid rules. A “small” effect in medicine might save thousands of lives. A “large” effect in education might be driven by a poorly designed study. Always interpret effect sizes in context.

9.7.2 Partial Eta-Squared (η_p^2)

When your ANOVA includes more than one factor (for example, music condition and time of day), **partial eta-squared** becomes the preferred measure. It represents the proportion of variance accounted for by a particular factor, after removing the variance accounted for by other factors.

$$\eta_p^2 = \frac{SS_{\text{effect}}}{SS_{\text{effect}} + SS_{\text{error}}}$$

In a one-way ANOVA with a single factor, partial eta-squared and eta-squared are identical. The distinction only matters when you have multiple factors, which is the setting of factorial ANOVA, typically covered in a second statistics course. For now, just know that partial eta-squared exists and is what most software reports by default in multi-factor designs.

9.7.3 Cramér's V for Categorical Comparisons

Eta-squared applies when your outcome variable is numerical and you are comparing means. But what if both your grouping variable and your outcome are categorical? The Bertrand and Mullainathan study is a good example. The grouping variable is the name type (White-sounding vs. Black-sounding), and the outcome is whether the resume received a callback (yes or no). Both are categorical. Here you would use a chi-square test (as you learned in Chapter 8), and the appropriate effect size is **Cramér's V**.

Cramér's V is defined as

$$V = \sqrt{\frac{\chi^2}{n \times (m - 1)}}$$

where χ^2 is the chi-square test statistic, n is the total sample size, and m is the smaller of the number of rows or the number of columns in the contingency table.

Cramér's V ranges from 0 to 1. A value of 0 means no association between the variables. A value of 1 means a perfect association. For a 2-by-2 table, Cramér's V is identical to the absolute value of the phi coefficient (ϕ).

Guidelines for interpreting Cramér's V depend on the size of the table. For a 2-by-2 table

Cramér's V	Interpretation
0.10	Small association
0.30	Medium association
0.50	Large association

For larger tables, the thresholds are lower because the degrees of freedom increase.

9.8 When to Use Which Test: A Decision Framework

At this point, you have encountered several tests for comparing groups. Here is a framework to help you choose the right one.

Step 1: What type is your outcome variable?

If your outcome is numerical (continuous or discrete quantities like test scores, incomes, wait times), you reach for *t*-tests or ANOVA.

If your outcome is categorical (counts or proportions, like callback vs. no callback, party affiliation, disease status), you reach for chi-square tests.

Step 2: How many groups are you comparing?

If you are comparing exactly two groups on a numerical outcome, use a **two-sample *t*-test** (or its Welch variant if variances are unequal).

If you are comparing three or more groups on a numerical outcome, use a **one-way ANOVA**, followed by post-hoc tests if the *F*-test is statistically meaningful.

If you are comparing the distribution of a categorical outcome across two or more groups, use a **chi-square test of independence** (or Fisher's exact test if expected cell counts are very small).

Step 3: Are your groups independent or related?

If the groups contain different individuals (as in a between-subjects experiment or comparing different populations), use the independent-samples versions of these tests.

If the groups contain the same individuals measured at different times or under different conditions, you need the paired or repeated-measures versions (paired *t*-test, repeated-measures ANOVA). These are not the focus of this chapter but are important to recognize.

Here is the framework in table form.

Outcome type	Number of groups	Groups independent?	Test
Numerical	2	Yes	Two-sample <i>t</i> -test
Numerical	2	No (paired)	Paired <i>t</i> -test
Numerical	3 or more	Yes	One-way ANOVA
Numerical	3 or more	No (repeated)	Repeated-measures ANOVA
Categorical	2 or more	Yes	Chi-square test of independence

This table covers the most common scenarios. Real research sometimes calls for more specialized methods, but if you can navigate this decision tree, you can handle the vast majority of group comparison problems.

A few additional considerations when choosing your test.

Sample size matters. With very small samples (fewer than 5 per group), the assumptions of ANOVA and the chi-square test become harder to justify. For small-sample ANOVA with non-normal data, the Kruskal-Wallis test (Chapter 8) is the rank-based alternative. For chi-square tests with small expected cell counts (typically less than 5), **Fisher's exact test** is the recommended substitute. Fisher's exact test computes the p -value directly from the hypergeometric distribution rather than relying on the chi-square approximation, which is why it remains valid in small-sample situations where the chi-square approximation would not be. R's `fisher.test()` produces it in one line.

Ordinal outcomes fall in a gray area. If your outcome is ordinal, like a satisfaction rating on a 1-to-5 scale, you have a choice. Many researchers treat ordinal data as numerical and use ANOVA, especially when the scale has five or more levels and the data are reasonably symmetric. Others prefer nonparametric methods like the Kruskal-Wallis test, which makes fewer assumptions about the measurement scale. There is no universal consensus. The choice depends on the specific data, the number of scale points, and how comfortable you are treating the distances between levels as approximately equal.

The relationship between the t -test and ANOVA. When you have exactly two groups, a one-way ANOVA and a two-sample t -test give identical results. The F -statistic equals the square of the t -statistic ($F = t^2$), and the p -values are the same. ANOVA generalizes the t -test to more than two groups. If someone asks you whether to use a t -test or ANOVA for two groups, the honest answer is that it does not matter, you will get the same answer either way.

9.9 Beyond the Résumé Study: ANOVA in Practice

Let us work through a more extended example to see how all these pieces fit together in practice.

A public health researcher wants to know whether awareness of nutrition labeling differs across four regions of the United States (Northeast, Southeast, Midwest, and West). The researcher surveys a random sample of adults from each region and administers a nutrition knowledge quiz scored from 0 to 100. The results produce the following data.

Region	n	Mean score	Standard deviation
Northeast	45	68.3	12.1
Southeast	52	61.7	14.5
Midwest	48	64.9	11.8
West	50	70.2	13.0

Total $N = 195$. Number of groups $k = 4$.

Step 1: Check assumptions.

Independence. The participants were randomly sampled from each region independently. This assumption appears met.

Normality. With group sizes ranging from 45 to 52, the Central Limit Theorem provides adequate coverage even if the underlying distributions are somewhat non-normal. The researcher checks histograms and finds no severe skewness.

Equal variances. The standard deviations range from 11.8 to 14.5. The ratio of the largest to the smallest is $14.5/11.8 = 1.23$, well within the guideline of 2. The equal variance assumption appears reasonable.

Step 2: Compute the grand mean.

$$\bar{x} = \frac{45(68.3) + 52(61.7) + 48(64.9) + 50(70.2)}{195} = \frac{3073.5 + 3208.4 + 3115.2 + 3510.0}{195} = \frac{12907.1}{195} =$$

Step 3: Compute the between-group sum of squares.

$$\begin{aligned} SS_{\text{between}} &= 45(68.3 - 66.19)^2 + 52(61.7 - 66.19)^2 + 48(64.9 - 66.19)^2 + 50(70.2 - 66.19)^2 \\ &= 45(4.4521) + 52(20.1601) + 48(1.6641) + 50(16.0801) \\ &= 200.34 + 1048.33 + 79.88 + 804.01 = 2132.56 \end{aligned}$$

Step 4: Compute the within-group sum of squares.

$$\begin{aligned}SS_{\text{within}} &= 44(12.1)^2 + 51(14.5)^2 + 47(11.8)^2 + 49(13.0)^2 \\&= 44(146.41) + 51(210.25) + 47(139.24) + 49(169.00) \\&= 6442.04 + 10722.75 + 6544.28 + 8281.00 = 31990.07\end{aligned}$$

Step 5: Compute mean squares and the F -statistic.

$$\begin{aligned}MS_{\text{between}} &= \frac{2132.56}{4 - 1} = \frac{2132.56}{3} = 710.85 \\MS_{\text{within}} &= \frac{31990.07}{195 - 4} = \frac{31990.07}{191} = 167.49 \\F &= \frac{710.85}{167.49} = 4.24\end{aligned}$$

Step 6: Find the p -value. With $df_1 = 3$ and $df_2 = 191$, an F -statistic of 4.24 yields a p -value of approximately 0.006.

Step 7: Interpret. At $\alpha = 0.05$, we reject the null hypothesis. The data provide convincing evidence that mean nutrition knowledge scores differ across the four regions.

Step 8: Post-hoc comparisons. Tukey's HSD reveals that the Southeast differs from both the Northeast ($p = 0.042$) and the West ($p = 0.003$). No other pairwise differences are statistically distinguishable at $\alpha = 0.05$.

Step 9: Effect size.

$$\eta^2 = \frac{2132.56}{2132.56 + 31990.07} = \frac{2132.56}{34122.63} = 0.062$$

About 6% of the variability in nutrition knowledge scores is associated with region. By Cohen's guidelines, this is a medium effect. Region plays a measurable but modest role in explaining differences in nutrition knowledge.

Notice the interplay between statistical meaning and practical meaning in this example. The p -value of 0.006 tells us we have strong evidence that the regional means are not all equal. The η^2 of

0.062 tells us that region explains about 6% of the variation, which leaves 94% explained by other factors, things like individual education, socioeconomic status, access to healthcare, cultural food practices, and personal interest in nutrition. A statistically detectable difference is not necessarily a large or important one. The effect size helps you keep perspective.

9.10 The Ethics of Comparing Groups

Every time you compare groups, you make a choice about framing. How you describe the groups, what language you use, and what explanations you offer for any differences you find all carry ethical weight. Nowhere is this more important than when the groups are defined by race, gender, socioeconomic status, disability, or other dimensions of identity.

! Ethics Moment: Deficit Framing

Deficit framing occurs when researchers describe group differences in a way that locates the problem within the group that scores lower, rather than in the systems and structures that produce the disparity.

Consider the National Assessment of Educational Progress (NAEP), sometimes called “The Nation’s Report Card.” NAEP data consistently show that Black and Hispanic students score lower on average than White students in reading and mathematics. This pattern has been described for decades as the “achievement gap.”

That term, “achievement gap,” is a deficit frame. It centers the analysis on what certain groups of students have not achieved. It subtly implies that the explanation lies within the students or their families or their cultures. It directs attention away from the systems that produced the disparity, including unequal school funding, residential segregation, differential access to experienced teachers, and the long-term consequences of discriminatory policies in housing, lending, and employment.

An alternative framing is the “opportunity gap” or the “education debt,” a term used by education scholar Gloria Ladson-Billings. This framing asks not “Why are these students scoring lower?” but “What resources, opportunities, and supports have been systematically withheld from these students?” The data are the same. The statistical analysis is the same. The interpretation shifts from blaming the group to examining the system.

The same issue arises in public health. The CDC Wonder database shows large disparities in life expectancy, infant mortality, and chronic disease prevalence across racial groups. Reporting that “Black Americans have higher rates of hypertension” is a factual statement, but

without context it can slide into deficit framing. The context matters. Decades of research have linked these disparities to residential segregation, environmental racism, unequal access to healthcare, the physiological effects of chronic stress from discrimination, and the historical exclusion of Black Americans from wealth-building opportunities like homeownership and higher education.

When you analyze group comparisons, ask yourself these questions.

- Am I describing what the data show, or am I implying a cause that the data cannot support?
- Am I locating the “problem” within the group that scores lower, or am I examining the structures that produce the disparity?
- Would my framing read as fair to a careful reader who belongs to the group I am describing?
- Am I comparing groups to a “default” that I have left unexamined?

The goal is not to avoid making group comparisons. Group comparisons are essential for identifying disparities, documenting differential treatment, and holding institutions accountable. The Bertrand and Mullainathan study is a powerful example of a group comparison that surfaced evidence of disparate treatment. The goal is to make those comparisons with precision, honesty, and an awareness that your framing shapes how your audience understands the world.

9.11 Connecting ANOVA to the Résumé Study

The B&M study itself does not call for ANOVA — its outcome (callback or no callback) is categorical, so the right tool is a chi-square test or a two-proportion z-test. The chapter methods apply once we extend the design.

Imagine the same experiment run with three name categories — White-sounding, Black-sounding, and Hispanic-sounding — and the same yes/no callback outcome. A chi-square test of independence on the resulting 3×2 table would tell you whether callback rates differ across the three name types. If the omnibus p -value is small, follow up with pairwise chi-square tests on 2×2 subtables under a Bonferroni-corrected significance level to identify which pairs of name types differ.

Now suppose the outcome were the *number of days until a callback* — a numerical variable — rather than a yes/no. With three name-type groups, that is exactly the setup for a one-way ANOVA, followed by Tukey’s HSD to identify specific group differences. Same study, different measurement, different tool.

The choice of test always follows from the type of outcome variable and the number of groups. The research question determines the groups; the measurement determines the test.

 AI Reality Check: The Multiple Subgroup Problem at Machine Scale

Modern analytical tools and AI systems can slice data into hundreds or thousands of subgroups and test for differences across all of them in seconds. A machine learning pipeline might automatically compare outcomes across every combination of age group, gender, region, income bracket, and education level. With enough subgroups, some comparisons will produce small p -values purely by chance.

This is the multiple comparisons problem on steroids. A human researcher running six pairwise comparisons might recognize the need for a Bonferroni correction or Tukey's HSD. An automated system running 500 comparisons might flag dozens of "differences" that are nothing but noise, and present them in a dashboard with the same visual weight as genuine findings.

The problem is compounded by the fact that automated systems do not understand context. An algorithm does not know that a 0.3-point difference in customer satisfaction between two regions is practically meaningless, even if it is statistically distinguishable with a large enough sample. It does not know that a disparity in loan approval rates between racial groups demands a fundamentally different kind of attention than a disparity in ice cream flavor preferences between age groups.

When you encounter AI-generated analyses that compare many subgroups, ask the following.

- How many comparisons were tested? If the report highlights 5 differences, were 5 comparisons run or 500?
- Was any correction applied for multiple testing? If not, expect a high proportion of false positives.
- Are the reported effect sizes practically meaningful, or just statistically detectable?
- Is the tool treating all group differences as equivalent, or does it distinguish between differences that matter and differences that do not?

The ability to run comparisons at scale is powerful. Without the statistical discipline to control error rates, it is also dangerous.

9.12 A Note on the Chi-Square Test

We introduced the chi-square concept in Chapter 8; here we apply it to comparing groups across categories.

While the focus of this chapter is ANOVA, the decision framework above includes the chi-square test, and you should understand how it connects. The **chi-square test of independence** evaluates whether two categorical variables are associated. You construct a contingency table showing the observed counts, calculate the expected counts under the assumption that the variables are independent, and compute

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where O is the observed count in each cell and E is the expected count. The sum is taken over all cells in the table. Large values of χ^2 indicate that the observed data deviate from what would be expected if the variables were independent.

For the résumé study, the contingency table would look like this.

	Callback	No callback	Total
White-sounding names	235	2,200	2,435
Black-sounding names	157	2,278	2,435
Total	392	4,478	4,870

The expected count for each cell is $\frac{\text{row total} \times \text{column total}}{\text{grand total}}$. The chi-square statistic tests whether the callback rates differ between the two name types. In this case, the answer is a clear yes.

The chi-square test generalizes to tables with more than two rows or columns. If you had three name types and two outcomes (callback vs. not), you would have a 3-by-2 table, and the chi-square test would tell you whether callback rates differ across the three name types simultaneously, controlling for multiple comparisons in the same way that ANOVA tests all group means at once.

The chi-square test has its own assumptions. The expected count in each cell of the contingency table should be at least 5 for the chi-square approximation to be reliable. If some expected counts are smaller than 5, you can combine categories if it makes substantive sense, or you can use Fisher's exact test, which does not rely on the chi-square approximation. The observations must also be independent, meaning each case contributes to only one cell in the table.

One more parallel worth noting. Just as the F -test in ANOVA is an omnibus test that tells you something differs but not what, the chi-square test with more than two groups tells you that the distribution of outcomes is not the same across all groups, but it does not pinpoint which groups differ from which. For that, you need follow-up pairwise comparisons, typically chi-square tests on 2-by-2 subtables with a Bonferroni correction to the significance level. The logic is exactly the same as with ANOVA post-hoc testing. The omnibus test opens the door, and the pairwise comparisons tell you which rooms to look in.

9.13 Pulling It All Together

This chapter has covered a lot of ground. Let us step back and see the big picture.

You started with a simple problem: comparing groups. When you have two groups and a numerical outcome, a t -test does the job. When you have more than two groups, running multiple t -tests inflates your Type I error rate. ANOVA solves this by testing all groups simultaneously with a single F -test.

The F -statistic compares how much the group means differ from each other (between-group variability) to how much individual observations differ within their groups (within-group variability). A large F means the group means are spread out relative to the within-group noise, which is evidence that at least some groups are different.

If the F -test yields a small p -value, post-hoc tests like Tukey's HSD identify which specific pairs of groups differ, while controlling the overall error rate. Effect sizes like eta-squared quantify how much of the total variability is explained by group membership, giving you a measure of practical importance beyond the binary of "statistically different or not."

When your outcome is categorical rather than numerical, the chi-square test of independence serves a parallel role, and Cramér's V provides the effect size.

Throughout all of this, the ethical imperative is clear. Group comparisons are tools for understanding the world, not for ranking people. The same analysis that documents discrimination can, if framed carelessly, reinforce the stereotypes that produce it. Your job as a statistical thinker is to let the data speak honestly while being honest about what the data cannot say.

The Bertrand and Mullainathan study did this effectively. It did not compare Black and White job applicants. It compared identical résumés with different names. The experimental design isolated the variable of interest, perceived race, from all confounding factors. The analysis documented a

disparity. The interpretation correctly placed the source of that disparity not in the applicants but in the employers who evaluated them.

That is what good group comparison looks like.

One final thought before we move to the exercises. The methods in this chapter, ANOVA, chi-square tests, post-hoc comparisons, and effect sizes, are among the most widely used tools in applied statistics. They appear in medical research comparing treatments, in education research comparing curricula, in business analytics comparing customer segments, in public policy research comparing program outcomes. If you understand the logic of comparing group means and proportions, you understand a framework that extends to hundreds of practical applications.

But tools are only as good as the person using them. An ANOVA can tell you that three groups have different averages. It cannot tell you why. It cannot tell you what to do about it. It cannot tell you whether the groups were fairly defined, whether the measurements were valid, or whether your interpretation serves truth or merely confirms what you already believed. Those are human judgments, and they matter as much as the math.

In the next chapter, we move from comparing groups to modeling relationships between variables — the subject of regression. But regression, as you will see, rests on the same foundation as ANOVA. The idea that you can partition variability into explained and unexplained components, and use that partition to draw inferences, carries straight through from this chapter into the next. The tools grow more flexible, but the logic stays the same.

Try It Online

Open the **ANOVA Visualizer** on the companion website. Adjust group means and within-group variability and watch the F -statistic update in real time. Try making all group means identical and see what happens to F . Then spread the means apart and see how quickly it climbs.

9.14 Looking Ahead

This chapter has given you tools for comparing groups: ANOVA for means, chi-square tests for proportions, post-hoc tests for identifying which groups differ, and effect sizes for quantifying how much they differ. But all of these methods treat the groups as the explanatory variable and ask whether the response variable differs across them. In the next chapter, we shift to a different kind of question. Instead of comparing discrete groups, we model the relationship between two continuous

variables. If spending per student varies across states, and test scores vary across states, how do they vary together? How much does one variable change, on average, when the other changes by one unit? That is the subject of regression analysis, and it turns out to be a natural extension of the ANOVA framework. The idea that you can partition variation into explained and unexplained components carries straight through.

9.15 Further Reading and References

The following works are cited in this chapter or provide valuable additional context for the ideas covered here.

On the résumé audit study: Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, 94(4), 991–1013. (The widely cited study that provided this chapter’s opening example, demonstrating racial discrimination in callback rates using a randomized field experiment.)

On multiple comparisons and the false discovery problem: Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1), 289–300. (The foundational paper introducing the false discovery rate procedure, an alternative to the Bonferroni correction that is less conservative and widely used in genomics, neuroscience, and other fields with large numbers of comparisons.)

On deficit framing in research: Tuck, E. (2009). Suspending damage: A letter to communities. *Harvard Educational Review*, 79(3), 409–428. (A foundational critique of “damage-centered research” that frames marginalized communities primarily in terms of their deficits, arguing for approaches that also document resilience, strength, and aspiration.)

On effect sizes and their interpretation: Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates. (The classic reference for effect size measures including Cohen’s d , η^2 , and the benchmarks for “small,” “medium,” and “large” effects that are widely used across disciplines.)

For further reading on comparing groups: Maxwell, S. E., Delaney, H. D., & Kelley, K. (2017). *Designing experiments and analyzing data: A model comparison perspective* (3rd ed.). Routledge. (A comprehensive treatment of ANOVA, factorial designs, repeated measures, and the relationship between ANOVA and regression.)

Ladson-Billings, G. (2006). From the achievement gap to the education debt: Understanding achievement in U.S. schools. *Educational Researcher*, 35(7), 3–12. (The paper that introduced the “education debt” framework discussed in this chapter’s Ethics Moment, arguing that framing disparities as “gaps” locates the problem in communities rather than in the systems that produced the disparity.)

9.16 Key Terms

- **Analysis of Variance (ANOVA):** A statistical method for testing whether the means of three or more groups are equal, using the F -statistic to compare between-group and within-group variability.
- **Between-group variability (SS_{between}):** The sum of squared differences between each group mean and the grand mean, weighted by group size. Measures how much the group means differ from each other.
- **Bonferroni correction:** A method for adjusting the significance level in multiple comparisons by dividing α by the number of tests performed.
- **Chi-square test of independence:** A test for whether two categorical variables are associated, based on comparing observed cell counts to the counts expected under independence.
- **Cramér’s V:** An effect size measure for the association between two categorical variables, derived from the chi-square statistic. Ranges from 0 to 1.
- **Deficit framing:** Describing group differences in a way that locates the problem within the lower-performing group, rather than examining the structures and systems that produce the disparity.
- **Degrees of freedom:** In ANOVA, the between-group degrees of freedom are $k - 1$ and the within-group degrees of freedom are $N - k$, where k is the number of groups and N is the total sample size.
- **Effect size:** A measure of the magnitude of a difference or association, independent of sample size. Common effect sizes for group comparisons include eta-squared, partial eta-squared, and Cramér’s V.
- **Eta-squared (η^2):** The proportion of total variability in the outcome that is accounted for by group membership. Calculated as $SS_{\text{between}}/SS_{\text{total}}$.
- **F-distribution:** The probability distribution of the F -statistic under the null hypothesis, characterized by two degrees-of-freedom parameters.

- ***F*-statistic**: The ratio of between-group mean square to within-group mean square. Large values provide evidence that the group means are not all equal.
 - **Family-wise error rate**: The probability of making at least one Type I error across a set of multiple comparisons.
 - **Grand mean**: The overall mean of all observations across all groups, denoted $\bar{x}$.
 - **Homogeneity of variance (homoscedasticity)**: The assumption that all groups have the same population variance. Checked informally by comparing group standard deviations or formally using Levene's test.
 - **Kruskal-Wallis test**: A nonparametric alternative to one-way ANOVA that does not assume normality. Based on ranks rather than raw values.
 - **Levene's test**: A formal test for the equality of variances across groups.
 - **Mean square**: A sum of squares divided by its degrees of freedom. Used to compute the *F*-statistic.
 - **Multiple comparisons problem**: The inflation of the overall Type I error rate that occurs when multiple hypothesis tests are conducted simultaneously.
 - **One-way ANOVA**: ANOVA with a single grouping variable (factor).
 - **Partial eta-squared (η_p^2)**: The proportion of variance accounted for by a factor after removing variance explained by other factors. Identical to eta-squared in one-way ANOVA.
 - **Post-hoc test**: A pairwise comparison conducted after a statistically meaningful *F*-test to determine which specific groups differ. Examples include Tukey's HSD, Bonferroni, and Scheffé's method.
 - **Sum of squares**: The sum of squared deviations used to measure variability. In ANOVA, the total sum of squares is partitioned into between-group and within-group components.
 - **Tukey's Honestly Significant Difference (HSD)**: A post-hoc test that compares all pairs of group means while controlling the family-wise error rate.
 - **Welch's ANOVA**: A modification of standard ANOVA that does not assume equal variances across groups.
 - **Within-group variability (SS_{within})**: The sum of squared differences between each observation and its group mean. Measures the natural spread of observations within groups.
-

9.17 Exercises

9.17.1 Check Your Understanding

1. Explain in your own words why running multiple t -tests to compare several group means inflates the Type I error rate. What happens to the probability of at least one false positive as the number of comparisons increases?
2. A one-way ANOVA produces an F -statistic of 1.03 with a p -value of 0.38. What does this tell you about the relationship between the between-group variability and the within-group variability? What do you conclude about the group means?
3. Describe the null and alternative hypotheses for a one-way ANOVA with five groups. Be precise about what the alternative hypothesis does and does not claim.
4. Why is the F -statistic always positive? Can it ever be less than 1? If so, what would that mean?
5. A researcher runs an ANOVA and obtains a p -value of 0.002. The researcher immediately concludes that every group mean is different from every other group mean. What is wrong with this conclusion?
6. What are the three main assumptions of one-way ANOVA? For each, explain what happens to the trustworthiness of the results when the assumption is violated, and name one way to check whether the assumption is met.
7. Explain the difference between SS_{between} and SS_{within} in your own words. What does each one measure, and how do they combine to form SS_{total} ?
8. Why do we need post-hoc tests after a statistically meaningful F -test? Why not just look at the group means and conclude that the highest and lowest means are different?
9. A study reports $\eta^2 = 0.03$ for the effect of teaching method on exam scores. The p -value is 0.01. Is the effect statistically detectable? Is it practically large? Explain the distinction.
10. When is a chi-square test of independence the appropriate choice instead of ANOVA? Give an example of a research question where you would use each test.

9.17.2 Apply It

(See Appendix B for complete variable descriptions for all datasets used in these exercises.)

The following exercises reference the `resume-callbacks.csv` dataset. This dataset contains real data from the field experiment by Bertrand and Mullainathan (2004), “Are Emily and Greg More Employable Than Lakisha and Jamal?”, sourced via the `OpenIntro` package. Each row represents one resume sent to an employer in response to help-wanted ads in Boston and Chicago. Variables in-

clude `resume_id` (a unique identifier), `name_type` (White-sounding or Black-sounding), `gender` (Male or Female, based on the first name assigned), `resume_quality` (High or Low), `years_experience` (a numerical variable), `education` (the level of education listed on the resume: Bachelor or High School), and `callback` (Yes if the resume received a callback, No otherwise).

1. Using the `resume-callbacks.csv` dataset, construct a contingency table showing callback counts by `name_type`. Compute the callback rate for White-sounding names and for Black-sounding names. What is the difference in callback rates?
2. Conduct a chi-square test of independence to test whether callback status is associated with `name_type`. State the null and alternative hypotheses, report the chi-square statistic and p -value, and interpret the result. Calculate Cramér's V and comment on the effect size.
3. Construct a contingency table showing callback counts by education level. Is there evidence that callback rates differ across education levels? Conduct a chi-square test and report your findings.
4. Now consider `years_experience` as the outcome variable. Compute the mean and standard deviation of `years_experience` separately for résumés that received a callback and those that did not. Conduct a two-sample t -test to determine whether the mean years of experience differs between the two callback groups. Report the test statistic, p -value, and your conclusion.
5. Compute the mean and standard deviation of `years_experience` separately for each combination of `name_type` and `gender` (i.e., four groups: White-sounding Female, White-sounding Male, Black-sounding Female, Black-sounding Male). Create a table showing these summary statistics. Based on a visual inspection, do the means appear to differ across the four groups?
6. Conduct a one-way ANOVA to test whether the mean `years_experience` differs across the four `name_type`-by-`gender` groups you created in the previous problem. Report the ANOVA table, including SS_{between} , SS_{within} , degrees of freedom, mean squares, the F -statistic, and the p -value.
7. If your ANOVA in the previous problem produced a statistically meaningful result, conduct Tukey's HSD post-hoc test. Which pairs of groups have statistically distinguishable mean years of experience? Present the results in a table. If the F -test was not statistically meaningful, explain why post-hoc tests should not be conducted.
8. Calculate η^2 for the ANOVA from problem 6. What proportion of the variability in years of experience is accounted for by the four-group classification? How would you characterize this effect size using Cohen's guidelines?
9. The dataset includes both `name_type` and `resume_quality`. Construct a 2-by-2-by-2 contingency table (or two separate 2-by-2 tables) to examine callback rates by `name_type` separately for low-quality and high-quality résumés. Does the gap in callback rates between name types appear

to change depending on resume quality? What does this suggest about how perceived race interacts with qualifications in the labor market?

10. Conduct a chi-square test of independence to test whether callback rates differ by gender. Report your findings. Then conduct separate chi-square tests for the association between `name_type` and callback within each gender group. Are the patterns consistent across gender groups, or does the magnitude of discrimination appear to differ?
11. Re-run the four-group ANOVA on `years_experience` from problem 6, but this time use **Welch's ANOVA** (which does not assume equal variances across groups). In R, `oneway.test(years_experience ~ name_gender, data = ..., var.equal = FALSE)` produces it. Compare the resulting F -statistic and p -value to your earlier classical ANOVA result. Then compute the standard deviation of `years_experience` within each of the four groups. Are the variances close enough that the equal-variance assumption seems reasonable, or does the Welch correction change the conclusion in a substantive way?

9.17.3 Think Deeper

1. The Bertrand and Mullainathan study used names as a proxy for perceived race. Some critics argued that the names might signal socioeconomic status rather than race, and that employers might be responding to perceived social class rather than perceived race. How does this critique relate to the concept of confounding? Does it invalidate the finding that identical résumés receive different treatment based on the name? How would you design a study to disentangle the effects of perceived race and perceived socioeconomic status?
2. Suppose a school district uses ANOVA to compare standardized test scores across five elementary schools and finds a statistically meaningful F -test. The superintendent publicly ranks the schools from “best” to “worst” based on their mean scores. What information is the superintendent ignoring? How might this ranking example illustrate the deficit framing discussed in this chapter? What additional data would you want before drawing conclusions about school quality?
3. An AI-powered hiring platform automatically screens résumés and reports that callback recommendations differ across 47 demographic subgroups at the $p < 0.05$ level for 6 of those subgroups. The company claims that their algorithm has identified 6 areas where bias exists. Using what you learned about multiple comparisons, evaluate this claim. How many “discoveries” would you expect by chance alone? What additional analysis would you recommend before concluding that the algorithm is biased in these 6 areas?
4. A pharmaceutical company tests a new drug in five countries and runs a separate clinical trial in each. The drug shows a statistically meaningful effect in two of the five countries but not

- the other three. The company wants to market the drug only in the two countries where the results were positive. Discuss the statistical and ethical problems with this strategy. How does this relate to the multiple comparisons problem? What would a responsible analysis look like?
5. Return to the NAEP data discussed in the Ethics Moment box. Suppose you are writing a report for a state education department that includes ANOVA results comparing test scores across racial and ethnic groups. Write two versions of a single paragraph interpreting the finding that scores differ across groups. In the first version, use deficit framing (even if it feels uncomfortable). In the second version, reframe the same finding in terms of systemic factors. Reflect on how the two framings might lead to different policy responses.

Chapter 10

Simple Linear Regression

10.1 The Scatter Plot That Launched a Thousand Debates

In 2016, the National Education Association published its annual report on education spending across the fifty states. The numbers invited an obvious question, one that parents, politicians, and school board members have argued about for decades. Does spending more money on education lead to better outcomes?

On one side of the argument, advocates for increased school funding pointed to states like Massachusetts and New Jersey, which spent well above the national average per student and consistently produced some of the highest standardized test scores in the country. On the other side, skeptics pointed to states like Utah, which spent among the least per student yet still posted respectable scores, and the District of Columbia, which spent more per student than nearly anywhere else in the nation yet ranked near the bottom on many achievement measures.

Someone, as someone always does, made a scatter plot. State-level per-pupil spending on the horizontal axis, average test scores on the vertical axis, one dot for each state. And that scatter plot became a Rorschach test for anyone with an opinion about education policy.

If you squinted one way, you could see a general upward trend. More spending, higher scores. If you squinted another way, you could see enormous variability, states that spent similar amounts producing wildly different results. A few outliers complicated both stories. The relationship was there, sort of, but it was messy. It was noisy. It was, in a word, real.

The question is not whether the scatter plot showed a pattern. The question is how to describe that pattern precisely enough that reasonable people can agree on what the data says, even if they

disagree about what to do about it. That is the work of regression analysis, and it is what this chapter is about.

Regression is a family of methods for describing how one variable depends on another. The simplest version, the subject of this chapter, is **simple linear regression**: one explanatory variable, one response variable, and the assumption that the relationship is well summarized by a straight line. The line is a model of the relationship. Other forms of regression handle multiple predictors (Chapter 11), curves rather than lines, categorical outcomes, and many other situations, but the logic in this chapter is the foundation that the rest build on.

For the simple-line case, regression gives us a way to draw a single line through the cloud of points — a line that captures the overall direction of the relationship between two variables while acknowledging the scatter around it. It tells us how much one variable tends to change when the other changes by one unit. It tells us how much of the variation in outcomes is accounted for by the predictor. And it tells us, through the scatter that remains, how much we still do not know.

This is the first chapter where we build an explicit predictive model. The probability distributions in Chapters 5 and 6 and the null distributions of Chapter 8 were models of *uncertainty*; what changes here is that we use the data to fit a model that maps an input variable to a predicted output. If you have followed the progression of this book from descriptions of data to probability to inference, regression is the natural next step. It is where those ideas come together.

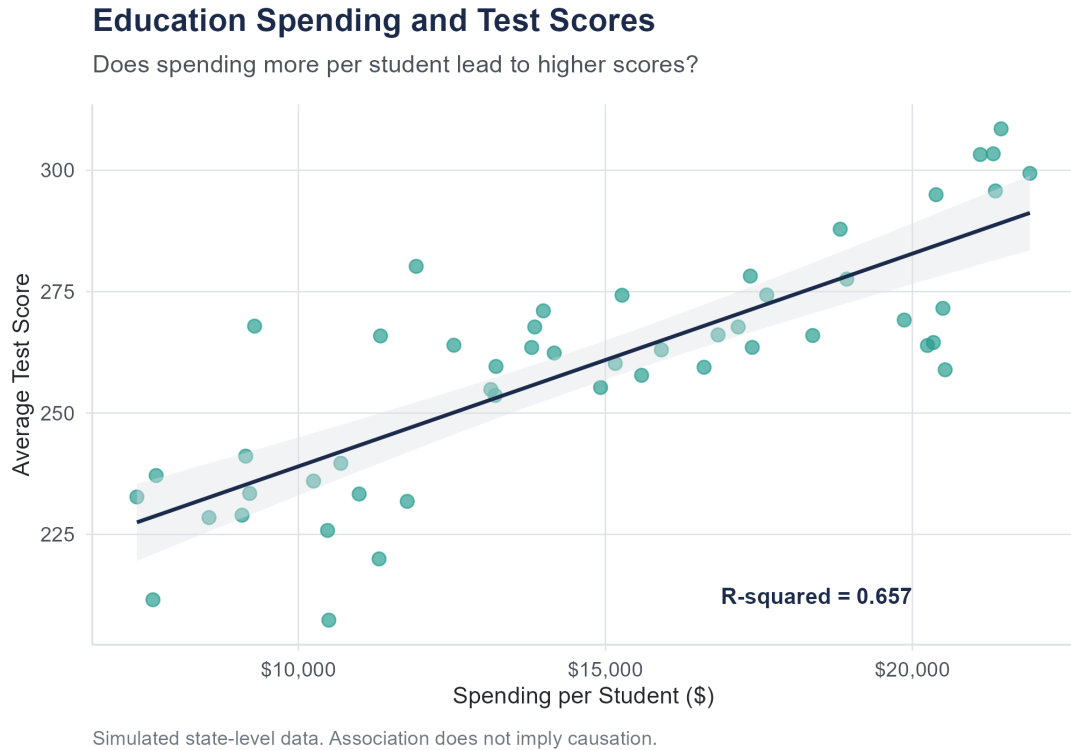


Figure 10.1: Scatter plot of state-level education spending per student versus average test scores, with a least-squares regression line and confidence band. The positive slope suggests higher spending is associated with higher scores, but the substantial scatter around the line shows that spending alone does not determine outcomes.

10.2 One Variable to Predict Another

Up to this point, we have mostly examined one variable at a time. What is the average? How spread out is the data? Is the population mean different from some hypothesized value? These are all questions about a single variable.

But many practical questions involve relationships between variables. Does studying more hours lead to higher exam scores? Does a company's advertising budget predict its quarterly sales? Does the temperature on a given day predict how much electricity a city will consume?

In each of these cases, we have two variables. One of them is the **explanatory variable** (also called the independent variable or predictor), which is the variable we think might help explain or predict the other. The other is the **response variable** (also called the dependent variable or outcome), which is the variable we are trying to understand or predict.

The terminology matters because the roles are not symmetric. We are asking whether X helps us predict Y , not whether Y helps us predict X . In the education example, per-pupil spending is the explanatory variable and test scores are the response. In a study of exercise and blood pressure, minutes of weekly exercise is the explanatory variable and systolic blood pressure is the response.

This does not mean that X causes Y . That distinction is crucial. Regression describes the association between two variables. It quantifies how much Y tends to change when X changes. But association is not causation, and we have spent enough time in this book on that point that you know why. A regression line through the education spending data tells us what the data shows. Whether spending more money would actually improve scores requires a much deeper analysis involving experimental design, confounders, and all the cautions of Chapter 2.

Consider a less politically charged example. If you regress ice cream sales (X) against drowning incidents (Y) for each month of the year, you will find a positive slope. Months with higher ice cream sales tend to have more drownings. The regression line is real. The pattern is real. But nobody believes ice cream causes drowning. The lurking variable is temperature. Hot months drive people to both buy ice cream and swim more often. Regression faithfully describes the pattern in the data. It takes a human being, not a formula, to figure out what that pattern means.

10.3 The Regression Line

Imagine you have a scatter plot with dozens of data points showing the relationship between per-pupil spending (X) and average test scores (Y). The points form a cloudy, upward-drifting swarm. Your goal is to draw a single straight line through this cloud that captures the overall trend as well as possible.

The equation for any straight line is

$$\hat{y} = b_0 + b_1x$$

where $\hat{y}$ (read “y-hat”) is the predicted value of Y for a given value of X , b_0 is the **y-intercept** (the predicted value of Y when $X = 0$), and b_1 is the **slope** (how much $\hat{y}$ changes for each one-unit increase in X).

That hat on the y is important. It signals that $\hat{y}$ is a prediction, not an observed value. Real data points sit at various distances above and below the line. The hat reminds us that the line gives us an estimate, not the truth.

But which line? You could draw infinitely many straight lines through a scatter plot. Some would cut through the middle of the cloud. Others would miss it entirely. We need a criterion for choosing the best one.

10.3.1 The Least Squares Criterion

Here is the idea. For each data point, the line makes a prediction. The difference between the actual y value and the predicted $\hat{y}$ value is called a **residual**.

$$\text{residual}_i = y_i - \hat{y}_i$$

A positive residual means the point is above the line. A negative residual means it is below. A residual of zero means the point sits exactly on the line.

Some residuals will be positive, some negative. If you simply add them up, the positives and negatives tend to cancel each other out, which is not helpful. So instead, we square each residual to make them all positive, and then add up all the squared residuals.

$$\text{Sum of squared residuals} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

The **least squares regression line** is the line that makes this sum as small as possible. Out of all possible straight lines, it is the one that minimizes the total squared distance between the observed data points and the line’s predictions. You will also encounter this method under the name **Ordinary Least Squares** (OLS) in econometrics and machine learning textbooks. The two names refer to the same thing.

Why squares rather than absolute values? Partly for mathematical convenience, the calculus works out more cleanly. But there is also a practical reason. Squaring gives extra weight to large residuals. A point that is far from the line contributes much more to the sum of squares than a point that is

close. This means the least squares line is particularly motivated to avoid making big errors, which is usually what we want.

Think of it this way. You are stretching a rubber band from each data point to the line, and each rubber band has a tension proportional to the square of its length. The least squares line is the position that minimizes the total tension. It is the line that balances all those pulls as efficiently as possible.

There is another way to think about it that some students find helpful. Imagine you are placing a straight plank on a lumpy surface. The lumps are your data points, sticking up at various heights. You want the plank to settle into the position where it is, on average, closest to all the lumps. The least squares criterion is the mathematical rule that finds that position. Every other line you could draw through the scatter plot would produce a larger total squared distance from the data. This is the best line, in the specific sense that no other straight line does a better job of staying close to the points overall.

Note

The least squares method dates back to the early 1800s, when both Carl Friedrich Gauss and Adrien-Marie Legendre claimed credit for developing it. Gauss used it to predict the orbit of the dwarf planet Ceres (then classified as an asteroid) after it disappeared behind the Sun, successfully locating it when other astronomers could not. The method was born from astronomy, but it has since found a home in everything from economics to epidemiology to machine learning. Two centuries later, least squares remains a workhorse method across statistics.

10.4 Fitting the Line

The formulas for the slope and intercept of the least squares line are derived using calculus, but you do not need to know the derivation to use them or understand them. Here they are.

10.4.1 The Slope

$$b_1 = r \frac{s_y}{s_x}$$

where r is the correlation between X and Y , s_y is the standard deviation of Y , and s_x is the standard deviation of X .

This formula tells you something important. The slope is directly related to the correlation. If r is positive, the slope is positive, meaning Y tends to increase as X increases. If r is negative, the slope is negative. If r is zero, the slope is zero, and the best-fitting line is flat, meaning X tells you nothing about Y .

The slope is also scaled by the ratio of the standard deviations. This ratio converts the unitless correlation into something with real units. If X is measured in thousands of dollars of spending and Y is measured in test score points, then b_1 tells you how many test score points Y changes, on average, for each additional thousand dollars of spending.

There is an equivalent formula you will sometimes see written as

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

This computes the same number. The numerator is the sum of the cross-products of deviations (how much each x and y deviate from their means, multiplied together). The denominator is the sum of squared deviations in x . The ratio tells you the average amount of change in y per unit change in x , weighted by how far each x is from the mean.

Interpreting the slope. Suppose you fit a regression line to the full 50-state education data and get $b_1 = 0.000635$. This means that for each additional dollar spent per student, the model predicts average test scores to be about 0.000635 points higher, or equivalently, for each additional \$1,000 spent per student, scores are predicted to be about 0.64 points higher. This is a statement about the overall pattern in the data, not a guarantee about any individual state. Some states will score higher than the line predicts and some will score lower. The slope describes the trend, not the exceptions.

10.4.2 The Intercept

$$b_0 = \bar{y} - b_1 \bar{x}$$

where $\bar{y}$ is the mean of Y and $\bar{x}$ is the mean of X .

This formula has a nice geometric consequence. The point $(\bar{x}, \bar{y})$ always lies exactly on the least squares line. In other words, the regression line passes through the center of the data. This makes intuitive sense. A line that summarizes the overall trend should at least get the center right.

Interpreting the intercept. The intercept b_0 is the predicted value of Y when $X = 0$. Sometimes this has a meaningful interpretation. If X is hours studied and Y is exam score, then b_0 is the predicted score for a student who studied zero hours, the baseline score.

Other times, the intercept makes no practical sense. If X is per-pupil spending, then b_0 is the predicted test score when spending is zero dollars, which is not a realistic scenario. No state spends zero dollars on education. In cases like this, the intercept is a mathematical necessity for defining the line, but its value does not carry a useful interpretation on its own. It simply anchors the line so that the slope can do its work.

This comes up more often than you might expect. In a regression predicting adult height from shoe size, the intercept would be the predicted height of someone with a shoe size of zero. In a regression predicting crop yield from inches of rainfall, the intercept would be the predicted yield with no rain at all. These are not useful numbers for making decisions. They exist because the algebra demands a starting point. When someone presents a regression equation, it is usually the slope that carries the interesting information. The intercept is the scaffolding.

10.4.3 A Worked Example

Suppose we have data from 10 states. After computing the necessary summaries, we find $\bar{x} = 12.4$ (in thousands of dollars per pupil), $\bar{y} = 265$ (average test score), $s_x = 2.8$, $s_y = 14.2$, and $r = 0.68$.

Step 1. Compute the slope.

$$b_1 = r \frac{s_y}{s_x} = 0.68 \times \frac{14.2}{2.8} = 0.68 \times 5.071 = 3.45$$

Step 2. Compute the intercept.

$$b_0 = \bar{y} - b_1 \bar{x} = 265 - 3.45 \times 12.4 = 265 - 42.78 = 222.22$$

Step 3. Write the equation.

$$\hat{y} = 222.22 + 3.45x$$

Interpretation. For every additional \$1,000 in per-pupil spending, the regression model predicts test scores to be about 3.45 points higher, on average. A state that spends \$12,400 per student would have a predicted score of $222.22 + 3.45(12.4) = 265$, which is the mean test score, confirming that the line passes through the point $(\bar{x}, \bar{y})$.

10.5 R-Squared: How Much Does the Model Explain?

The regression line gives us a prediction for each observation, but how good are those predictions? We need a measure of how well the line fits the data. That measure is R^2 , called the **coefficient of determination**.

The idea behind R^2 starts with a simple question. The values of Y vary. Some of that variation might be related to X , and some of it might not. How much of the total variation in Y can be accounted for by the linear relationship with X ?

To answer this, we decompose the variation in Y into two parts.

Total variation. How much do the observed y values vary around their mean? We measure this with the total sum of squares.

$$SS_{Total} = \sum_{i=1}^n (y_i - \bar{y})^2$$

Explained variation. How much of that variation is captured by the regression line? This is the regression sum of squares.

$$SS_{Regression} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

Unexplained variation. How much variation is left over after we account for the regression? This is the residual sum of squares, the same quantity we minimized when fitting the line.

$$SS_{Residual} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

These three quantities are related by

$$SS_{Total} = SS_{Regression} + SS_{Residual}$$

The total variation equals the explained variation plus the unexplained variation. And R^2 is simply the fraction of the total variation that is explained.

$$R^2 = \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{SS_{Residual}}{SS_{Total}}$$

In simple linear regression (one predictor), R^2 equals the square of the correlation coefficient r . For the full 50-state education spending dataset, the correlation between per-pupil spending and average test scores is $r = 0.434$. If $r = 0.434$, then $R^2 = 0.434^2 = 0.188$. This means that about 19% of the variation in test scores across states is associated with differences in per-pupil spending.

10.5.1 Interpreting R-Squared

An R^2 of 0.188 means something concrete. If you knew nothing about spending levels, you would predict every state's test score to be the mean, and all the variation would be unexplained. By knowing each state's spending level, you can explain about 19% of the differences in scores. The remaining 81% is due to other factors, things like demographics, teacher quality, curriculum choices, poverty rates, and countless other variables that we have not included in this simple model.

People sometimes ask, "What is a good R^2 ?" The answer depends entirely on the context. In the physical sciences, where relationships between variables can be very tight, an R^2 below 0.90 might be disappointing. In the social sciences, where human behavior is noisy and influenced by hundreds of factors, an R^2 of 0.20 or 0.30 might be perfectly respectable. In education research, explaining 19% of the variation with a single predictor tells you the relationship is real but that most of the story lies elsewhere.

A useful analogy: R^2 is like the volume knob on a radio that is tuned between stations. An R^2 of 0 is pure static, no signal at all. An R^2 of 1 is perfect clarity, all signal and no noise. Most real data in the social sciences falls somewhere in the range of a scratchy broadcast where you can make out parts of the melody but plenty of static remains. You know something is playing. You just cannot catch all the lyrics. The question is not whether the signal is perfect. The question is whether there is enough signal to be useful, and that depends on what you are trying to do with it.

10.5.2 Limitations of R-Squared

R^2 is useful, but it has some important limitations that trip people up.

A high R^2 does not mean the model is correct. You can get a high R^2 from a linear model even when the true relationship is curved. The line might pass near the data for the wrong reasons. Always check the residual plots (more on this soon).

A low R^2 does not mean the relationship is unimportant. Even a small R^2 can represent a relationship that matters practically. If a medication explains only 5% of the variation in blood pressure, that might still be a clinically meaningful effect. A variable can be a weak predictor overall but still produce changes that have real consequences.

R^2 says nothing about causation. This should sound familiar by now. A high R^2 between two variables means they are closely associated. It does not tell you that changes in X cause changes in Y .

R^2 cannot tell you whether you have the right predictor. Maybe spending is correlated with test scores, but the real driver is something else, like median household income, which is correlated with both spending and scores. R^2 tells you how much variation is accounted for, not whether your explanation is the right one.

10.6 Is the Slope Real? Inference for Regression

The regression output for the 50-state education spending data gives a slope of 0.000635; test scores increase by about 0.64 points for every additional \$1,000 in per-pupil spending. But that number came from a sample. A different sample would give a different slope. Maybe a little higher, maybe a little lower, maybe even negative. The question is whether the slope you observed reflects a genuine relationship in the population, or whether it could have arisen by chance from data with no real pattern.

This is the same question we asked about means in Chapter 8 and about proportions in Chapter 7. The tools are the same: standard errors, test statistics, p -values, and confidence intervals. The application is new.

10.6.1 Standard Error of the Slope

The slope b_1 is a sample statistic. Like any sample statistic, it has a sampling distribution: if you could repeat the study with a new random sample each time, you would get a different slope each time. The **standard error of the slope**, denoted $SE(b_1)$, measures how much the slope would vary across those repeated samples.

Conceptually:

$$SE(b_1) = \frac{s_e}{\sqrt{\sum(x_i - \bar{x})^2}}$$

where s_e is the **residual standard error**, the standard deviation of the residuals (roughly, how far typical observations fall from the regression line). You do not need to compute this by hand, but software does it automatically, and the formula reveals three important facts about what makes a slope estimate more or less precise:

1. **Less scatter around the line** (smaller s_e) means a smaller standard error. When the data hugs the regression line tightly, the slope is estimated more precisely.
2. **More spread in the X values** (larger $\sum(x_i - \bar{x})^2$) means a smaller standard error. Data points spread across a wide range of X give the line more to “anchor” on, making the slope more stable.
3. **Larger sample size** contributes to both effects: more data points means more information and typically more spread in X .

The residual standard error s_e itself is computed as:

$$s_e = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n - 2}}$$

We divide by $n - 2$ because we estimated two parameters (slope and intercept) from the data, leaving $n - 2$ degrees of freedom. This is analogous to dividing by $n - 1$ when computing a sample standard deviation, and the logic is the same: we lose a degree of freedom for each parameter we estimate.

10.6.2 Testing Whether the Slope Is Zero

The hypothesis test for the slope asks: could this slope have come from a population where the true slope is zero?

$$H_0 : \beta_1 = 0 \quad (\text{no linear relationship})$$

$$H_a : \beta_1 \neq 0 \quad (\text{a linear relationship exists})$$

The test statistic follows the same logic as every other t -test you have encountered:

$$t = \frac{b_1 - 0}{SE(b_1)} = \frac{b_1}{SE(b_1)}$$

This t -statistic follows a t -distribution with $n - 2$ degrees of freedom. A large $|t|$ means the observed slope is many standard errors away from zero, which is strong evidence against the null hypothesis.

Worked example. Returning to the education spending data with $n = 50$ states, the regression output gives $b_1 = 0.000635$, $SE(b_1) = 0.000191$. Then:

$$t = \frac{0.000635}{0.000191} = 3.33$$

With $df = 48$, this gives a two-tailed p -value of approximately 0.002. At any conventional significance level, we reject the null hypothesis. The data provides strong evidence of a positive linear relationship between per-pupil spending and average test scores.

10.6.3 Confidence Interval for the Slope

Just as we constructed confidence intervals for means and proportions, we can construct one for the slope:

$$b_1 \pm t^* \times SE(b_1)$$

where t^* is the critical value from the t -distribution with $n - 2$ degrees of freedom.

For a 95% confidence interval with $df = 48$, $t^* \approx 2.011$:

$$0.000635 \pm 2.011 \times 0.000191 = 0.000635 \pm 0.000384 = (0.000252, 0.001019)$$

Interpretation: We are 95% confident that the true slope is between 0.00025 and 0.00102. For each additional dollar of per-pupil spending, test scores increase by somewhere between 0.00025

and 0.00102 points, on average. Since the interval does not contain zero, this confirms the slope is statistically significant at the 5% level.

10.6.4 Reading Regression Output

Every statistics package produces a regression output table with the same basic columns. Here is what a typical table looks like:

	Coefficient	Std. Error	t value	p-value
Intercept	271.67	3.534	76.87	< 0.001
spending_per_student	0.000635	0.000191	3.33	0.002

Each row represents a coefficient. The columns tell you:

- **Coefficient:** the estimated value (b_0 or b_1).
- **Std. Error:** the standard error of that estimate.
- **t value:** the coefficient divided by its standard error.
- **p-value:** the probability of observing a t -statistic this extreme if the true coefficient were zero.

This is exactly the table that R's `summary(lm(...))` produces. The slope row is the one that usually matters most. If its p -value is small, you have evidence that the predictor is linearly related to the response.

10.6.5 Statistical vs. Practical Significance in Regression

A slope can be statistically significant but practically trivial. A p -value of 0.001 tells you the slope is almost certainly not zero. It does not tell you the slope is large enough to care about.

In the education example, the slope of 0.00182 means each additional dollar of spending is associated with 0.00182 more points on a standardized test. That is less than two-thousandths of a point per dollar. Even \$10,000 of additional spending is associated with about 18 additional points. Whether that is a meaningful improvement depends on the scale of the test, the cost of the spending, and the goals of the policy. The p -value cannot answer those questions. The slope itself, combined with context and judgment, is what tells you whether the effect matters in practice.

Always report both the coefficient and its p -value. The coefficient tells you how big the effect is. The p -value tells you how sure you are that it is not zero. You need both.

10.7 Residuals: What the Model Misses

Every data point has a residual, the difference between what was actually observed and what the model predicted.

$$e_i = y_i - \hat{y}_i$$

Residuals are the leftovers. They are what remains after the model has done its best to explain the data. And they carry a great deal of information, because they tell you where the model works well and where it falls short.

If a state has a positive residual, it scored higher than the regression line predicted given its level of spending. Something about that state is producing better outcomes than spending alone would suggest. Maybe it has an unusually effective curriculum, or a high proportion of well-trained teachers, or a demographic profile that favors higher test performance.

If a state has a negative residual, it scored lower than predicted. Something about that state is dragging outcomes down relative to what spending alone would predict.

Large residuals deserve attention. They might indicate unusual cases that do not fit the general pattern, and investigating them often reveals interesting features of the data that the model cannot capture.

10.7.1 Properties of Residuals

The residuals from a least squares regression have two mathematical properties that are always true.

1. The residuals sum to zero. $\sum_{i=1}^n e_i = 0$. Positive and negative residuals perfectly cancel out.
2. The residuals are uncorrelated with the predicted values $\hat{y}$. This means there is no leftover linear pattern between the predictions and the residuals.

These properties are consequences of the least squares fitting procedure, not assumptions. They hold for every least squares regression, regardless of whether the data actually follows a linear pattern. They are mathematical facts, not things you need to check.

10.7.2 Signal, Noise, and the Population Model

A useful way to think about regression is in terms of *signal* and *noise*. The line is what regression treats as the systematic pattern — the signal. Everything that is not on the line is treated as random scatter — the noise. The residuals are the realized noise in the sample at hand.

At the population level, this is sometimes written

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where β_0 and β_1 are the unknown population intercept and slope (the parameters we are estimating with b_0 and b_1), and ε is the random error term — what is left over once the linear signal has been removed. Each observation's residual $e_i = y_i - \hat{y}_i$ is the sample-level estimate of this otherwise unobservable ε . We will see this notation again in Chapter 11, where multiple predictors enter the model and the bookkeeping of signal and noise becomes more involved. For now, the takeaway is that the residuals are not bookkeeping leftovers — they are the empirical signature of the noise, and the LINE assumptions in the next section are statements about how that noise is supposed to behave.

What you do need to check is whether the residuals behave in the ways that the regression model assumes. And that is our next topic.

10.8 The Assumptions Behind Regression

A regression line can always be computed. You can always fit a least squares line to a scatter plot. The software will never refuse. But the line is only meaningful, and the inferences drawn from it are only valid, if certain conditions are met.

These conditions are often summarized by the acronym **LINE**, which stands for Linearity, Independence, Normality, and Equal variance. Let us walk through each one.

10.8.1 L: Linearity

The relationship between X and Y should be approximately linear. This sounds obvious for a method called “linear regression,” but it is the assumption most often violated in practice.

If the true relationship between X and Y is curved, fitting a straight line will systematically over-predict in some regions of X and underpredict in others. The line might still pass near the data in some average sense, but the predictions will be biased, too high in some places and too low in others.

How to check it. Make a scatter plot of X versus Y before you fit the regression. Look for curvature. If the data bends, a straight line is not the right tool. You might need to transform the data (taking logarithms is a common approach) or use a more flexible model.

10.8.2 I: Independence

The observations should be independent of one another. The value of one observation should not provide information about the value of another.

Independence is violated when there is some kind of structure in how the data was collected. Time series data, where observations are recorded sequentially, often shows dependence because today's value is related to yesterday's. Clustered data — students nested within schools, patients nested within hospitals — can also violate independence, since observations from the same cluster tend to be more similar to one another than to observations from a different cluster.

How to check it. This is less about a plot and more about understanding how the data was collected. If the data was gathered by random sampling from a population, independence is usually reasonable. If the data has a time component or a grouping structure, be cautious.

10.8.3 N: Normality of Residuals

The residuals should be approximately normally distributed. This assumption matters most for inference, specifically for the confidence intervals and hypothesis tests that are built on top of the regression. The point estimates of slope and intercept do not require normality. But if you want to construct confidence intervals or test whether the slope is different from zero, you need the residuals to be roughly bell-shaped.

How to check it. Make a histogram of the residuals and see if it looks roughly symmetric and bell-shaped. Better yet, make a **normal probability plot** (also called a QQ plot), which plots the residuals against the values they would take if they were perfectly normal. If the points fall close to a straight diagonal line, the normality assumption is reasonable. Deviations from the line, especially in the tails, suggest non-normality.

For large samples, the normality assumption becomes less critical thanks to the Central Limit Theorem, which makes the sampling distributions of b_0 and b_1 approximately normal even when the residuals are not perfectly so. But for small samples, non-normal residuals can make your confidence intervals and p -values unreliable. This is worth emphasizing: the LINE assumptions are not just about making the line fit well; they are the foundation for all the inference tools (standard errors, t -tests, confidence intervals, p -values) described in the section on inference for regression.

10.8.4 E: Equal Variance (Homoscedasticity)

The spread of the residuals should be roughly the same across all values of X . This is the same **homoscedasticity** assumption introduced in the ANOVA discussion in Chapter 9, applied here to the residuals around a regression line rather than to multiple groups. Another name for it is **constant variance**. The opposite, where the spread of residuals changes across X , is called **heteroscedasticity**.

Why does this matter? The standard errors used in regression inference assume that the variability of Y around the line is the same everywhere. If the residuals fan out (getting wider as X increases, for example), the standard errors will be wrong. They might be too small in some regions and too large in others, leading to confidence intervals and hypothesis tests that cannot be trusted.

How to check it. Make a scatter plot of the residuals versus the fitted values ($\hat{y}$). If the assumption holds, the residuals should form a horizontal band of roughly constant width. If the band widens or narrows, or shows a funnel shape, the equal variance assumption is violated. We come back to this plot, and to the QQ plot for normality, in detail later in the chapter.

Note

The word “homoscedasticity” comes from the Greek *homos* (same) and *skedasis* (scattering). Its opposite, “heteroscedasticity,” means “different scattering.” The vocabulary is intimidating, but the concept is simple. Does the scatter in the residuals stay the same across the range of predictions, or does it change? If it stays the same, you have homoscedasticity. If it changes, you have heteroscedasticity. The vocabulary is hard. The idea is not.

10.9 Checking Assumptions with Residual Plots

In practice, you check most of the LINE assumptions using two plots.

10.9.1 Residuals vs. Fitted Values Plot

This is the workhorse diagnostic. Plot the residuals (e_i) on the vertical axis against the fitted values ($\hat{y}_i$) on the horizontal axis. A clean plot — random scatter, centered on zero, with roughly constant spread from left to right — is consistent with both the linearity (L) and the equal-variance (E) assumptions discussed earlier.

The patterns to watch for are:

- A curved shape (U, frown, wave) → linearity is in trouble.
- A funnel that opens or closes → equal variance is in trouble.
- Distinct subgroups → an important grouping variable is missing from the model.
- A single point sitting far from the rest → an influential observation pulling the line.

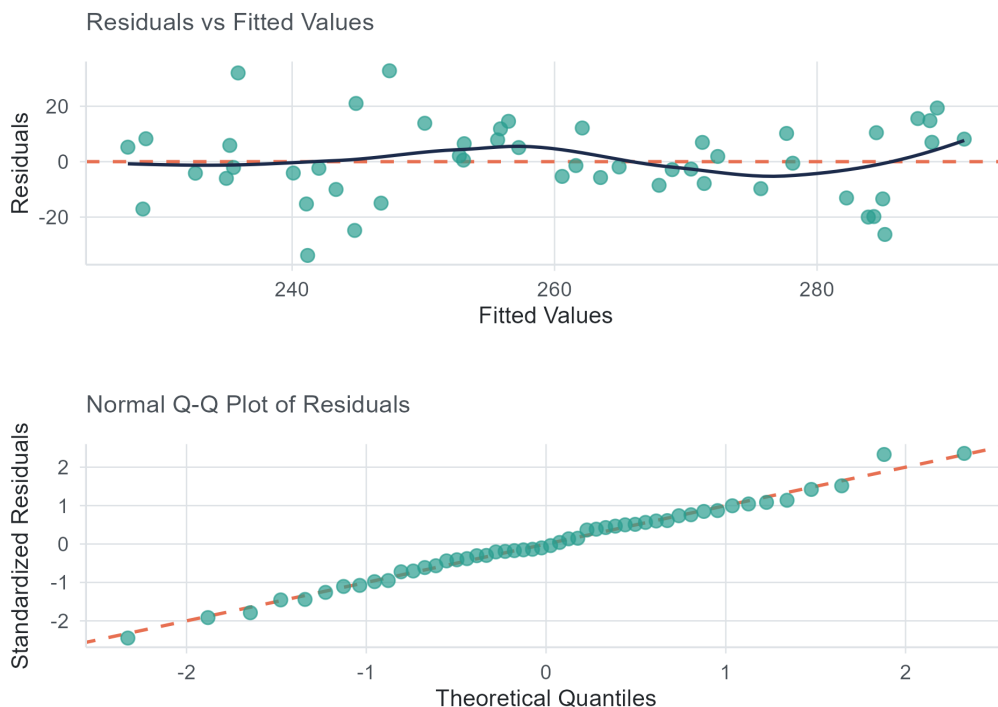
10.9.2 Normal QQ Plot

Plot the ordered residuals against the theoretical quantiles of a normal distribution. This is the same QQ-plot reading practice introduced in Chapter 6: points close to the diagonal line are consistent with the normality (N) assumption, an S-shape signals heavy or light tails, and a frown or smile signals skew. Tail wobble in small samples is common and usually fine.

These diagnostic plots do not produce a yes-or-no answer. There is no test that says “your assumptions are perfectly met.” You are looking for gross violations — patterns obvious enough to call the inferential results into question. Minor wiggles are normal. Clear patterns are problems.

Checking Regression Assumptions: Residual Plots

Education spending vs test scores model diagnostics



Top: look for random scatter around zero (no patterns). Bottom: points should follow the diagonal line.

Figure 10.2: Two residual diagnostic plots for the education spending regression model. The top panel shows residuals versus fitted values, where random scatter around zero indicates the linearity and constant variance assumptions are reasonable. The bottom panel shows a normal QQ plot, where points following the diagonal line confirm the normality assumption.

Beginning analysts sometimes freeze at this point, worried that they are not experienced enough to tell a “minor wiggle” from a “clear pattern.” This is a reasonable concern, and the honest answer is that it takes practice. But here is a useful benchmark: if you showed the residual plot to three reasonable people and all three noticed the same pattern without prompting, it is probably a real problem. If only one of them sees something, and only after squinting, it is probably fine. Statistics involves judgment, more than formulas, and this is one of those places where judgment matters.

 Try It Online

Open the **Regression Explorer** on the companion website. Add data points, fit the model, and check residuals. Try adding a single outlier far from the trend and watch how much the slope shifts. Then remove it and see the line snap back.

10.10 Prediction and the Danger of Extrapolation

One of the most common uses of regression is prediction. Given a value of X , what does the model predict for Y ? The equation makes this straightforward. Plug in the value of X , compute $\hat{y}$, and you have your prediction.

If per-pupil spending is \$14,000, and our regression equation is $\hat{y} = 271.67 + 0.000635x$ (where x is in dollars), then the predicted test score is

$$\hat{y} = 271.67 + 0.000635(14,000) = 271.67 + 8.89 = 280.56$$

This prediction comes with the usual caveat that it is an average prediction. It does not mean a specific state spending \$14,000 per student will have a score of exactly 280.56. It means that, based on the pattern observed in the data, 280.56 is our best guess for the average score among states spending that amount.

10.10.1 Interpolation vs. Extrapolation

Predictions are most trustworthy when they are made within the range of X values that were used to build the model. If our data includes states spending between about \$10,000 and \$33,000 per student, then predicting the score for a state spending \$14,000 is **interpolation**. We are predicting within the observed range. The model has data to support this prediction.

Extrapolation is predicting outside the observed range. What if someone asks you to predict the test score for a state spending \$40,000 per student, or \$3,000 per student? The model will happily give you a number. The formula does not care whether you plug in 14,000 or 40,000 or 3,000. But the prediction is unreliable because you have no data to support it.

Why is extrapolation dangerous? Because the linear pattern you observed within one range may not hold outside that range. Maybe the relationship between spending and scores is linear between

\$10,000 and \$33,000, but it levels off above \$33,000 because there are diminishing returns to additional spending. Maybe it drops off below \$5,000 because there is a minimum level of funding needed for schools to function. The straight line knows nothing about these possibilities. It just extends forever in both directions, regardless of whether the real relationship does the same.

A widely cited example of extrapolation gone wrong involves predicting athletic performance. If you plot world record times in the men's 100-meter dash over the twentieth century, you see a clear downward trend. A regression line fit to this data would predict that, at some point in the future, humans will run the 100 meters in zero seconds. Then in negative seconds. The linear model does not know about the physical limits of human speed. It just extends the trend, and the trend eventually becomes absurd.

The rule of thumb is simple. Use regression for interpolation with reasonable confidence. Use it for modest extrapolation with caution. Use it for aggressive extrapolation at your peril.

A subtler form of extrapolation catches people off guard. Sometimes the range of X in your data is technically broad enough, but the prediction involves an unusual combination of circumstances. Suppose you build a regression predicting home sale prices from square footage, using data from a suburban neighborhood where most homes are between 1,200 and 3,000 square feet. Predicting the price of a 2,000-square-foot home is straightforward interpolation. But what about a 2,800-square-foot home that happens to be on a flood plain, or one that was built in 1890 when the rest of the neighborhood dates from the 1990s? The square footage falls within range, but the home itself is unlike anything in the dataset. This is a reminder that extrapolation is not only about the numerical range of X . It is about whether the new observation resembles the observations that built the model.

10.11 A Complete Example

Let us walk through a full regression analysis, from scatter plot to diagnostics, using a concrete scenario.

A researcher collects data on 15 metropolitan areas, recording the average daily commute time in minutes (X) and the average stress level on a 1-to-100 scale from a survey of residents (Y). Here are the summary statistics.

$$n = 15, \quad \bar{x} = 28.6, \quad \bar{y} = 54.3, \quad s_x = 8.4, \quad s_y = 12.7, \quad r = 0.72$$

Step 1. Compute the regression equation.

$$b_1 = r \frac{s_y}{s_x} = 0.72 \times \frac{12.7}{8.4} = 0.72 \times 1.512 = 1.089$$

$$b_0 = \bar{y} - b_1 \bar{x} = 54.3 - 1.089 \times 28.6 = 54.3 - 31.15 = 23.15$$

$$\hat{y} = 23.15 + 1.089x$$

Interpretation. For each additional minute of average commute time, the model predicts stress levels to increase by about 1.09 points on the 100-point scale. A city with zero commute time would have a predicted stress level of 23.15, though this extrapolation should be taken with caution since no city in the dataset has a zero commute time.

Step 2. Compute R^2 .

$$R^2 = r^2 = 0.72^2 = 0.5184$$

About 52% of the variation in stress levels across these metropolitan areas is accounted for by differences in average commute time. The other 48% is due to other factors we have not measured, things like cost of living, population density, climate, access to green space, or local culture.

Step 3. Check assumptions. In a real analysis, you would now produce a residuals versus fitted values plot and a QQ plot. You would look for curvature in the residuals plot (checking linearity), a fan shape (checking equal variance), and departures from the diagonal in the QQ plot (checking normality). With only 15 observations, you would be especially attentive to any single point that seems to pull the line disproportionately.

Step 4. Make a prediction. What is the predicted stress level for a city with an average commute of 35 minutes?

$$\hat{y} = 23.15 + 1.089(35) = 23.15 + 38.12 = 61.27$$

Since 35 minutes falls within the range of commute times in the dataset, this is interpolation, and the prediction is reasonable within the limits of the model.

10.12 Before You Regress: Examine Your Variables

It is tempting to jump straight into regression the moment you have two numerical variables. Resist that temptation. Before fitting a line, spend time getting to know your variables and the relationship between them.

Start with a scatter plot. Is the relationship approximately linear? Is it curved? Is it strong or weak? Are there outliers that might dominate the fit? A scatter plot answers these questions at a glance, and the answers determine whether simple linear regression is appropriate.

Also consider looking at the correlation between your variables and, if you have more than two variables in your dataset, at the correlations between all pairs. A correlation matrix can reveal which variables are most strongly associated with each other, which might help you choose the best predictor for your regression, and which might alert you to potential confounders.

Try It Online

Return to the **Regression Explorer**. Load a dataset with multiple variables and explore pairwise scatter plots before building your model. Identify which predictors show the strongest linear relationship with your outcome before adding them to the regression.

10.13 Outliers and Influential Points

Not all data points contribute equally to the regression line. A few points can exert outsized influence, especially if they sit far from the bulk of the data on the X axis.

A **high-leverage point** is an observation with an X value far from $\bar{x}$. Because the regression line must pass through $(\bar{x}, \bar{y})$ and is anchored there, points far from the center on the horizontal axis act like a long lever arm. They have the potential to swing the line more than points near the center.

An **influential point** is one whose removal would substantially change the regression line. Not all high-leverage points are influential. If a high-leverage point falls right on the trend established by the other data, it reinforces the line rather than pulling it somewhere new. But a high-leverage point that is also far from the trend line can dramatically change the slope and intercept.

When you find an influential point, the question is not “should I delete it?” The question is “why is it unusual?” Maybe it represents a data entry error that should be corrected. Maybe it represents

an unusual case that has a reasonable explanation. Maybe it reveals that the data comes from two different populations that should be analyzed separately.

Deleting a point simply because it is inconvenient is not good practice. Investigating it almost always is.

A good example comes from public health. In studies of air pollution and mortality rates across cities, researchers occasionally find a city whose mortality rate is far lower than its pollution level would suggest. The temptation is to remove it as an outlier, since it makes the regression line less steep and the story less dramatic. But investigating that city might reveal something valuable. Maybe it has an unusually extensive public transit system that reduces individual exposure despite high ambient pollution levels, or maybe its population skews younger than average. The outlier, rather than being a problem to eliminate, becomes a clue pointing toward factors the model has not captured. Some of the most interesting findings in research have come from taking outliers seriously instead of deleting them.

A practical guideline is to fit the regression with and without the suspicious point. If the slope, intercept, and R^2 barely change, the point is unusual but not influential, and you can proceed with it included. If removing it dramatically changes the results, you need to understand why before deciding how to handle it. Report both analyses if the point is ambiguous. Transparency about influential observations is always better than silent deletion.

! Ethics Moment

Regression results shape policy, and the level of aggregation can change the story in ways that mislead. The education spending data in this chapter is measured at the state level: each data point represents one state's average spending and one state's average test scores. A positive slope tells you that states spending more tend to have higher average scores. It does not tell you that an individual student in a higher-spending school will score higher.

This distinction matters because it is routinely ignored. State-level regression results have been used to justify funding formulas that affect millions of students, despite the fact that the relationship between spending and outcomes at the individual or school level may look very different from the state-level pattern. This error, drawing individual-level conclusions from group-level data, is called the **ecological fallacy**, and it has a long, damaging history in social science and policy.

The same issue arises in health research (countries with higher fat consumption have higher cancer rates, but that does not mean the person eating butter is at higher risk), in criminal justice (neighborhoods with higher poverty have higher crime, but that does not mean a

given poor person is more likely to commit a crime), and in business (regions with higher advertising spend have higher sales, but the individual-level effect of an ad may be small). When you use regression to inform decisions about people, always ask whether the data was measured at the same level as the decisions you are making. If the data is aggregated and the decisions are individual, be cautious about generalizing.

10.14 Correlation and Regression: Close Relatives

Correlation was introduced in Chapter 4, where Pearson's r , scatter plots, and Anscombe's Quartet provided the visual and numerical foundation for thinking about how two numerical variables move together. With regression now in hand, it is worth being explicit about how the two methods are related, because they are close relatives but play different roles.

Correlation (r) measures the strength and direction of the linear relationship between two variables. It is a unitless number between -1 and $+1$. It treats X and Y symmetrically. The correlation between X and Y is the same as the correlation between Y and X .

Regression uses one variable to predict another. It is asymmetric. The regression of Y on X gives a different equation than the regression of X on Y . The slope of the regression has units (points per thousand dollars, stress per minute, kilograms per centimeter) and a directional interpretation that correlation lacks.

The two are connected mathematically. The slope of the regression line is $b_1 = r \frac{s_y}{s_x}$, and $R^2 = r^2$. But they answer different questions. Correlation answers, "How tightly do X and Y move together?" Regression answers, "By how much does Y change, on average, when X changes by one unit?"

You need both. Correlation tells you whether a relationship is worth modeling. Regression models it.

10.15 What Regression Cannot Do

Before we close, it is worth stating plainly what simple linear regression cannot do.

It cannot establish causation. We have said this, but it bears repeating. A regression line between education spending and test scores does not mean that increasing spending will increase scores. There may be confounding variables that explain both.

It cannot capture nonlinear relationships. If the scatter plot bends, a straight line misses the story. More advanced techniques (polynomial regression, transformations, nonlinear models) are needed.

It cannot handle multiple predictors. Simple linear regression uses one explanatory variable. If test scores are influenced by spending, poverty rates, teacher salaries, and class sizes, you need multiple regression, which is the subject of the next chapter.

It cannot rescue bad data. If the data was collected with bias, if the sample is unrepresentative, if the measurements are unreliable, regression will faithfully produce an answer that inherits all of those flaws.

Regression is a powerful tool. But like any tool, it is only as good as the hands that wield it and the materials it is given.

There is one more limitation worth mentioning, because it trips up beginners and experts alike. Simple linear regression assumes that the relationship you care about can be summarized by a single straight line for the entire dataset. But sometimes the relationship differs across subgroups. The effect of study hours on exam scores might be steep for students who are studying effectively and flat for students who are just rereading the same highlighted passages. The regression line, averaged across everyone, might show a moderate positive slope that accurately describes nobody. When you suspect that different groups in your data have different relationships between X and Y , a single regression line is papering over important differences. This is another reason to look at your scatter plot carefully before fitting a line, and to consider whether color-coding the points by a grouping variable reveals patterns that a single line would miss.

AI Reality Check

Give an AI tool two columns of numbers and ask it to “analyze the relationship,” and you will almost certainly get a regression line. The slope, the intercept, the R^2 , and a p -value will appear in seconds. What you will almost never get, unless you explicitly ask, is a residual plot.

This is a problem, because the regression output can look perfectly reasonable, statistically significant slope, respectable R^2 , while the residual plot reveals that the model is fundamentally wrong. A curved pattern in the residuals means the linear model is missing the real relationship. A funnel shape means the inference (standard errors, p -values, confidence intervals) cannot be trusted. A single influential point dragging the line means the slope describes one outlier rather than the data.

AI tools will also extrapolate without warning. Ask for a prediction at a value of X far outside the observed range and the tool will compute $\hat{y}$ without hesitation, even when the prediction is absurd. It has no concept of “the data doesn’t go there.”

The fix is simple in principle: always plot your data before fitting a model, and always check the residual plots after. But it requires the discipline to ask the AI for diagnostics rather than accepting the first output it produces. The AI is doing the arithmetic. You are doing the thinking.

10.16 Looking Ahead

Simple linear regression models the relationship between one explanatory variable and one response variable. But most outcomes in the real world are influenced by more than one factor. Test scores depend on spending, yes, but also on demographics, teacher quality, class size, and a dozen other variables. Salary depends on experience, but also on education, occupation, industry, and geography. The next chapter extends regression to handle multiple predictors simultaneously. Multiple regression lets you ask, “What is the relationship between X and Y , after accounting for the effects of other variables?” That question, and the careful interpretation it demands, is where regression becomes both one of the workhorses of applied statistics and one of the methods most often misused.

10.17 Further Reading and References

The following works are cited in this chapter or provide valuable additional context for the ideas covered here.

On the history of regression: Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press. (Covers the competing claims of Gauss and Legendre over the invention of the least squares method, the historical episode described in this chapter’s note on the prediction of Ceres.)

On education spending and student outcomes: Jackson, C. K., Johnson, R. C., & Persico, C. (2016). The effects of school spending on educational and economic outcomes: Evidence from school finance reforms. *Quarterly Journal of Economics*, 131(1), 157–218. (A rigorous causal analysis showing that increased school spending, when targeted at low-income districts, leads to

improved educational attainment and adult earnings, using methods that go beyond the simple correlations discussed in this chapter.)

On the ecological fallacy: Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3), 351–357. (The classic paper demonstrating that correlations at the aggregate level can differ dramatically from correlations at the individual level, a warning relevant to the state-level education spending analysis in this chapter.)

On regression diagnostics: Fox, J. (2016). *Applied regression analysis and generalized linear models* (3rd ed.). Sage Publications. (A comprehensive treatment of regression diagnostics, influence measures, and assumption checking, going well beyond the introduction provided in this chapter.)

For further reading on regression: Gelman, A., Hill, J., & Vehtari, A. (2020). *Regression and other stories*. Cambridge University Press. (A modern, applied approach to regression that emphasizes interpretation, assumptions, and causal reasoning, written for a broad audience of data analysts and researchers.)

10.18 Key Terms

- **Explanatory variable (independent variable, predictor):** The variable used to predict or explain the response. Plotted on the horizontal axis. Denoted X .
- **Response variable (dependent variable, outcome):** The variable being predicted or explained. Plotted on the vertical axis. Denoted Y .
- **Regression line:** A straight line of the form $\hat{y} = b_0 + b_1x$ that summarizes the linear relationship between X and Y in a dataset.
- **Least squares regression:** The method of fitting the regression line by minimizing the sum of the squared residuals.
- **Slope (b_1):** The amount by which $\hat{y}$ changes for each one-unit increase in X . Computed as $b_1 = r \frac{s_y}{s_x}$.
- **Standard error of the slope ($SE(b_1)$):** A measure of how much the estimated slope would vary across repeated samples. Used to construct confidence intervals and hypothesis tests for the slope.

- **t-test for the slope:** A hypothesis test of $H_0 : \beta_1 = 0$, testing whether the observed slope is significantly different from zero. The test statistic is $t = b_1 / SE(b_1)$ with $n - 2$ degrees of freedom.
- **Intercept (b_0):** The predicted value of Y when $X = 0$. Computed as $b_0 = \bar{y} - b_1 \bar{x}$.
- **Residual (e_i):** The difference between an observed value and its predicted value, $e_i = y_i - \hat{y}_i$. Positive residuals indicate the observation is above the line, negative residuals indicate it is below.
- **Sum of squared residuals (SSR or $SS_{Residual}$):** The total of all squared residuals, $\sum (y_i - \hat{y}_i)^2$. The quantity minimized by the least squares method.
- **Coefficient of determination (R^2):** The proportion of the total variability in Y that is explained by the linear relationship with X . Equals r^2 in simple linear regression.
- **Ecological fallacy:** The error of drawing conclusions about individuals from data measured at the group or aggregate level. A positive relationship at the state level does not necessarily hold at the individual level.
- **Interpolation:** Predicting Y for an X value within the range of the observed data.
- **Extrapolation:** Predicting Y for an X value outside the range of the observed data. Generally unreliable because the linear pattern may not hold beyond the observed range.
- **Linearity assumption:** The condition that the relationship between X and Y is approximately a straight line.
- **Independence assumption:** The condition that the observations are independent of one another.
- **Normality assumption:** The condition that the residuals are approximately normally distributed. Most important for inference (confidence intervals, hypothesis tests) rather than for the point estimates of slope and intercept.
- **Equal variance (homoscedasticity):** The condition that the spread of residuals is roughly the same across all values of X .
- **Heteroscedasticity:** The condition where the spread of residuals changes across values of X , violating the equal variance assumption.
- **High-leverage point:** A data point with an X value far from $\bar{x}$, giving it the potential to strongly influence the regression line.
- **Influential point:** A data point whose removal would substantially change the fitted regression line.
- **Normal probability plot (QQ plot):** A diagnostic plot that compares the distribution of the residuals to a normal distribution. Points should fall along a straight diagonal line if the normality assumption is satisfied.

- **Residual standard error** (s_e): The standard deviation of the residuals, measuring the typical distance of observed values from the regression line. Computed as $s_e = \sqrt{SS_{Residual}/(n-2)}$.
-

10.19 Exercises

10.19.1 Check Your Understanding

1. In the regression equation $\hat{y} = b_0 + b_1x$, explain what the hat symbol over y means and why it is important.
2. In your own words, explain the least squares criterion. What is being minimized, and why do we square the residuals instead of just adding them up?
3. A regression analysis of hours spent exercising per week (X) and resting heart rate in beats per minute (Y) produces the equation $\hat{y} = 82.4 - 1.3x$. Interpret the slope in the context of this problem. Then interpret the intercept.
4. Explain what R^2 measures. If $R^2 = 0.64$, what does this tell you? What does it not tell you?
5. What is a residual? If a data point has a residual of -7.2 , what does that tell you about how the model's prediction compares to the actual observation?
6. State the four assumptions of simple linear regression using the LINE acronym. For each, explain in one or two sentences what the assumption means.
7. Describe what you would look for in a residuals vs. fitted values plot if you were checking the linearity and equal variance assumptions. What patterns would concern you?
8. Explain the difference between interpolation and extrapolation. Why is extrapolation risky even when the regression model fits the observed data well?
9. A researcher computes $r = 0.85$ between two variables and reports that X causes changes in Y because the regression has a high R^2 . What is wrong with this reasoning?
10. Can you always compute a least squares regression line for any scatter plot, even one with a curved pattern? If so, why is it a problem to do so?
11. A regression of monthly sales (Y) on advertising spending (X) produces a slope of $b_1 = 2.4$ with a standard error of $SE(b_1) = 0.8$ and a sample size of $n = 30$. Compute the t -statistic for the slope. With $df = 28$, would this slope be statistically significant at $\alpha = 0.05$? Explain what this test tells you.

12. A regression output shows a slope of 0.03 with a p -value of 0.001 and an R^2 of 0.02. A classmate says, “The p -value is tiny, so this must be a strong and important relationship.” Do you agree? Explain the difference between statistical significance and practical significance in this context.

10.19.2 Apply It

(See Appendix B for complete variable descriptions for all datasets used in these exercises.)

Use the dataset **education-spending.csv** for the following problems. This is a simulated dataset, mathematically modeled to reflect the kinds of state-level patterns reported in the NCES Digest of Education Statistics, NAEP composite test scores, and the Census Bureau’s Small Area Income and Poverty Estimates (SAIPE). The 50 rows represent the 50 U.S. states. Variables include `state`, `spending_per_student` (per-pupil spending, dollars), `avg_score` (an average composite test score in NAEP-style units), `median_household_income` (dollars), `pct_poverty`, `student_teacher_ratio`, and `pct_free_lunch`. The correlations and orders of magnitude in the simulated data are calibrated to the relationships observed in the underlying real sources, but no individual state’s row should be read as that state’s actual reported value.

1. Create a scatter plot of per-pupil spending (X) versus average test score (Y). Describe the overall pattern. Does the relationship appear approximately linear? Are there any obvious outliers?
2. Compute the correlation between per-pupil spending and average test score. Based on this value, describe the strength and direction of the linear relationship.
3. Using the summary statistics provided with the dataset (or computed by your software), calculate the slope and intercept of the least squares regression line for predicting test score from per-pupil spending. Write the equation in the form $\hat{y} = b_0 + b_1x$.
4. Interpret the slope of the regression line in the context of education spending and test scores. Be specific about units.
5. Interpret the intercept of the regression line. Does the intercept have a meaningful interpretation in this context? Explain why or why not.
6. Compute R^2 and interpret it in context. What percentage of the variability in test scores is explained by per-pupil spending? What might account for the unexplained variability?
7. Using the regression equation, predict the average test score for a state that spends \$11,500 per pupil. Is this interpolation or extrapolation? How do you know?

8. Using the regression equation, predict the average test score for a state that spends \$40,000 per pupil (note that this is above the maximum spending level in the dataset). Why should you be cautious about this prediction?
9. Identify the state with the largest positive residual and the state with the largest negative residual in the dataset. For each, calculate the residual and explain what it tells you about that state's performance relative to what the model predicts.
10. Create a residuals vs. fitted values plot and a QQ plot for the residuals. Based on these plots, assess whether the LINE assumptions appear to be met. Describe any patterns you observe and what they suggest about the model's adequacy.
11. Using your regression model from problem 3, examine the regression output from your software. Report the standard error of the slope, the t -statistic, and the p -value. Is the slope statistically significant at $\alpha = 0.05$? Construct a 95% confidence interval for the slope and interpret it in the context of education spending and test scores.
12. Identify any high-leverage or influential observations in your regression model. In R, `cooks.distance(model)` produces **Cook's distance** for each observation; a common rule of thumb flags any observation with Cook's distance greater than $4/n$ as worth a second look. Plot Cook's distance against observation index, mark the threshold, and identify which states (if any) cross it. Then refit the regression with those states removed. How much do the slope, intercept, and R^2 change? Should those states be excluded from the analysis, or do they represent genuine variation in the population? Justify your answer.

10.19.3 Think Deeper

1. The opening story of this chapter describes the scatter plot of education spending vs. test scores. Suppose someone uses this regression to argue that the state legislature should increase per-pupil spending by \$2,000, predicting it will raise test scores by a specific amount. What assumptions is this person making beyond what the regression supports? What confounding variables might complicate this claim? What study design would give stronger evidence about the causal effect of spending?
2. Two researchers analyze the same education dataset. Researcher A fits a regression of test scores on per-pupil spending and gets $R^2 = 0.19$. Researcher B fits a regression of test scores on median household income and gets $R^2 = 0.25$. Researcher B concludes that income "matters more" than spending for education outcomes. Evaluate this conclusion. What is right about it? What is misleading? What additional analysis would you want before drawing conclusions about which variable matters more?

3. A school district collects data over 20 years on its annual budget (adjusted for inflation) and average student test scores. They fit a regression and find a positive slope with $R^2 = 0.72$. A board member argues this proves that their funding increases have been effective. Discuss at least two threats to the validity of this conclusion. Consider the independence assumption, the possibility of confounders that change over time, and the difference between cross-sectional and longitudinal data.
4. Consider the residuals from a regression of test scores on per-pupil spending. Suppose the residuals vs. fitted values plot shows a clear funnel shape, with residuals becoming more spread out as fitted values increase. What does this violation of the equal variance assumption mean in practical terms for the education data? Which predictions from the model should you trust more, and which less? What could you do to address this problem?
5. A data journalist publishes an article with the headline, “States That Spend More on Education See Only Modest Test Score Gains, Regression Shows.” The regression has a small positive slope and an R^2 of 0.15. A reader comments that the analysis is misleading because it ignores the differences between states. Another reader comments that the low R^2 proves spending does not matter. Evaluate both the headline and the two reader comments. Who is closest to the truth, and what nuance is everyone missing?

Chapter 11

Multiple Regression

11.1 The Question That Launched a Thousand Arguments

In the 1985 Current Population Survey data used throughout this chapter, the average hourly wage for women working full-time was about 78.8% of men's wages. This means that for every dollar earned by the average man, the average woman earned about 79 cents.

That number, the raw gender wage gap, is among the most-argued statistics in American public life. The arguments almost always take the same form.

One side says the gap is evidence of discrimination. The other side says that once you “control for” occupation, experience, hours worked, education, and industry, the gap shrinks to a few cents or disappears entirely. The first side counters that you cannot control for occupation and still call the analysis fair, because the very occupations women end up in are shaped by discrimination. The second side says you cannot attribute to discrimination what is explained by choices.

Around and around it goes. And what almost nobody in the argument stops to do is ask the question that actually matters, the question this chapter will teach you to answer. When someone says they are “controlling for” a variable, what, precisely, are they doing? What does the math look like? What assumptions is the analysis making? And when are those assumptions reasonable versus dangerously misleading?

The tool at the center of this debate is **multiple regression**. It is widely used across empirical social science, business analytics, public health research, and increasingly, everyday journalism. It lets you model the relationship between an outcome and several predictors simultaneously, which

is what people mean when they say they are “controlling for” things. Understanding what it can and cannot do is a foundational form of statistical literacy.

Let us build that understanding from the ground up.

11.2 From One Predictor to Many

In Chapter 10, we explored simple linear regression. We modeled the relationship between a single predictor variable X and an outcome variable Y using the equation

$$Y = b_0 + b_1X + e$$

where b_0 is the intercept, b_1 is the slope, and e is the sample-level residual — the part of Y that the fitted line does not explain. The corresponding population-level equation, introduced in the *Signal, Noise, and the Population Model* section of Chapter 10, replaces b_0, b_1, e with $\beta_0, \beta_1, \varepsilon$, where ε is the unobservable random error. The same convention carries through this chapter: Greek letters for unknown population parameters, Latin for the sample estimates we actually compute.

If X is years of work experience and Y is annual salary, then b_1 tells you how much salary tends to change for each additional year of experience. The slope b_1 has units: dollars per year of experience. This will matter shortly when we discuss what coefficients mean.

But salary is not determined by experience alone. Education matters. Occupation matters. Geographic location matters. Union membership and other factors matter. If you want to understand salary, a model with one predictor is going to miss a lot.

Multiple regression addresses this by allowing more than one predictor variable in the model. The general form is

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \cdots + b_kX_k + e$$

where k is the number of predictors. Each b_j is a **regression coefficient** that tells you the estimated relationship between predictor X_j and the outcome Y , while accounting for all the other predictors in the model. Each coefficient carries units: b_j is in *units of Y per unit of X_j* . If Y is annual salary in dollars and X_j is years of experience, b_j is in dollars per year. If X_j is a dummy variable (which is unitless, just 0 or 1), then b_j is in units of Y — for example, dollars. The unitless quantity from

Chapter 4, the correlation r , is a different beast; raw regression coefficients keep the units of the variables they relate.

Suppose you model salary as a function of experience, education (measured in years of schooling), and weekly hours worked.

$$\text{Salary} = b_0 + b_1(\text{Experience}) + b_2(\text{Education}) + b_3(\text{Hours}) + e$$

If you fit this model to data and find $b_1 = 1,850$, that means each additional year of experience is associated with an estimated \$1,850 increase in salary, *after accounting for differences in education and hours worked*. This is the “holding all other variables constant” interpretation, and it is both the great power and the great peril of multiple regression. We will dig into what it means, and what it does not mean, shortly.

11.2.1 How the Coefficients Are Estimated

The estimation procedure for multiple regression follows the same logic as simple regression. We find the set of coefficients $b_0, b_1, \dots, b_k$ that minimizes the sum of squared residuals

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

where $\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \dots + b_k X_{ki}$ is the predicted value for observation i . This is still ordinary least squares (OLS, introduced in Chapter 10), just extended to more dimensions. The mathematics involves matrix algebra rather than the simple formulas from Chapter 10, but the principle is identical. Find the line, or more accurately the hyperplane, that makes the predictions as close to the actual values as possible, on average.

In simple regression, you are fitting a line through a cloud of points in two-dimensional space. In multiple regression with two predictors, you are fitting a plane through a cloud in three-dimensional space. With more predictors, you are fitting a surface in higher-dimensional space that you cannot visualize but that the math handles without complaint.

11.3 “Holding All Other Variables Constant”

This phrase is the most important thing to understand about multiple regression, and it is also the thing most likely to be misunderstood.

When we say that the coefficient on experience is \$1,850 “holding education and hours constant,” we mean something very specific. We are comparing people who have the same education level and the same weekly hours, and asking how salary differs between those who have, say, 10 years of experience versus 11 years. Among people with equal education and equal hours, each extra year of experience is associated with about \$1,850 more in salary.

This is not the same as the simple relationship between experience and salary. In a simple regression of salary on experience alone, you might get a coefficient of \$2,400. The reason it differs is that experience is correlated with other variables. People with more experience often have more education. People with more experience might work longer hours. The simple regression lumps all of these effects together. Multiple regression tries to separate them.

The word “tries” is doing heavy lifting in that sentence. Multiple regression separates the effects *to the extent that the model’s assumptions are met and the right variables are included*. It cannot separate what it cannot see. If there is an important variable that affects salary and is correlated with experience, but you left it out of the model, then the coefficient on experience will absorb some of that missing variable’s effect. This is called **omitted variable bias**, and we will come back to it later in this chapter because it is a foundational concept for thinking critically about regression results.

For now, keep this in mind. “Holding all other variables constant” means “holding all other *included* variables constant.” It says nothing about variables not in the model.

Note

The phrase “controlling for” has become so common in popular writing that people use it almost casually. “After controlling for income, there is no difference in health outcomes between the two groups.” Statements like this can be accurate, but they are only as good as the model that produced them. When you see “after controlling for” in a news article or a research summary, your first question should be, “What exactly was controlled for, and what was left out?” Your second question should be, “Are any of the control variables themselves affected by the thing we are studying?” We will explore this second question in the ethics section of this chapter.

11.4 Categorical Predictors and Dummy Variables

So far, we have talked about predictors that are numerical, like years of experience or hours worked. But many central predictors in social science and business are categorical. Gender. Race. Region of the country. Occupation. Job title. Department.

You cannot plug “female” into the equation $Y = b_0 + b_1X_1$ as a number. So how do we include categorical variables in a regression model?

The answer is **dummy variables**, also called **indicator variables**. A dummy variable is a variable that takes the value 1 if a condition is true and 0 otherwise. To include gender (coded as male and female) in a regression model, you create a variable

$$\text{Female} = \begin{cases} 1 & \text{if the person is female} \\ 0 & \text{if the person is male} \end{cases}$$

A note on this coding before we go further. Treating gender as a binary variable here reflects how it was recorded in the 1985 Current Population Survey (and in most administrative data of that period), not a claim that gender is binary in fact. When working with a richer gender variable that includes nonbinary identities or self-described categories, you would create $k - 1$ dummy variables in the same way the next subsection describes for any categorical variable with k levels. The mechanics of the regression do not change. Only the categories — and the care with which you describe them — do.

Then you include this variable in the model like any other predictor.

$$\text{Salary} = b_0 + b_1(\text{Experience}) + b_2(\text{Education}) + b_3(\text{Female}) + e$$

If $b_3 = -4,200$, that means women in this sample earn an estimated \$4,200 less than men, after accounting for differences in experience and education. The group coded as 0, in this case men, is called the **reference category**. The coefficient on the dummy variable always represents the difference between the coded group and the reference group.

11.4.1 Categorical Variables with More Than Two Categories

What if you want to include a variable like occupation, with six categories: Manufacturing, Tech, Education, Healthcare, Finance, and Retail? You cannot represent this with a single variable coded

1 through 6, because that would imply that the occupations have a natural ordering and that the distance between Manufacturing and Tech is the same as between Tech and Education. That makes no sense.

Instead, you create $k-1$ dummy variables, where k is the number of categories. For six occupation categories, you create five dummies.

$$\begin{aligned} \text{Tech} &= \begin{cases} 1 & \text{if Tech} \\ 0 & \text{otherwise} \end{cases} & \text{Education} &= \begin{cases} 1 & \text{if Education} \\ 0 & \text{otherwise} \end{cases} & \text{Healthcare} &= \begin{cases} 1 & \text{if Healthcare} \\ 0 & \text{otherwise} \end{cases} \\ & & \text{Finance} &= \begin{cases} 1 & \text{if Finance} \\ 0 & \text{otherwise} \end{cases} & \text{Retail} &= \begin{cases} 1 & \text{if Retail} \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Someone in Manufacturing would have all five dummies equal to 0. Manufacturing becomes the reference category. The coefficient on Tech tells you the average difference in the outcome between people in Tech and people in Manufacturing, holding everything else constant. The coefficient on Education tells you the Education-versus-Manufacturing difference. And so on.

Why $k-1$ dummies instead of k ? Because if you know that someone is not in Tech, not in Education, not in Healthcare, not in Finance, and not in Retail, you already know they are in Manufacturing. The sixth dummy is redundant, and including it would cause a mathematical problem called perfect multicollinearity (the columns of the design matrix would be linearly dependent, and the estimation procedure would break down). Most software handles this automatically, but it helps to understand why one category always goes missing from the output.

11.4.2 Choosing the Reference Category

The choice of reference category does not change the substance of the model. It only changes which comparisons show up directly in the coefficients. If you make Education the reference category instead of Manufacturing, the coefficients on the other occupation categories change, but the predicted values for every observation remain exactly the same.

That said, it is a good practice to choose a reference category that makes interpretation easy. Common choices include the most common group, the group that serves as a natural baseline, or the group that your audience is most likely to be comparing against.

11.5 Interaction Effects

So far, every model we have written assumes that the effect of each predictor is the same regardless of the values of the other predictors. Experience adds \$1,850 to salary whether you are male or female, in Manufacturing or Finance, in Alabama or Alaska. That is a strong assumption, and it is often wrong.

An **interaction effect** occurs when the relationship between one predictor and the outcome depends on the value of another predictor. In the wage gap context, suppose the salary premium for each additional year of experience is larger for men than for women. That would mean gender and experience *interact* in their effect on salary.

To model this, you create an **interaction term** by multiplying the two predictors together and adding the product as a new variable.

$$\text{Salary} = b_0 + b_1(\text{Experience}) + b_2(\text{Female}) + b_3(\text{Experience} \times \text{Female}) + e$$

Now the effect of experience is not a single number. For men (Female = 0), each additional year of experience is associated with a b_1 increase in salary. For women (Female = 1), each additional year is associated with a $b_1 + b_3$ increase. The coefficient b_3 tells you how much the experience effect *differs* between men and women. If b_3 is negative, the returns to experience are smaller for women. If b_3 is positive, the returns are larger.

This matters for interpretation. A model without the interaction might show a \$4,000 gender gap that looks constant across experience levels. A model with the interaction might reveal that the gap is small for recent graduates and grows with each year of experience, a very different story with very different policy implications.

A practical caution: interaction terms multiply the number of coefficients and can introduce **multicollinearity** (defined formally in the next major section), especially when the interacted variables are correlated. Include interactions when you have a substantive reason to believe the effect of one predictor depends on another. Do not include them routinely in every model.

11.6 R-Squared vs. Adjusted R-Squared

In Chapter 10, we introduced R^2 as a measure of how well the regression model fits the data. It tells you what proportion of the variance in the outcome is explained by the predictors. An R^2

of 0.45 means the model explains 45% of the variability in the outcome, with the remaining 55% unexplained.

Multiple regression introduces a subtlety. When you add a new predictor to a regression model, R^2 will go up, or at worst stay the same, *even if the new predictor has no real relationship with the outcome*. This is built into the algebra — every new predictor gives the model one more degree of freedom to chase the noise in the data, and R^2 rewards that chase. If you added random numbers as a predictor, R^2 would still go up (by a tiny amount).

This means R^2 is a poor guide for comparing models with different numbers of predictors. A model with 20 predictors will almost always have a higher R^2 than a model with 5 predictors, but that does not mean the 20-predictor model is better. Some of those 15 extra predictors might be adding noise, not signal.

Adjusted R^2 fixes this by penalizing the model for each additional predictor. The formula is

$$R_{adj}^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}$$

where n is the sample size and k is the number of predictors. The penalty increases as you add predictors. If a new predictor does not improve the model enough to overcome the penalty, adjusted R^2 will actually go down. This makes adjusted R^2 a better tool for comparing models of different sizes.

A practical example. You model salary using experience alone and get $R^2 = 0.32$ and $R_{adj}^2 = 0.319$. You add education and hours worked. Now $R^2 = 0.47$ and $R_{adj}^2 = 0.465$. Both R^2 and adjusted R^2 went up, suggesting the additional predictors provide real improvement. You then add 15 more variables, including shoe size, number of vowels in the person's last name, and birth month. Now $R^2 = 0.49$ but $R_{adj}^2 = 0.441$. The adjusted R^2 went down, telling you that the additional variables are adding complexity without enough explanatory value to justify it.

Neither R^2 nor adjusted R^2 should be the sole criterion for model selection. They say nothing about whether the model's assumptions are met, whether the coefficients are interpretable, or whether important variables were omitted. But adjusted R^2 is a useful early diagnostic for deciding whether a more complex model is pulling its weight.

11.7 Multicollinearity

Multiple regression works by separating the individual contributions of each predictor. But what happens when two or more predictors are highly correlated with each other? This situation is called **multicollinearity**, and it creates real problems for interpretation.

11.7.1 What It Is

Consider a model predicting salary from both “years of experience” and “age.” These two variables are highly correlated. Older people generally have more experience. When you put both in a regression, the model has a hard time telling which variable is driving the relationship with salary. Is salary going up because of experience, or because of age, or some mixture of both?

The math still works. The model will produce estimates. But the coefficients become unstable. Small changes in the data can cause large swings in the estimated coefficients. The standard errors of the coefficients become inflated, making it harder to detect statistically significant relationships. And the individual coefficients become difficult to interpret, because “holding age constant while increasing experience” is nearly impossible in reality, the two move together.

11.7.2 The Extreme Case

Perfect multicollinearity, where one predictor is an exact linear function of others, causes the estimation to fail entirely. The classic example is including dummy variables for all categories of a categorical variable (including the reference), which we discussed above. Another example would be including both “income in dollars” and “income in thousands of dollars” as separate predictors. Software will refuse to estimate the model or will drop one variable automatically.

11.7.3 Detecting Multicollinearity with VIF

The most common diagnostic for multicollinearity is the **Variance Inflation Factor** (VIF). The VIF for a predictor X_j is calculated by regressing X_j on all the other predictors and computing

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the R-squared from that auxiliary regression. A VIF of 1 means the predictor is not correlated with the others at all. A VIF of 5 means the variance of the coefficient is inflated by a factor of 5 due to correlation with other predictors. A VIF of 10 means tenfold inflation.

Common rules of thumb suggest concern when VIF exceeds 5, and serious concern when it exceeds 10. But these are guidelines, not laws. In some contexts, even a VIF of 3 can cause problems. In others, a VIF of 8 might be tolerable if you are not interested in interpreting the individual coefficient.

11.7.4 What To Do About It

If multicollinearity is a problem, your options include dropping one of the correlated predictors, combining correlated predictors into a single measure (for instance, creating an index), collecting more data (larger samples reduce the instability, though they do not eliminate it), or simply acknowledging that the individual coefficients for the correlated predictors are unreliable while the overall model may still predict well.

The key distinction is between **prediction** and **interpretation**. Multicollinearity can severely undermine interpretation (you cannot tell which predictor matters) while leaving prediction largely unaffected (the combined effect of the correlated predictors is still captured). If your goal is prediction, multicollinearity may not matter much. If your goal is understanding which individual factors drive the outcome, multicollinearity is a serious obstacle.

Note

Here is a useful analogy for multicollinearity. Imagine you are trying to figure out whether a basketball team's success is due to the point guard or the shooting guard. If the two players always play together, you never see what happens when one plays without the other. You can measure the team's performance, but you cannot separate the individual contributions. The only way to isolate their effects would be to observe games where one plays and the other does not. Multicollinearity is the statistical version of this problem. When two predictors move together, the model cannot separate their effects.

Try It Online

Open the **Regression Explorer** on the companion website. Add and remove predictor variables from a multiple regression model, watch how the coefficients and standard errors

change, and try toggling on a predictor that is correlated with another (the app's preset X_5 is correlated with X_1) to see multicollinearity in action through the inflated standard errors. In R, `car::vif()` computes the variance inflation factors directly for any fitted `lm()` model.

11.8 When Regression Can and Cannot Support Causal Claims

This is the section where the rubber meets the road. Multiple regression is used every day to make claims that sound causal. “Education increases earnings.” “Smoking causes lung cancer.” “After controlling for confounders, the treatment improved outcomes.” But whether regression can actually support such claims depends on conditions that are much harder to satisfy than most people realize.

11.8.1 The Problem of Confounding

A **confounder** is a variable that is associated with both the predictor you care about and the outcome, and that is not on the causal pathway between them. If you are studying whether exercise reduces blood pressure, a potential confounder is overall health consciousness. People who exercise more might also eat better, sleep more, and manage stress better. If you just regress blood pressure on exercise, the coefficient on exercise captures the combined effect of exercise and all the healthy behaviors that go along with it.

Multiple regression addresses confounding by including confounders as control variables. If you add diet quality, sleep hours, and stress level to the model alongside exercise, the coefficient on exercise (in theory) captures the effect of exercise that is not attributable to those other factors.

11.8.2 The Catch

But this only works if you have measured and included *all* the relevant confounders. If there is an unmeasured confounder, a variable that affects both the predictor and the outcome but that you did not include in the model, then the coefficient on your predictor of interest is biased. It captures some of the unmeasured confounder's effect along with the predictor's own effect.

This is a core limitation of regression with observational data. You can control for what you measure. You cannot control for what you do not measure. And you can never be certain you have

measured everything that matters. There could always be another confounder lurking outside your model.

11.8.3 When Causal Claims Are Defensible

Regression-based causal claims are most defensible when several conditions hold simultaneously.

All important confounders are measured and included. This is the hardest condition to verify, because it requires subject-matter knowledge about what all the relevant confounders are, and it requires that you actually have data on each one.

The confounders are measured accurately. Even if you include the right variables, if they are measured with error, the “control” they provide is incomplete. This is sometimes called residual confounding. If you control for “education” using a crude measure like “has a college degree or not,” you are not fully accounting for differences between someone with a degree from a community college and someone with a PhD.

None of the control variables are themselves caused by the predictor of interest. This is the trickiest condition, and we will return to it in the ethics section. If exercise causes people to eat better, and you control for diet quality, you have actually removed part of the effect of exercise that operates through diet. This is called controlling for a **mediator**, and it biases the estimated effect of the predictor.

There is no reverse causation. If the outcome actually causes the predictor, or if both are caused by something else in a way that creates a feedback loop, regression coefficients do not have a causal interpretation.

11.8.4 The Gold Standard

The cleanest way to establish causation is a randomized experiment. When participants are randomly assigned to treatment and control groups, randomization ensures that confounders, both measured and unmeasured, are approximately balanced across groups. This is why randomized controlled trials sit at the top of the evidence hierarchy in medicine, and why economists and social scientists invest enormous effort in designing natural experiments and quasi-experiments that approximate random assignment.

When you cannot randomize, multiple regression is one of the best tools available. But it requires more caution, more humility, and more careful thinking about what could be missing from the model.

11.9 Omitted Variable Bias

Omitted variable bias (OVB) is what happens when you leave an important variable out of your regression model. It is far from a theoretical concern. It is a frequent source of misleading results in applied regression analysis.

11.9.1 The Intuition

Suppose you regress salary on gender. You find that women earn \$4,233 less than men. But you did not include occupation or human capital variables in the model. If women and men tend to work in different occupations (they do), and those occupations pay differently (they do), then the coefficient on gender is picking up two things: the salary difference between men and women within the same occupation, and the salary difference across the occupations that men and women tend to hold. These are very different things, and the coefficient from the simple model mixes them together.

Now suppose you add experience and education to the model. The coefficient on gender actually increases to \$4,676. This can happen when the omitted variables work in the opposite direction from what you might expect, in this case, women in the sample having slightly more education on average. Add occupation to the model, and the gender coefficient drops to \$3,930 as occupational sorting accounts for some of the gap.

Each time you add a relevant variable, the coefficient on gender changes. This is omitted variable bias at work. The coefficient on any predictor is biased when a relevant variable is left out of the model, *as long as the omitted variable is correlated with both the predictor and the outcome.*

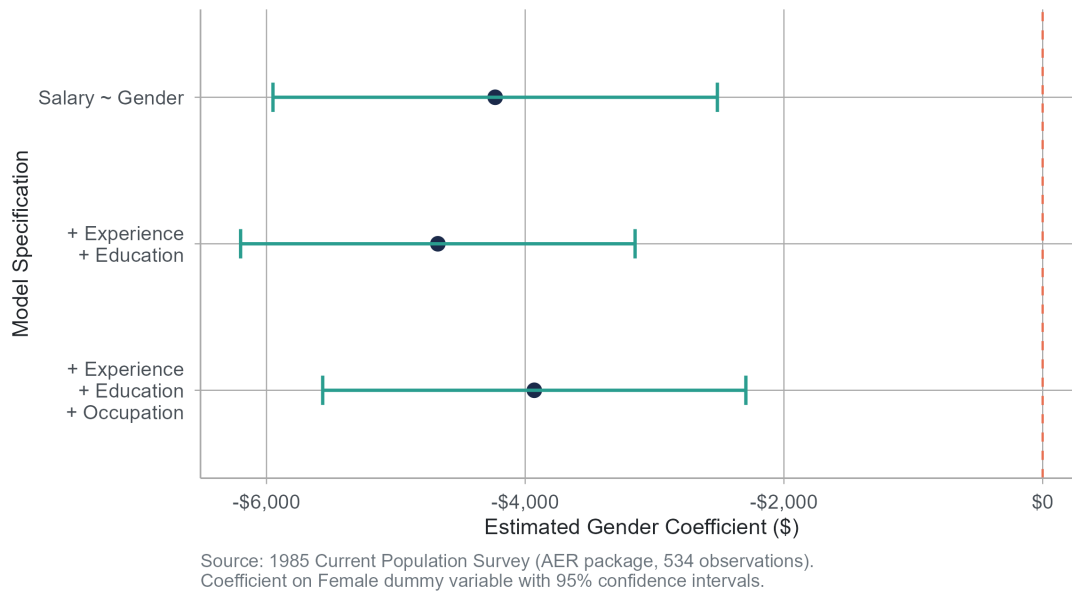


Figure 11.1: Gender wage coefficient across progressively controlled regression models. Source: 1985 Current Population Survey (534 observations, AER package).

11.9.2 The Formal Condition

Omitted variable bias occurs when two conditions are met simultaneously.

1. The omitted variable is correlated with the included predictor of interest.
2. The omitted variable is related to the outcome.

If either condition fails, there is no bias. If an omitted variable affects salary but is completely uncorrelated with gender (say, a random employee ID number that happens to correlate with salary due to some payroll system quirk), leaving it out does not bias the gender coefficient. And if a variable correlates with gender but has no relationship with salary (say, the brand of shampoo the person uses), omitting it does not cause bias either.

11.9.3 The Direction of Bias

You can even predict the direction of the bias. If the omitted variable is positively correlated with both the predictor and the outcome, then leaving it out inflates the coefficient on the predictor (positive bias). If the correlations go in opposite directions, the bias works the other way.

Returning to our wage gap example, experience is positively correlated with salary (more experience, higher pay) and let us suppose it is also positively correlated with being male (men in the sample tend to have more experience). If we leave experience out, the coefficient on the “male” dummy will be too large, overstating the male earnings advantage, because it absorbs some of the experience effect. Adding experience to the model removes that upward bias from the gender coefficient.

11.10 Simpson's Paradox

Simpson's Paradox is a common illustration of how aggregated data can mislead, and it connects directly to the themes of this chapter. It describes a situation where a trend that appears in aggregated data reverses when you look at subgroups. It is not a paradox in the logical sense — it is a consequence of confounding — but it is counterintuitive the first time you encounter it.

11.10.1 The Berkeley Admissions Case

In the fall of 1973, the University of California, Berkeley faced concerns about gender bias in graduate admissions. The aggregate numbers told a stark story: about 44% of male applicants were admitted, compared to only 35% of female applicants. That is a nine-point gap, and it favored men.

But when researchers Peter Bickel, Eugene Hammel, and J. William O'Connell examined admissions at the department level, the picture reversed. In most individual departments, women were admitted at equal or slightly higher rates than men. In some departments, women actually had a higher admission rate.

How is this possible? How can women be disadvantaged overall but advantaged within most departments?

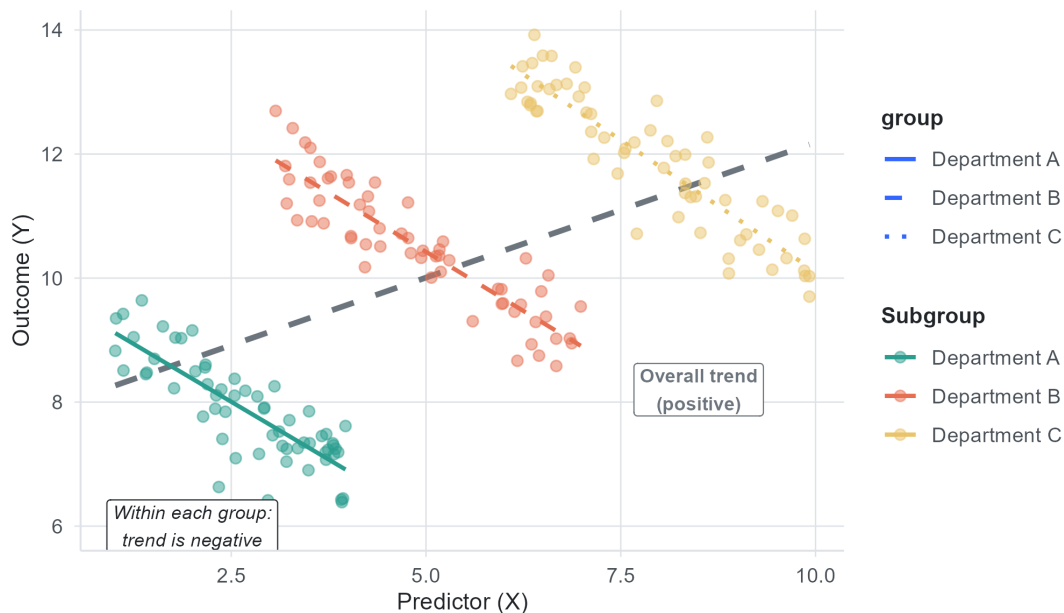
The answer lies in the pattern of applications. Women tended to apply to departments with low overall admission rates, competitive humanities and social science programs that rejected most applicants regardless of gender. Men tended to apply to departments with high overall admission rates, often in engineering and the sciences. When you aggregate across departments, the low admission rates in the departments where women concentrated drag down the overall female admission rate, even though within each department, women did fine.

The confounding variable is department. Department is associated with both gender (men and women applied to different departments) and the outcome (departments had different admission

rates). Ignoring department produced a misleading picture. Including it, the equivalent of “controlling for” department in a regression, revealed a very different story.

Simpson's Paradox: When the Whole Contradicts Its Parts

The overall positive trend reverses within each subgroup



Simulated data. Ignoring the grouping variable leads to the wrong conclusion.

Figure 11.2: Scatter plot demonstrating Simpson’s Paradox with three subgroups. The overall trend across all data points is positive (dashed line), but within each subgroup the trend is negative (solid lines). Ignoring the grouping variable leads to the opposite conclusion from the one supported by the subgroup-level data.

11.10.2 What Simpson’s Paradox Teaches Us

The lesson is not that you should always look at subgroups. Sometimes the aggregated picture is the right one. The lesson is that the relationship between two variables can change, even reverse, when you account for a third variable. And that means you need to think carefully about which variables to include in your analysis and why. The statistics alone cannot tell you which analysis is “correct.” That requires judgment about the causal structure of the problem, which variables are

confounders that should be controlled for, which are mediators that should not be, and which are irrelevant.

This is exactly where the wage gap debate lives. Whether the “raw” gap or the “controlled” gap is the more informative number depends on what question you are trying to answer. Both numbers are correct. They just answer different questions.

11.11 The Ethics of “Controlling Away” Discrimination

! Ethics Moment

One consequential misuse of multiple regression is what scholars call “controlling away” discrimination. It works like this.

You start with a raw wage gap between men and women (or between racial groups, or any other comparison). Then you add control variables: occupation, industry, education, experience, hours worked, geographic location. With each variable you add, the gap shrinks. Eventually, someone declares that “after controlling for relevant factors, the gap disappears,” and concludes that there is no discrimination.

The logical error here is subtle but consequential. Many of the variables being “controlled for” are themselves shaped by the very discrimination you are trying to measure. If women are steered away from high-paying occupations by cultural expectations, biased mentoring, hostile workplace cultures, or outright discrimination, then occupation is not an independent factor that explains the gap. It is a *mechanism* through which discrimination operates. Controlling for it does not eliminate the effect of discrimination. It hides it.

Consider a concrete example. Suppose an employer pays men and women equally within every job title. No discrimination within roles, right? But the employer systematically promotes men to senior roles faster than equally qualified women. If you control for job title in your regression, the gender coefficient goes to zero, and the analysis says there is no pay gap. But the discrimination is real. It just operates through promotion rather than through within-role pay.

This does not mean you should never include occupation or job title in a wage gap analysis. Sometimes the within-occupation gap is exactly the question you want to answer. The point is that the choice of control variables is not a neutral, technical decision. It is a substantive choice that determines what kind of discrimination your analysis can detect and what kind

it will miss. An analysis that “controls away” a gap has not demonstrated the absence of discrimination. It has narrowed its gaze to one specific pathway and ignored the others. When you read or conduct regression analyses involving equity and fairness, ask yourself what the control variables are, whether any of them could be contaminated by the outcome you are studying, and what kinds of disparities the analysis is structurally unable to detect.

This issue extends well beyond gender. In studies of racial disparities in the criminal justice system, researchers sometimes control for “criminal history” when estimating the effect of race on sentencing. But if policing itself is biased, if certain communities are policed more intensively, leading to higher arrest and conviction rates for equivalent behavior, then criminal history is contaminated by the very racial bias the study is trying to measure. Controlling for it masks rather than clarifies the role of race.

The general principle is this. Before you include a variable as a control, ask whether it could be on the causal pathway between the predictor of interest and the outcome. If it is, controlling for it can absorb part or all of the effect you are trying to estimate, leading you to wrongly conclude that no effect exists.

This is not a reason to abandon multiple regression. It is a reason to use it thoughtfully. Regression is a tool, and like any tool, it does what you tell it to, not what you intend. The choice of what to control for is where the intelligence lives.

11.12 The Bias-Variance Trade-Off in Model Building

Building a good multiple regression model requires navigating a fundamental tension. On one side, you want to include enough predictors to capture the real relationships in the data and avoid omitted variable bias. On the other, you want to keep the model parsimonious enough to avoid multicollinearity, overfitting, and uninterpretable results.

This tension is sometimes called the **bias-variance trade-off**. A model that is too simple (too few predictors) may be biased because it omits important variables. A model that is too complex (too many predictors) may have high variance, with coefficients that jump around wildly from sample to sample and do not generalize well to new data.

There is no formula for resolving this trade-off. It requires judgment. Good practice involves starting with predictors that theory and subject-matter knowledge suggest are important, adding variables incrementally while watching for signs of multicollinearity and overfitting, checking

whether adjusted R^2 improves with each addition, and resisting the temptation to throw every available variable into the model just because the data exists.

AI Reality Check

Automated feature selection algorithms, built into many statistical and machine learning software packages, promise to solve the model-building problem for you. Stepwise regression, LASSO, elastic net, and similar methods use mathematical criteria to decide which predictors to include. Machine learning pipelines in business settings routinely run through hundreds or thousands of potential predictors and select the combination that optimizes some fit metric.

These tools can be useful, but they can also produce models that are technically optimized and substantively absurd. An automated procedure might find that zip code is the most powerful predictor of loan default, but if the reason zip code predicts default is that it is a proxy for race, the model has baked racial discrimination into its predictions without anyone intending it. A stepwise regression might select predictors that have a spurious correlation with the outcome in your particular sample but no real-world connection to it.

The fundamental problem is that algorithms optimize for a statistical objective (minimizing prediction error, maximizing R^2 , or similar) without understanding what the variables mean. They do not know that including “number of hospitals within 10 miles” in a model predicting individual health outcomes is capturing geography and socioeconomic status, not some direct effect of hospital proximity. They do not know that selecting “applicant’s first name” as a predictor of job performance is likely encoding racial and ethnic bias.

Use automated feature selection as a tool for exploration, not as a replacement for thinking. After the algorithm suggests a model, ask yourself whether the predictors make substantive sense, whether any of them might be proxies for protected characteristics, and whether the model would hold up with new data from a different context. If you cannot explain why a predictor is in the model, be suspicious.

11.13 Putting It All Together with the Wage Gap

Let us return to where we started and trace through the wage gap analysis step by step, applying everything we have covered.

Step 1, the raw gap. Regress salary on gender (a single dummy variable where Female = 1). You find $b_1 = -4,233$. Women earn \$4,233 less on average. This is a descriptive fact about the data.

It does not tell you why. This model has an R^2 of just 0.042, meaning gender alone explains only about 4% of the variation in salaries.

Step 2, adding human capital variables. Add years of experience and years of education to the model. The coefficient on Female changes to $-4,676$, while experience contributes \$226 per year and education contributes \$1,880 per year. The R^2 jumps to 0.253. Interestingly, the gender gap actually widens slightly after accounting for experience and education, suggesting that women in this sample have somewhat more education or experience than men on average, and that the raw gap was actually understating the within-group difference. Whether that portion reflects discrimination or other factors is a question regression cannot answer.

Step 3, adding occupation. The coefficient on Female changes to $-3,930$, with experience at \$206 per year and education at \$1,432 per year. The R^2 rises to 0.302. Some of the gap is accounted for by the fact that men and women cluster in different occupations, and those occupations pay differently. The question of whether this clustering is itself a product of discrimination is outside the scope of the regression.

Step 4, interpretation. After controlling for experience, education, and occupation, an estimated \$3,930 gap remains. This is sometimes called the “unexplained” or “adjusted” gap. Some people interpret it as a lower bound on discrimination (it captures within-category disparities that cannot be attributed to the control variables). Others point out that unmeasured factors could explain even this residual. Both arguments have merit. What neither argument can do is use the regression alone to settle the debate. The regression tells you what the data says under a set of assumptions. Whether those assumptions are reasonable requires thinking about the world, beyond the math.

This is what statistical literacy looks like. Not blind trust in the numbers. Not blanket skepticism either. Just careful, honest engagement with what the analysis can and cannot tell you.

11.14 Assumptions and Diagnostics

Multiple regression, like any statistical method, rests on assumptions. When those assumptions are badly violated, the results can be misleading. Here are the key assumptions and how to check them.

Linearity. The relationship between each predictor and the outcome (after accounting for other predictors) should be approximately linear. Check this by plotting residuals against each predictor. If you see a clear curve, the linearity assumption is violated. Solutions include adding polynomial terms or transforming the variable. A common transformation is the **logarithm**. When salary

data is right-skewed (a few very high earners pulling the distribution), regressing $\log(\text{salary})$ on the predictors often produces better-behaved residuals and more interpretable results. In a log-transformed model, coefficients approximate percentage changes rather than dollar changes; each additional year of experience is associated with, say, a 3% increase in salary rather than a flat dollar amount. Log transformations are also useful for predictors. Taking the log of a highly skewed predictor like company revenue or city population can linearize its relationship with the outcome.

Independence of residuals. The residual for one observation should not be correlated with the residual for another. This assumption is often violated with time-series data (today's stock return is correlated with yesterday's) or clustered data (students within the same classroom are more similar to each other than to students in other classrooms). Violations require specialized methods like time-series models or multilevel models.

Homoscedasticity. The spread of the residuals should be roughly constant across all levels of the predicted values. If the residuals fan out (getting more spread as $\hat{Y}$ increases), you have **heteroscedasticity**, and the standard errors of the coefficients may be wrong. A residual-versus-fitted-values plot is the standard diagnostic. If heteroscedasticity is present, you can use robust standard errors or transform the outcome variable.

Normality of residuals. For inference (hypothesis tests and confidence intervals about the coefficients), the residuals should be approximately normally distributed. This is the least important assumption for large samples, because the Central Limit Theorem provides coverage. For small samples, check with a Q-Q plot or a histogram of the residuals. Severe skewness or heavy tails may call for transformations or non-parametric alternatives.

No perfect multicollinearity. As discussed, the predictors should not be perfectly linearly dependent. Check VIF values and correlation matrices (the latter introduced in Chapter 4 as a way to look at all pairwise correlations among numerical variables at once).

No real-world dataset satisfies all these assumptions perfectly. The question is whether violations are severe enough to undermine the conclusions. A mild departure from normality in a sample of 500 is nothing to worry about. Severe heteroscedasticity in a sample of 50 is a real problem. Developing a feel for what matters and what does not is part of the art of regression analysis.

11.15 A Note on Statistical Significance of Individual Predictors

When software reports regression output, each coefficient comes with a standard error, a t -statistic, and a p -value. The p -value tests the null hypothesis that the corresponding population coefficient is zero, that is, that the predictor has no linear relationship with the outcome after accounting for the other predictors.

A small p -value is treated as evidence against the null hypothesis, leading to a conclusion that the predictor has a statistically distinguishable relationship with the outcome. As discussed in Chapter 8, the conventional 0.05 threshold is exactly that — a convention — and a p -value of 0.04 is not categorically different from one of 0.06. A large p -value means you cannot rule out the possibility that the predictor's apparent relationship is due to sampling variability.

But remember from Chapter 8 that statistical significance is not the same as practical importance. A coefficient of \$50 on some predictor might be statistically significant with a large enough sample, but \$50 per unit change might be meaningless in context. Conversely, a coefficient of \$5,000 might not reach statistical significance in a small sample, even though a \$5,000 difference would be quite meaningful if real.

Report coefficients and their confidence intervals alongside p -values. The coefficient tells you how big the estimated effect is. The confidence interval tells you how uncertain you are. The p -value tells you whether the data are consistent with no effect at all. All three pieces of information are needed for a complete picture.

When the Outcome Is Yes or No

Everything in this chapter assumes the outcome variable is numerical: salary, test scores, blood pressure. But many important outcomes in business and social science are binary: Did the customer buy or not? Did the patient survive or not? Did the loan default or not? Was the applicant hired or not?

Linear regression is not appropriate for binary outcomes. If you try to predict a 0/1 variable with ordinary least squares, the model can produce predicted values below 0 or above 1, which make no sense as probabilities. The solution is **logistic regression**, which models the *probability* of the outcome occurring rather than the outcome itself. Instead of a straight line, it fits an S-shaped curve (the logistic function) that keeps predicted probabilities between 0 and 1.

Logistic regression is beyond the scope of this book, but it uses the same logic of multiple predictors, coefficients, and “holding other variables constant” that you learned here. If your

next course or job requires you to model binary outcomes, you already have the conceptual foundation. The leap from linear to logistic is smaller than it looks.

11.16 Looking Ahead

This chapter has taken you from simple regression, one predictor and one outcome, to the far richer world of multiple regression, where several predictors work together (and sometimes against each other) to explain variation in an outcome. You have learned what “controlling for” really means, why it can go wrong, and how to think critically about which variables belong in a model and which do not. With this foundation, you now have the core toolkit of applied statistical analysis: description, probability, inference, and modeling. The final chapter steps back from the mechanics and looks forward. What lies beyond introductory statistics? Where do these methods lead? And what responsibilities come with the ability to analyze data? Those are the questions we close with.

11.17 Further Reading and References

The following works are cited in this chapter or provide valuable additional context for the ideas covered here.

On the gender wage gap: Blau, F. D., & Kahn, L. M. (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*, 55(3), 789–865. (A comprehensive review of the wage gap literature, including the role of occupation, human capital, and discrimination, directly relevant to this chapter’s opening example.)

On the Berkeley admissions case: Bickel, P. J., Hammel, E. A., & O’Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, 187(4175), 398–404. (The original analysis that uncovered Simpson’s Paradox in Berkeley’s graduate admissions, one of the most cited examples of how aggregated data can mislead.)

On causal inference with regression: Angrist, J. D., & Pischke, J.-S. (2014). *Mastering ‘metrics: The path from cause to effect*. Princeton University Press. (An accessible introduction to regression as a tool for causal inference, with a clear treatment of omitted variable bias and the conditions under which regression coefficients can be interpreted causally.)

Cunningham, S. (2021). *Causal inference: The mixtape*. Yale University Press. (Available free online, covers similar ground with more examples and a conversational style. An excellent next step if this chapter's treatment of causality left you wanting more.)

On the ethics of “controlling away” discrimination: Goldin, C. (2014). A grand gender convergence: Its last chapter. *American Economic Review*, 104(4), 1091–1119. (A comprehensive analysis showing that the remaining gender wage gap is driven less by occupation-level sorting and more by within-occupation differences in how work is structured and rewarded, challenging the assumption that controlling for occupation eliminates discrimination.)

For further reading on multiple regression: Gelman, A., Hill, J., & Vehtari, A. (2020). *Regression and other stories*. Cambridge University Press. (A modern treatment of regression that emphasizes interpretation, causal reasoning, and practical diagnostics, building naturally on the concepts introduced in this chapter.)

11.18 Key Terms

- **Multiple regression:** A statistical model that relates an outcome variable to two or more predictor variables. The general form is $Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k + e$.
- **Regression coefficient (b_j):** The estimated change in the outcome associated with a one-unit increase in predictor X_j , holding all other predictors in the model constant.
- **Dummy variable (indicator variable):** A variable that takes the value 1 if a condition is true and 0 if it is false, used to include categorical predictors in a regression model.
- **Reference category:** The category of a categorical variable that is coded as 0 on all dummy variables. Coefficients on the dummy variables represent comparisons to this baseline group.
- **R-squared (R^2):** The proportion of the variance in the outcome that is explained by the regression model. Ranges from 0 to 1, but always increases (or stays the same) when a predictor is added.
- **Adjusted R-squared (R_{adj}^2):** A modified version of R^2 that includes a penalty for the number of predictors, making it more appropriate for comparing models of different sizes.
- **Multicollinearity:** A condition in which two or more predictor variables in a regression model are highly correlated with each other, making it difficult to separate their individual effects on the outcome.
- **Variance Inflation Factor (VIF):** A diagnostic measure for multicollinearity. VIF values above 5 or 10 indicate concerning levels of correlation among predictors.

- **Omitted variable bias:** The bias in a regression coefficient that occurs when a relevant variable is left out of the model and that variable is correlated with both the included predictor and the outcome.
- **Confounder (confounding variable):** A variable that influences both the explanatory variable and the response variable, creating a misleading association between them if not accounted for.
- **Mediator:** A variable that lies on the causal pathway between the predictor and the outcome. Controlling for a mediator removes part of the predictor's effect, which may or may not be desirable.
- **Simpson's Paradox:** A phenomenon in which a trend that appears in aggregated data reverses or disappears when the data is separated into subgroups. It arises from confounding.
- **Heteroscedasticity:** A condition in which the spread of the residuals varies across levels of the predicted values, violating one of the assumptions of ordinary least squares regression.
- **Homoscedasticity:** The assumption that the variance of the residuals is constant across all levels of the predicted values. The opposite of heteroscedasticity.
- **Interaction effect:** A situation in which the relationship between one predictor and the outcome depends on the value of another predictor.
- **Interaction term:** A variable created by multiplying two predictors together, included in the model to capture an interaction effect.
- **Logistic regression:** A regression method for binary outcomes (yes/no, 0/1) that models the probability of the outcome using an S-shaped curve. Beyond the scope of this book but uses the same logic of multiple predictors and "holding constant" developed here.
- **Controlling for (a variable):** Including a variable in a regression model so that the coefficients on other predictors represent relationships that hold after accounting for that variable. Only as good as the model's assumptions and the variables included.
- **Bias-variance trade-off:** The tension between using a simple model (which may be biased due to omitted variables) and a complex model (which may have high variance and overfit the data).

11.19 Exercises

11.19.1 Check Your Understanding

1. Write out the general form of a multiple regression model with three predictors. Label each component and explain what it represents.
2. In a regression of salary on experience and education, the coefficient on experience is \$1,850. A student says, “Each year of experience increases your salary by \$1,850.” What is missing from this interpretation? Rewrite it more carefully.
3. Explain why you need $k - 1$ dummy variables to represent a categorical variable with k categories. What happens if you include all k dummy variables instead?
4. A researcher fits two models. Model A has three predictors and an R^2 of 0.42. Model B has ten predictors and an R^2 of 0.47. The researcher concludes that Model B is better because it explains more variance. What is wrong with this reasoning? What measure should the researcher use instead?
5. Define multicollinearity in your own words. Give an example of two predictor variables that you would expect to be highly collinear if both were included in a regression model predicting household spending.
6. A regression model has a VIF of 8.3 for one of its predictors. What does this tell you? What are your options for addressing the problem?
7. Explain the difference between a confounder and a mediator. Why does the distinction matter for deciding which variables to include in a regression model?
8. State the two conditions that must be met for omitted variable bias to occur. For each condition, explain what happens if that condition is not met (that is, why there is no bias in that case).
9. Describe Simpson’s Paradox in your own words. Give an example (different from the Berkeley admissions case discussed in the chapter) where aggregated data might reverse a trend visible in subgroups.
10. A news headline reads, “After controlling for education, experience, and occupation, the racial wage gap in tech is only 1.2%.” A reader concludes that there is very little discrimination in tech. What important question should the reader ask before accepting this conclusion?

11.19.2 Apply It

(See Appendix B for complete variable descriptions for all datasets used in these exercises.)

Use the `wage-gap.csv` dataset for the following problems. This dataset contains real data from the 1985 Current Population Survey (CPS), sourced from the AER R package. Note that salaries are in 1985 dollars. The dataset contains 534 observations with the following variables.

Variable	Description
<code>employee_id</code>	Unique identifier for each worker
<code>salary</code>	Annual salary in dollars (1985 dollars)
<code>gender</code>	“Male” or “Female”
<code>years_experience</code>	Years of work experience
<code>education_years</code>	Years of formal education
<code>occupation</code>	Occupation category (six levels: “Worker,” “Technical,” “Services,” “Office,” “Sales,” “Management”)
<code>region</code>	Geographic region (“South” or “Other”)
<code>union_member</code>	Whether the worker is a union member
<code>married</code>	Marital status
<code>ethnicity</code>	Worker’s ethnicity

1. Fit a simple regression model predicting salary from gender alone (with “Male” as the reference category). Report the intercept and the coefficient on Female. Interpret both values in context. What does this model tell you about the raw wage gap in these data?
2. Now fit a multiple regression model predicting salary from gender, `years_experience`, and `education_years`. Report all coefficients. How does the coefficient on Female change compared to the simple regression in Problem 1? What does this change suggest about the role of experience and education in explaining the raw wage gap?
3. Create dummy variables for occupation (using “Manufacturing” as the reference category) and add them to the model from Problem 2. Report the coefficient on Female and the coefficients on each occupation dummy. Which occupation has the highest estimated salary relative to Manufacturing, holding other variables constant? How did adding occupation affect the gender gap coefficient?
4. Add `region` to the model (using “Other” as the reference category). Report the full regression equation with all coefficients. How does the estimated salary in the South compare to the Other region? Is the difference statistically significant at the 0.05 level?
5. Add `union_member`, `married`, and `ethnicity` to the model. Report the new coefficient on Female and the adjusted R^2 for the model with and without these additional variables. Did the addition

of these variables improve the model according to adjusted R^2 ? How did they affect the gender gap coefficient?

6. For your full model (with all predictors), calculate and report the VIF for each predictor. Are any predictors showing signs of concerning multicollinearity? If so, which ones and what would you recommend?
7. Compare the R^2 and adjusted R^2 for the following three models: (a) salary on gender alone, (b) salary on gender, years_experience, and education_years, (c) the full model with all predictors. For each step, note whether adjusted R^2 increased and by how much. What do these comparisons tell you about the relative importance of different groups of predictors?
8. Using your full model, compute the predicted salary for a female worker with 10 years of experience, 16 years of education, in the Technical occupation in the South, who is not a union member, is married, and of Hispanic ethnicity. Then compute the predicted salary for a male worker with identical characteristics. What is the difference? Does this match the coefficient on Female? Explain why or why not.
9. Fit two models: one predicting salary from years_experience alone, and another predicting salary from both years_experience and education_years. Compare the coefficient on years_experience in the two models. Did it increase, decrease, or stay about the same? Using the concept of omitted variable bias, explain why the coefficient changed in the direction it did. (Hint: think about the correlation between experience and education, and the relationship between education and salary.)
10. Create a residual plot for your full model (residuals on the vertical axis, predicted values on the horizontal axis). Does the plot suggest any violations of the assumptions of linear regression? Specifically, comment on whether the residuals appear to have constant variance (homoscedasticity) and whether there are any obvious patterns suggesting non-linearity.
11. **Interaction effect.** Fit a model predicting salary from years_experience, gender, and the interaction years_experience $\times$ gender. In R, `lm(salary ~ years_experience * gender, data = ...)` includes both main effects and the interaction in one shot. Report the coefficients on years_experience, genderFemale (assuming Male is the reference), and the interaction term years_experience:genderFemale. Interpret each coefficient. According to this model, is the dollar return to an additional year of experience the same for men and women, or different — and by how much? Sketch (or actually plot) the two predicted regression lines, one for men and one for women, on the same axes. Why is this different information from a model that just adds gender as a control without the interaction term?

11.19.3 Think Deeper

1. Return to the wage gap analysis from this chapter. Suppose you are advising a policymaker who asks you, “What is the gender wage gap?” How would you explain that there is not a single answer to this question, but rather multiple answers depending on what you control for and why? What question does the raw gap answer? What question does the fully controlled gap answer? Why might a policymaker want both numbers?
2. A company analyzes its pay data and finds that after controlling for job title, department, and years of service, there is no statistically significant pay gap between male and female employees. The CEO announces that the analysis proves the company has pay equity. Write a memo to the CEO explaining why this conclusion may be premature. Be specific about which forms of potential discrimination the analysis can detect and which it cannot.
3. A health insurance company builds a regression model to predict healthcare costs. The model includes age, BMI, smoking status, number of chronic conditions, and zip code. A data scientist notices that zip code is one of the strongest predictors. The company plans to use the model to set premium rates. Discuss the ethical implications. What might zip code be capturing that makes it such a powerful predictor? What are the potential consequences of using zip code in rate-setting? Should the company remove zip code from the model, even if doing so makes the predictions less accurate?
4. A researcher studying the effect of class size on student test scores obtains data from 500 schools. The researcher regresses average test scores on average class size and finds a negative relationship: smaller classes are associated with higher scores. The researcher concludes that reducing class size improves learning. A colleague points out that wealthier school districts tend to have smaller classes and also tend to have students who score higher for many other reasons (parental involvement, private tutoring, better facilities). The colleague says the result is driven by omitted variable bias. The researcher responds by adding average household income as a control variable. Is this sufficient to address the concern? What other variables might still be missing? Is there a study design that would more convincingly establish a causal effect of class size?
5. Consider a criminal justice dataset where researchers examine the relationship between race and sentencing length. They control for “severity of the offense” and “prior criminal record.” After adding these controls, the racial gap in sentencing shrinks considerably. A commentator writes, “Once you account for the nature of the crime and the defendant’s history, race plays almost no role.” Critically evaluate this claim. Under what circumstances could “severity of the offense” and “prior criminal record” themselves be influenced by racial bias in the justice system? What does controlling for them potentially mask?

6. In the Berkeley admissions example discussed in this chapter, the aggregated data showed women admitted at lower rates than men, but department-level data showed women admitted at equal or higher rates. Now consider a hypothetical in the opposite direction. Suppose aggregated data shows equal admission rates for men and women, but at the department level, men are admitted at higher rates in every single department. Construct a numerical example with at least three departments showing how this could happen. (Hint: think about which departments have high admission rates and which have low ones, and how men and women distribute across them.) What lesson does this reinforce about the importance of disaggregating data?

Chapter 12

Where Do You Go from Here?

12.1 The Map So Far

You have spent eleven chapters learning to think with data. You know how to design a study that can actually answer the question you care about. You know how to summarize information with numbers and pictures that do not mislead. You understand probability well enough to know that your intuitions about it are frequently wrong, and you know how to use the normal distribution and the central limit theorem to bridge the gap between a sample and a population. You can construct confidence intervals, test hypotheses, compare groups, and build regression models that separate signal from noise.

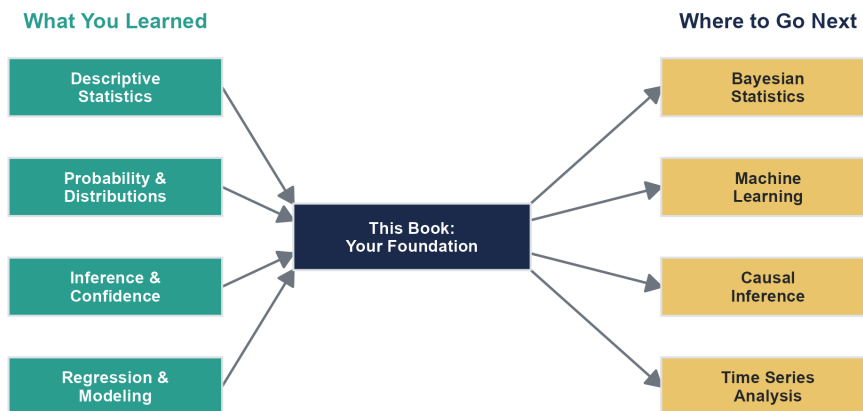
That is not a small thing. Most people who make decisions with data, in business, in government, in medicine, in everyday life, do not have even this much training. You are better equipped than you were twelve chapters ago, and better equipped than a large share of the people who will hand you charts and claim to know what the numbers mean.

But here is the honest truth. What you have learned is a foundation. A strong one, and one that will serve you well in a wide range of situations. But the field of statistics and data science extends far beyond what a single introductory course can cover. This chapter is about what lies ahead, not to overwhelm you, but to give you a sense of the terrain so you know where to go when you need more.

Think of it this way. You have learned to read. This chapter is a visit to the library.

Your Statistical Journey: Where to Go Next

The foundations in this book open doors to advanced topics



Each advanced topic builds on the core ideas of inference, modeling, and uncertainty

Figure 12.1: A roadmap showing the topics covered in this book on the left (descriptive statistics, probability and distributions, inference and confidence, regression and modeling) feeding into advanced directions on the right (Bayesian statistics, machine learning, causal inference, time series analysis). Each advanced topic builds on the foundational ideas developed in earlier chapters.

12.2 Patterns Over Time

Every dataset we have worked with in this book treats each observation as independent, as if the order does not matter. Shuffle the rows of any dataset we have analyzed, and the results would come out the same. But in many real-world problems, order is everything. Stock prices tomorrow depend on stock prices today. Monthly sales figures follow seasonal rhythms. The number of emergency room visits fluctuates with day of the week, time of year, and whether there is a pandemic happening. Rearranging these observations would destroy the very structure that makes them informative.

Time series analysis is the branch of statistics that takes the clock seriously. Instead of treating time as just another variable, time series methods are built around the idea that what happened recently tells you something about what will happen next. The core concepts are trend (the long-run

direction), seasonality (repeating patterns tied to calendar cycles), and noise (the unpredictable fluctuation that remains after you account for the other two). Models like ARIMA and exponential smoothing decompose a time series into these components and use the structure to generate forecasts.

If you have ever seen a weather forecast, an economic projection, or a demand planning model for a retail chain preparing for the holiday season, you have seen the output of time series analysis. The math gets involved, but the intuition is approachable. Past patterns, when understood carefully, help you anticipate the future.

To make this concrete, consider the problem of staffing an emergency department. Hospitals need to decide how many doctors, nurses, and support staff to schedule for each shift, every day, months in advance. Too few staff and patients wait dangerously long for care. Too many and the hospital wastes resources it cannot afford to waste. The solution is time series forecasting applied to patient arrival data. A good model captures the fact that Mondays are busier than Wednesdays, that flu season produces a predictable winter surge, that holidays bring spikes in certain types of injuries, and that a slow upward trend in the surrounding population means that last year's staffing levels will not be enough this year. Each of those patterns (weekly seasonality, annual seasonality, holiday effects, long-run trend) can be modeled and projected forward. The forecasts are never perfect, because the noise component is real and irreducible. But they are far better than guessing, and in emergency medicine, the difference between a good forecast and a bad one is measured in lives.

The tricky part, and it is a hard part, is distinguishing patterns that will persist from patterns that are about to break. The financial crisis of 2008 was, among many other things, a failure of time series models that assumed housing prices would keep following the trend they had followed for decades. They did not. The models built on that assumption were not wrong about the past. They were catastrophically wrong about the future. The COVID-19 pandemic broke time series models across nearly every domain. Demand forecasts for retail, travel projections for airlines, energy consumption models for utilities, all of them relied on historical patterns that suddenly stopped applying. The models had no way to anticipate a once-in-a-century event because their entire logic depends on the future resembling the past. Understanding when to trust a forecast and when to question it is one of the deepest challenges in this area, and one of the reasons that human judgment remains essential even when the models are sophisticated.

12.3 Causation Gets Serious

In earlier chapters, we spent time on a distinction you will carry with you for the rest of your analytical life. Correlation is not causation. Observing that two variables move together does not mean one causes the other. We talked about confounders, omitted variables, and the dangers of drawing causal conclusions from observational data.

But here is the thing. Sometimes you need to know about causation, and you cannot run an experiment. You cannot randomly assign some cities to raise the minimum wage and others not to. You cannot randomly give some people health insurance and deny it to others, at least not without serious ethical guardrails. And yet policymakers need to know whether raising the minimum wage increases unemployment, and whether expanding health insurance improves health outcomes. The decisions will be made regardless. The question is whether they will be informed by credible evidence or by ideology and gut feeling.

A set of clever methods has emerged over the past few decades to tackle exactly this problem. They go by names like **difference-in-differences**, **instrumental variables**, and **regression discontinuity**, and they all share a common logic. If you cannot create a randomized experiment, look for situations in the real world that approximate one.

Difference-in-differences works by comparing two groups over two time periods. One group experiences some change (a new policy, a new program, an external shock) and the other does not. By comparing the before-and-after difference in the treatment group to the before-and-after difference in the comparison group, you can isolate the effect of the change from everything else that was happening at the same time. The classic example is the work of David Card and Alan Krueger on the minimum wage. In 1992, New Jersey raised its minimum wage while neighboring Pennsylvania did not. Card and Krueger surveyed fast-food restaurants on both sides of the border before and after the change, using Pennsylvania as a natural comparison group. Their analysis found no evidence that the minimum wage increase reduced employment, a finding that challenged decades of conventional economic wisdom and reshaped the entire debate. The study was controversial, and the debate continues, but the method itself opened a door that has never closed. David Card was awarded the Nobel Prize in Economics in 2021, in part for this and related empirical contributions to labor economics.

Regression discontinuity takes advantage of sharp cutoffs. If students who score above 80 on a test get a scholarship and students who score below 80 do not, then comparing students who scored 79 to students who scored 81 gives you something close to a randomized experiment, because the

difference between those students is essentially random. The method is elegant precisely because it uses the arbitrariness of a threshold to create comparability.

Instrumental variables use a third variable to isolate variation in a cause that is unrelated to confounders. The logic is indirect and can be hard to grasp at first, but the method has been used to study questions ranging from the effect of education on earnings to the effect of military service on lifetime outcomes.

Another influential study in causal inference is the Oregon Health Insurance Experiment. In 2008, Oregon expanded its Medicaid program but had more applicants than slots, so it used a lottery to decide who received coverage. That lottery created something close to a randomized experiment at enormous scale, allowing researchers to study the causal effects of health insurance on health outcomes, financial security, and health care utilization with a rigor that observational data alone could never provide. The results were nuanced, showing clear improvements in financial security and mental health, but less clear effects on measured physical health outcomes, which itself became an important finding for the policy debate.

These methods are not magic. Each one relies on assumptions that must be argued for, not just asserted. Difference-in-differences requires that the two groups would have followed parallel trends without the intervention. Regression discontinuity requires that people cannot precisely manipulate their score to land on one side of the cutoff. Instrumental variables require that the instrument affects the outcome only through its effect on the treatment. Violate any of these assumptions, and the method falls apart. But if you find yourself in a career where you need to evaluate the effect of a policy, a program, or an intervention, this is the toolkit you will want next.

12.4 When Prediction Is the Point

Everything in this book has been oriented toward **inference**, toward understanding relationships, testing claims, and drawing conclusions about populations. But a parallel tradition in data analysis cares less about understanding and more about predicting. That tradition is **machine learning**.

The difference is subtle but important. In a regression model, you care about what the coefficients mean. Is the relationship between advertising spending and sales positive? How large is the effect? Is it statistically distinguishable from zero? In machine learning, you often do not care about interpreting individual parameters at all. You care about whether the model can accurately predict outcomes it has never seen before. A model that uses 500 variables in ways no human could interpret might be perfectly useful if your goal is to predict which customers will churn next month, which emails are spam, or which credit card transactions are fraudulent.

Machine learning encompasses a wide family of algorithms, from decision trees and random forests to neural networks and the deep learning models that power large language models like the one described in the AI Reality Checks throughout this book. Some of these methods have roots in the statistical methods you already know. Logistic regression, which you would encounter in a second statistics course, is also a core machine learning classifier. The lasso and ridge regression are extensions of the multiple regression you learned in Chapter 11, modified to handle situations with many more variables than observations.

One of the key ideas in machine learning is the distinction between training data and test data. When you fit a regression model in this book, you used all your data to estimate the coefficients. In machine learning, you typically split your data into two parts. You train the model on one part and evaluate its performance on the other part, data the model has never seen. This protects against **overfitting**, the problem of building a model that captures not just the real patterns in your data but also the noise, the random fluctuations that will not repeat in new data. A model that overfits looks great on the data it was trained on and performs terribly on new observations. If you have ever met someone who can explain everything that happened in the past but cannot predict anything about the future, you have met the human equivalent of an overfit model.

Here is a concrete example of how this plays out. In 2006, Netflix offered a million-dollar prize to anyone who could improve their movie recommendation algorithm by at least 10%. Thousands of teams competed over three years, building increasingly complex models that combined collaborative filtering (finding users who liked similar movies), matrix factorization, and ensemble methods that blended dozens of individual models together. The winning team, BellKor's Pragmatic Chaos, achieved the 10% improvement target, but Netflix never fully implemented the winning solution. The model was too complex and too computationally expensive for production use, and by the time the prize was awarded, Netflix's business had shifted from DVD rentals to streaming, which changed the prediction problem entirely. The competition remains a touchstone in the history of machine learning, not because the winning model was deployed, but because it demonstrated both the power and the practical limits of optimizing for prediction accuracy alone.

The tension between prediction and inference is a productive debate in modern data science. In many applications, you need both. You want a model that predicts well, and you want to understand why it predicts what it predicts. A hospital might use a machine learning model to predict which patients are at high risk of readmission, but doctors will not trust the model unless they can understand the factors driving those predictions. The field of **interpretable machine learning** has grown rapidly in response to this tension, developing tools that help explain the predictions of complex models. If this interests you, it is a productive area to explore next.

12.5 Updating Your Beliefs

Throughout this book, we have used the framework of frequentist statistics. Confidence intervals, hypothesis tests, p -values. These tools answer questions in a particular way. They ask, “If the null hypothesis were true, how surprising would this data be?” They do not directly tell you the probability that the null hypothesis is true, even though that is usually what you want to know.

Bayesian statistics offers a different framework, one that many people find more intuitive once they get past the initial unfamiliarity. The Bayesian approach starts with a **prior**, your best belief about a parameter before you see any data. Then you collect data, and the data updates your belief using Bayes’ theorem, producing a **posterior**, your revised belief that combines what you knew before with what the data tells you.

Here is a simple example. Suppose you are considering whether a coin is fair. Before flipping it, you believe it probably is, because most coins are fair. That is your prior. Then you flip it 20 times and get 15 heads. A frequentist approach would compute a p -value for the null hypothesis that the coin is fair. A Bayesian approach would combine your prior belief (probably fair) with the evidence (15 out of 20 heads) and give you a posterior distribution that shows how your belief has shifted. If your prior was strong, 20 flips might not move you much. If your prior was weak, the data will dominate.

Now consider a medical example. A diagnostic test for a rare disease comes back positive. If the disease affects 1 in 10,000 people and the test has a 5% false positive rate, then most positive results are false positives, a fact that remains stubbornly counterintuitive to most people, including many doctors. We covered this with Bayes’ theorem in the probability chapter. Bayesian statistics takes that same logic and extends it to every parameter you want to estimate, every hypothesis you want to evaluate, every prediction you want to make. It is Bayes’ theorem scaled up to the full range of statistical problems.

The Bayesian framework handles many problems more naturally than the frequentist one, especially problems involving sequential updating (you get more data over time and want to keep revising your estimate), incorporating expert knowledge, and making direct probability statements about hypotheses. It also avoids some of the interpretive headaches of p -values and confidence intervals. Instead of saying “we are 95% confident the parameter is in this interval” while carefully not saying “there is a 95% probability the parameter is in this interval,” a Bayesian analysis gives you a 95% **credible interval** that means exactly what you always wanted the confidence interval to mean. There is a 95% probability, given the data and the prior, that the parameter falls in this range. The trade-off is that it requires you to specify a prior, and people can disagree about what

the right prior is, which introduces a layer of subjectivity that frequentist methods avoid, at least on the surface.

Consider a more consequential example. During the search for Malaysia Airlines Flight 370, which disappeared over the Indian Ocean in March 2014, analysts used Bayesian methods to guide the physical search. They started with a prior distribution over possible crash locations based on the flight's last known position, fuel capacity, and satellite communication data. As search teams covered areas of ocean floor and found nothing, that absence of evidence became evidence itself. The posterior distribution was updated to shift probability away from searched areas and toward unsearched ones, focusing resources where the plane was most likely to be found given everything known so far. This sequential updating, where each new piece of information (including the information that a search area came up empty) refines the estimate, is Bayesian reasoning in its most natural form. The same approach had been used decades earlier to locate a lost hydrogen bomb off the coast of Spain in 1966 and to find the wreckage of the SS *Central America* in 1988. In each case, the Bayesian framework provided a principled way to combine incomplete information from multiple sources and update the search strategy as new data arrived.

Bayesian methods have become computationally practical thanks to advances in algorithms and computing power. Modern software like Stan and the brms package in R can fit complex Bayesian models that would have been impossible to compute a generation ago. If you continue in statistics or data science, you will eventually encounter Bayesian thinking, and when you do, you will find that the foundational concepts from this book, sampling distributions, estimation, the logic of evidence, translate directly.

12.6 The Responsibility You Carry

Let us turn from methods to something harder to put in a formula.

You now have tools that many people do not have. You can look at a dataset and see things that are invisible to someone without training. You can evaluate claims, check assumptions, identify misleading presentations of evidence, and construct analyses that separate real patterns from noise. These are powerful capabilities, and like all powerful capabilities, they can be used well or badly.

Data analysis is not neutral. Every choice an analyst makes, which variables to include, which observations to drop, which model to fit, which results to emphasize, shapes the conclusions that emerge. Two analysts working with the same data can reach different conclusions, not because one of them made an error, but because they made different defensible choices along the way. This is

not a flaw in statistics. It is a feature of any discipline that requires judgment. But it means that the analyst's integrity carries real weight.

Consider a few scenarios. A pharmaceutical company tests a drug in twelve subgroups and finds it works in one of them. They publish the positive result and do not mention the eleven failures. A consulting firm runs a regression with thirty potential control variables and reports the specification that makes the client's product look best. A social media platform adjusts its engagement algorithm based on metrics that maximize time spent on the platform without measuring effects on users' mental health. None of these involve fabricating data. All of them involve choices that distort the picture that data presents.

You will face situations where the data says something inconvenient. Where the results do not support the conclusion your boss wants, or the narrative your organization has already committed to, or the hypothesis you spent six months developing. In those moments, you will be tempted to go back and try different specifications, drop different outliers, reframe the question until the numbers cooperate. Some of that is legitimate exploratory analysis. Some of it is p-hacking — the practice the economist Ronald Coase called “torturing the data until it confesses” — which we discussed earlier in this book.

The line between legitimate exploration and data manipulation is not always bright, but here is a useful test. Would you be comfortable explaining every analytical choice you made to a skeptical but fair-minded colleague? If the answer is yes, you are probably on solid ground. If the answer involves phrases like “well, technically” or “it depends on how you look at it,” slow down.

This question of responsibility extends well beyond the classroom. Consider how it plays out in something as everyday as a lending algorithm. Having the technical ability to build a model does not, by itself, tell you whether the model should be built, or how it should be used, or who bears the cost when it is wrong. A lending algorithm that maximizes profit for the lender while systematically disadvantaging certain communities might score well on traditional accuracy metrics and still be doing harm. The statistical question, “does this model predict accurately?”, is necessary but not sufficient. You also need to ask who benefits, who is harmed, and whether the trade-offs are justified.

Those are questions that require judgment and an awareness of how the model's outputs land on the people they describe. Analyzing data is one skill. Asking whether an analysis is being used responsibly is a second skill, and the second is at least as worth developing as the first. You are now in a position to do both.

AI Reality Check

We have included AI Reality Checks throughout this book, and this one is the last, so it is worth making count.

Artificial intelligence, and large language models in particular, will keep getting better at computation. They can already clean data, run regressions, generate visualizations, write code, and summarize findings faster than any human. By the time you read this, they will probably be even better at all of those things. The trajectory is clear, and it would be foolish to pretend otherwise.

But computation is not the hard part. The hard part is deciding what computation to run. The hard part is knowing whether a regression model is appropriate or whether the data violates the assumptions in ways that make the output misleading. The hard part is recognizing when a question sounds statistical but is actually ethical, or political, or deeply contextual in ways that no model can resolve on its own. The hard part is having the judgment to say “this analysis answers the wrong question” when everyone around you is excited about the results. AI will not replace the person who can think critically about data. It will replace the person who cannot. If all you know is how to plug numbers into formulas and report the output, you are competing with software that does it faster and cheaper. If you know how to frame questions, evaluate assumptions, interpret results in context, communicate uncertainty honestly, and exercise judgment about what the numbers mean for real people, you are the person the AI needs on the other end of the conversation.

Here is a concrete way to think about it. An AI can build you a regression model in seconds. It can test every possible combination of variables, try dozens of functional forms, and report the one with the best fit statistics. What it cannot do, at least not yet, is sit with a community group and understand why a particular variable should not be used as a predictor even though it improves accuracy. It cannot navigate the trade-off between a model that is slightly less predictive and one that is fairer. It cannot decide whether a finding should be published or whether doing so would cause more harm than good. Those decisions require something the model does not have.

Your job, going forward, is not to be the computer. Your job is to be the person who asks whether the computation was the right one to run.

12.7 The Statistical Mindset

Before we turn to resources and reading lists, I want to pause and talk about something that does not fit neatly into any single chapter but runs through all of them. Call it the statistical mindset. It is the collection of habits, instincts, and reflexes that you have been building throughout this book, and it is arguably more valuable than any individual technique you have learned.

The statistical mindset begins with a comfort with uncertainty. Most of the decisions you will face in your life, in your career, in your community, in your personal choices, will be made with incomplete information. You will never have the whole picture. You will never be able to eliminate all the confounders, survey the entire population, or predict the future with certainty. The statistical mindset does not pretend otherwise. Instead, it asks a different question. Given what I know (and what I do not know), what is the most reasonable conclusion? How confident should I be? What would change my mind?

That last question is particularly important. One of the most underrated skills in analytical thinking is the ability to specify, in advance, what evidence would cause you to revise your position. If the answer is “nothing,” you are not reasoning. You are advocating. There is a place for advocacy, but it is not the same thing as analysis, and confusing the two is a common mistake people make when working with data.

The statistical mindset also involves a habit of asking “compared to what?” Every number, every result, every trend exists in context. A company reporting 10% revenue growth sounds impressive until you learn that the industry grew 25% over the same period. A medical treatment that reduces the risk of a disease by 50% sounds like a breakthrough until you learn that the risk went from 2 in 10,000 to 1 in 10,000. An absolute difference that small might not justify the cost and side effects of the treatment. You learned about this distinction between relative and absolute differences earlier in the book, and it is an idea worth keeping close. Once you start asking “compared to what?” you cannot stop. It changes how you read news, how you evaluate marketing claims, and how you think about your own experiences.

Another piece of the statistical mindset is a respect for variation. Before this book, you might have thought of an average as a fact, a single number that captures the truth about a group. Now you know better. An average is a summary, and like all summaries, it leaves things out. The distribution behind the average matters. The spread matters. The shape matters. Two groups can have the same average and look completely different. A single outlier can drag a mean in a misleading direction. When someone gives you an average without telling you about the variation, you now know to be suspicious, and that suspicion is well-calibrated.

There is also the habit of thinking about how data was generated. This is one of the first things a trained analyst does when encountering a new dataset, and it is one of the last things an untrained one thinks to ask. Where did this data come from? Who collected it? How were the participants or observations selected? What is missing? The answers to these questions determine what conclusions the data can support, regardless of how fancy the analysis is. A machine learning model trained on biased data will produce biased predictions, and no amount of algorithmic sophistication will fix a sampling problem. You learned this in Chapter 2 when we discussed study design, and it applies to every analysis you will ever encounter.

Finally, the statistical mindset includes a willingness to say “I do not know.” This might be the hardest habit to develop, because the world rewards confidence and penalizes uncertainty. People who say “the data is inconclusive” get less attention than people who say “the data proves X.” Analysts who report wide confidence intervals feel less useful than analysts who report narrow ones. But honest uncertainty is more valuable than false precision. A confidence interval that is wide because the sample was small is telling you something important. Ignoring that message, or hiding it behind false certainty, leads to bad decisions. The statistical mindset treats uncertainty not as a failure but as information, something to be quantified, communicated, and respected.

You may not realize it yet, but these habits have already started to change how you think. The next time someone shows you a chart without axes, you will notice. The next time someone claims a causal relationship based on a correlation, you will hesitate. The next time you hear a news story about a study that “proves” something, you will want to know the sample size, the confidence interval, and whether the result has been replicated. These are more than academic reflexes. They are practical tools for navigating a world that produces more data than understanding.

The formal techniques in this book, the confidence intervals and hypothesis tests and regression models, are important. But they are tools. The statistical mindset is what tells you which tool to pick up, when to put it down, and what to do when no tool quite fits the problem in front of you. Carry it with you. It will serve you well in places you do not yet expect.

If you have made it this far, through all twelve chapters, through the formulas and the examples and the careful distinctions between what data can and cannot tell you, then you have done something that matters. You have learned to think more carefully about evidence. You have built a set of tools for reasoning under uncertainty. And you have, I hope, developed a healthy respect for both the power and the limits of statistical analysis.

The world is full of numbers. They show up in news headlines, in business dashboards, in medical studies, in policy debates, in the algorithms that shape what you see and buy and believe. Some of those numbers are presented honestly. Some are not. Some are analyzed well. Some are not.

Before this book, you might not have been able to tell the difference. Now you can, or at least you can start to, and that changes things.

I want to leave you with one thought. Statistics is often taught as if it were a finished product, a set of techniques to be memorized and applied. It is not. It is a living discipline, constantly evolving, constantly arguing with itself about the right way to reason from evidence. The debates between frequentists and Bayesians, between prediction and inference, between complexity and interpretability, are not signs of a field in trouble. They are signs of a field that takes its own questions seriously. As you move forward, I hope you will engage with those questions, beyond being a consumer of statistical results, as someone who contributes to better thinking about data in whatever context you find yourself.

12.8 Where to Keep Learning

This book has given you a foundation, but a foundation is only useful if you build on it. Here are resources for the directions discussed in this chapter, drawn from books that have shaped how working analysts and researchers think about each topic.

For machine learning. *An Introduction to Statistical Learning* by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani is the best bridge between the statistics you have learned here and the world of machine learning. It is rigorous without being inaccessible, and it is available as a free PDF from the authors' website. The book covers classification, resampling methods, tree-based models, support vector machines, and unsupervised learning, all building on the regression foundation you developed in Chapters 10 and 11. What makes this book special is that it was written by statisticians for a broad audience, so the emphasis is on concepts and applications rather than mathematical proofs. It has become a standard textbook in data science.

For causal inference. Two outstanding books are available for free online. *Causal Inference: The Mixtape* by Scott Cunningham covers the methods I described earlier in this chapter, difference-in-differences, instrumental variables, regression discontinuity, and more, with clear explanations and code examples in both Stata and R. Cunningham writes with personality and wit, which makes the sometimes-dense material more approachable than you might expect. *The Effect: An Introduction to Research Design and Causality* by Nick Huntington-Klein takes a similar approach with a slightly different emphasis, excellent visualizations, and code in R, Python, and Stata. Huntington-Klein is particularly good at using diagrams and directed acyclic graphs (DAGs) to make causal reasoning visual and intuitive. Either one would be a natural next step from this book.

For Bayesian statistics. *Statistical Rethinking* by Richard McElreath is the book I would recommend starting with. McElreath writes with unusual clarity, and the accompanying lecture videos (freely available online) are some of the best statistics teaching I have encountered in any format. The book takes a fundamentally Bayesian approach to statistical modeling and will change how you think about many of the methods covered in this textbook. For a more comprehensive and advanced treatment, *Bayesian Data Analysis* by Andrew Gelman, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin is the standard reference. I would recommend tackling Gelman after you have taken at least one more statistics course, but if Bayesian thinking intrigues you, even reading the first few chapters of either book will deepen your understanding of how statistical reasoning works and why the frequentist-Bayesian debate matters.

For time series. *Forecasting: Principles and Practice* by Rob Hyndman and George Athanasopoulos is available free online and is, in my opinion, one of the best-written statistics textbooks in any subfield. Hyndman is one of the leading researchers in forecasting, and the book covers everything from simple exponential smoothing to modern methods like ARIMA and dynamic regression. The R code is integrated throughout, so you can work through the examples as you read. If the time series section of this chapter sparked your curiosity, this is the place to go.

For critical thinking about data. *Calling Bullshit: The Art of Skepticism in a Data-Driven World* by Carl Bergstrom and Jevin West is the book I wish everyone would read. It is not a statistics textbook. It is a book about how data, statistics, and quantitative arguments are used to mislead, sometimes intentionally and sometimes not, and how to defend yourself against it. The authors cover misleading visualizations, selection bias in big data, the misuse of algorithms, and the way that quantitative claims acquire unearned authority simply by being quantitative. It pairs beautifully with the skills you have developed in this book, and it is the rare book that will make you both smarter and more humble about how much you do not know.

12.9 Parting Words

We started this book in Flint, Michigan, where the water was poisoned and the people who said so were told they were wrong. Marc Edwards and Mona Hanna-Attisha did not uncover the Flint water crisis with fancy algorithms or advanced methods. They used the same basic tools you have spent this book learning. Sampling. Distributions. Comparisons between groups. Confidence in their data and the courage to present it clearly when the people in power did not want to hear it.

What made the difference in Flint was not the sophistication of the statistics. It was the fact that someone knew how to ask the right question, collect the right data, analyze it honestly, and stand

behind the results. Edwards looked at the distribution of lead levels across hundreds of homes and saw a pattern that could not be explained away. Hanna-Attisha compared blood lead levels before and after the water switch and found a difference that was too large and too consistent to be coincidence. These are not exotic techniques. They are the methods you learned in the first half of this book.

That is both the humbling and the hopeful message of this book. The tools are not complicated. What is rare is the willingness to use them honestly, the discipline to follow the evidence where it leads, and the willingness to stand behind findings even when they are inconvenient. Flint needed someone who could look at data and see what it contained. Your community, your workplace, your corner of the world, needs people like that too.

You now have the foundation to be one of them. Not because you have memorized every formula in this book (you will forget most of them, and that is fine, because you can always look them up). But because you understand what the formulas are for. You understand that data does not speak for itself. It speaks through the people who collect it, analyze it, and interpret it with care. You understand that uncertainty is not weakness but honesty, that a well-constructed confidence interval is more valuable than a false certainty, and that the gap between what the data says and what someone claims the data says is often where the most important work happens.

It has been a privilege to guide you through this material. Whatever direction you go from here, you carry with you a set of ideas that will help you see the world more clearly, ask better questions, and reach better-supported conclusions. That was always the point.

Thank you for reading.

12.10 References

Bergstrom, C. T., & West, J. D. (2020). *Calling bullshit: The art of skepticism in a data-driven world*. Random House.

Card, D., & Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review*, 84(4), 772-793.

Cunningham, S. (2021). *Causal inference: The mixtape*. Yale University Press.

Davey, S. J., Gordon, N., Holland, I. D., Rutten, M., & Williams, J. (2016). *Bayesian methods in the search for MH370*. Springer. (A detailed account of how Bayesian updating guided the physical search for Malaysia Airlines Flight 370, the application described in this chapter's section on Bayesian statistics.)

- Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., Baicker, K., & Oregon Health Study Group. (2012). The Oregon Health Insurance Experiment: Evidence from the first year. *Quarterly Journal of Economics*, 127(3), 1057-1106.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Chapman and Hall/CRC.
- Huntington-Klein, N. (2021). *The effect: An introduction to research design and causality*. Chapman and Hall/CRC.
- Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer.
- McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (2nd ed.). CRC Press.
- McGrayne, S. B. (2011). *The theory that would not die: How Bayes' rule cracked the Enigma code, hunted down Russian submarines, and emerged triumphant from two centuries of controversy*. Yale University Press. (A narrative history of Bayesian reasoning covering the Palomares hydrogen bomb search and the SS *Central America* discovery discussed in this chapter, accessible to a general audience.)

Chapter 13

Interactive Apps

Interactive Apps

A set of free, browser-based Shiny apps that accompany the book. Each one is referenced from a **Try It Online** callout in the chapter that introduces the underlying idea. No account or installation is required.

If an app’s link ever stops working, this is the canonical page to check — every chapter callout points here, so updates land in one place.

13.1 Chapter 1 — Why Statistics Matters Now

13.1.1 Variable Type Explorer

Classify variables from a hospital dataset as nominal, ordinal, discrete, or continuous. The “What Goes Wrong” tab shows what happens when you apply the wrong summary statistic; the “Numbers Trap” tab tests whether you can spot variables that contain digits but are not numerical.

Launch the Variable Type Explorer

13.2 Chapter 2 — Asking Good Questions: Research Design

13.2.1 Sampling Explorer

Compare simple random, stratified, cluster, and systematic sampling side by side. Draw repeated samples and watch how tightly the estimates cluster across methods. The same app supports the sample-size and margin-of-error sliders used in Chapter 7.

Launch the Sampling Explorer

13.2.2 Total Survey Error Explorer

A clickable taxonomy of every survey error type, with definitions, practitioner examples, and notes on which errors larger samples actually fix.

Launch the Total Survey Error Explorer

13.3 Chapter 3 — Summarizing Data with Numbers

13.3.1 Datasaurus Explorer

Twelve datasets with identical means, standard deviations, and correlations — yet wildly different shapes. The visual argument for why summaries alone are not enough.

Launch the Datasaurus Explorer

13.4 Chapter 4 — Summarizing Data with Pictures

13.4.1 Correlation Game

See a scatter plot, estimate r , then reveal the answer. Builds visual intuition for what different correlation strengths actually look like.

Launch the Correlation Game

The Datasaurus Explorer (Chapter 3) is also reused in Chapter 4 to make the case that pictures reveal what tables hide.

13.5 Chapter 5 — Probability

13.5.1 Probability Rules Sandbox

Set $P(A)$, $P(B)$, and $P(A \cap B)$ with sliders and watch every probability rule update live on a Venn diagram, a unit-square view, and a tree diagram. Force-toggles for independence and mutual exclusivity.

Launch the Probability Rules Sandbox

13.5.2 Base Rate Lab

The COVID-test base-rate problem made visceral. Drag the prevalence slider and watch the positive predictive value collapse even though the test never changes.

Launch the Base Rate Lab

13.5.3 Binomial Explorer

Set n and p with sliders and watch the binomial PMF and CDF update. Overlay the normal approximation and see when it succeeds and when it fails.

Launch the Binomial Explorer

13.6 Chapter 6 — Normal Distribution and the Central Limit Theorem

13.6.1 Normal Curve Explorer

Set μ and σ , pick a region (less than a , between a and b , within $\pm k$ standard deviations), and watch the shaded area and probability update live.

Launch the Normal Curve Explorer

13.6.2 Distribution Explorer

The Central Limit Theorem in action. Pick a skewed population, set the sample size, and watch the sampling distribution of the mean transform from skewed to normal as n grows.

Launch the Distribution Explorer

13.7 Chapter 7 — Confidence Intervals

13.7.1 Confidence Interval Simulator

Generate hundreds of confidence intervals from random samples and watch the capture rate. Switch between 90%, 95%, and 99% confidence and see the miss rate change.

Launch the Confidence Interval Simulator

The Sampling Explorer (Chapter 2) is reused here for sample-size and margin-of-error sliders.

13.8 Chapter 8 — Hypothesis Testing

13.8.1 Hypothesis Testing Playground

Set α , sample size, and effect size, then watch Type I error, Type II error, and power respond. Run repeated experiments under a true null and see how often you reject.

Launch the Hypothesis Testing Playground

13.8.2 P-Hacking Simulator

A dataset with no real relationships and a menu of analysis choices. See how quickly you can reach $p < 0.05$ — and what that implies for published research.

Launch the P-Hacking Simulator

13.9 Chapter 9 — Comparing Groups

13.9.1 ANOVA Visualizer

Adjust group means and within-group variability with sliders and watch the F -statistic and effect size update in real time. Set all means equal to see what happens to F .

Launch the ANOVA Visualizer

13.10 Chapters 10 & 11 — Simple and Multiple Regression

13.10.1 Regression Explorer

Add data points, fit the model, and check residuals. Drop in an outlier and watch the slope shift. Used in Chapter 10 for simple regression and again in Chapter 11 to add and remove predictors and see multicollinearity inflate the standard errors.

Launch the Regression Explorer

Chapter 14

Datasets

Datasets

All datasets used in *Margin of Error* are available for download as CSV files. See **Appendix B** of the book for complete variable descriptions, source documentation, and notes on simulation methodology.

14.1 Download

Dataset	Chapter	Rows	Description	Download
flint-water-lead.csv	1	271	Lead levels from the Virginia Tech Flint Water Study (2015)	Download
sampling-demo.csv	2	200	Simulated student population for sampling-method demos	Download
acs-household-income.csv	3	1,200	U.S. household income from the 2022 American Community Survey 1-Year PUMS	Download
housing-redlining.csv	4	551	HOLC redlining grades linked to modern demographics	Download
spurious-correlations.csv	4	10	Pairs of variables with high correlation but no causal link	Download

Dataset	Chapter	Rows	Description	Download
covid-testing.csv	5	2,000	Simulated COVID test results with true infection status	Download
birth-weights.csv	6	944	Birth weights and maternal characteristics (OpenIntro births14)	Download
polling-data.csv	7	50	Real U.S. election polls with margins of error (FiveThirtyEight)	Download
energy-reports.csv	8	200	Home energy report experiment (treatment vs. control)	Download
resume-callbacks.csv	9	4,870	Bertrand & Mullainathan (2004) résumé audit study	Download
education-spending.csv	10	50	State-level education spending and test scores	Download
wage-gap.csv	11	534	CPS 1985 wage data with gender, experience, education, and occupation	Download

14.2 Loading Data in R

```
library(tidyverse)

# Option 1: Download the CSV and read from your local working directory
flint <- read_csv("flint-water-lead.csv")

# Option 2: Read directly from the companion site
flint <- read_csv("https://stats.marginoferrormedia.com/datasets/flint-water-lead.csv")
```

14.3 Data Sources

The datasets are a mix of real public data and carefully simulated data calibrated to published research. Real datasets include the Virginia Tech Flint Water Study, FiveThirtyEight's redlining and polling collections, the Bertrand & Mullainathan résumé audit, OpenIntro's `births14`, and the CPS 1985 wage data from the AER package. Simulated datasets reproduce key statistical features from published studies while keeping individual-level data freely available for teaching.

See Appendix B for full source citations, variable codebooks, and simulation notes.

Chapter 15

Get Margin of Error

Get Margin of Error

The full text of *Margin of Error* is available free here on this site. If you would like a print copy or a Kindle edition, here is where to get one.

15.0.1 Print Edition

Paperback designed for the desk, the classroom, and the commute. 6" × 9", professionally typeset, with figures optimized for print.

\$39.99 on Amazon

[Buy Print Edition](#)

15.0.2 Kindle Edition

Read on any device. Searchable, highlightable, and always with you.

\$14.99 on Amazon

[Buy Kindle Edition](#)

15.1 What you get on this companion site

- Twelve full chapters and four appendices, free to read in your browser
- Twelve downloadable datasets — real and calibrated
- A set of interactive Shiny apps, referenced from chapter callouts

- A printable PDF and EPUB of the entire book

 For Instructors

Teaching with *Margin of Error*? Exam banks, sample syllabi, and other instructor materials are available on request. Contact patilv@gmail.com.

 Found an error?

Please report it on the errata tracker or email patilv@gmail.com. Corrections are merged into the next render.

Appendix A: Statistical Tables

These tables are provided as a reference. In practice, statistical software computes exact values. These tables remain useful for building intuition, checking work by hand, and exam settings where software is not available.

Standard Normal (Z) Table

The table gives the area to the LEFT of z , that is, $P(Z \leq z)$.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

t-Distribution Critical Values

The table gives the critical value t^* such that $P(T > t^*) = \alpha$ for a one-tailed test, or $P(|T| > t^*) = \alpha$ for a two-tailed test.

df	$\alpha = 0.10$ (two-tail)	$\alpha = 0.05$ (two-tail)	$\alpha = 0.02$ (two-tail)	$\alpha = 0.01$ (two-tail)
1	6.314	12.706	31.821	63.657
2	2.920	4.303	6.965	9.925
3	2.353	3.182	4.541	5.841
4	2.132	2.776	3.747	4.604
5	2.015	2.571	3.365	4.032
6	1.943	2.447	3.143	3.707

df	$\alpha = 0.10$ (two-tail)	$\alpha = 0.05$ (two-tail)	$\alpha = 0.02$ (two-tail)	$\alpha = 0.01$ (two-tail)
7	1.895	2.365	2.998	3.499
8	1.860	2.306	2.896	3.355
9	1.833	2.262	2.821	3.250
10	1.812	2.228	2.764	3.169
11	1.796	2.201	2.718	3.106
12	1.782	2.179	2.681	3.055
13	1.771	2.160	2.650	3.012
14	1.761	2.145	2.624	2.977
15	1.753	2.131	2.602	2.947
16	1.746	2.120	2.583	2.921
17	1.740	2.110	2.567	2.898
18	1.734	2.101	2.552	2.878
19	1.729	2.093	2.539	2.861
20	1.725	2.086	2.528	2.845
25	1.708	2.060	2.485	2.787
30	1.697	2.042	2.457	2.750
40	1.684	2.021	2.423	2.704
50	1.676	2.009	2.403	2.678
60	1.671	2.000	2.390	2.660
80	1.664	1.990	2.374	2.639
100	1.660	1.984	2.364	2.626
120	1.658	1.980	2.358	2.617
∞	1.645	1.960	2.326	2.576

Chi-Square Distribution Critical Values

The table gives the critical value χ^2 such that $P(\chi^2 > \text{value}) = \alpha$.

df	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.005$
1	2.706	3.841	5.024	6.635	7.879
2	4.605	5.991	7.378	9.210	10.597
3	6.251	7.815	9.348	11.345	12.838

df	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.005$
4	7.779	9.488	11.143	13.277	14.860
5	9.236	11.070	12.833	15.086	16.750
6	10.645	12.592	14.449	16.812	18.548
7	12.017	14.067	16.013	18.475	20.278
8	13.362	15.507	17.535	20.090	21.955
9	14.684	16.919	19.023	21.666	23.589
10	15.987	18.307	20.483	23.209	25.188
11	17.275	19.675	21.920	24.725	26.757
12	18.549	21.026	23.337	26.217	28.300
13	19.812	22.362	24.736	27.688	29.819
14	21.064	23.685	26.119	29.141	31.319
15	22.307	24.996	27.488	30.578	32.801
16	23.542	26.296	28.845	32.000	34.267
17	24.769	27.587	30.191	33.409	35.718
18	25.989	28.869	31.526	34.805	37.156
19	27.204	30.144	32.852	36.191	38.582
20	28.412	31.410	34.170	37.566	39.997
25	34.382	37.652	40.646	44.314	46.928
30	40.256	43.773	46.979	50.892	53.672

F-Distribution Critical Values ($\alpha = 0.05$)

The table gives the critical value F such that $P(F > \text{value}) = 0.05$.

Rows are denominator df. Columns are numerator df.

df2	df1	1	2	3	4	5	6	8	10	20
5		6.608	5.786	5.409	5.192	5.050	4.950	4.818	4.735	4.558
6		5.987	5.143	4.757	4.534	4.387	4.284	4.147	4.060	3.874
7		5.591	4.737	4.347	4.120	3.972	3.866	3.726	3.637	3.445
8		5.318	4.459	4.066	3.838	3.687	3.581	3.438	3.347	3.150
9		5.117	4.256	3.863	3.633	3.482	3.374	3.230	3.137	2.936
10		4.965	4.103	3.708	3.478	3.326	3.217	3.072	2.978	2.774

df2	df1	1	2	3	4	5	6	8	10	20
12		4.747	3.885	3.490	3.259	3.106	2.996	2.849	2.753	2.544
15		4.543	3.682	3.287	3.056	2.901	2.790	2.641	2.544	2.328
20		4.351	3.493	3.098	2.866	2.711	2.599	2.447	2.348	2.124
25		4.242	3.385	2.991	2.759	2.603	2.490	2.337	2.236	2.007
30		4.171	3.316	2.922	2.690	2.534	2.421	2.266	2.165	1.932
40		4.085	3.232	2.839	2.606	2.449	2.336	2.180	2.077	1.839
60		4.001	3.150	2.758	2.525	2.368	2.254	2.097	1.993	1.748
120		3.920	3.072	2.680	2.447	2.290	2.175	2.016	1.910	1.659
∞		3.841	2.996	2.605	2.372	2.214	2.099	1.938	1.831	1.571

For other significance levels or degrees of freedom not listed here, use statistical software (R, Python, or online calculators).

Appendix B: Dataset Descriptions

All datasets used in the book are available for download from the companion website. Each is provided as a CSV file. Where publicly available, we use real data from the original studies or authoritative government sources. Where real microdata is unavailable or the study design requires it, we use simulated data calibrated to published parameters. Each dataset’s status is clearly marked below.

Chapter 1: Why Statistics Matters Now

flint-water-lead.csv

Status: Real data **Source:** Virginia Tech Flint Water Study. Data collected by Marc Edwards’s research team during the Flint water crisis and made publicly available through the study’s GitHub repository. **Rows:** 271 | **Variables:** 5

Variable	Type	Description
sample_id	Nominal	Unique identifier for each sample
zip_code	Nominal	ZIP code of the home
ward	Nominal	Ward (geographic district) within Flint (0–9)
lead_ppb	Continuous	Lead concentration in parts per billion
notes	Nominal	Flags for special cases (e.g., homes sampled twice)

No missing values. The dataset contains first-draw water samples from 271 Flint homes across 10 wards. Lead levels range from 0.34 to 158 ppb, with a median of 3.5 ppb and a mean of 10.65

ppb. Approximately 16.6% of homes exceeded the EPA action level of 15 ppb, consistent with the crisis-level contamination documented by the Virginia Tech team.

Chapter 2: Research Design

sampling-demo.csv

Status: Simulated (seed = 202) **Based on:** Fictional small university with realistic demographic and academic distributions. Designed for practicing sampling methods (simple random, stratified, cluster, systematic). **Rows:** 200 | **Variables:** 7

Variable	Type	Description
student_id	Nominal	Unique identifier (S001–S200)
major	Nominal	Academic major (Business/Engineering/Liberal Arts/Nursing/Sciences)
gpa	Continuous	Grade point average (0–4 scale)
year	Ordinal	Class year (Freshman/Sophomore/Junior/Senior)
commute_distance_miles	Continuous	Distance of commute in miles (right-skewed)
housing	Nominal	Housing type (On-campus/Off-campus)
works_part_time	Nominal	Whether the student works part-time (Yes/No)

Why simulated: The scenario is fictional by design. No specific real university study is being replicated; the dataset is constructed to provide a realistic population for demonstrating sampling techniques.

Chapter 3: Summarizing Data with Numbers

acs-household-income.csv

Status: Real data **Source:** U.S. Census Bureau, 2022 American Community Survey 1-Year Public Use Microdata Sample. Sampled to represent the national distribution of income and educational attainment across all 50 states. **Rows:** 1,200 | **Variables:** 6

Variable	Type	Description
household_id	Nominal	Unique identifier
state	Nominal	U.S. state
region	Nominal	Geographic region (Northeast/Southeast/Midwest/Southwest/West)
household_income	Continuous	Annual household income in 2022 dollars
education_level	Ordinal	Highest education level attained by head of household (High School or Less, Some College or Associate, Bachelor, Master or Professional, Doctorate)
household_size	Discrete	Number of people in the household

The dataset contains approximately 1,200 household records drawn from the American Community Survey, representing all 50 states. The income distribution exhibits the characteristic right skew of U.S. household income, with education-income gradients and regional variation consistent with published Census summaries.

Chapter 4: Summarizing Data with Pictures

housing-redlining.csv

Status: Real data **Source:** FiveThirtyEight analysis of Home Owners' Loan Corporation (HOLC) neighborhood grades linked to modern U.S. Census demographic data. Originally published as part of FiveThirtyEight's reporting on the lasting effects of redlining. **Rows:** 551 | **Variables:** 9

Variable	Type	Description
neighborhood_id	Nominal	Unique identifier
metro_area	Nominal	Metropolitan area name
holc_grade	Ordinal	Historical HOLC grade (A = “Best”, B = “Still Desirable”, C = “Definitely Declining”, D = “Hazardous”)
total_population	Discrete	Total population of the neighborhood/tract
pct_white	Continuous	Percentage of white residents
pct_black	Continuous	Percentage of Black residents
pct_hispanic	Continuous	Percentage of Hispanic residents
pct_asian	Continuous	Percentage of Asian residents
pct_minority	Continuous	Percentage of minority (non-white) residents

The data covers 551 neighborhoods across 138 metropolitan areas. The clear demographic gradient by HOLC grade—neighborhoods graded “A” average 73.8% white, while those graded “D” average only 39.4% white—demonstrates the persistent legacy of 1930s redlining on neighborhood racial composition.

spurious-correlations.csv

Status: Mixed (two pairs from verified public data, one pair simulated) **Inspired by:** Tyler Vigen’s Spurious Correlations project (tylervigen.com). **Rows:** 10 | **Variables:** 7

Variable	Type	Description
year	Discrete	Year (2000–2009)
nicolas_cage_films	Discrete	Number of Nicolas Cage film appearances
pool_drownings	Discrete	Number of pool drownings in the U.S.
cheese_consumption_lbs	Continuous	Per capita cheese consumption (lbs)
bedsheet_deaths	Discrete	Deaths by becoming tangled in bedsheets
margarine_consumption_lbs	Continuous	Per capita margarine consumption (lbs)

Variable	Type	Description
maine_divorce_rate	Continuous	Maine divorce rate per 1,000 population

Three pairs of variables with high correlations but no causal relationship: Nicolas Cage films vs. pool drownings ($r = 0.66$), cheese consumption vs. bedsheet deaths ($r = 0.95$), and margarine consumption vs. Maine divorce rate ($r = 0.99$). The cheese/bedsheet and margarine/divorce pairs use verified data from USDA and CDC sources. The Cage/drownings pair is simulated to approximate the correlation reported by Vigen, as the original source data is no longer publicly accessible.

Chapter 5: Probability

covid-testing.csv

Status: Simulated (seed = 50003) **Based on:** Published sensitivity and specificity estimates for rapid antigen and PCR tests during the COVID-19 pandemic. **Rows:** 2,000 | **Variables:** 5

Variable	Type	Description
patient_id	Nominal	Unique identifier
truly_infected	Nominal	True infection status (Yes/No, ~5% prevalence)
test_result	Nominal	Test result (Positive/Negative)
age_group	Ordinal	Age group category (18–29/30–44/45–59/60–74/75+)
risk_category	Nominal	Risk category (Low/Medium/High)

Why simulated: Individual-level COVID testing data with verified true infection status is not publicly available. The dataset is generated with sensitivity of approximately 91% and specificity of approximately 95%, consistent with published performance estimates for rapid antigen tests. The 5% prevalence reflects a moderate community transmission scenario.

Chapter 6: The Normal Distribution and CLT

birth-weights.csv

Status: Real data **Source:** OpenIntro births14 dataset, a sample from the 2014 CDC National Vital Statistics natality data. **Rows:** 944 | **Variables:** 6

Variable	Type	Description
baby_id	Nominal	Unique identifier
birth_weight_g	Continuous	Birth weight in grams (mean = 3,266, SD = 593)
gestational_weeks	Continuous	Gestational age at birth in weeks
mother_age	Continuous	Mother's age in years
prenatal_visits	Discrete	Number of prenatal care visits
smoking_status	Nominal	Whether the mother smoked during pregnancy (Yes/No)

The dataset includes both full-term and preterm births, which is why the standard deviation (593g) is larger than the approximately 450g typically reported for full-term-only samples. The slight left skew (skewness = -1.0) reflects the inclusion of preterm, low-birth-weight infants. Among full-term births only (gestational weeks ≥ 37 , $n = 829$), the distribution is more symmetric with mean = 3,373g and SD = 464g.

Chapter 7: Confidence Intervals

polling-data.csv

Status: Real data **Source:** FiveThirtyEight pollster ratings dataset (raw_polls). Polls are from real U.S. election polling conducted by established polling organizations. **Rows:** 50 | **Variables:** 8

Variable	Type	Description
poll_id	Nominal	Unique identifier

Variable	Type	Description
pollster	Nominal	Polling organization name (31 unique pollsters)
date	Date	Date the poll was conducted
sample_size	Discrete	Number of respondents (range: 200–22,170)
candidate_a_pct	Continuous	Percentage supporting Candidate A
candidate_b_pct	Continuous	Percentage supporting Candidate B
margin_of_error	Continuous	Reported margin of error (percentage points)
confidence_level	Continuous	Confidence level used (all 95%)

The margins of error in this dataset are computed from the actual sample sizes using the standard formula for a 95% confidence interval for a proportion, allowing students to verify the relationship between sample size and margin of error.

Chapter 8: Hypothesis Testing

energy-reports.csv

Status: Simulated (seed = 808) **Based on:** Allcott, H. (2011). “Social Norms and Energy Conservation.” *Journal of Public Economics*, 95(9–10), 1082–1095. **Rows:** 200 | **Variables:** 7

Variable	Type	Description
household_id	Nominal	Unique identifier
group	Nominal	Experimental group (Treatment/Control, 100 each)
energy_kwh_before	Continuous	Monthly energy usage before intervention (kWh)
energy_kwh_after	Continuous	Monthly energy usage after intervention (kWh)
home_size_sqft	Continuous	Home size in square feet
num_occupants	Discrete	Number of household occupants

Variable	Type	Description
region	Nominal	Geographic region (Northeast/Southeast/Midwest/Southwest/West)

Why simulated: The Allcott (2011) study involved over 600,000 households across multiple utility companies, and the individual-level data is not publicly available. This simulated dataset reproduces the study’s key finding: treatment households (who received home energy reports comparing their usage to neighbors) reduced consumption by approximately 1.6%, while control households showed a slight increase. The treatment effect is statistically significant ($p < 0.001$).

Chapter 9: Comparing Groups

resume-callbacks.csv

Status: Real data **Source:** Bertrand, M. & Mullainathan, S. (2004). “Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination.” *American Economic Review*, 94(4), 991–1013. Data accessed via the OpenIntro R package. **Rows:** 4,870 | **Variables:** 7

Variable	Type	Description
resume_id	Nominal	Unique identifier
name_type	Nominal	Name category (White-sounding/Black-sounding, 2,435 each)
gender	Nominal	Gender (Male/Female)
resume_quality	Ordinal	Résumé quality level (Low/High)
years_experience	Discrete	Years of work experience
education	Nominal	Education level (Bachelor/High School)
callback	Nominal	Whether the applicant received a callback (Yes/No)

This is the complete dataset from the Bertrand and Mullainathan field experiment. The researchers sent 4,870 fictitious résumés to real job postings in Boston and Chicago, randomly assigning either

White-sounding or Black-sounding names. The callback rate for White-sounding names (9.65%) was approximately 50% higher than for Black-sounding names (6.45%), providing direct experimental evidence of racial discrimination in hiring.

Chapter 10: Simple Linear Regression

education-spending.csv

Status: Simulated, calibrated to patterns from real sources. **Sources of calibration:** National Center for Education Statistics (NCES) Digest of Education Statistics (spending, test scores, student-teacher ratios, free lunch eligibility) and U.S. Census Bureau Small Area Income and Poverty Estimates (SAIPE) for median household income and poverty rates. The simulated rows preserve the orders of magnitude and approximate correlations of these real sources but do not reproduce any individual state's reported figures. **Rows:** 50 | **Variables:** 7

Variable	Type	Description
state	Nominal	U.S. state (all 50 states)
spending_per_student	Continuous	Annual spending per student in constant dollars
avg_score	Continuous	Average NAEP (National Assessment of Educational Progress) mathematics score
median_household_income	Continuous	State median household income
pct_poverty	Continuous	State poverty rate (percentage)
student_teacher_ratio	Continuous	Students per teacher
pct_free_lunch	Continuous	Percentage of students eligible for free/reduced-price lunch

Each row represents one U.S. state. The correlation between per-student spending and average NAEP math score is $r = 0.43$, indicating a moderate positive association. Additional variables allow students to explore confounders and build toward multiple regression concepts in Chapter 11.

Chapter 11: Multiple Regression

wage-gap.csv

Status: Real data **Source:** Current Population Survey (CPS) May 1985, distributed in the AER (Applied Econometrics with R) package as CPS1985, originally from Berndt’s *The Practice of Econometrics* (1991). This is a widely used teaching dataset in labor economics. **Rows:** 534 | **Variables:** 10

Variable	Type	Description
employee_id	Nominal	Unique identifier
salary	Continuous	Full-time-equivalent annual salary in 1985 dollars (computed as hourly wage $\times$ 2,000 hours/year)
gender	Nominal	Gender (Male/Female) — recorded as binary in the original CPS instrument
years_experience	Continuous	Years of work experience
education_years	Continuous	Years of education
occupation	Nominal	Occupation category (Worker/Technical/Services/Office/Sales/Management)
region	Nominal	Geographic region (South/Other)
union_member	Nominal	Union membership status (Yes/No)
married	Nominal	Marital status (Yes/No)
ethnicity	Nominal	Ethnicity (cauc / hispanic / other, as coded in the source)

Note on construction: The CPS1985 source records wage as an hourly rate. To support the chapter’s narrative in dollar-per-year terms (so coefficients read as “dollars more per year”), the CSV reports $\text{salary} = \text{wage} \times 2,000$, the standard 50-week $\times$ 40-hour full-time benchmark. All chapter coefficients reconcile exactly with this transformation: the raw Female coefficient is approximately $-\$4,233/\text{year}$ ($R^2 \approx 0.042$), narrowing to $-\$3,930/\text{year}$ after controlling for experience, education, and occupation ($R^2 \approx 0.302$). The progression — from a simple bivariate regression to a multiple regression that accounts for confounders — is the pedagogical core of the chapter, illustrating both controlling-for logic and the limits of “controlling away” discrimination.

Appendix C: Glossary of Key Terms

Adjusted R-squared. A modified version of R-squared that adjusts for the number of predictors in a regression model, penalizing unnecessary complexity. Unlike R-squared, adjusted R-squared can decrease when a new predictor does not improve the model enough to justify the added complexity.

Alternative hypothesis (H_a or H_1). The claim that there is an effect, a difference, or a relationship. The hypothesis the researcher typically hopes to find evidence for.

ANOVA (Analysis of Variance). A statistical method for comparing means across three or more groups using the F-statistic. Tests whether at least one group mean differs from the others.

Bar chart. A visualization for categorical data where the height (or length) of each bar represents the count or proportion in each category.

Bayes' theorem. A formula for updating the probability of an event based on new evidence. Relates conditional probabilities: $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$.

Bias. Systematic error in sampling or measurement that causes results to deviate from the truth in a consistent direction.

Binomial distribution. The probability distribution for the number of successes in a fixed number of independent trials, each with the same probability of success.

Bonferroni correction. A method for adjusting significance levels when performing multiple hypothesis tests simultaneously. The corrected significance level is alpha divided by the number of tests, reducing the risk of Type I errors from multiple comparisons.

Box plot. A visualization showing the five-number summary (minimum, Q1, median, Q3, maximum) of a distribution, with outliers plotted individually.

Categorical variable. A variable whose values represent group membership rather than measured quantities. Includes nominal and ordinal types.

Central Limit Theorem (CLT). The result that the sampling distribution of the sample mean approaches a normal distribution as the sample size increases, regardless of the shape of the population distribution.

Chi-square test. A hypothesis test for categorical data. The goodness-of-fit test compares observed frequencies to expected frequencies. The test of independence tests whether two categorical variables are associated.

Coefficient of determination (R^2). The proportion of variability in the response variable that is explained by the regression model. Ranges from 0 to 1.

Cohen's d. An effect size measure for the difference between two group means, expressed in standard deviation units. Small: 0.2, Medium: 0.5, Large: 0.8.

Conditional probability. The probability of an event occurring given that another event has already occurred. Written $P(A|B)$.

Confidence interval. A range of plausible values for a population parameter, constructed from a sample. A 95% confidence interval means that 95% of intervals constructed this way would contain the true parameter.

Confidence level. The long-run proportion of confidence intervals that capture the true population parameter when the sampling and construction process is repeated. Common levels are 90%, 95%, and 99%.

Confounding variable. A variable that influences both the explanatory variable and the response variable, creating a misleading association between them.

Continuous variable. A numerical variable that can take any value within a range, including decimals.

Correlation coefficient (Pearson's r). A measure of the strength and direction of the linear relationship between two numerical variables. Ranges from -1 to +1.

Critical value. The threshold value of a test statistic that determines the boundary of the rejection region.

Cramér's V . A measure of the strength of association between two categorical variables, based on the chi-square statistic. It ranges from 0 (no association) to 1 (perfect association) and is appropriate for tables larger than 2x2.

Degrees of freedom (df). A parameter that depends on the sample size and determines the shape of certain distributions (t, chi-square, F). Generally related to the number of independent pieces of information in the data.

Density plot. A smoothed version of a histogram that estimates the probability density function of a continuous variable.

Dependent variable. See response variable.

Discrete variable. A numerical variable that takes on countable values, usually integers.

Dummy variable. A binary (0/1) variable used to represent a category in regression. A categorical variable with k categories requires $k - 1$ dummy variables.

Effect size. A quantitative measure of the magnitude of a phenomenon. Unlike p-values, effect sizes are not influenced by sample size.

Empirical Rule (68-95-99.7 rule). For a normal distribution, approximately 68% of observations fall within 1 standard deviation of the mean, 95% within 2, and 99.7% within 3.

Eta-squared (η^2). An effect size measure for ANOVA, representing the proportion of total variance explained by the grouping variable.

Expected value. The long-run average of a random variable. For a discrete random variable, $E(X) = \sum x_i \cdot P(x_i)$.

Experiment. A study in which the researcher deliberately assigns treatments to subjects and compares outcomes.

Explanatory variable. A variable used to explain or predict changes in the response variable. Also called independent variable or predictor.

Extrapolation. Using a regression model to predict outcomes for values of the explanatory variable outside the range of the data used to build the model. Generally unreliable.

F-statistic. The test statistic for ANOVA, computed as the ratio of between-group variance to within-group variance.

F-test. A statistical test that compares the variability between group means to the variability within groups. Used in ANOVA to determine whether at least one group mean differs significantly from the others.

Five-number summary. The minimum, first quartile (Q1), median, third quartile (Q3), and maximum of a dataset.

Frequency. The count of observations falling into a particular category or interval.

Histogram. A visualization for the distribution of a numerical variable, where bars represent the frequency of observations in each interval (bin).

Hypothesis test. A formal procedure for using sample data to decide between two competing claims (hypotheses) about a population parameter.

Independence. Two events are independent if the occurrence of one does not affect the probability of the other. $P(A \cap B) = P(A) \cdot P(B)$.

Independent variable. See explanatory variable.

Interaction effect. A condition in regression where the effect of one predictor on the response depends on the value of another predictor. Modeled by including the product of the two predictors as an additional term.

Intercept (b_0). In regression, the predicted value of the response variable when all explanatory variables equal zero.

Interquartile range (IQR). The difference between the third quartile and the first quartile: $IQR = Q3 - Q1$. A resistant measure of spread.

Kruskal-Wallis test. A nonparametric alternative to one-way ANOVA that compares three or more groups using ranks rather than raw values. Does not assume normality.

Least squares. The method for fitting a regression line by minimizing the sum of the squared residuals. Also called ordinary least squares (OLS) when the assumptions of standard linear regression are met.

Lurking variable. See confounding variable.

Margin of error. The maximum expected difference between a sample statistic and the population parameter, given a specified confidence level. Equal to the critical value multiplied by the standard error.

Mean. The arithmetic average of a set of values. The sum divided by the count.

Median. The middle value when observations are arranged in order. Resistant to outliers.

Mediator variable. A variable that lies on the causal pathway between an independent variable and a dependent variable, helping to explain how or why the effect occurs. Contrasted with a confounding variable, which creates a spurious association.

Missing data. Values absent from a dataset. The pattern and mechanism of missingness affect how (and whether) the data can be analyzed validly.

Mode. The most frequently occurring value in a dataset.

Multicollinearity. A condition in multiple regression where two or more predictor variables are highly correlated, making it difficult to isolate the individual effect of each predictor.

Multiple regression. A regression model with two or more explanatory variables.

Nominal variable. A categorical variable with no natural ordering among categories.

Nonparametric test. A hypothesis test that does not assume the data come from a specific distribution. Rank-based tests (Wilcoxon signed-rank, Wilcoxon rank-sum / Mann-Whitney U, Kruskal-Wallis) are common nonparametric alternatives to t-tests and ANOVA.

Normal distribution. A symmetric, bell-shaped distribution completely described by its mean (μ) and standard deviation (σ).

Null hypothesis (H_0). The claim of no effect, no difference, or no relationship. The default assumption that the hypothesis test seeks to find evidence against.

Numerical variable. A variable whose values represent measured quantities where arithmetic is meaningful.

Observational study. A study in which researchers observe subjects without intervening or assigning treatments.

Omitted variable bias. Bias in regression coefficients that occurs when an important confounding variable is not included in the model.

Ordinal variable. A categorical variable with a natural ordering, but where the distances between categories are not necessarily equal.

Ordinary Least Squares (OLS). The standard method for fitting a linear regression by minimizing the sum of squared residuals. Identical in spirit to “least squares” as introduced in Chapter 10 and extended to multiple predictors in Chapter 11.

Outlier. An observation that falls far from the rest of the data. The IQR method identifies observations below $Q1 - 1.5 \cdot IQR$ or above $Q3 + 1.5 \cdot IQR$ as outliers.

P-hacking. The practice of manipulating data analysis (choosing variables, subsets, or tests) until a statistically significant result is obtained, without accounting for the multiple tests performed.

P-value. The probability of obtaining a test statistic as extreme as, or more extreme than, the one observed, assuming the null hypothesis is true. It is NOT the probability that the null hypothesis is true.

Paired data. Data where each observation in one group is naturally linked to a specific observation in another group (e.g., before-and-after measurements on the same subjects).

Percentile. The value below which a given percentage of observations fall. The 75th percentile means 75% of values are at or below that point.

Point estimate. A single value calculated from a sample and used to estimate a population parameter (e.g., the sample mean $\bar{x}$ estimates the population mean μ).

Population. The entire group of interest about which conclusions are to be drawn.

Post-hoc test. A follow-up analysis conducted after a significant ANOVA result to determine which specific group means differ from each other. Common methods include Tukey's HSD and Bonferroni-corrected pairwise comparisons.

Power. The probability of correctly rejecting the null hypothesis when it is false. Equal to $1 - \beta$, where β is the probability of a Type II error.

Prediction interval. An interval estimate for an individual future observation, wider than a confidence interval for the mean because it accounts for individual variability.

Proportion. The fraction of observations falling into a particular category. Ranges from 0 to 1.

QQ plot (quantile-quantile plot). A graphical tool for assessing whether data follow a particular distribution (often normal). Points falling along the reference line suggest the assumed distribution is appropriate.

Quartiles. Values that divide the data into four equal parts. Q1 (25th percentile), Q2 (median, 50th percentile), Q3 (75th percentile).

Random variable. A variable whose value is determined by the outcome of a random process.

Randomized controlled experiment. An experiment where subjects are randomly assigned to treatment and control groups.

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

Reference category. In a regression model with categorical predictors, the category coded as zero on all dummy variables. Coefficients on the dummy variables represent comparisons against this baseline group.

Regression line. The line that best fits the relationship between an explanatory and response variable, determined by least squares.

Replication crisis. The finding that many published scientific results fail to reproduce when the original studies are repeated.

Residual. The difference between an observed value and the value predicted by a model. $e_i = y_i - \hat{y}_i$.

Response variable. The variable being predicted or explained in a study. Also called dependent variable or outcome variable.

Sample. A subset of a population from which data is actually collected.

Sampling distribution. The probability distribution of a sample statistic (such as the sample mean) computed from repeated random samples of the same size from the same population.

Scatter plot. A visualization for the relationship between two numerical variables, where each point represents one observation.

Selection bias. Systematic error arising from how subjects are selected for inclusion in a study or sample.

Shapiro-Wilk test. A formal hypothesis test for normality. A small p-value suggests the data deviate from a normal distribution.

Simple linear regression. A regression model with exactly one explanatory variable.

Simpson's Paradox. A phenomenon where a trend that appears in aggregated data reverses or disappears when the data is separated into subgroups.

Slope (b_1). In simple regression, the predicted change in the response variable for a one-unit increase in the explanatory variable.

Standard deviation. A measure of the spread of data around the mean. The square root of the variance. For a sample: $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$.

Standard error. The standard deviation of a sampling distribution. Measures the variability of a sample statistic from sample to sample. For the sample mean: $SE = \frac{s}{\sqrt{n}}$.

Statistic. A numerical summary computed from a sample (e.g., sample mean, sample proportion).

Statistical significance. A result is statistically significant at level α if the p-value is less than α . Does not necessarily imply practical importance.

Stratified sampling. A sampling method where the population is divided into strata (subgroups) and random samples are drawn from each stratum.

Survivorship bias. A form of selection bias that occurs when analysis is limited to subjects that survived a selection process, ignoring those that did not.

T-distribution. A symmetric, bell-shaped distribution similar to the normal but with heavier tails. Used for inference about means when the population standard deviation is unknown.

Test statistic. A numerical value calculated from sample data that is used to decide whether to reject the null hypothesis.

Tukey HSD (Honestly Significant Difference). A post-hoc test used after ANOVA to identify which specific group pairs differ, while controlling the family-wise error rate.

Type I error. Rejecting the null hypothesis when it is actually true (a false positive). The probability is denoted α .

Type II error. Failing to reject the null hypothesis when it is actually false (a false negative). The probability is denoted β .

Variance. A measure of spread computed as the average of the squared deviations from the mean. For a sample: $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$.

Variance Inflation Factor (VIF). A diagnostic for multicollinearity in regression. VIF values above 5 or 10 suggest problematic collinearity.

Welch's t-test. A modification of the independent-samples t-test that does not assume equal variances in the two groups. The recommended default for two-sample comparisons.

Wilcoxon rank-sum test (Mann-Whitney U test). A rank-based alternative to the independent two-sample t-test, evaluating whether two groups tend to produce systematically larger or smaller values.

Wilcoxon signed-rank test. A rank-based alternative to the one-sample or paired t-test, evaluating whether the median (or median within-pair difference) differs from a hypothesized value.

Z-score. A standardized value indicating how many standard deviations an observation is above or below the mean: $z = \frac{x - \mu}{\sigma}$.

Appendix D: Guide to the Companion Resources

The full text of *Margin of Error* is freely available online at stats.marginoferrormedia.com, where it is hosted alongside everything in this appendix. The companion site is the book — searchable, with figures rendered crisp on a screen, and updated whenever errata or new exercises are merged. The print and Kindle editions are produced from the same source.

Interactive Applications

A set of free, browser-based Shiny apps lets you explore the ideas in the book hands-on. They run in any modern browser, with no account or installation required, and each one is referenced from a “Try It Online” callout in the chapter that introduces the underlying idea. The complete list, organized by chapter, lives at stats.marginoferrormedia.com/apps.html.

App	Chapters
Variable Type Explorer	1
Sampling Explorer	2, 7
Total Survey Error Explorer	2
Datasaurus Explorer	3, 4
Correlation Game	4
Probability Rules Sandbox	5
Base Rate Lab	5
Binomial Explorer	5
Normal Curve Explorer	6
Distribution Explorer (Sampling Distributions / CLT)	6

App	Chapters
Confidence Interval Simulator	7
Hypothesis Testing Playground	8
P-Hacking Simulator	8
ANOVA Visualizer	9
Regression Explorer	10, 11

Datasets

All twelve datasets used in the book are available for download as CSV files. The download page is at stats.marginofferrormedia.com/datasets.html; see Appendix B for complete variable descriptions, sources, and notes on construction.

Errata and Feedback

If you find an error, have a suggestion, or want to flag something that could be explained better, open an issue at github.com/patilv/margin-of-error-feedback or email the author at patilv@gmail.com.

Instructor Resources

In development. Resources will include a solutions manual covering all three exercise tiers across the chapters with exercises, sample 15-week and 8-week syllabi, and exam templates for midterm (Chapters 1–8) and final (Chapters 1–12). Contact patilv@gmail.com for the current state and access.