

OpenIntro Statistics

Fourth Edition

David Diez

Data Scientist

OpenIntro

Mine Çetinkaya-Rundel

Associate Professor of the Practice, Duke University

Professional Educator, RStudio

Christopher D Barr

Investment Analyst

Varadero Capital

Editions 1, 2, and 3 can be found in the book's extra files,
which also include tablet-friendly versions of the latest edition.

Copyright © 2019. Fourth Edition.
Updated: April 12th, 2022.

This book may be downloaded as a free PDF at openintro.org/os. This textbook is also available under a Creative Commons license, with the source files hosted on Github.

Table of Contents

1	Introduction to data	7
1.1	Case study: using stents to prevent strokes	9
1.2	Data basics	12
1.3	Sampling principles and strategies	22
1.4	Experiments	32
2	Summarizing data	39
2.1	Examining numerical data	41
2.2	Considering categorical data	61
2.3	Case study: malaria vaccine	71
3	Probability	79
3.1	Defining probability	81
3.2	Conditional probability	95
3.3	Sampling from a small population	112
3.4	Random variables	115
3.5	Continuous distributions	125
4	Distributions of random variables	131
4.1	Normal distribution	133
4.2	Geometric distribution	144
4.3	Binomial distribution	149
4.4	Negative binomial distribution	158
4.5	Poisson distribution	163
5	Foundations for inference	168
5.1	Point estimates and sampling variability	170
5.2	Confidence intervals for a proportion	181
5.3	Hypothesis testing for a proportion	189
6	Inference for categorical data	206
6.1	Inference for a single proportion	208
6.2	Difference of two proportions	217
6.3	Testing for goodness of fit using chi-square	229
6.4	Testing for independence in two-way tables	240
7	Inference for numerical data	249
7.1	One-sample means with the t -distribution	251
7.2	Paired data	262
7.3	Difference of two means	267
7.4	Power calculations for a difference of means	278
7.5	Comparing many means with ANOVA	285

8	Introduction to linear regression	303
8.1	Fitting a line, residuals, and correlation	305
8.2	Least squares regression	317
8.3	Types of outliers in linear regression	328
8.4	Inference for linear regression	331
9	Multiple and logistic regression	341
9.1	Introduction to multiple regression	343
9.2	Model selection	353
9.3	Checking model conditions using graphs	358
9.4	Multiple regression case study: Mario Kart	365
9.5	Introduction to logistic regression	371
A	Exercise solutions	384
B	Data sets within the text	403
C	Distribution tables	408

Preface

OpenIntro Statistics covers a first course in statistics, providing a rigorous introduction to applied statistics that is clear, concise, and accessible. This book was written with the undergraduate level in mind, but it's also popular in high schools and graduate courses.

We hope readers will take away three ideas from this book in addition to forming a foundation of statistical thinking and methods.

- Statistics is an applied field with a wide range of practical applications.
- You don't have to be a math guru to learn from real, interesting data.
- Data are messy, and statistical tools are imperfect. But, when you understand the strengths and weaknesses of these tools, you can use them to learn about the world.

Textbook overview

The chapters of this book are as follows:

- 1. Introduction to data.** Data structures, variables, and basic data collection techniques.
- 2. Summarizing data.** Data summaries, graphics, and a teaser of inference using randomization.
- 3. Probability.** Basic principles of probability.
- 4. Distributions of random variables.** The normal model and other key distributions.
- 5. Foundations for inference.** General ideas for statistical inference in the context of estimating the population proportion.
- 6. Inference for categorical data.** Inference for proportions and tables using the normal and chi-square distributions.
- 7. Inference for numerical data.** Inference for one or two sample means using the t -distribution, statistical power for comparing two groups, and also comparisons of many means using ANOVA.
- 8. Introduction to linear regression.** Regression for a numerical outcome with one predictor variable. Most of this chapter could be covered after Chapter 1.
- 9. Multiple and logistic regression.** Regression for numerical and categorical data using many predictors.

OpenIntro Statistics supports flexibility in choosing and ordering topics. If the main goal is to reach multiple regression (Chapter 9) as quickly as possible, then the following are the ideal prerequisites:

- Chapter 1, Sections 2.1, and Section 2.2 for a solid introduction to data structures and statistical summaries that are used throughout the book.
- Section 4.1 for a solid understanding of the normal distribution.
- Chapter 5 to establish the core set of inference tools.
- Section 7.1 to give a foundation for the t -distribution
- Chapter 8 for establishing ideas and principles for single predictor regression.

Examples and exercises

Examples are provided to establish an understanding of how to apply methods

EXAMPLE 0.1

This is an example. When a question is asked here, where can the answer be found?

The answer can be found here, in the solution section of the example!

When we think the reader should be ready to try determining the solution to an example, we frame it as Guided Practice.

GUIDED PRACTICE 0.2

The reader may check or learn the answer to any Guided Practice problem by reviewing the full solution in a footnote.¹

Exercises are also provided at the end of each section as well as review exercises at the end of each chapter. Solutions are given for odd-numbered exercises in Appendix A.

Additional resources

Video overviews, slides, statistical software labs, data sets used in the textbook, and much more are readily available at

openintro.org/os

We also have improved the ability to access data in this book through the addition of Appendix B, which provides additional information for each of the data sets used in the main text and is new in the Fourth Edition. Online guides to each of these data sets are also provided at openintro.org/data and through a companion R package.

We appreciate all feedback as well as reports of any typos through the website. A short-link to report a new typo or review known typos is openintro.org/os/typos.

For those focused on statistics at the high school level, consider *Advanced High School Statistics*, which is a version of *OpenIntro Statistics* that has been heavily customized by Leah Dorazio for high school courses and AP® Statistics.

Acknowledgements

This project would not be possible without the passion and dedication of many more people beyond those on the author list. The authors would like to thank the OpenIntro Staff for their involvement and ongoing contributions. We are also very grateful to the hundreds of students and instructors who have provided us with valuable feedback since we first started posting book content in 2009.

We also want to thank the many teachers who helped review this edition, including Laura Action, Matthew E. Aiello-Lammens, Jonathan Akin, Stacey C. Behrensmeyer, Juan Gomez, Jo Hardin, Nicholas Horton, Danish Khan, Peter H.M. Klaren, Jesse Mostipak, Jon C. New, Mario Orsi, Steve Phelps, and David Rockoff. We appreciate all of their feedback, which helped us tune the text in significant ways and greatly improved this book.

¹Guided Practice problems are intended to stretch your thinking, and you can check yourself by reviewing the footnote solution for any Guided Practice.