Contents

Welcome

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

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

IMPORTANT NOTE: This book is currently in active development, and is planned for launch in late 2020 as the 2nd edition of OpenIntro Statistics - Introduction to Statistics with Randomization and Simulation. The 1st edition of the book can be accessed at openintro.org/book/isrs/.

Textbook overview

1. **Introduction to data.** Data structures, variables, summaries, graphics, and basic data collection techniques.
2. **Exploratory data analysis.** Data visualization and summarisation.
3. **Correlation and regression.** Visualising relationships between many variables and descriptive summaries for quantifying the relationship between two variables.
4. **Multiple regression.** Descriptive summaries for quantifying the relationship between two variables.
5. **Foundations for inference.** Case studies are used to introduce the ideas of statistical inference with randomization and simulations.
6. **Inference for categorical data.** Inference for proportions using simulation and randomization techniques as well as the normal and chi-square distributions.
7. **Inference for numerical data.** Inference for one or two sample means using simulation and randomization techniques as well as the normal and F distributions.

8. **Inference for regression.** Extending inference techniques presented thus-far to regression settings.
9. **Appendix: Probability.** An introduction to probability is provided as an optional reference. Exercises and additional probability content may be found in Chapter~3 of OpenIntro Statistics at openintro.org/book/os.

Examples, exercises, and additional appendices

Examples and guided practice exercises throughout the textbook may be identified by their distinctive bullets:

[MCR-TODO: Need to update language below]

OpenIntro, online resources, and getting involved

OpenIntro is an organization focused on developing free and affordable education materials.

We encourage anyone learning or teaching statistics to visit [openintro.org] (http://www.openintro.org) and get involved. We also provide many free online resources, including free course software. Students can test their knowledge with practice quizzes for each chapter or try an application of concepts learned using real data.

Data sets for this textbook are available on the website and through a companion R package, **openintro**.[1] All of these resources are free, and we want to be clear that anyone is welcome to use these online tools and resources with or without this textbook as a companion.

We value your feedback. If there is a part of the project you especially like or think needs improvement, we want to hear from you. You may find our contact form at openintro.org.

Acknowledgements

[MCR-TODO: Will need to update language and people mentioned here.]

This project would not be possible without the dedication and volunteer hours of all those involved. No one has received any monetary compensation from this

[1] Mine Çetinkaya-Rundel, David Diez, Andrew Bray, Albert Kim, Ben Baumer, Chester Ismay and Christopher Barr (2019). openintro: Data Sets and Supplemental Functions from 'OpenIntro' Textbooks and Labs. R package version 2.0.1. https://github.com/OpenIntroStat/openintro-r-package.

project, and we hope you will join us in extending a *thank you* to all those who volunteer with OpenIntro.

The authors would especially like to thank Andrew Bray and Meenal Patel for their involvement and contributions to this textbook. We are also grateful to Andrew Bray, Ben Baumer, and David Laffie for providing us with valuable feedback based on their experiences while teaching with this textbook, and to the many teachers, students, and other readers who have helped improve OpenIntro resources through their feedback.

The authors would like to specially thank George Cobb of Mount Holyoke College and Chris Malone of Winona State University. George has spent a good part of his career supporting the use of nonparametric techniques in introductory statistics, and Chris was helpful in discussing practical considerations for the ordering of inference used in this textbook. Thank you, George and Chris!

About the authors

Mine Çetinkaya-Rundel mine@openintro.org University of Edinburgh, Duke University, RStudio

Johanna Hardin jo@openintro.org Pomona College

David Diez david@openintro.org Google/YouTube

others...

Copyright

Copyright © 2020. Second Edition.

This textbook is available under a Creative Commons license. Visit openintro.org for a free PDF, to download the textbook's source files, or for more information about the license.

Chapter 1

Introduction to data

Chapter 2

Exploratory Data Analysis

2.1 Cat vs. cat - segmented plots / contingency tables

- Conditional probability from contingency tables
- Bayes Theorem (law of total probability?)

2.2 Num vs. cat - side-by-side box plots / comparing distributions

- Mention univariate - center, skew, shape, spread
- Mention conditional probabilities as well

Chapter 3

Correlation and Regression

3.1 Visual summaries of data: scatterplot, side-by-side boxplots, histogram, density plot, box plot (lead out with multivariate, follow with univariate)

3.2 Describing distributions: correlation, central tendency, variability, skew, modality

3.3 Num vs. num - SLR

- correlation
- Line fitting, residuals, and correlation
- Fitting a line by least squares regression
- Types of outliers in linear regression

Chapter 4

Multiple Regression

4.1 Num vs. whatever - MLR

- Introduction to multiple regression

4.2 Parallel slopes

4.3 Hint at interaction, planes, and parallel planes but not quantify

- Visualization of higher-dimensional models (rgl demo)

4.4 Logistic regression

- Binary vs. num/whatever
- Three scales interpretation (e.g. probability, odds, log-odds)
- "parallel" logistic curves?

Chapter 5

Foundations of inference

Chapter 6

Inference for categorical data

6.1 Inference for a single proportion

- Simulation
- Exact (if we include course on probability)
- CLT and Normal approximation

6.2 Difference of two proportions

6.3 Testing for goodness of fit using chi-square (special topic, include simulation version)

6.4 Testing for independence in two-way tables (special topic)

Chapter 7

Inference for numerical data

7.1 One-sample means

- Bootstrap (for means, medians)
- t-distribution

7.2 Paired data

7.3 Difference of two means

7.4 Comparing many means with ANOVA (special topic, include simulation version)

Chapter 8

Inference for regression

8.1 Inference for linear regression

- Bootstrap for regression coefficients
- t-distribution for regression coefficients
- Model Comparison: Occam's Razor and R^2 > R^2_adj

8.2 Checking model assumptions using graphs

- L-I-N-E

8.3 Inference for multiple regression

- residuals vs. fitted instead of residuals vs. x

8.4 Inference for logistic regression

Chapter 9

Appendix: Probability

(Keep same content as before, minus the bit of probability that got moved to categorical EDA)