Contents

```
setwd("~/gitRepos/DataCamp/randomization-and-simulation")
bookdown::render_book("index.Rmd", output_dir = "_book")
```

```
#install_github("openintrostats/openintro-r-package")
library(openintro)
library(tidyverse)
library(knitr)
```

```
# move to _common.R based on R4DS: https://github.com/hadley/r4ds/blob/master/_common.
set.seed(25)
options(digits = 3)

knitr::opts_chunk$set(
  comment = "#>",
  collapse = TRUE,
```

```
  cache = TRUE,
  echo = FALSE, # hide code unless otherwise noted in chunk options
  out.width = "70%",
  fig.align = 'center',
  fig.width = 6,
  fig.asp = 0.618,  # 1 / phi
  fig.show = "hold"
)

options(dplyr.print_min = 6, dplyr.print_max = 6)
```

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

Chapter 1

Introduction to data

Scientists seek to answer questions using rigorous methods and careful observations. These observations – collected from the likes of field notes, surveys, and experiments – form the backbone of a statistical investigation and are called . Statistics is the study of how best to collect, analyze, and draw conclusions from data, and in this first chapter, we focus on both the properties of data and on the collection of data.

1.1 Case Study: using stents to prevent strokes

Section [INSERT REFERENCE] introduces a classic challenge in statistics: evaluating the efficacy of a medical treatment. Terms in this section, and indeed much of this chapter, will all be revisited later in the text. The plan for now is simply to get a sense of the role statistics can play in practice.

In this section we will consider an experiment that studies effectiveness of stents in treating patients at risk of stroke. Stents are devices put inside blood vessels that assist in patient recovery after cardiac events and reduce the risk of an additional heart attack or death. Many doctors have hoped that there would be similar benefits for patients at risk of stroke. We start by writing the principal question the researchers hope to answer:

> Does the use of stents reduce the risk of stroke?

The researchers who asked this question conducted an experiment with 451 at-risk patients. Each volunteer patient was randomly assigned to one of two groups:

- **Treatment group**. Patients in the treatment group received a stent and medical management. The medical management included medications, management of risk factors, and help in lifestyle modification.

Table 1.1: Results for five patients from the stent study.

patient	group	0-30 days	0-365 days
1	treatment	no event	no event
2	control	no event	stroke
3	control	no event	no event
4	control	no event	stroke
5	control	no event	no event

Table 1.2: Descriptive statistics for the stent study.

group	0-30 days_stroke	0-30 days_no event	0-365 days_stroke	0-365 days_no event
treatment	33	191	45	179
control	13	214	28	199

- **Control group**. Patients in the control group received the same medical management as the treatment group, but they did not receive stents.

Researchers randomly assigned 224 patients to the treatment group and 227 to the control group. In this study, the control group provides a reference point against which we can measure the medical impact of stents in the treatment group.

Researchers studied the effect of stents at two time points: 30 days after enrollment and 365 days after enrollment. The results of 5 patients are summarized in Figure 1.1. Patient outcomes are recorded as `stroke` or `no event`, representing whether or not the patient had a stroke at the end of a time period.

Considering data from each patient individually would be a long, cumbersome path towards answering the original research question. Instead, performing a statistical data analysis allows us to consider all of the data at once. Figure **??** summarizes the raw data in a more helpful way. In this table, we can quickly see what happened over the entire study. For instance, to identify the number of patients in the treatment group who had a stroke within 30 days, we look on the left-side of the table at the intersection of the treatment and stroke: 33.

[MCR-TODO: Figure out LaTeX exercise counter stuff that's differently formatted for in text exercises.] **Exercise: Of the 224 patients in the treatment group, 45 had a stroke by the end of the first year. Using these two numbers, compute the proportion of patients in the treatment group who had a stroke by the end of their first year. (Please note: answers to all Guided Practice exercises are provided using footnotes.)[1]

We can compute summary statistics from the table. A is a single number summarizing a large amount of data. For instance, the primary results of the study

[1] The proportion of the 224 patients who had a stroke within 365 days: $45/224 = 0.20$.

after 1 year could be described by two summary statistics: the proportion of people who had a stroke in the treatment and control groups.

- Proportion who had a stroke in the treatment (stent) group: $45/224 = 0.20 = 20\%$.
- Proportion who had a stroke in the control group: $28/227 = 0.12 = 12\%$.

These two summary statistics are useful in looking for differences in the groups, and we are in for a surprise: an additional 8% of patients in the treatment group had a stroke! This is important for two reasons. First, it is contrary to what doctors expected, which was that stents would **reduce** the rate of strokes. Second, it leads to a statistical question: do the data show a "real" difference between the groups?

This second question is subtle. Suppose you flip a coin 100 times. While the chance a coin lands heads in any given coin flip is 50%, we probably won't observe exactly 50 heads. This type of fluctuation is part of almost any type of data generating process. It is possible that the 8% difference in the stent study is due to this natural variation. However, the larger the difference we observe (for a particular sample size), the less believable it is that the difference is due to chance. So what we are really asking is the following: is the difference so large that we should reject the notion that it was due to chance?

While we don't yet have our statistical tools to fully address this question on our own, we can comprehend the conclusions of the published analysis: there was compelling evidence of harm by stents in this study of stroke patients.

Be careful: Do not generalize the results of this study to all patients and all stents. This study looked at patients with very specific characteristics who volunteered to be a part of this study and who may not be representative of all stroke patients. In addition, there are many types of stents and this study only considered the self-expanding Wingspan stent (Boston Scientific). However, this study does leave us with an important lesson: we should keep our eyes open for surprises.

1.2 Taxonomy of Data

1.3 Overview of data collection principles

1.4 Observational studies and sampling strategies

1.5 Experimental design and causality

1.6 Revisit case study with new terminology we learned

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)