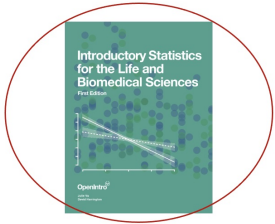
OpenIntro Biostatistics: developing open source materials for teaching and learning applied statistics with R

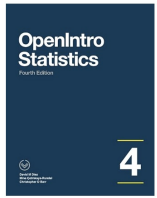
Dave Harrington and Julie Vu

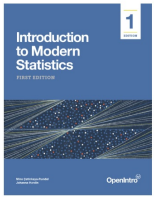
12 October 2023

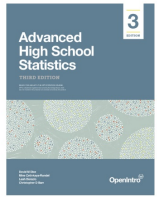
Introduction

The OpenIntro Family of Texts









Design Goals for OI Biostat

Intended student population

- Planning to work in health or life sciences
- Motivated more by application than theory
- No previous statistics courses
- Familiarity with basic algebra
- No or limited experience with computing

Pedagogical goals

- Introduce statistics using examples from important issues in medical science and biology
- Provide opportunity to work with data from published studies
- Introduce computing with the open source language R
- Use the text as a core component of an active learning environment
- Introduce ideas of reproducible research
  - Labs, problem sets, and exams contain code, output, and analysis narrative

Availability

- Text PDF freely available
- All source files (LaTeXand R) available on GitHub
- Text can be modified for local environments
- Sample labs, problem sets, exams on the OpenIntro site
- Datasets in an R package

Teaching with *OpenIntro Biostatistics*

Course structure for STAT 102 and STAT S-100

STAT 102 and STAT S-100

- Audience: almost exclusively students without prior stats/coding experience
- Varying course size
- STAT 102 is a life sciences audience, STAT S-100 is more general (includes students in master's degree programs)
- STAT S-100 taught online

Course Structure

- Recommended pre-reading before class
- Class consists of brief lecture, followed by active learning labs
- Optional: TA-led discussion sections
- Assessments: weekly problem sets, open-note/open-book exams

Why active learning (labs)?

"Isn't lecturing more efficient than active learning?"

Motivation: students struggle to apply content from lecture when faced with a problem to solve

- Active learning labs provide immediate feedback for both students and instructors.
  - Avenue for eliciting questions
  - Facilitates identifying common areas of understanding in class, rather than after assessments are graded
- Active learning labs promote an open, collaborative learning environment.
  - Sets up the expectations that everyone will participate
  - Produces a more congenial atmosphere conducive to questions and discussion
- Completed labs become useful reference material for students.

What does a lab look like?

- Labs consist of three files: handout, template, solutions.
- During class, students work in an R Markdown template. Question statements are most readable in the PDF handout.
- After class, students can review solutions in more depth than discussed in class.
- Note: Students access R Studio in their browser via Posit Cloud.

````
50  1. Run the following code to take a random sample of 30 individuals from
    `yrbss_complete` and store the subset as `yrbss_sample`. The code includes a line
    that removes any rows from `yrbss` where no data has been recorded for `weight`;
    this version of `yrbss` is named `yrbss_complete`.
51
52  ```{r}
53  #load the data
54  library(oibiostat)
55  data("yrbss")
56
57  #remove rows with missing values
58  yrbss_complete <- yrbss[complete.cases(yrbss$weight), ]
59
60  #set parameters
61  sample_size <- 30
62
63  #obtain random sample of row numbers
64  set.seed(5011)
65  sample_rows <- sample(1:nrow(yrbss_complete), sample_size)
66
67  #create yrbss_sample
68  yrbss_sample <- yrbss_complete[sample_rows, ]
69  ```
70
71  a) Calculate $\mu_{weight}$, the mean weight in the `yrbss` population.
72
73  ```{r}
74
75  ```
76
77  ***
78
79  Your answer here.
80
81  ***
````

Three key features of labs

1. Scaffolding

- Provide students with an initial framework.
- Break down problems/concepts into digestible parts.
- Suitable for students studying the material independently.

2. Mammograms incorrectly suggest breast cancer is present approximately 5% of the time when a woman does *not* have cancer. Suppose a clinic conducts approximately 50 mammograms in a week. If none of the women tested in a week actually have breast cancer, what is the probability that at least 1 woman will test positive?

   a) The event of interest is that at least 1 woman will test positive out of 50 women. What is the complement of this event?

   b) What is the probability of receiving a true negative result; i.e., testing negative when one does not have breast cancer?

   c) Is it reasonable to assume that the 50 mammograms conducted in a week are independent; i.e., the test result that one woman receives provides no information about the test result another woman receives?

   d) Based on the set-up done in parts a) - c), compute the probability that at least 1 woman will test positive if none of the women have breast cancer.

Three key features of labs. . .

2. Comprehension Check

- Questions inspired by common points of confusion.
- Goal: build skills that students will be expected to demonstrate on assessments.

3. Conduct a hypothesis test to assess the strength of evidence for whether 5-bedroom houses in Saratoga County have mean living area different from 2,700 square feet. Let $\alpha = 0.05$.
   a) State the null and alternative hypotheses.
   b) Compute the test statistic.
   c) Compute the p-value.
   d) Draw a conclusion, including interpreting the p-value in the context of the data.

4. Suppose that the alternative hypothesis had been $H_A : \mu < 2700$ sq. ft. Would you expect this p-value to be smaller or larger than the p-value from Question 3? Explain your reasoning.

5. Suppose that the alternative hypothesis had been $H_A : \mu > 2700$ sq. ft. Would you expect this p-value to be smaller or larger than the p-value from Question 3? Explain your reasoning.

Three key features of labs...

3. R Coding

- R coding taught on an as-needed basis.
- Approach: prompting students to explore code that already works.
- Supplemented by the Lab Notes: reference guide specifically for R syntax.
- Lab Notes for Probability unit included in talk materials.

3. The following code illustrates the use of the `sample()` function to simulate the result for one set of 5 coin tosses.

```
#define parameters
prob.heads <-
number.tosses <-

#simulate the coin tosses
outcomes <- sample(c(0, 1), size = number.tosses,
                   prob = c(1 - prob.heads, prob.heads), replace = TRUE)

#view the results
table(outcomes)

#store the results as a single number
total.heads <- sum(outcomes)
total.heads
```

a) Using the information given about the experiment, assign values to `prob.heads` and `number.tosses` then run the code chunk.

b) To generate the vector `outcomes`, the `sample()` function draws from the values `0` and `1` with probabilites corresponding to those specified by the argument `prob`. Which number corresponds to heads, and which corresponds to tails?

c) Why is it necessary to sample with replacement from the vector `(0, 1)`?

d) Run the code chunk again to simulate another set of 5 coin tosses. Is it reasonable to expect that the results might differ from the first set of tosses? Explain your answer.

e) If we instead sampled from a vector `(T, H)`, how would we need to modify the argument of `sum()` to count the number of heads?

Types of labs

- Exploration-oriented: explore a new concept
  - Constructing a sampling distribution from a "hypothetical" population: visualize how the distribution of $\overline{X}$ changes with different n, understand why a larger sample can be helpful in a realistic setting when only one sample is observed
- Practice-oriented: immediately practice new material
  - Interpreting regression models: after adjusting for age, is there evidence that statin use is associated with lower cognitive function?
- Application-oriented: synthesize several ideas and illustrate a workflow
  - California Department of Developmental Services (DDS) case study: exploratory data analysis, importance of multivariate analysis, confounding and Simpson's Paradox[1]

[1] From Taylor and Mickel (2017)

Suggestions for using active learning labs

- Implement a structure to the lab time.
- Plan timing carefully.
- What works in my experience:
  - Walking the room while the labs are happening.
  - Checking in with students even if they aren't asking for help.
  - Acknowledging to students that active learning is challenging.

Resources

Useful information

Feel free to contact us with questions

- Julie Vu, julievu@g.harvard.edu
- Dave Harrington, davidharrington@g.harvard.edu

On workshop web site

- These slides
- Zip archive with sample exams, labs and lecture slides (source and pdf)
- Important links to text and associated files