A Minimal Book Example

Yihui Xie

2019-10-12

Chapter 1

Prerequisites

This is a *sample* book written in **Markdown**. You can use anything that Pandoc's Markdown supports, e.g., a math equation $a^2 + b^2 = c^2$. A

The **bookdown** package can be installed from CRAN or Github:

```
install.packages("bookdown")
# or the development version
# devtools::install_github("rstudio/bookdown")
```

Remember each Rmd file contains one and only one chapter, and a chapter is defined by the first-level heading `#`.

To compile this example to PDF, you need XeLaTeX. You are recommended to install TinyTeX (which includes XeLaTeX): https://yihui.name/tinytex/.

Chapter 2

Sampling, Experiments, and Exploratory Data Analysis

Data in the Wild

Words

Experiment Background

Marketing example. Goal to describe the customers, how they tend to purchase/shop, and maybe find some shared qualities in order to adverstise curated packages to folks.

Define basic things like population, parameters, statistics, and sample.

Discuss conceptual vs actual populations and when we might care about one or the other. Our "sample" is really a bit of data from the conceptual population. Or we could consider it as the population and we just want to describe it.

Selecting Response Variables

Marketing example with data such as Clicks, Impressions, Total Revenue, Total Spent, Average Order Value, Sport, Time of visit/purchase, Campaigns running, etc.

Identifying Sources of Variation

Consider variables linked to the user. Age, other accounts, etc.

Choose an Experimental Design

Discuss our "sampling" scheme vs a random sample. This seems like a case where we aren't doing a "good" scheme but not much else could be done...

Maybe talk about how in the future you could do alternate email ads or something and do an AB type study.

Peform the Test

Get the data from google analytics or whatever, have a plan for updating each month?

Look at the Data

Careful discussion of not selecting a modeling technique based on this unless it is a pilot study or an exploratory study else we have increased our nominal type I error rate...

(sometimes EDA sometimes data validation only/cleaning - more formal experiments)

Spend a lot of time here talking about graphs of different types. Sample means, sample variances, etc.

Discuss population curves vs sample histograms and the relationship.

Statistically Analyze the Data

New variables as functions of old?

Not a formal test here but comparisons of interest etc.

Draw conclusions

What actionable things have we found? Likely some trends to investigate further. Perhaps run an experiment to formally see if some alteration can be effective.

What can we conclude realistically from this data? To what population are we talking?

Statistical Testing Ideas

Experiment Background

This example would lend itself to a reasonably easy randomization test or simulation based test. Maybe an AB type study where we swap labels and do that with a nice visual.

Maybe third example with simulation test.

Selecting Response Variables

Identifying Sources of Variation

Choose an Experimental Design

Good discussion of what makes a good sampling design. Maybe a statified example like the river and selecting houses example as a quick expose of the issues with not doing a truly random sampling technique.

Basics of experimental design (randomization, replication, error control ideas).

Recap benefits of doing an experiment vs an observational study.

Peform the Test

Explore the Data

NHST paradigm with false discovery?

Statistically Analyze the Data

Draw conclusions

Chapter 3

Point Estimates

Learning objectives for this lesson: - How to estimate a mean - Definition of "convenience sample" - Definition of "systematic sample" - Benefits/drawbacks to both approaches - Understand how to estimate a mean - Understand how to estimate a quantile - Understand implicit assumptions for these approaches

Estiamte with means

Experiment background

Someone wants to know how much of something they need to satisfy some population To get a good estimate of this, we can use the average amount for each one and then multiply by the whole population

Estimate with quantiles

Experiment background

Big Deborah's is making new packaging for their cookies. The engineer responsible for the new desing needs to make sure that the packaging fits the new cookies. While the cookie manufacturing process is standardized, there's inevitably some degree of variation in cookie size. After discussing the issue with corporate, the engineer decides that a the new cookie sleeves should be large enough to fit 95% of cookies that are baked. (The largest five percent will be marketed separately as "JUMBO" cookies.)

Define the object of the experiment

The Engineer is tasked with determining how large the cooke sleeve needs to be. There's no way for her to know the size of every cookie that Big Deborah's has made (or will make going forward!), so she'll need to collect data on existing cookies to inform her cookie sleeve size determination.

Select appropriate response variables

If the maximum distance from any one point on the (round) cookie's perimeter to any other point is smaller than the diameter of the cookie sleeve, then the cookie will fit. This makes "cookie diameter" a good measure for this test. It is easy to measure for each cookie and is directly relevant to the experiment's objective.

[probably have something in here about]

IDentify sources of variation

While the manufacturing process is standardized, there is variation in size from one cookie to the next. This is one source of variation. The engineer isn't sure of any others. However, she knows that cookies are made in multiple factories, and that each factory has multiple ovens. Ovens and factories could also be sources of variation.

Choose an experimental design

The Engineer knows that she needs to look at multiple cookies, since she knows that there is variation in diameter from one cookie to the next. One option would be to just use the remaining cookies in the box she has in her office (22 of the 25-count box remain). [something about convenience sample] However, she knows that cookies from the same oven are typically packaged together. If there is variation from one oven to the next, looking at the cookies she has in her office may not tell the whole story.

Instead, she chooses to take every 20th cookie manufactured off the assembly line until she gets 500 cookies. [something about systematic sample]

Perform the test

The day of the test comes, and the Engineer starts collecting cookies. However, problems arise! The plan has to shut down half-way through, so she only gets 431 cookies instead of the 500 she thought she would. However, she measures the diameters of each cookie and records the data in a spreadsheet.

Statistically analyze the data

The initial plan had been to rank-order the 500 cookies and estimate the 95th percentile using the diamter of the 475th largets cookie. Since we didn't get all of our data, we have to improvise. 431 doesn't neatly yield a value such that exactly 95% are less than or equal and 5% are greater than or equal. One option is to choose the 410th largest cookie to estimate our percentile. Slightly more than 95% of cookies will have smaller diameters than this. Alternatively, we could interpolate between the 409th and 410th cookies. [reasons and logic and math for each of these]

Draw conclusions

Based on this study, the Engineer concludes that a cookie sleeve large enough for a cookie of diameter XX will be big enough to contain 95% of Big Deborah cookies.

Discussion

- pros and cons to the approach chosen
- generalizing to other types of point estimates

Chapter 4

Accounting for Uncertainty

Some *significant* applications are demonstrated in this chapter.

Example one

Example two

Chapter 5

Inference via Hypothesis Tests for One Sample

We have finished a nice book.

Chapter 6

Inference via Confidence Intervals for One Sample

We have finished a nice book.

Chapter 7

Inference for Two Categorical Variables

We have finished a nice book.

Chapter 8

One-Way ANOVA

We have finished a nice book.

Chapter 9

Multi-way ANOVA

We have finished a nice book.

Chapter 10

Block Designs

We have finished a nice book.

Chapter 11

Regression Models

We have finished a nice book.

Chapter 12

The General Linear Model

We have finished a nice book.

Chapter 13

Mixed Models

We have finished a nice book.

Chapter 14

Split Plot and Repeated Measures Designs

We have finished a nice book.

Chapter 15

Logistic Regression and Generalized Linear Models

Stuff here

We have finished a nice book.

Chapter 16

Generalized Linear Mixed Models

We have finished a nice book.

Chapter 17

Appendix - Learning Objectives

Book-level

After reading this book you will be able to:

- identify relevent sources of variability for a potential study
  - covariates
  - noise variables
  - random effects
- utilize principles of design to plan a reasonable experiment to help answer questions of interest
  - randomization
  - systematic variation of factors/covariates
  - factor identifiability
- compare and contrast methods for designing an experiment when the goal of a study is prediction versus when the goal is statistical inference
  -
- articulate the scope of inferential conclusions in light of the method of data collection, the experimental design used, and the statistical analysis applied
  - limitations due to sampling/sample frame
  - missing data
  - modeling assumptions
  - sampling assumptions
  - requirements for causal inference
- choose appropriate numerical summaries and graphical displays for a set of data and create these using software
  - when to use tables vs. a picture
  - types of graphical displays
    - bar charts
    - pie charts
    - plotting data vice just predictions/conclusions
    - when to include uncertainty bounds
    - five-number summaries
    - means vs. medians

    - general plotting recommendations
    - use of colors in you plots (discrete vs. divergent vs. continuous color scales, gray-scale, color-blind-friendly scales)
  - use of annotations
  - general graphical design philosophy (building a chart to illustrate a conclusion)
  - trade-offs between detail and interpretability
  - not screwing up your axes
- explain the general concept of point estimation and how to account for sampling variability
  - definition
  - identify the right point estimate for your response variable of interest
  - estimating uncertainty for point estimates
    - normal approximation
    - bootstrap CI
    - others?
- explain the importance of statistical distributions when conducting statistical inference
  - normal distribution and approximations plus properties
    - robustness
    - generality
    - CLT
  - costs and benefits of using nonparametric approaches
- describe the fundamental inferential techniques of hypothesis testing and confidence intervals
  - identify a null and alternative for a given problem
  - interpret hypotheses
  - characterize the test statistic under the null
  - explain what a rejection region and be able to identify one
  - ????
- compare and contrast the use of and interpretations from hypothesis tests and confidence intervals
  - identify when using a CI and NSHT will result in the same conclusion
  - explain when you can use a confidence interval to test for differences (e.g., comparing a single point estimate to a threshold) and when you can't (e.g., when you have CIs for two different means)
- fit statistical models in software and interpret their output
  - Which PROCs from SAS? REG, GLM, MIXED, GLIMMIX, others??
  - `lm()`, `glm()`, `anova()` .... `broom`? `modlr`? `ciTools`?
  - p-values, point estimates, standard errors, f-statistics, chi-square-statistics, degrees of freedom, SS/MS, residual plots
- Something about understanding the relationships between the models (linear model framework)
  - Write statistical models using matrix representaiton
  - identify models written in matrix representation with their representation in software
  - identify when models written in different notation are the same or different
  - describe when specific models will give you the same results
    - ANOVA w/ 2 factors and a t-test or a SLR
    - ANCOVA and MLR
    - random effects vs. fixed effects
    - split plots vs. more general mixed models
    - logistic regression w/ categorical factors vice contingency table analysis
  - discuss differences in assumptions associated with ANOVA vice SLR/MLR
- Maybe another bit about data types you see, how the assumptions we make impact things

Topic-level

Chapter 2 - Sampling, Design, and Exploratory Data Analysis

Chapter 3 - Point Estimation

Chapter 4 - Accounting for Uncertainty in Estimation

Chapter 5 - Inference via Hypothesis Testing for a Proportion or Mean

Chapter 6 - Inference via Confidence Intervals for a Proportion or Mean

Chapter 7 - Inference on Two Categorical Variables

Chapter 8 - Inference for Multiple Means

Chapter 9 - Multiway ANOVA

Chapter 10 - Block Designs

Chapter 11 - Regression

Chapter 12 - The General Linear Model

Chapter 13 - Mixed Models

Chapter 14 - Repeated Measures and Split Plot Designs

Chapter 15 - Logistic Regression and Generalized Linear Models

Chapter 16 - Generalized Linear Mixed Models

From ST512

WE NEED TO ORGANIZE THESE UNDER DIFFERENT CHAPTERS AT SOME POINT Learning Objectives

1. Recognize a completely randomized design with one treatment factor and write the corresponding one-way analysis of variance model, with assumptions
2. Estimate treatment means
3. Estimate the variance among replicates within a treatment
4. Construct the analysis of variance table for a one factor analysis of variance, including computing degrees of freedom, sums of squares, mean squares, and F-ratios
5. Interpret results and draw conclusions from a one-factor analysis of variance
6. Estimate differences between two treatment means in a one factor analysis of variance
7. Test differences between two treatment means in a one factor analysis of variance

8. Construct a contrast to estimate or test a linear combination of treatment means
9. Estimate the standard error of a linear combination of treatment means
10. Make inferences about linear combinations of treatment means, including contrasts.
11. Obtain and understand SAS output for linear combinations of treatment means, including contrasts.
12. Explain when and why corrections for multiple comparisons are needed
13. Know when and how to use Tukey's correction for all pairwise comparisons
14. Compute Bonferroni confidence intervals
15. Create and interpret orthogonal contrasts.
16. Define main effects and interactions
17. Write contrasts to estimate main effects and interactions
18. Estimate these contrasts and their standard errors
19. Compute sums of squares associated with these contrasts
20. Test hypotheses about the main effects and interactions.
21. Identify and define simple effects.
22. Identify and define interaction effects.
23. Identify and define main effects.
24. Understand when to use simple, interaction, and main effects when drawing inferences in a two-way ANOVA.
25. Write the analysis of variance model and SAS code for a completely randomized design with two factors
26. Test hypotheses and interpret the analysis of variance for a factorial experiment.
27. Explain the appropriate use of correlations and compute the correlation coefficient
28. Read and interpret a scatterplot and guess the correlation coefficient by examination of a scatter plot
29. Interpret the strength and direction of association indicated by the correlation coefficient and judge when a correlation coefficient provides an appropriate summary of a bivariate relationship
30. Test the hypothesis that the correlation coefficient is zero using either a t-test or the Fisher z transformation, Compute confidence intervals using Fisher's z transformation
31. Write a statistical model for a straight line regression or a multiple regression and explain what all the terms of the model represent
32. Explain the assumptions underlying regression models, evaluate whether the assumptions are met
33. Estimate the intercept, slope and variance for a simple linear regression model
34. Fit a multiple regression model in SAS and interpret the output, use the coefficient of determination to evaluate model fit
35. Use a regression model to predict Y for new values of X
36. Estimate the variance and standard error of parameters in regression models, test hypotheses about the parameters, and construct confidence intervals for the parameters.

37. Explain the difference between a confidence interval and a prediction interval and know when to use each of them
38. Construct a confidence interval for the expected value of Y at a given value of X
39. Construct a prediction interval for a new value of Y at a given value of X
40. Write a linear model in matrix notation
41. Find the expectation and variance of a linear combination of random variables, a'Y
42. Set up the expressions to calculate parameter estimates and predicted values using the matrix form of the model
43. Estimate standard errors for parameter estimates and predicted values
44. Use extra sums of squares to test hypotheses about subsets of parameters
45. Construct indicator variables for including categorical regressor variables in a linear model
46. Understand how to interpret parameters of a general linear model with indicator variables
47. Estimate contrasts of treatment means and their standard errors using the general linear model notation and matrix form of the model
48. Compare nested models with a lack of fit test to select a model
49. Explain what a covariate is and how they are used
50. Explain the assumptions of the analysis of covariance model and determine when these assumptions are met
51. Fit an analysis of covariance model in SAS and conduct appropriate tests for treatment effects
52. Estimate and interpret treatment means and their standard errors adjusted for covariates using SAS, Construct confidence intervals for adjusted treatment means
53. Construct and estimate contrasts of treatment means adjusted for covariates and estimate the standard errors and confidence intervals of such contrasts.

Analysis of variance and design of experiments Recognize each of the following types of experimental designs and determine when each type would be advantageous. 1. completely randomized design 2. randomized complete block design 3. split plot design Recognize whether factors should be considered fixed effects or random effects and explain the scope of inference for each case. Recognize whether factors are crossed or nested. For all of the designs listed and for experiments with crossed and/or nested fixed factors, random factors, or a combination of fixed and random effects, be able to 1. Write the corresponding analysis of variance model, with assumptions, and define all terms 2. Estimate treatment means and their standard errors 3. Construct the analysis of variance table, including computing degrees of freedom, sums of squares, mean squares, and F-ratios 4. Determine whether the assumptions of the model are satisfied 5. Interpret results and draw conclusions 6. Construct and estimate linear combinations of treatment means and their standard errors 7. Test hypotheses and construct confidence intervals about linear combinations of treatment means 8. Explain when and why corrections for multiple comparisons are needed, know when and how to use Tukey's correction for all pairwise comparisons, compute Bonferroni confidence intervals 9. Create and interpret orthogonal contrasts. 10. Define and interpret main effects, simple effects and interactions 11. Use a table of expected mean squares to estimate variance components and determine appropriate F-statistics for testing effects in the analysis of variance 12. Interpret variance components and estimate and interpret the intraclass correlation coefficient. Regression and correlation Explain the appropriate use of correlations and compute the correlation coefficient, read and interpret a scatterplot and guess the correlation coefficient by examination of a scatter plot, test the hypothesis that the correlation coefficient is zero using either

a t-test or the Fisher z transformation, compute confidence intervals using Fisher's z transformation You should be able to do the following for fitting models to describe the relationships of one or several variables to a response variable. The regressor variables may be continuous or categorical or a mix of the two (e.g., analysis of covariance models) 1. Write a general linear model, including assumptions, in standard or matrix notation, and explain what all the terms and assumptions represent. Be able to handle models that contain interaction terms, polynomial terms, and dummy variables. 2. Evaluate whether the model assumptions are met 3. Fit a general linear model in SAS and interpret the output 4. Work with the general linear model in matrix form, including finding the expectation and variance of a linear combination of regression coefficients or treatment means 5. Test hypotheses and construct confidence intervals for linear combinations of the parameters 6. Construct and interpret a confidence interval for the expected value of Y at a given value of X 7. Construct and interpret a prediction interval for a new value of Y at a given value of X 8. Use extra sums of squares to test hypotheses about subsets of parameters. 9. Explain what a covariate is and how covariates are used

For Point Estimates Chapter

- Definitions for Mean, Median, Quantile, Percentile
- Explain uses for the above
- Identify the correct point estimate to use for a given test
- Define Systematic Random Sample and Convenience Sample
- Explain strengths and weaknesses of each
- Identify conditions when Systematic and Convenience Sampling may not provide representitive samples

Chapter 18

Appendix - Notation

Standard notation

Vectors of variables are denoted with Roman letters, such as x and Y. Capital letters denote random variables while lower case letters denote fixed variables. Note that these vectors may be of length 1 depending on context. Bolded values (x) denote matrices, and in the case of Y, possibly single-column matrices.

Unknown parameters are denoted with Greek letters, with boldface font indicating matrices.

In most models, Y will denote the univariate response, x will describe a matrix of predictor variables, and E a vector of random errors. The Greek letter β will be commonly used for regression parameters (either with subscripts for each values as in $\beta_0 + \beta_1 X_1$ or as a vector (as in $X\beta$). The letters i, j, k, and l will be most commonly used as subscripts or indices. N will typically denote a sample size (not a random vector), with subscripted versions (n_i) describing the number of observations in a group, and p describing the number of parameters in a model beyond the intercept.

We may therefore describe a simple linear regresion model as:

$$Y = x\beta + E$$

In this model, Y is a $N \times 1$ random vector, x is a $N \times (p+1)$ matrix of fixed values, and E is a $N \times 1$ vector.

π is typically used to describe probability parameters, as in Bernoulli or binomial random variables.

Mixed models

Still need to add something for this

Effects model representation

In the effects formulation of ANOVA models, additional greek letters (α, γ, etc.) will appear as parameter effects, as will μ, which will typically represent the grand mean. Group-specfic means will be denoted via subscripts: μ_{ij}. When using this representation, it is convenient to describe a single observation as Y_{ijk}, which is the kth observation from the group with with the ith level of the first factor and the jth level of the second factor. In the main effects version of this model, we have:

$$Y_{ijk} = \mu + \alpha_i + \gamma_j + E_{ijk}$$

We can therefore estimate μ_{ij} as $\hat{\mu}_{ij} = \frac{1}{n}\sum_{k=1}^{n} Y_{ijk} = \bar{Y}_{ij\cdot}$. This "dot" notation can be extended to any subscript and indicates summing over the index that has been replaced by the dot. Further note that the "hat" over a paremeter value denotes the estimator for that parameter value, and the "bar" indicates an average. These features are used generally throughout this book.

Estimators vs. Estimates

If we want to get pedantic, we can differentiate between estimates and estimators in our notation. Estimators are functions of random variables used to estimate parameters. Estimates are realized values of estimators. To differentiate these, we use Roman letters with hats to represent estimators ($\hat{B} = (x'x)^{-1}x'Y$) and Greek letters with hats to represent estimates ($\hat{\beta} = 1.52$).