Sports Analytics

Aaron Nielsen, Department of Statistics, Colorado State University

2022-05-24

Contents

About **5**

Current Tasks **7**

1 Exploratory Data Analysis **9**
1.1 Installing R 9
1.2 Getting Started With R 10
1.3 Definitions 11
1.4 Descriptive Statistics 12
1.5 Baseball 13
1.6 Football 13
1.7 Basketball 13
1.8 Soccer 13
1.9 Volleyball 13
1.10 Hockey 13

2 Probability **15**
2.1 Definitions and Axioms 15
2.2 Theorems and Laws 15
2.3 Random Variables 15

3 Simulation **17**

4 Statistical Inference **19**
4.1 One Sample and Two Sample t-tests and confidence intervals . . 19

5 Correlation **21**

6 Linear Regression **23**

7 Data Scraping **25**

8 Principal Component Analysis **27**

9 Clustering **29**

10 Classification **31**

11 Decision Trees **33**
11.1 Random Forests 33
11.2 Gradient Boosting 33

12 Non-parametric Statistics **35**

13 Baseball **37**

14 Football **39**

15 Basketball **41**

16 Soccer **43**

17 Hockey **45**

18 Volleyball **47**
18.1 Resources 47

19 Other Sports **49**

20 Ellie's stuff **51**

21 Levi's stuff **53**

22 Isaac's stuff **55**

23 Aaron's stuff **57**
23.1 Notes for Chapter 2 (Probability) 57
23.2 Suggested Readings 57
23.3 Notes for Chapter 4 (Simulation) 58

Reference: Blocks **59**
23.4 Equations 59
23.5 Theorems and proofs 59
23.6 Callout blocks 59

Reference: Footnotes and citations **61**
23.7 Footnotes 61
23.8 Citations 61

24 References **63**

About

This book serves as the course textbook for the following courses at Colorado State University:

- STAT 351 (Sports Statistics and Analytics 1)
- STAT 451 (Sports Statistics and Analytics 2)

CSU students contributed to the creation of this book. Many thanks to the following student collaborators:

- Levi Kipp
- Ellie Martinez
- Isaac Moorman

Current Tasks

Updated: "2022-05-24"

Team Tasks and Tips

1. Find datasets from various sports to use as examples for EDA and later chapters
2. Show how to get basic sumamry statistics from these datasets using dplyr, tidy
3. Describe and calculate useful team and individual (descriptive statistics). Example: Baseball: calculate AVG, OBP, OPS, WOBA
4. (High quality) Visualizations using ggplot
5. Look for relevant "sports" R packages
6. Include examples from CSU and Colorado sports teams when possible
7. Sports to be included: Baseball/Softball, Football, Basketball, Soccer, Hockey, Volleyball
8. Sports to be potentially included: Lacrosse, Cricket, Handball,

---

Aaron:

Sports:

Chapters: Currently working to add content to chapters 1-4

Ellie:

Sports: Soccer, Volleyball

Chapters: EDA, Probability

Levi:

Sports: Basketball, Hockey

Chapters: EDA, Probability

Isaac:

Sports: Baseball, Football, Tennis

Chapters: EDA, Scraping

Chapter 1

Exploratory Data Analysis

1.1 Installing R

For this class, you will be using R Studio to complete statistical analyses on your computer.

To begin using R Studio, you will need to install "R" first and then install "R Studio" on your computer.

Step 1: Download R
(a) Visit https://www.r-project.org/
(b) Click **CRAN** under **Download** (c) Select any of the mirrors
(d) Click the appropriate link for your type of system (Mac, Windows, Linux)
(e) Download R on this next page.
(For Windows, this will say **install R for the first time**. For Mac, this will be under **Latest release** and will be something like **R-4.1.0.pkg** – the numbers may differ depending on the most recent version)
(f) Install R on your computer

Step 2: Download R Studio
(a) Visit https://www.rstudio.com/products/rstudio/download/#download
(b) Click to download
(c) Install R Studio on your computer

Step 3: Verify R Studio is working
(a) Open R Studio
(b) Let's enter a small dataset and calculate the average to make sure everything is working correctly.

(c) In the console, type in the following dataset of Sammy Sosa's season home run totals from 1998–2002:

```
sosa.HR <- c(66,63,50,64,49)
```

(d) In the console, calculate the average season home run total for Sammy Sosa between 1998–2002:

```
mean(sosa.HR)
```

```
## [1] 58.4
```

(e) Did you find Slammin' Sammy's average home run total was 58.4? If so, you should be set up correctly!

1.2 Getting Started With R

For the following examples, let's consider Peyton Manning's career with the Denver Broncos. In his four seasons with the Broncos, Manning's passing yard totals were: 4659, 5477, 4727, 2249. Let's enter this data into R. To enter a vector of data, use the **c()** function.

```
peyton <- c(4659, 5477, 4727, 2249)
```

To look at the data you just put in the variable *peyton*, type *peyton* into the console and press enter.

```
peyton
```

```
## [1] 4659 5477 4727 2249
```

Some basic function for calculating summary statistics include **mean()**, **median()**, **var()**, and **sd()**.

```
mean(peyton)
```

```
## [1] 4278
```

```
median(peyton)
```

```
## [1] 4693
```

```
var(peyton)
```

```
## [1] 1967068
```

```
sd(peyton)
```

```
## [1] 1402.522
```

R allows you to install additional packages (or functions) that aren't offered in the base version of R. To install a package, use **install.packages()** and to load a package, use **library()**.

One package that we will use frequently is **tidyverse**. This package includes several other packages and functions such as **ggplot** (plotting function), **dplyr** (data manipulation package), and **stringr** (string manipulation package).

```
install.packages("tidyverse")
library("tidyverse")
```

1.3 Definitions

Definition 1.1. A ***population*** is a well-defined complete collection of objects.

Definition 1.2. A ***sample*** is a subset of the population.

Example 1.1. Suppose we are interested in studying Peyton's Manning's season passing yards totals. What would be the population and what is one possible sample?

Example 1.2. Suppose we are interested in studying Peyton's Manning's season passing yards totals. What would be the population and what is one possible sample?

Definition 1.3. ***Quantitative data*** is numeric data or numbers. It can be broken into two further categories: discrete and continuous data.

Definition 1.4. ***Discrete data*** is quantitative data with a finite or countably infinite number of values.

Definition 1.5. ***Continuous data*** is quantitative data with an uncountably infinite number of values or data taken from an interval.

Example 1.3. What are possible discrete and continuous data associated with Peyton Manning?

Definition 1.6. ***Qualitative data*** refers to names, categories, or descriptions. It can also be broken down into two further categories, nominal data and ordinal data.

Definition 1.7. ***Nominal data*** is qualitative data with no natural ordering.

Definition 1.8. ***Ordinal data*** is qualitative data with a natural ordering.

Example 1.4. What are possible nominal and ordinal data associated with Peyton Manning?

1.4 Descriptive Statistics

While we will learn about some descriptive statistics that are unique to specific sports, there are some descriptive statistics that are frequently used in many applications.

To begin, let $x_1, x_2, \ldots, x_n$ represent a sample of size n, where x_i is the i^{th} value in the dataset.

Definition 1.9. The ***sum*** of the data values is given by: $\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$

Definition 1.10. The ***sample mean*** of the dataset is given by $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Example 1.5. Recall that Peyton Manning's season passing yards total while with the Broncos were: 4659, 5477, 4727, 2249. Calculate the sample mean of these values.

NFL Team Passing Statistics, 2021

1.5 Baseball

1.6 Football

1.7 Basketball

1.8 Soccer

1.9 Volleyball

1.10 Hockey

Chapter 2

Probability

2.1 Definitions and Axioms

2.2 Theorems and Laws

2.3 Random Variables

Chapter 3

Simulation

Chapter 4

Statistical Inference

4.1 One Sample and Two Sample t-tests and confidence intervals

Chapter 5

Correlation

Chapter 6

Linear Regression

Chapter 7

Data Scraping

Chapter 8

Principal Component Analysis

Chapter 9

Clustering

Chapter 10

Classification

Chapter 11

Decision Trees

11.1 Random Forests

11.2 Gradient Boosting

Chapter 12

Non-parametric Statistics

Chapter 13

Baseball

Chapter 14

Football

Chapter 15

Basketball

Chapter 16

Soccer

Chapter 17

Hockey

Chapter 18

Volleyball

18.1 Resources

Women's Volleyball D1 Statistics

Chapter 19

Other Sports

Chapter 20

Ellie's stuff

Chapter 21

Levi's stuff

Chapter 22

Isaac's stuff

Chapter 23

Aaron's stuff

23.1 Notes for Chapter 2 (Probability)

Axioms of Probability:

1. $P(A) \geq 0$
2. $P(\Omega) = 1$
3. If $A_1, A_2, \ldots, A_n$ are disjoint events, then $P(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$

Theorem 23.1 (Bayes theorem)**.** *Let A and B be events in Ω such that $P(B) > 0$. Then we have the following:*

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

23.2 Suggested Readings

23.2.1 Moneyball

Moneyball, Chapter 2, How to Find a Ballplayer (Lewis, 2004)

Near the end of the chapter (page 40), Michael Lewis give a list of players the Oakland Athletics hoped to draft. How did these players turn out? Find the WAR for each of the players in their pre-free agency years and compare it against the Rockies draft picks in the same rounds from the same draft.

23.2.2 Future Value

Future Value, Chapter 7, How to Scout (Longenhagen and McDaniel, 2020)

If a player receives a running grade of 40, approximately what proportion of MLB players have a lower have a lower running grade?

For a given tool, about 95% of all player grades fall between what two bounds? (Consider the middle 95% of the distribution of grades.)

23.3 Notes for Chapter 4 (Simulation)

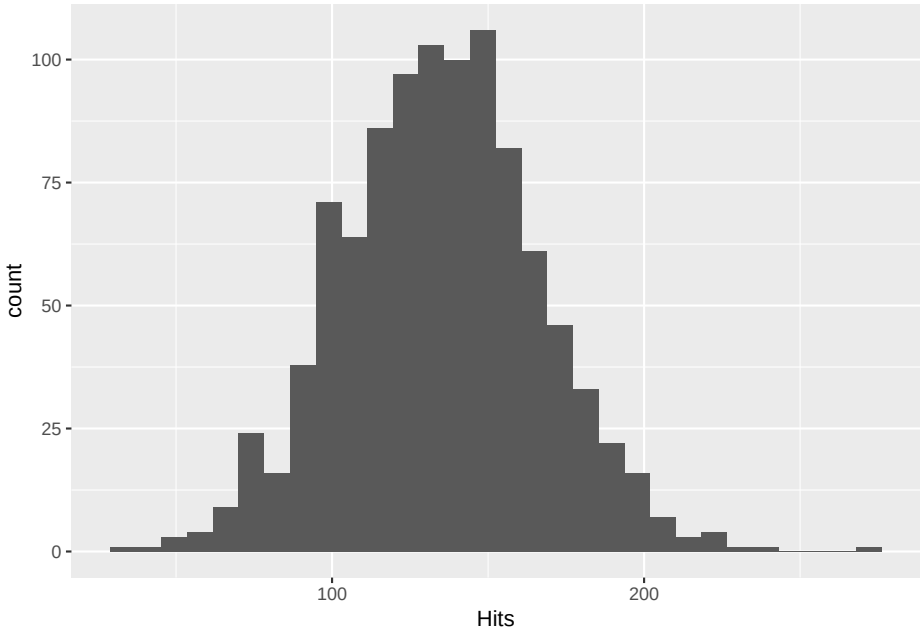
23.3.1 Baseball Simulation Example

```
library(tidyverse)
```

This is a baseball example for chapter 4.

```
set.seed(2022)
n.sims <- 1000
hits <- rep(0,n.sims)
avg <- 0.300
atbats.mean <- 450
atbats.sd <- 100
sim.atbats <- round(rnorm(n.sims,atbats.mean,atbats.sd))

for(i in 1:n.sims){
  sim.hits <- rbinom(1,sim.atbats[i],avg)
  hits[i] = sim.hits
}
hits.df <- data.frame(Hits=hits)
hits.df %>% ggplot(aes(x=Hits)) + geom_histogram()
```



Reference: Blocks

23.4 Equations

Here is an equation.

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{23.1}$$

You may refer to using `\@ref(eq:binom)`, like see Equation (23.1).

23.5 Theorems and proofs

Labeled theorems can be referenced in text using `\@ref(thm:tri)`, for example, check out this smart theorem 23.2.

Theorem 23.2. *For a right triangle, if c denotes the* length *of the hypotenuse and a and b denote the lengths of the* ***other*** *two sides, we have*

$$a^2 + b^2 = c^2$$

Read more here https://bookdown.org/yihui/bookdown/markdown-extensions-by-bookdown.html.

23.6 Callout blocks

The R Markdown Cookbook provides more help on how to use custom blocks to design your own callouts: https://bookdown.org/yihui/rmarkdown-cookbook/custom-blocks.html

Reference: Footnotes and citations

23.7 Footnotes

Footnotes are put inside the square brackets after a caret `^[]`. Like this one [1].

23.8 Citations

Reference items in your bibliography file(s) using `@key`.

For example, we are using the **bookdown** package (Xie, 2016) (check out the last code chunk in index.Rmd to see how this citation key was added) in this sample book, which was built on top of R Markdown and **knitr** (Xie, 2015) (this citation was added manually in an external file book.bib). Note that the `.bib` files need to be listed in the index.Rmd with the YAML `bibliography` key.

The RStudio Visual Markdown Editor can also make it easier to insert citations: https://rstudio.github.io/visual-markdown-editing/#/citations

[1] This is a footnote.

Chapter 24

References

Bibliography

Lewis, M. (2004). *Moneyball: The art of winning an unfair game*. WW Norton & Company.

Longenhagen, E. and McDaniel, K. (2020). *Future Value: The battle for baseball's soul and how teams will find the next superstar*. Triumph Books.

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.

Xie, Y. (2016). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.3.9.