

Sports Analytics

Aaron Nielsen, Department of Statistics, Colorado State University

2022-06-01

Contents

About	7
Current Tasks	9
1 Exploratory Data Analysis	11
1.1 Getting Started With R	11
1.2 Descriptive Statistics	13
1.3 Visualizations	19
1.4 Baseball	29
1.5 Football	29
1.6 Basketball	29
1.7 Soccer	29
1.8 Volleyball	29
1.9 Hockey	29
2 Probability	31
2.1 Definitions and Axioms	31
2.2 Theorems and Laws	31
2.3 Random Variables	31
3 Simulation	33
4 Statistical Inference	35
4.1 One Sample and Two Sample t-tests and confidence intervals . . .	35
5 Correlation	37
6 Linear Regression	39
7 Data Scraping	41
8 Principal Component Analysis	43
9 Clustering	45

10 Classification	47
11 Decision Trees	49
11.1 Random Forests	49
11.2 Gradient Boosting	49
12 Non-parametric Statistics	51
13 Baseball	53
14 Football	55
15 Basketball	57
16 Soccer	59
17 Hockey	61
18 Volleyball	63
18.1 Resources	63
19 Other Sports	65
20 Ellie's stuff	67
21 Levi's stuff	69
21.1 Hockey - ggplot	69
21.2 Law of Total Probability - Hockey	71
21.3 Sets and Conditional Probability	72
21.4 Binomial Probability	72
21.5 Binomial Coefficient Symmetry	73
21.6 Binomials and Multinomials	74
21.7 Geometric (First Success) RVs	74
21.8 Geometric Distribution - Hockey	75
22 Isaac's stuff	77
23 Aaron's stuff	79
23.1 Notes for Chapter 2 (Probability)	79
23.2 Suggested Readings	79
23.3 Notes for Chapter 4 (Simulation)	80
Reference: Blocks	81
23.4 Equations	81
23.5 Theorems and proofs	81
23.6 Callout blocks	81

<i>CONTENTS</i>	5
Reference: Footnotes and citations	83
23.7 Footnotes	83
23.8 Citations	83
24 References	85

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: “2022-06-01”

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 summary statistics from these datasets using dplyr, tidy
3. Describe and calculate useful team and individual (descriptive statistics).
Example: Baseball: calculate AVG, OBP, OPS, WOB
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 Getting Started With R

1.1.1 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 from 1998–2002 was 58.4? If so, you should be set up correctly!

1.1.2 Some R Basics

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 **summary**, **mean()**, **median()**, **var()**, and **sd()**.

```
summary(peyton)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      2249    4056    4693    4278    4914    5477
```

```
mean(peyton)
```

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

```
sd(peyton)
```

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

R allows you to install additional packages (collections of 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")
```

You will also need to know how to load datasets from files. For this class, we will typically provide data files in .csv format.

Here is how to load a file:

```
# load readr package and load example dataset
library(readr)
NFL_2021_Team_Passing <- read_csv("data/NFL_2021_Team_Passing.csv")

# we can look at the header (first few entries) using "head()"
head(NFL_2021_Team_Passing)
```

```
## # A tibble: 6 x 25
##      Rk Tm      G  Cmp  Att `Cmp%`  Yds  TD `TD%`  Int `Int%`  Lng
##   <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1     1 Tampa Bay~  17  492  731   67.3  5229   43   5.9   12   1.6   62
## 2     2 Los Angel~  17  443  674   65.7  4800   38   5.6   15   2.2   72
## 3     3 Dallas Co~  17  444  647   68.6  4800   40   6.2   11   1.7   73
## 4     4 Kansas Ci~  17  448  675   66.4  4791   37   5.5   13   1.9   75
## 5     5 Los Angel~  17  406  607   66.9  4642   41   6.8   18    3   79
## 6     6 Las Vegas~  17  429  628   68.3  4567   23   3.7   14   2.2   61
## # ... with 13 more variables: `Y/A` <dbl>, `AY/A` <dbl>, `Y/C` <dbl>,
## #   `Y/G` <dbl>, Rate <dbl>, Sk <dbl>, SKYds <dbl>, `Sk%` <dbl>, `NY/A` <dbl>,
## #   `ANY/A` <dbl>, `4QC` <dbl>, GWD <dbl>, EXP <dbl>
```

1.2 Descriptive Statistics

1.2.1 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. How could you define 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 1.2. 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 1.3. What are possible nominal and ordinal data associated with Peyton Manning?

1.2.2 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.

1.2.2.1 Descriptive Statistics for Quantitative Data

There are different descriptive statistics depending on the type of data you are analyzing. We will begin by looking at descriptive statistics for quantitative data.

To begin, let x_1, x_2, \dots, x_n represent a numerical dataset with 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: $\sum_{i=1}^n x_i = x_1 + x_2 + \dots + x_n$

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

Definition 1.11. The **population mean** (or population average), μ , is the mean value for the entire population.

The mean can be thought of as a measure of center or more generally, a measure of location.

Example 1.4. 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.

```
# Calculate the sample of Peyton Manning's passing yards season totals with Colts
peyton.broncos <- c(4659, 5477, 4727, 2249)
mean(peyton.broncos)
```

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

In sports statistics, we often have to choose between using a descriptive statistic that summarizes a quantity versus a descriptive statistic that summarizes a rate. For instance, in basketball, we can compare two players based on how many points they score in a game (total quantity) or we can compare two players based on how many points per minute played (rate statistic). Many applications in sports analytics focus more on rate statistics rather than quantity statistics. Why?

We can measure the spread or variability of a dataset using *variance* and *standard deviation*.

Definition 1.12. The **sample variance**, s^2 , of the numerical dataset is a measure of spread and is given by $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$

Definition 1.13. The *sample standard deviation*, s , of the numerical dataset is a measure of spread and is given by $s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$

Definition 1.14. The *population variance*, σ^2 , is the variance for an entire population.

Definition 1.15. The *population standard deviation*, σ , is the standard deviation for an entire population.

We often prefer to work with standard deviations as a measure of spread as opposed to variance because standard deviations are given in our original units.

```
# Calculate the variance and standard deviation of Peyton Manning's passing yards season totals
var(peyton.broncos) # units: yards^2
```

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

```
sd(peyton.broncos) # units: yards
```

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

Definition 1.16. The *sample median*, \tilde{x} , of a numerical dataset is the middle value when the data are ordered from smallest to largest. In other words, let x_1, x_2, \dots, x_n be the (unordered) dataset and let $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ be the same dataset but ordered from smallest to largest. If n is odd, then $\tilde{x} = x_{(n+1)/2}$ and if n is even, then $\tilde{x} = \frac{1}{2} \cdot [x_{(n/2)} + x_{(n/2+1)}]$.

Example 1.5. Calculate the sample median of Peyton Manning's season passing yards total while with the Colts (3739, 4135, 4413, 4131, 4200, 4267, 4557, 3747, 4397, 4040, 4002, 4500, 4700).

Like sample mean, sample median is a measure of center. It gives you an idea of where the “middle” of your dataset is.

We can calculate sample mean and sample median in R as follows:

```
# Calculate the median of Peyton Manning's passing yards season totals with Broncos and Colts
peyton.colts <- c(3739, 4135, 4413, 4131, 4200, 4267, 4557, 3747, 4397, 4040, 4002, 4500, 4700)
median(peyton.colts)
```

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

```
median(peyton.colts)
```

```
## [1] 4200
```


Definition 1.17. A *percentile* is a measure of relative standing. The p^{th} percentile is the number where at least $p\%$ of the data values are less than or equal to this number.

Definition 1.18. A *quantile* is a measure of relative standing and are the cut points for breaking a distribution of values into equal sized bins.

Definition 1.19. A *quartile* is a measure of relative standing and are the cut points for breaking a distribution of values into four equal parts.

Calculate the 10th and 90th percentile of Peyton Manning's passing yards season totals with Colts

```
quantile(peyton.colts,0.10)
```

```
## 10%
## 3798
```

```
quantile(peyton.colts,0.90)
```

```
## 90%
## 4545.6
```

```
quantile(peyton.colts,c(0.1,0.9))
```

```
## 10% 90%
## 3798.0 4545.6
```

Special percentiles:

1. 25th percentile = 1st quartile = Q_1
2. 50th percentile = 2nd quartile = $Q_2 = \tilde{x}$
3. 75th percentile = 3rd quartile = Q_3

Definition 1.20. *Range* is a measure of spread, measures the full width of a dataset, and is given by: $Range = Max - Min$.

Definition 1.21. *Interquartile range* is a measure of spread, measures the width of the middle 50% of a dataset, and is given by: $IQR = Q_3 - Q_1$.

Definition 1.22. A *five number summary* describes the center, spread, and edges of a dataset and is given by: $(Min, Q_1, Q_2, Q_3, max)$.

```
summary(peyton.colts)
```

```
## Min. 1st Qu. Median Mean 3rd Qu. Max.
## 3739 4040 4200 4218 4413 4700
```

```
quantile(peyton.colts,c(0,0.25,0.5,0.75,1))
```

```
## 0% 25% 50% 75% 100%
## 3739 4040 4200 4413 4700
```

1.2.2.2 Descriptive Statistics for Qualitative Data

In sports statistics, we also encounter qualitative (categorical) data which is names or labels which has its own descriptive statistics.

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

Definition 1.23. The *proportion* of sampled data that fall into a category is given by: $p = \frac{\# \text{ in category}}{\# \text{ total}}$

“Proportion” and “Probability” are often used interchangeably. Both have a minimum value of 0 and a maximum value of 1.

Definition 1.24. The *percentage* of sampled data that fall into a category is given by: $P\% = 100 \cdot p = 100 \cdot \frac{\# \text{ in category}}{\# \text{ total}}$

Percentages in this context can have a minimum value of 0% and a maximum value of 100%.

Example 1.6. In 2014, Peyton Manning started as quarterback for the Denver Broncos. The result of the Broncos’ 16-game season was:

Win, Win, Loss, Win, Win, Win, Win, Loss, Win, Loss, Win, Win, Win, Win, Loss, Win

Calculate the proportion and percentage of Broncos’ winning games in 2014.

```
broncos2014 <- c("Win", "Win", "Loss", "Win", "Win", "Win", "Win", "Loss", "Win", "Loss", "Win", "Win", "Win", "Win", "Loss", "Win")
broncos.prop <- sum(broncos2014 == "Win")/length(broncos2014); broncos.prop
```

```
## [1] 0.75
```

```
broncos.perc <- 100*broncos.prop; broncos.perc
```

```
## [1] 75
```

We can also build a frequency table that summarizes the categories and their occurrences using **table()** in R. Note that **table()** works for quantitative and qualitative data.

```
table(broncos2014)
```

```
## broncos2014
## Loss Win
##      4  12
```

1.3 Visualizations

Conveying information visually is also an important part in providing a description of a dataset.

R provides some basic plotting functions such as **plot**, **hist**, and **barplot**. These plotting functions are simple and not always very clean looking.

In this class, we will use analogous plotting functions in **ggplot2** that are much improved plotting functions.

If you have already installed the **tidyverse** package, it should have also installed the **ggplot2** package.

```
# You have likely already installed the tidyverse package but if not, use the following command
# install.packages("tidyverse")

# Load the tidyverse package (which includes ggplot2)
library(tidyverse)
```

```
## -- Attaching packages ----- tidyverse 1.3.1 --

## v ggplot2 3.3.6      v dplyr   1.0.9
## v tibble  3.1.7      v stringr 1.4.0
## v tidyr   1.2.0      v forcats 0.5.1
## v purrr   0.3.4

## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

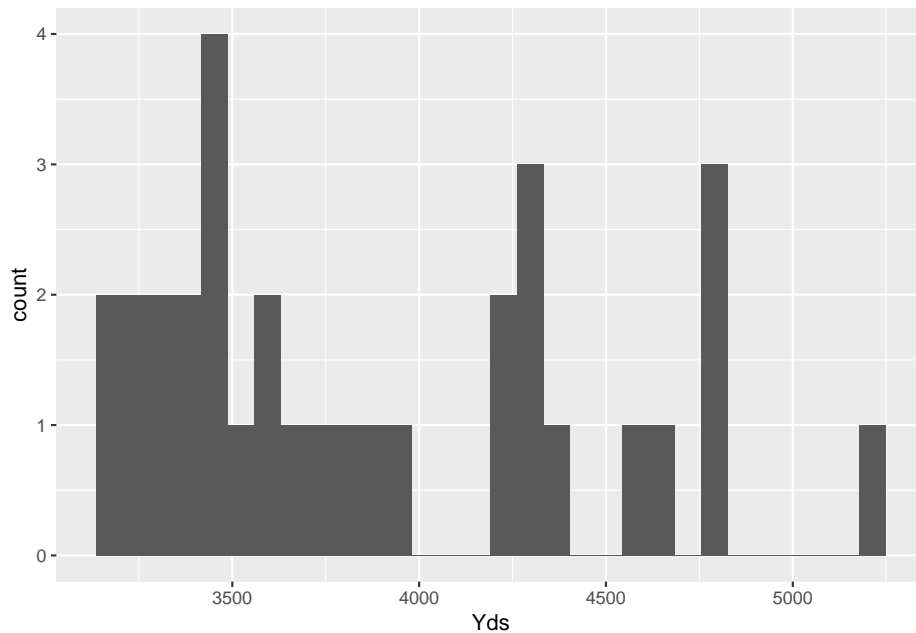
Let's load the file "NFL_2021_Team_Passing.csv" which contains NFL Team Passing Statistics, 2021

```
library(readr)
NFL_2021_Team_Passing <- read_csv("data/NFL_2021_Team_Passing.csv")
```

Histograms are one of the most common and basic ways to visualize a dataset's distribution of values. To make a histogram, you will use **ggplot** and **geom_histogram**.

Example 1.7. Create a histogram of the NFL Team Passing Yards in 2021.

```
NFL_2021_Team_Passing %>% ggplot(aes(x=Yds)) + geom_histogram()
```

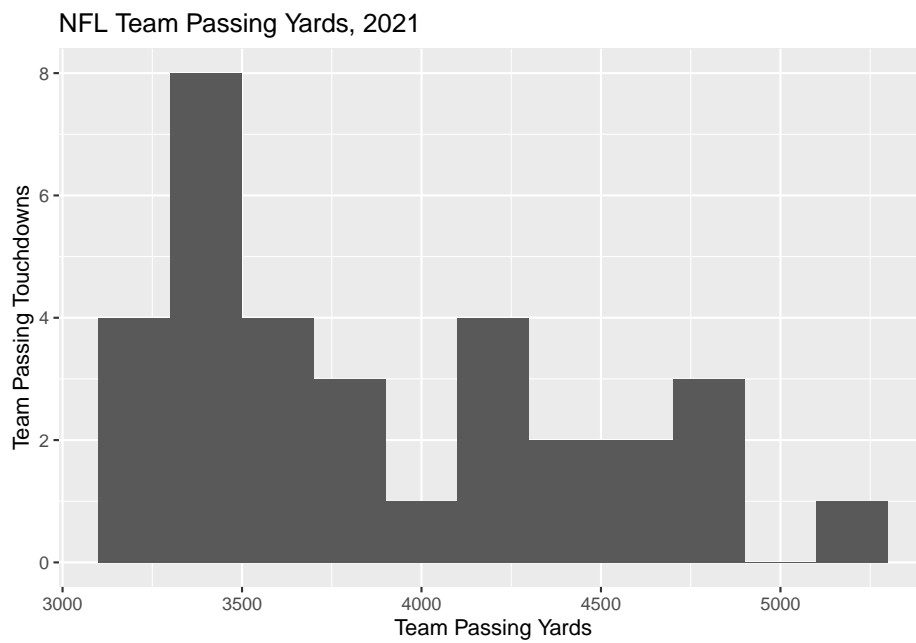


Notice how `%>%` is used to **pipe** the dataset into `ggplot`. This is using the pipe function from the **dplyr** package.

By default, `geom_histogram` uses 30 bins but this is customizable. Let's make the bins have a width of 200.

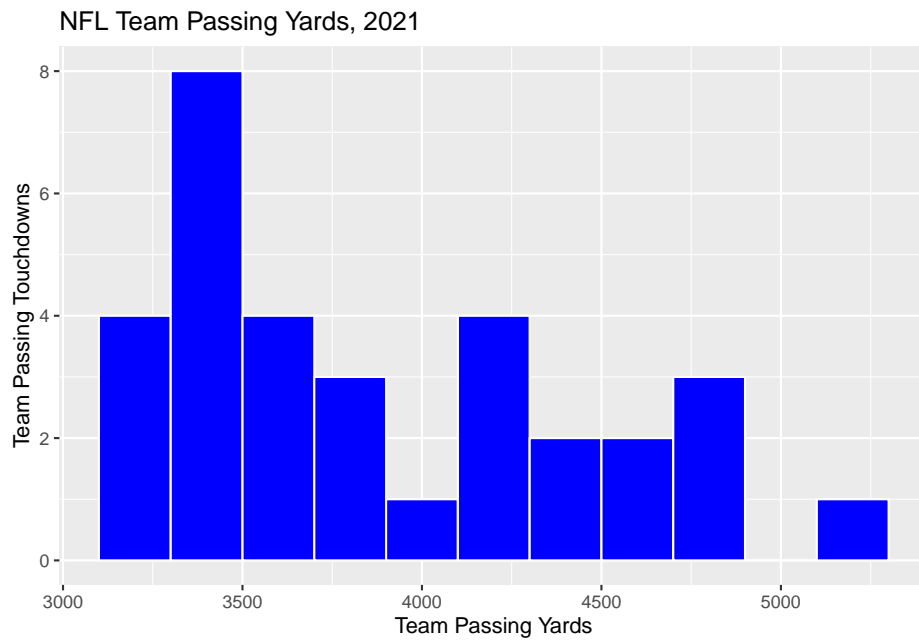
All good visualizations have good labels. Let's improve the axis labels and give the figure a title.

```
NFL_2021_Team_Passing %>% ggplot(aes(x=Yds)) +  
  geom_histogram(binwidth = 200) +  
  labs(x="Team Passing Yards", y="Team Passing Touchdowns", title="NFL Team Passing Yards")
```



We also have numerous options to change the appearance of plots when using **ggplot**. Let's change the bins color to *blue* and change the bin borders to *white*.

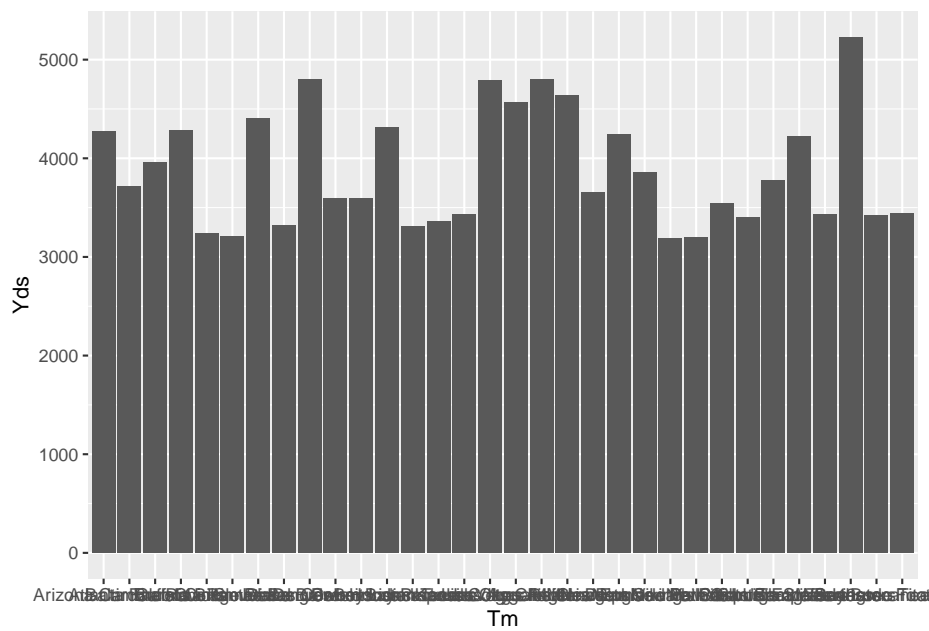
```
NFL_2021_Team_Passing %>% ggplot(aes(x=Yds)) +  
  geom_histogram(color = "white", fill = "blue", binwidth = 200) +  
  labs(x="Team Passing Yards",y="Team Passing Touchdowns",title="NFL Team Passing Yards, 2021")
```



We can also create bar plots using ggplot using the `geom_bar` function.

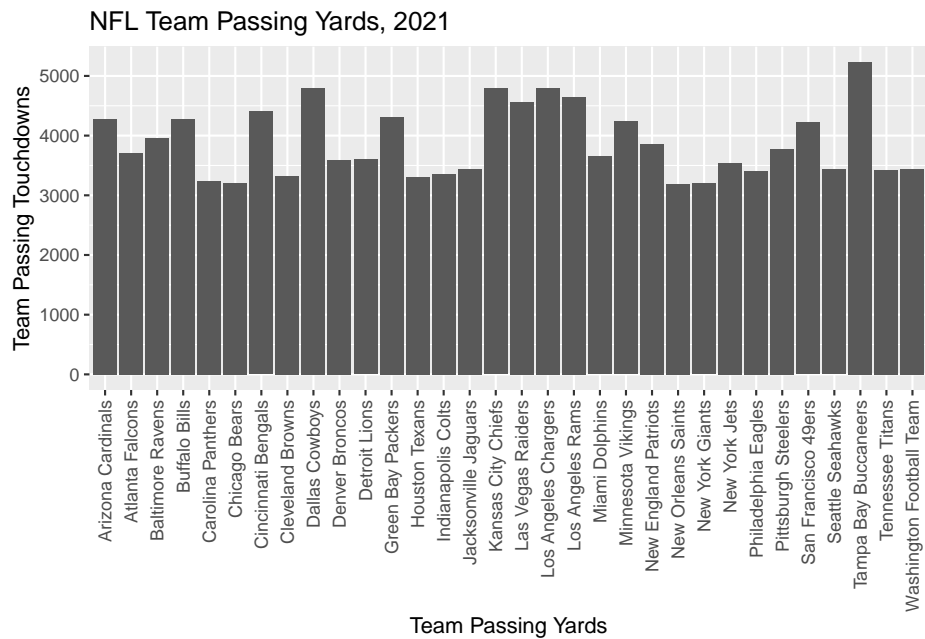
Example 1.8. Create a bar plot with teams on the horizontal axis and passing touchdowns on the vertical axis.

```
NFL_2021_Team_Passing %>% ggplot(aes(x=Tm,y=Yds)) +  
  geom_bar(stat="identity")
```



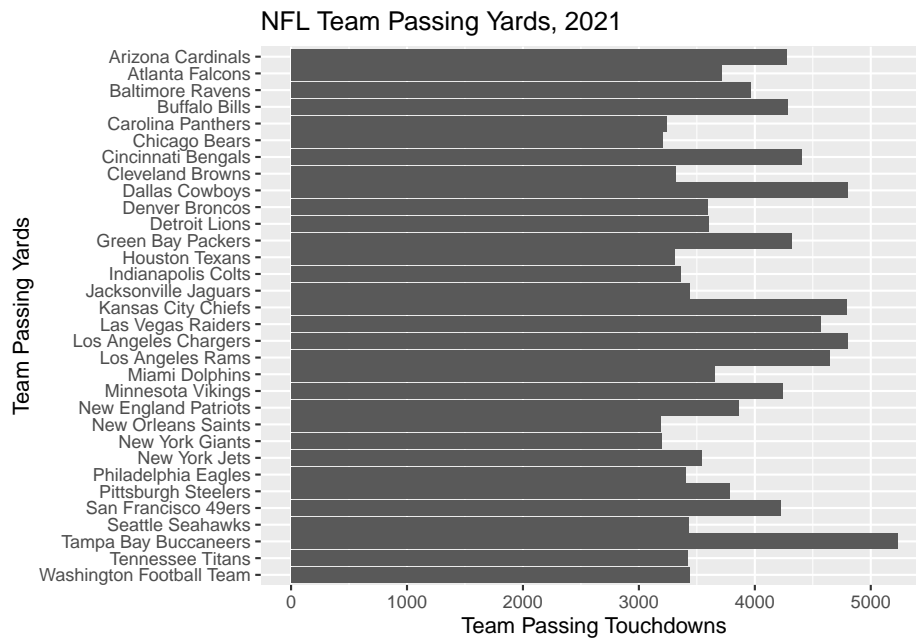
The team labels are a complete mess. Let's fix this and make some adjustments to the axis labels and figure title.

```
NFL_2021_Team_Passing %>% ggplot(aes(x=Tm,y=Yds)) +
  geom_bar(stat="identity") +
  labs(x="Team Passing Yards",y="Team Passing Touchdowns",title="NFL Team Passing Yards, 2021") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



We can flip this graph if we like as well. Note that when we flip the graph, our labels get in reverse ordering, so this can be fixed using `fct_rev()` which is part of the `forcats` package.

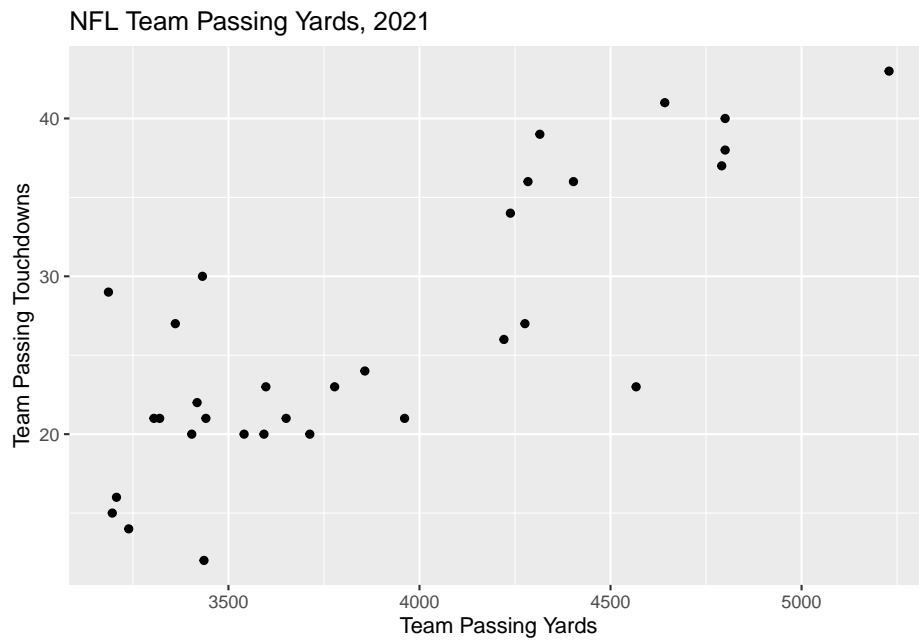
```
NFL_2021_Team_Passing %>%
  ggplot(aes(x=fct_rev(Tm),y=Yds)) +
  geom_bar(stat="identity") +
  labs(x="Team Passing Yards",y="Team Passing Touchdowns",title="NFL Team Passing Yards") +
  coord_flip()
```

Another common and useful visualization is a scatterplot which shows the relationship between two numeric variable. In ggplot, you use `geom_point()`.

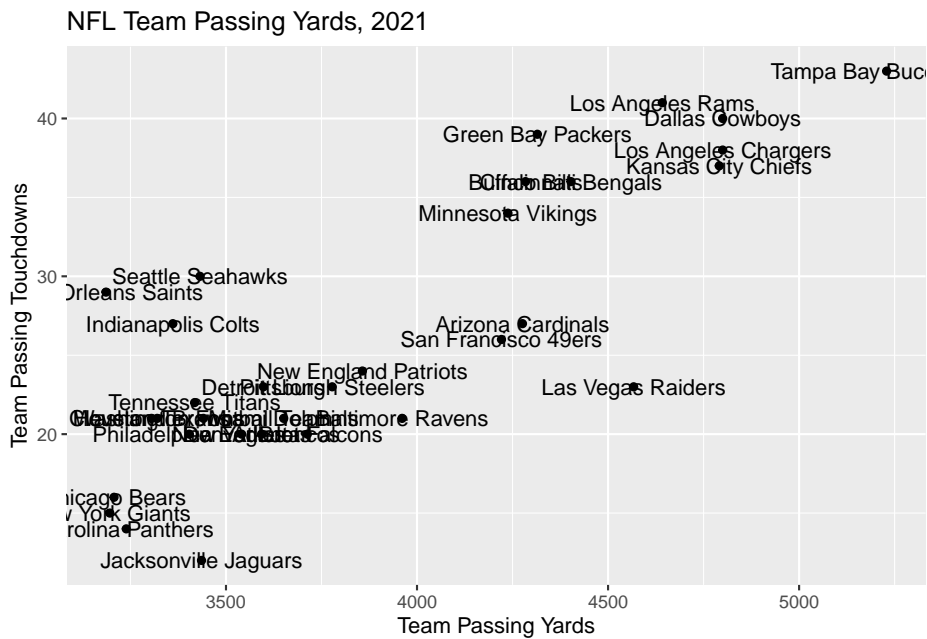
Example 1.9. Create a scatterplot of Team Passing Yards and Team Passing Touchdowns from the NFL 2021 dataset.

```
NFL_2021_Team_Passing %>%
  ggplot(aes(x=Yds,y=TD,label=Tm)) +
  geom_point() +
  labs(x="Team Passing Yards",y="Team Passing Touchdowns",title="NFL Team Passing Yards, 2021")
```



We may want to include team labels on this plot, however, it can get messy very quickly with a lot of points.

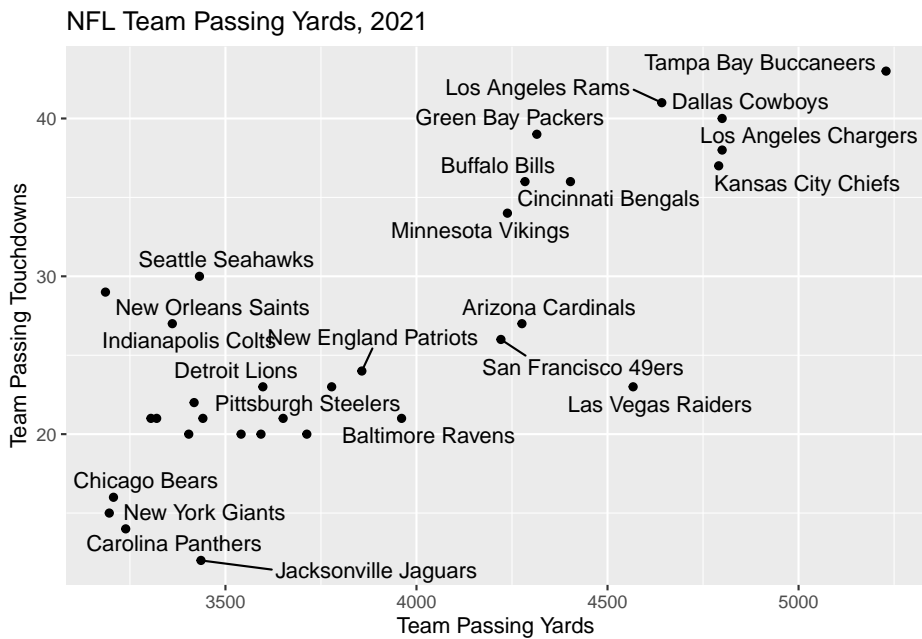
```
NFL_2021_Team_Passing %>%
  ggplot(aes(x=Yds,y=TD,label=Tm)) +
  geom_point() +
  labs(x="Team Passing Yards",y="Team Passing Touchdowns",title="NFL Team Passing Yards")
  geom_text()
```



Many sports leagues have around 30 teams, so a clean scatterplot with labels can be tricky to make. Here are some options below.

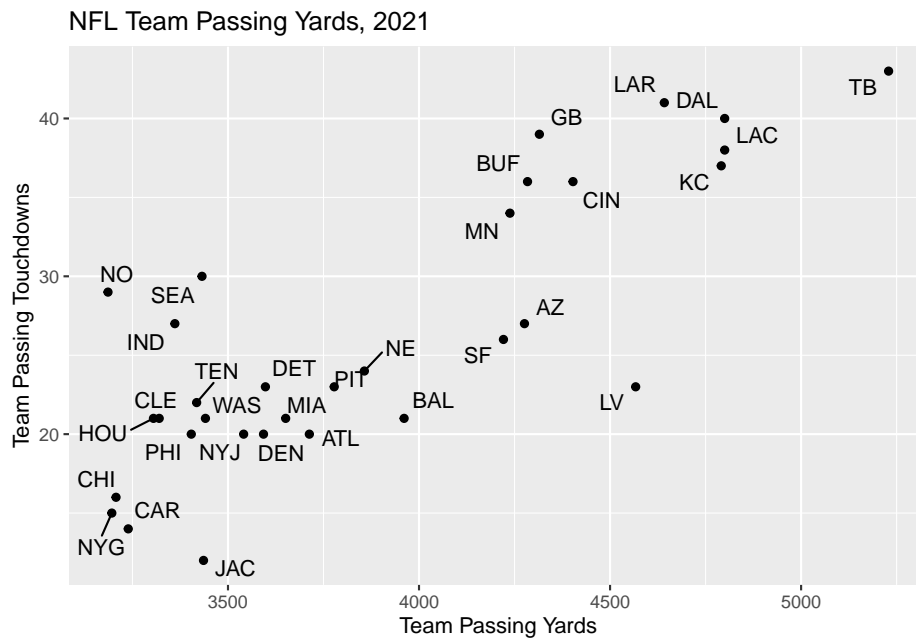
```
# install ggrepel package
library(ggrepel)
NFL_2021_Team_Passing %>%
  ggplot(aes(x=Yds,y=TD,label=Tm)) +
  geom_point() +
  labs(x="Team Passing Yards",y="Team Passing Touchdowns",title="NFL Team Passing Yards, 2021") +
  geom_text_repel()
```

```
## Warning: ggrepel: 9 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



```
NFL_2021_Team_Passing$Abbr <- c("TB","LAC","DAL","KC","LAR","LV","CIN","GB","BUF","AZ",
                                "BAL","NE","PIT","ATL","MIA","DET","DEN","NYJ","WAS","",
                                "TEN","PHI","IND","CLE","HOU","CAR","CHI","NYG","NO")

NFL_2021_Team_Passing %>%
  ggplot(aes(x=Yds,y=TD,label=Abbr)) +
  geom_point() +
  labs(x="Team Passing Yards",y="Team Passing Touchdowns",title="NFL Team Passing Yards") +
  geom_text_repel(box.padding = 0.3)
```



1.4 Baseball

1.5 Football

1.6 Basketball

1.7 Soccer

1.8 Volleyball

1.9 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

21.1 Hockey - ggplot

For this example, we'll use a set of NHL data from money puck.com. First, let's load the data into R and open the data frame.

```
nhl_2022_data <- read.csv("https://money puck.com/moneypuck/playerData/seasonSummary/2021/regular/")
View(nhl_2022_data)
```

This dataset includes a *lot* of covariates. It also splits these data by different game situations: even-strength (5 on 5), power play (5 on 4), etc. Let's subset the data to include all game situations.

Use the `nrow` command to check the number of columns in the new data frame. Check: Is it the same as the number of teams in the league for the 2021-2022 season?

```
nhl_data_all <- filter(nhl_2022_data, situation == "all")
View(nhl_data_all)

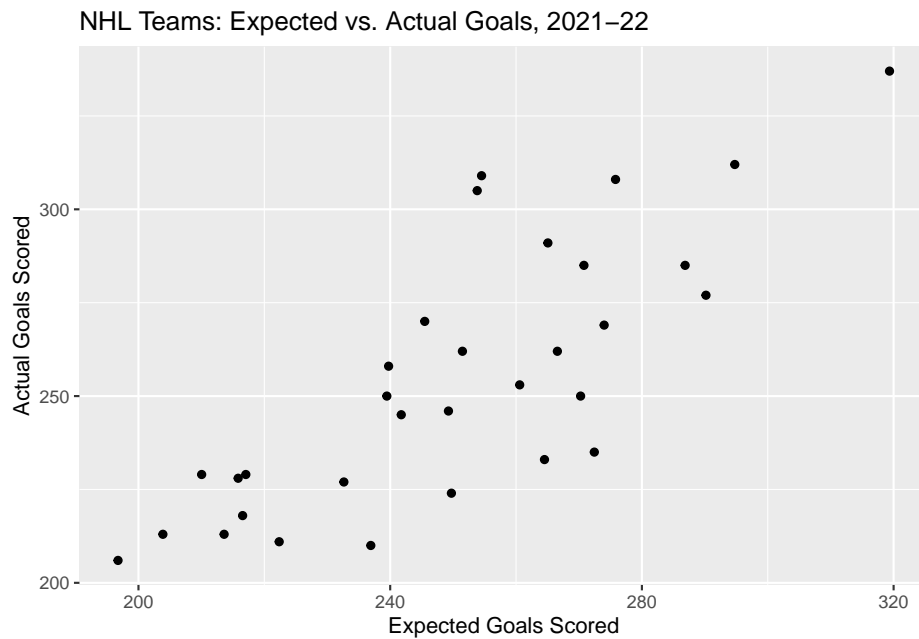
nrow(nhl_data_all)
```

```
## [1] 32
```

The dataset includes an Expected Goals statistic for each team in the `xGoalsFor` column. Let's plot this quantity against the team's actual number of goals scored; this is given by the `goalsFor` column.

(Remember to always have a good title and axis labels!)

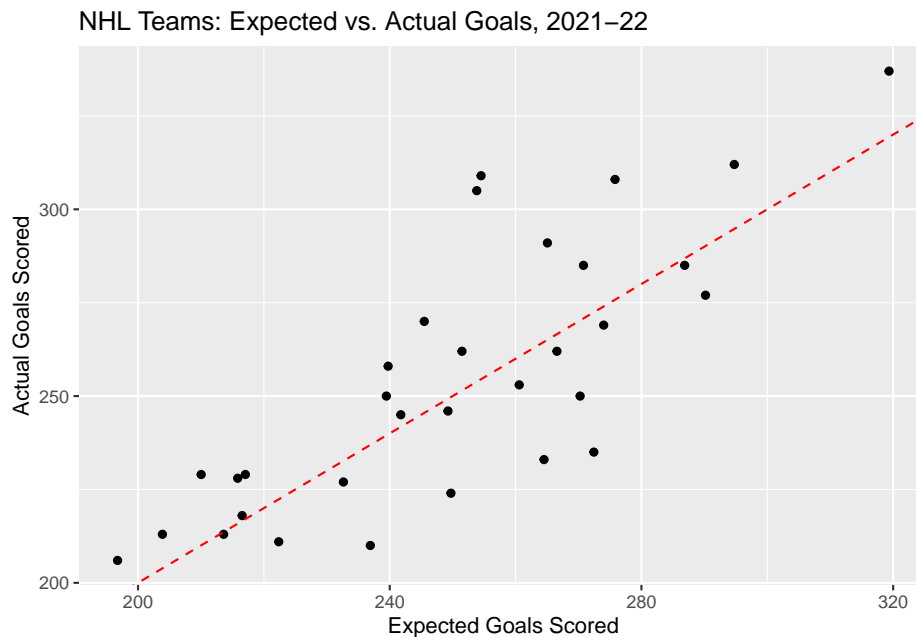
```
ggplot(data=nhl_data_all, aes(x=xGoalsFor, y=goalsFor)) + labs(x="Expected Goals Scored", y="Actual Goals Scored")
```



As expected, there is a general positive correlation between expected and actual goals ($r \approx 0.8$). However, there is some variability - for example, the Kings only scored 7 more actual goals than the Ducks, despite having 56.6 more expected goals.

Let's add a line to the graph using the `geom_abline` function corresponding to the line $y = x$, the line on which data points would fall if expected goals were equal to actual goals. We can also customize the line's color and type.

```
ggplot(data=nhl_data_all, aes(x=xGoalsFor, y=goalsFor)) + labs(x="Expected Goals Scored") + geom_abline(slope=1, intercept=0, color="red", linetype="dashed")
```



Note: A slope of 0 and an intercept of 1 are actually the default parameters for the function.

Q: What does it mean for a team's data point to fall below this line? Above it?

A: If the data point is below the line, it means the expected goals were greater than the actual goals; if the data point is above the line, it means the actual goals were greater than the expected goals.

Q: Do you think that a team's expected goals would be more likely to be closer to its actual goals for a ten-game stretch, an entire season, or five consecutive seasons? Why?

A: We would expect that as sample size increases, the result would become closer to expectation. So, actual goals would be most likely closer to expected goals over a span of five seasons.

21.2 Law of Total Probability - Hockey

Over the course of a season, a hockey player scored a goal 30% of the time during a home game, and $P(\text{player scores} \mid \text{away game}) = .18$. Assume all games are either home or away.

Q: What is the probability the player scored a goal in any game if there were an equal number of home and away games?

A: $P(\text{score}) = P(\text{score} \mid \text{home})P(\text{home}) + P(\text{score} \mid \text{away})P(\text{away}) = .3(.5) + .18(.5) = .24$

Q: What is the probability the player scored a goal in any game if there were twice as many home games as away games?

$$\text{A: } P(\text{score}) = P(\text{score}|\text{home})P(\text{home}) + P(\text{score}|\text{away})P(\text{away}) = .3\left(\frac{2}{3}\right) + .18\left(\frac{1}{3}\right) = .26$$

Q: What is the probability the player scored a goal in any game if the ratio of home games to away games is 2:3?

$$\text{A: } P(\text{score}) = P(\text{score}|\text{home})P(\text{home}) + P(\text{score}|\text{away})P(\text{away}) = .3\left(\frac{2}{5}\right) + .18\left(\frac{3}{5}\right) = .228$$

21.3 Sets and Conditional Probability

100 sports fans in Colorado were polled and it was found that 64 had attended either a Denver Nuggets or Colorado Avalanche game at Ball Arena (formerly Pepsi Center). 34 people had seen only a Nuggets game, while 17 had seen both a Nuggets and an Avalanche game.

Q: How many people saw an Avalanche game but not a Nuggets game?

$$\text{A: } 64 - 34 - 17 = 13$$

Q: What is the probability that a randomly selected person in the poll had been to a Nuggets game?

$$\text{A: } (34 + 17) / 100 = .51$$

Q: What is the probability that a randomly selected person that had been to a game at Ball Arena had been to a Nuggets game?

$$\text{A: } (34 + 17) / 64 = .797$$

Q: What is the probability that a randomly selected person had been to a Nuggets game given they had been to an Avalanche game?

$$\text{A: } 17 / (17 + 13) = .567$$

21.4 Binomial Probability

Two baseball teams are playing a 4-game series. The home team has a .65 probability of winning each game, and the away team a .35 probability. Assume each game is independent.

I used baseball in this example because it's the sport that most often has 4-game series, but it could easily be replaced by another sport.

Find the following probabilities.

- (a) The road team wins exactly 1 game.

$$\binom{4}{1} .65^3 .35^1 = \binom{4}{3} .65^3 .35^1 \approx .384$$

```
dbinom(1, 4, .35)
```

```
## [1] 0.384475
```

```
dbinom(3, 4, .65)
```

```
## [1] 0.384475
```

(b) The home team wins exactly 2 games.

$$\binom{4}{2} .65^2 .35^2 \approx .311$$

```
dbinom(2, 4, .65)
```

```
## [1] 0.3105375
```

```
dbinom(2, 4, .35)
```

```
## [1] 0.3105375
```

(c) The road team wins at least 2 games.

$$\binom{4}{2} .65^2 .35^2 + \binom{4}{3} .65^1 .35^3 + .35^4 = 1 - [.65^4 + \binom{4}{1} .65^3 .35^1] \approx .437$$

```
pbinom(1.9, 4, .35, lower.tail=F)
```

```
## [1] 0.4370187
```

```
pbinom(2, 4, .65, lower.tail=T)
```

```
## [1] 0.4370187
```

(d) The series ends in a sweep.

$$.65^4 + .35^4 \approx .194$$

```
dbinom(4, 4, .65) + dbinom(4, 4, .35)
```

```
## [1] 0.1935125
```

```
.65^4 + .35^4
```

```
## [1] 0.1935125
```

21.5 Binomial Coefficient Symmetry

Playoff series for a certain sports league are played as a best-of-seven series, with one team hosting four games and the opposing team hosing three. An executive for the league wishes to know the number of ways the home and away games can be assigned. (One such combination is A-A-B-B-A-B-A, the format used by the NBA and NHL for their best-of-seven series.) What is the total number of combinations?

Answer: Since there are a fixed number of games (seven) and a fixed number of games that must be given to the lower-seeded team (four), there are $\binom{7}{4} = \frac{7!}{4! \cdot (7-4)!} = 35$ ways to create a home-away pattern for the seven-game series.

However, instead of thinking about the number of ways to assign the games to the team that gets four home games, what if we thought about the number of ways to assign games to the team that gets three home games?

That would be $\binom{7}{3}$. We can use the `choose` command in R to find this quantity.

```
choose(7,3)
```

```
## [1] 35
```

It turns out that this binomial coefficient is also equal to 35.

Theorem: $\binom{n}{k} = \binom{n}{n-k}$

$$\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!}$$

$$\binom{n}{n-k} = \frac{n!}{(n-k)! \cdot (n-(n-k))!} = \frac{n!}{(n-k)! \cdot k!} = \binom{n}{k}$$

21.6 Binomials and Multinomials

Suppose we are curious about probabilities regarding the results of a soccer team's next five games.

Wait!!! A soccer game has three possible outcomes (win, lose, draw)! We can't use the binomial distribution, since it limits us to two possible outcomes!

It depends. If we are interested in the probability that a soccer team wins 2 of their next 5 games, we can use the binomial distribution. We can create the following partition of the sample space of outcomes: (Win) and (Win^C) , where the second set includes both losing and drawing.

Then, the formula would be represented as:

$$\binom{5}{2} P(Win)^2 P(Win^C)^{(5-2)}$$

If we are interested in the probability of the team winning two of the next five games, drawing two, and losing one, we cannot use the binomial theorem. That involves three outcomes, and would be represented as a multinomial.

21.7 Geometric (First Success) RVs

Caution: Some references parameterize the Geometric distribution based on the number of failures before the first success, rather than the trial on which the first success occurs. This changes the PMF, mean, and variance, so be careful.

```
set.seed(2022)
geometric <- rgeom(100, 1/3)
head(geometric, 20)
```

```
## [1] 2 5 1 3 12 7 1 4 2 2 1 1 1 2 0 0 0 4 3 0
```

Some of the values were 0, which could not happen if R was considering the number of the trial on which the first success occurred. You can add 1 to the values given by R to arrive at the First Success distribution.

```
first_success <- geometric + 1
head(first_success, 20)
```

```
## [1] 3 6 2 4 13 8 2 5 3 3 2 2 2 3 1 1 1 5 4 1
mean(first_success)
```

```
## [1] 3.03
```

The mean of this sample of variables is 3.03, which is close to the expected mean of $\frac{1}{p} = 3$.

21.8 Geometric Distribution - Hockey

Suppose the number of shots needed by a hockey team in order to score their first goal, X , is modeled by a Geometric($\frac{1}{10}$) random variable.

Q: What is the probability that it takes more than 3 shots to score the first goal?

A: $P(X > 3) = P(X = 4) + P(X = 5) + P(X = 6) + \dots$

This is an infinite series, so let's use the Law of Total Probability.

$$P(X > 3) = 1 - P(X \leq 3) = 1 - [P(X = 1) + P(X = 2) + P(X = 3)] = 1 - \left[\left(\frac{1}{10}\right) + \left(\frac{9}{10}\right)^1\left(\frac{1}{10}\right) + \left(\frac{9}{10}\right)^2\left(\frac{1}{10}\right)\right] = .729$$

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, \dots, 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)

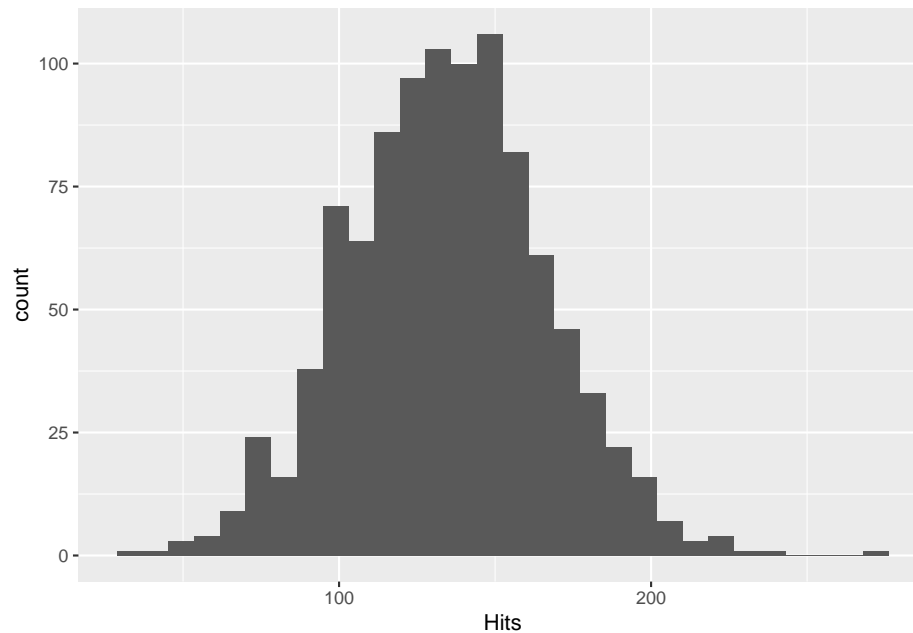
23.3.1 Baseball Simulation Example

```
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.

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

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>

¹This 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.