Chapter 2

- **In the needed packages we have not loaded *tibble,* but we are using it at Section 2.8, this will result in an error.**

Needed packages

Let's load all the packages needed for this chapter (this assumes you've already installed them). Read Section 1.3 for information on how to install and load R packages.

```
library(nycflights13)
library(ggplot2)
library(moderndive)
```

2.8 5NG#5: Barplots

Both histograms and boxplots are tools to visualize the distribution of numerical variables. Another commonly desired task is to visualize the distribution of a categorical variable. This is a simpler task, as we are simply counting different categories within a categorical variable, also known as the *levels* of the categorical variable. Often the best way to visualize these different counts, also known as *frequencies*, is with barplots (also called barcharts).

One complication, however, is how your data is represented. Is the categorical variable of interest "pre-counted" or not? For example, run the following code that manually creates two data frames representing a collection of fruit: 3 apples and 2 oranges.

```
fruits <- tibble(
  fruit = c("apple", "apple", "orange", "apple", "orange")
)
fruits_counted <- tibble(
  fruit = c("apple", "orange"),
  number = c(3, 2)
)
```

- **Two different names for the data frame.**

3.8.2 `rename` variables

Another useful function is `rename()`, which as you may have guessed changes the name of variables. Suppose we want to only focus on `dep_time` and `arr_time` and change `dep_time` and `arr_time` to be `departure_time` and `arrival_time` instead in the `flights_time` data frame:

```
flights_time_new <- flights %>%
   select(dep_time, arr_time) %>%
   rename(departure_time = dep_time, arrival_time = arr_time)
glimpse(flights_time_new)
```

Note that in this case we used a single `=` sign within the `rename()`. For example, `departure_time = dep_time` renames the `dep_time` variable to have the new name `departure_time`. This is because we are not testing for equality like we would using `==`. Instead we want to assign a new variable `departure_time` to have the same values as `dep_time` and then delete the variable `dep_time`. Note that new `dplyr` users often forget that the new variable name comes before the equal sign.

Chapter 4

- **In Chapter 4, LC4.3 the *airline_safety_smaller* data frame has 3 columns, but in the Appendix D, solution to learning checks, the *airline_safety_smaller* data frame has 7 columns.**

(LC4.3) Take a look at the `airline_safety` data frame included in the `fivethirtyeight` data package. Run the following:

```
airline_safety
```

After reading the help file by running `?airline_safety`, we see that `airline_safety` is a data frame containing information on different airline companies' safety records. This data was originally reported on the data journalism website, FiveThirtyEight.com, in Nate Silver's article, "Should Travelers Avoid Flying Airlines That Have Had Crashes in the Past?". Let's only consider the variables `airlines` and those relating to fatalities for simplicity:

```
airline_safety_smaller <- airline_safety %>%
  select(airline, starts_with("fatalities"))
airline_safety_smaller
```

```
# A tibble: 56 × 3
  airline              fatalities_85_99 fatalities_00_14
  <chr>                           <int>            <int>
1 Aer Lingus                          0                0
```

(LC4.3) Take a look the `airline_safety` data frame included in the `fivethirtyeight` data. Run the following:

```
airline_safety
```

After reading the help file by running `?airline_safety`, we see that `airline_safety` is a data frame containing information on different airlines companies' safety records. This data was originally reported on the data journalism website FiveThirtyEight.com in Nate Silver's article "Should Travelers Avoid Flying Airlines That Have Had Crashes in the Past?". Let's ignore the `incl_reg_subsidiaries` and `avail_seat_km_per_week` variables for simplicity:

```
airline_safety_smaller <- airline_safety %>%
  select(-c(incl_reg_subsidiaries, avail_seat_km_per_week))
airline_safety_smaller
```

```
# A tibble: 56 × 7
  airline              incidents_85_99 fatal…¹ fatal…² incid…³ fatal…⁴ fatal…⁵
  <chr>                          <int>   <int>   <int>   <int>   <int>   <int>
1 Aer Lingus                         2       0       0       0       0       0
```

Chapter 5

- **In Chapter 5, at figure 5.3, the title of the chart is different from the title in the code.**

```
ggplot(evals_ch5, aes(x = bty_avg, y = score)) +
  geom_jitter() +
  labs(x = "Beauty Score", y = "Teaching Score",
       title = "Scatterplot of relationship of teaching and beauty scores")
```

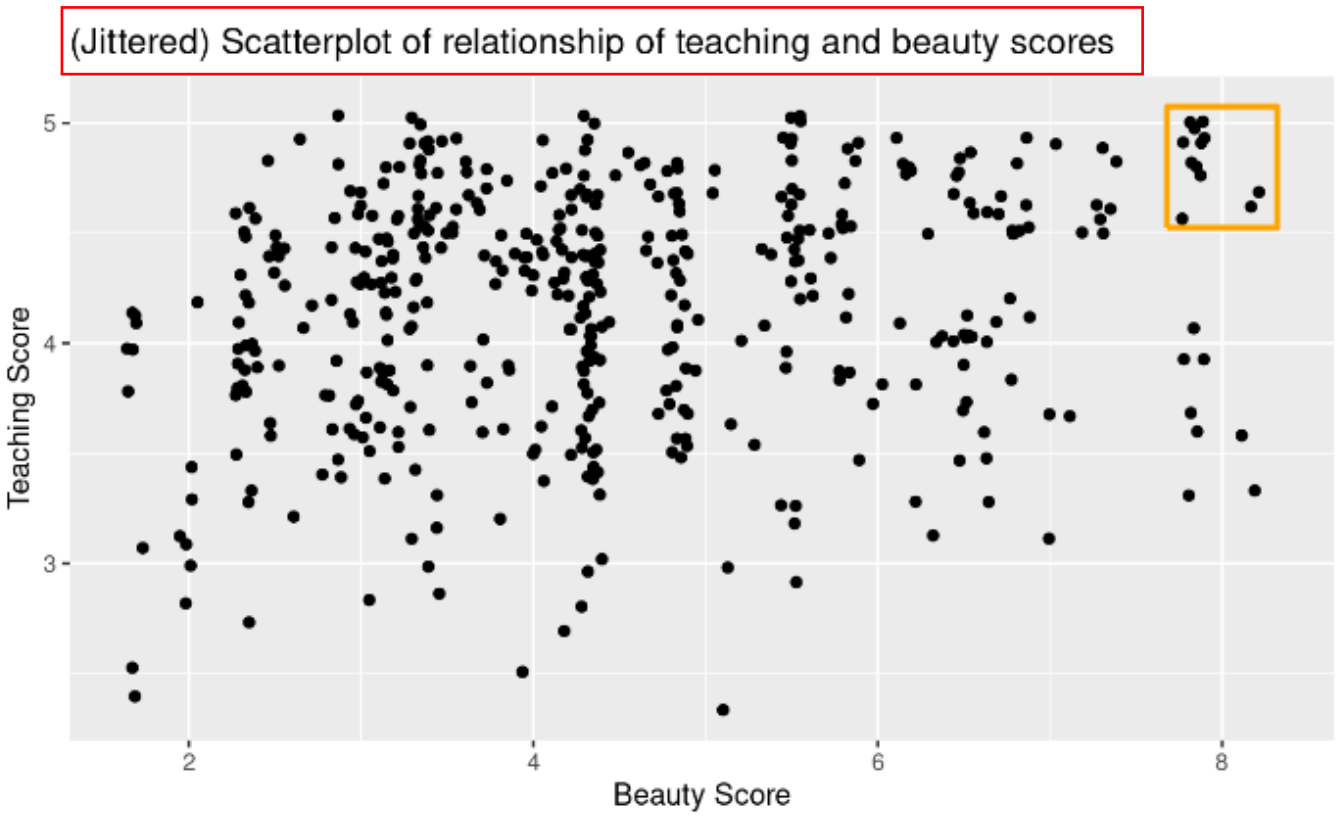

FIGURE 5.3: Instructor evaluation scores at UT Austin.

Chapter 6

- **We can add a reason why we used *as_tibble()* in the code when we can create a subset without it.**

6.2.1 Exploratory data analysis

Let's load the `Credit` dataset. To keep things simple let's `select()` the subset of the variables we'll consider in this chapter, and save this data in the new data frame `credit_ch6`. Notice our slightly different use of the `select()` verb here than we introduced in Subsection 3.8.1. For example, we'll select the `Balance` variable from `Credit` but then save it with a new variable name `debt`. We do this because here the term "debt" is easier to interpret than "balance."

```
library(ISLR)
credit_ch6 <- Credit %>% as_tibble() %>%
  select(ID, debt = Balance, credit_limit = Limit,
         income = Income, credit_rating = Rating, age = Age)
```

- **In the highlighted part, "we can apply the get_regression_points() function <u>on</u> our saved model,"**

6.3.2 Model selection using R-squared

At the end of the previous section in Figure 6.8 you compared an interaction model with a parallel slopes model, where both models attempted to explain y = the average math SAT score for various high schools in Massachusetts. In Tables 6.12 and 6.13, we observed that the interaction model was "more complex" in that the regression table had 6 rows versus the 4 rows of the parallel slopes model.

Most importantly however, when comparing the left and right-hand plots of Figure 6.8, we observed that the three lines corresponding to small, medium, and large high schools were not that different. Given this similarity, we stated it could be argued that the "simpler" parallel slopes model should be favored.

In this section, we'll mimic the model selection we just performed using the qualitative "eyeball test", but this time using a numerical and quantitative approach. Specifically, we'll use the R^2 summary statistic (pronounced "R-squared"), also called the "coefficient of determination". But first, we must introduce one new concept: the *variance* of a numerical variable.

We've previously studied two summary statistics of the *spread* (or *variation*) of a numerical variable: the standard deviation when studying the normal distribution in A.2 and the interquartile range (IQR) when studying boxplots in Section 2.7.1. We now introduce a third summary statistic of spread: the *variance*. The variance is merely the standard deviation squared and it can be computed in R using the `var()` summary function within `summarize()`. If you would like to see the formula, see A.1.3.

Recall that to get: 1) the observed values y, 2) the fitted values $\hat{y}$ from a regression model, and 3) the resulting residuals $y - \hat{y}$, we can apply the `get_regression_points()` function our saved model, in this case `model_2_interaction`:

- **In figure 6.11, the title is misspelled, it should be <u>credit card</u>**

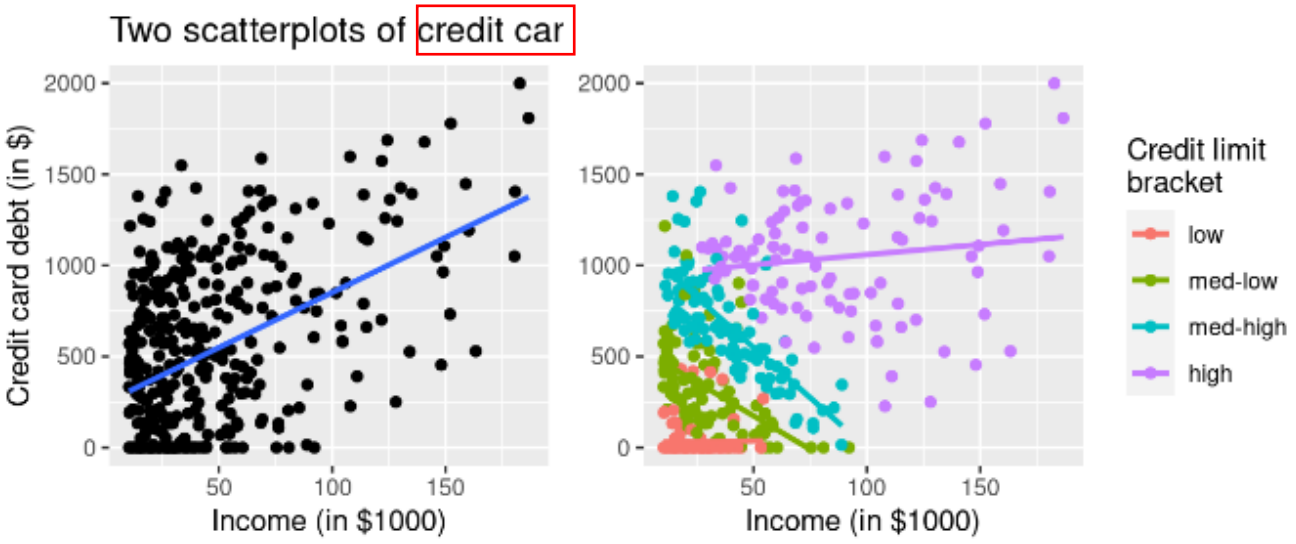

FIGURE 6.11: Relationship between credit card debt and income by credit limit bracket.

Chapter 8

- **In the highlighted part it is said that we have three pennies with the year 1999, but in figure 8.2, we have only one penny for 1999. Also, in the *pennies_sample* data frame and figure 8.2 we have 19 unique years not 26 as highlighted below.**

TABLE 8.1: Scenarios of sampling for inference

Scenario	Population parameter	Notation	Point estimate	Symbol(s)
1	Population proportion	p	Sample proportion	$\hat{p}$
2	Population mean	μ	Sample mean	$\overline{x}$ or $\hat{\mu}$

Going back to our 50 sampled pennies in Figure 8.2, the point estimate of interest is the sample mean $\overline{x}$ of 1995.44. This quantity is an *estimate* of the population mean year of *all* US pennies μ.

Recall that we also saw in Chapter 7 that such estimates are prone to *sampling variation*. For example, in this particular sample in Figure 8.2, we observed three pennies with the year 1999. If we sampled another 50 pennies, would we observe exactly three pennies with the year 1999 again? More than likely not. We might observe none, one, two, or maybe even all 50! The same can be said for the other 26 unique years that are represented in our sample of 50 pennies.

FIGURE 8.2: 50 US pennies labelled.

- **In Appendix D, learning check 8.2, misspelling of <u>and</u>.**

(LC8.2) Looking at the bootstrap distribution for the sample mean in Figure 8.14, between what two values would you say *most* values lie?

Solution:

Most values lie in 1990 amd 2000.

- **In Appendix D, learning check 8.4, in the highlighted part, there should be <u>68% confidence interval</u> instead of 95%.**

(LC8.4) Say we wanted to construct a 68% confidence interval instead of a 95% confidence interval for μ. Describe what changes are needed to make this happen. Hint: we suggest you look at Appendix A.2 on the normal distribution.

Solution:

Thus, using our 68% rule of thumb about normal distributions from Appendix A.2, we can use the following formula to determine the lower and upper endpoints of a 95% confidence interval for μ:

$$\overline{x} \pm 1 \cdot SE = (\overline{x} - 1 \cdot SE, \overline{x} + 1 \cdot SE)$$

- **In the following figure 8.27, we can use some different color instead of Grey color to make it more visible and to distinguish from the black lines.**

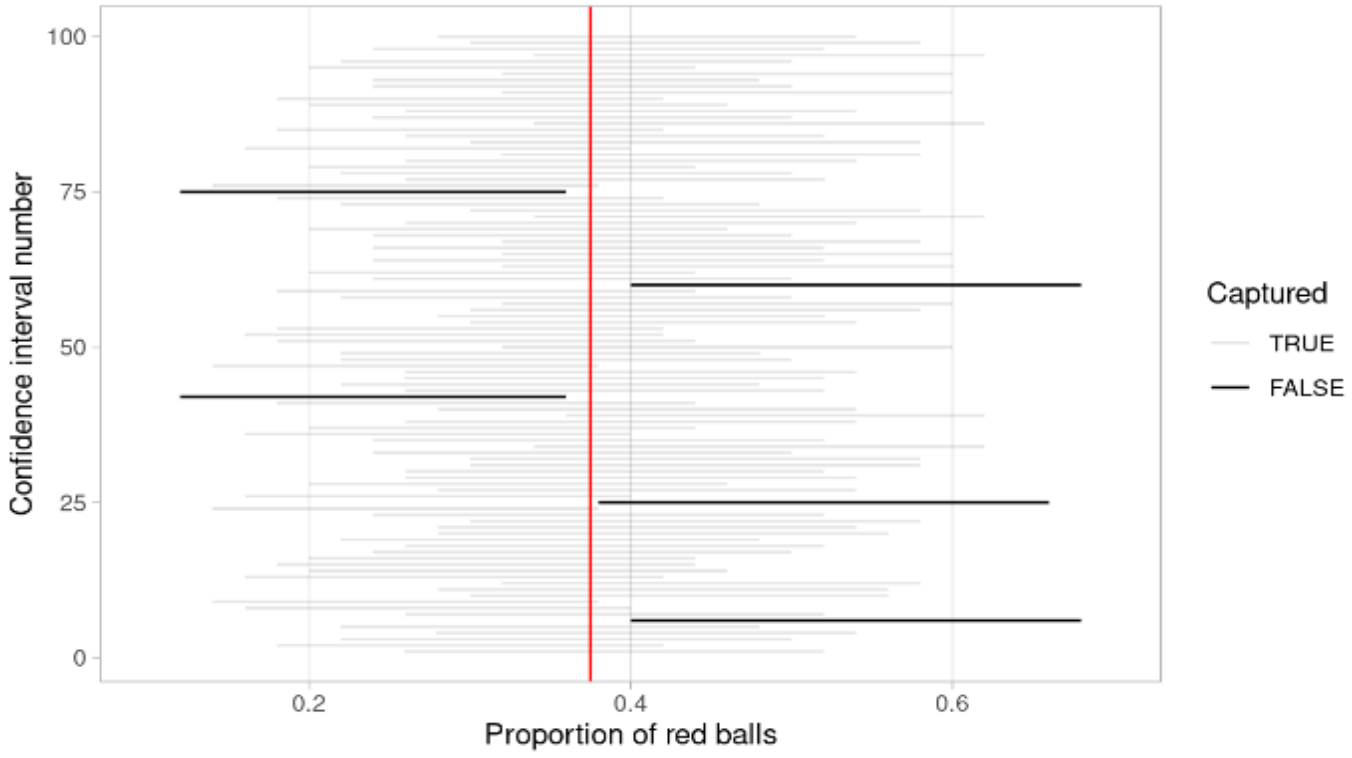

FIGURE 8.27: 100 percentile-based 95% confidence intervals for p.

- **In Chapter 8, after table 8.4, in the highlighted part, the result is coming negative, so in order to correct it, we can change the values in the numerator by 10 and 4 making the fraction 10/34 and 4/16 respectively. (Since we are looking for the difference in proportion of people who yawned in both seed and control group, so the fraction should be 10/34 and 4/16)**

TABLE 8.4: Scenarios of sampling for inference

Scenario	Population parameter	Notation	Point estimate	Symbol(s)
1	Population proportion	p	Sample proportion	$\hat{p}$
2	Population mean	μ	Sample mean	$\overline{x}$ or $\widehat{\mu}$
3	Difference in population proportions	$p_1 - p_2$	Difference in sample proportions	$\hat{p}_1 - \hat{p}_2$

This is known as a *two-sample* inference situation since we have two separate samples. Based on their two-samples of size n_{seed} = 34 and $n_{control}$ = 16, the point estimate is

$$\hat{p}_{seed} - \hat{p}_{control} = \frac{24}{34} - \frac{12}{16} = 0.04411765 \approx 4.4\%$$