PS2

BSS Stat 20

2022-06-26

NOTICE

This problem set is worth zero points; you do not have to turn it in. However, if you want to assess yourself fairly, go ahead and type your answers in an RMarkdown file, knit and save to an html or pdf. Feel free to collaborate or use any kind of help while completing this assignment.

This notice will be pasted at the beginning of every problem set.

Questions - Longform

Question 1

We will continue working with the contingency table from **Problem Set 1, Question 3:** that is, the table from `UCBAdmissions` dataset included in R. The entire dataset gives us admissions statisics in 1973 for six of Berkeley's top departments. Here is the contingency table for "Dept. A." that we were working with last week.

	Reported Gender	
Admit	Male	Female
Admitted	512	89
Rejected	313	19

---

part a

Among students who were admitted to Department A, what proportion were reported as female?

part b

Among students who were rejected from Department A, what proportion were reported as male?

part c

The following code creates a data frame out of the Department A data, `DeptA`, that can be used to complete this problem. It uses the `rep()` function, which comes with the base installation of R and is quite useful.

```
Admission_Status <- c( rep("Yes", times = 512 + 89),
                       rep("No", times = 313 + 19))

Reported_Gender <- c( rep("Male", times = 512 + 313),
                      rep("No", times = 89 + 19))

DeptA <- tibble(Admission_Status, Reported_Gender)
```

With the `DeptA` data set, create a stacked, normalized bar chart with admission status on the horizontal axis and each bar split by reported gender.

part d

Identify where the proportions you found in **part a** and **part b** are represented on the bar chart.

Question 2

We will continue to work with the data set from **Problem Set 1, Question 5**. This data was borne of the 2016 study published by Berkeley faculty titled *Student evaluations of teaching(mostly) do not measure teaching effectiveness*

In this publication, the authors argued that students' evaluations are biased by their own performance in the class and more notably, the gender of the instructor. The data can be found here.

The relevant background for this data-set can be found in the Problem Set 1 write-up.

You can load in the data and assign it to the object `SET` by running these lines:

```
library(tidyverse)
SET <- read_csv("https://www.dropbox.com/s/jog3lnqjinabe9s/set.csv?dl=1", show_col_types = FALSE)
```

You can think of the column `ta_gender_id` representing the gender that the TA actually identifies as ("*Actual*") and `ta_gender` as the gender that they presented themselves ("*Presented*") as to the class.

---

part a

Since all assignments were returned at the same time for each section of the course, regardless of each section's *Actual-Presented* combination of teachers, *we should expect that the distribution of the* ***`prompt`*** *variable for each combination should be roughly the same.*

See if this is the case by plotting a faceted bar chart (each combination has a subplot) of the distribution of `prompt` scores. You should have four different subplots in the end. Then make a conclusion as to what extent the above italicized assertion is true.

---

part b

Say that I wanted to visualize distributions of multiple SET sub-scores (i.e, distributions of the `prompt` score and the `responsive` scores) on the same plot. What property of the data as it is currently formatted prevents this?

part c

Reconfigure the data so that you can visualize the `prompt` score and `responsive` score on the same plot.

part d

Now, trim the data-set to just include rows where `ta_gender_id` (Actual gender) is "female". Create an overlayed density plot of the distribution of SET scores on score type (`prompt` or `responsive`).

Question 3

The hometown Golden State Warriors won another NBA Championship recently with their victory against the Boston Celtics. Some would say that the Warriors recent run of success is all the more impressive considering that the league that they play in during the season and during the playoffs, the Western Conference, has generally had the more talented teams than its counterpart, the Eastern Conference. However, the Eastern Conference has gotten much stronger in recent years and arguably now has the better players among the two leagues. How did the two conferences fare in this year's playoffs?

To think about this question, we can pull individual player per game statistics during the 2022 NBA playoffs from Basketball Reference (linked here). We can read the data in with the following code:

```
library(rvest)
url <- "https://www.basketball-reference.com/playoffs/NBA_2022_per_game.html"

NBA <- (read_html(url) %>% html_table())[[1]] %>% filter(Tm != "Tm")
```

The following line of code also adds a `Conference` column which takes the values of `Eastern` and `Western` depending on the team which each player is a part of.

```
NBA <- NBA %>% mutate(Conference =
                 ifelse(Tm %in% c("ATL","TOR", "MIA", "BOS", "CHI",
                                  "BRK", "MIL", "PHI"), "Eastern",
                                   "Western"))
```

Let's look at a few of Dean Oliver's famous "Four Factors" to see the state of parity between the two conferences. We will examine turnovers per game `TOV` and effective field goal percentage `eFG%`. The latter statistic accounts for the fact that a 3-point shot is worth more than a 2-point shot.

---

part a

First, trim the data to only include players who have started more than 67 percent of the games they played in. (Use the provided link to find the right columns to perform the trimming on if you are unsure just by looking at the data set which columns to work with). Save the result.

part b

Now, return a data frame with summary statistics for turnovers per game, by conference. Include the mean, median, IQR, standard deviation and median absolute deviation.

part c

Create a boxplot(s) of turnovers per game, separated by conference.

part d

Based off your results of *part b* and *part c*, what is your conclusion about the distributions of turnovers for game for players in the Eastern Conference versus those in the Western conference?

part e

Repeat parts b, c, and d for effective field goal percentage. **NOTE: you will need to surround `eFG%` with the back-tick when accessing it in your pipeline.** This is because `%` is a special character in R.

Questions - Shorter

Question 4

Imagine that we get 100 people in a room and ask them to replicate a painting. Each person does the best they can, and then judges assign a score from 0 to 10 for their efforts, with 0 being the worst and 10 being awarded to a perfect copy. Theorize about the shape of the distribution of drawing scores. What might it look like if:

part a

The painting is very easy to replicate for most people.

part b

The painting is of a normal difficulty to replicate for most people.

part c

The painting is very hard to replicate for most people.

part d

The painting is hard to replicate for some, easy to replicate for some, and of normal difficulty to replicate for some.

Now, assume the painting is of normal difficulty to replicate and assume coming in that students have some level of variation when it comes to previous painting skills.

part e

What happens to the shape of the distribution of scores if it turns out most people have very similar painting expertise?

part f

What happens to the shape of the distribution if it turns out people have even higher levels of variation of painting expertise than we originally thought?

Question 5 - True or False

Make sure you give an explanation as to your decision.

part a

True or False: A dot plot can be thought of as a type of histogram.

part b

True or False: For describing a roughly uniform distribution, the mode is a useful measure of center.

part c

True or False: `NA` values in a column do not affect calculations on that column, such as `mean()`, `sum()`, ...

part d

True or False: For determining whether there are outliers in the distribution of one numerical variable, the boxplot is preferred over a histogram.