PS1

BSS Stat 20

2022-06-14

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

For this week, try not to use the `%>%` operator where it might be applicable.

Question 1

This first question will center on the `iris` data set, which is included in R. You can access it just by typing `iris`. However, it might be useful for you to assign it to an object for the purposes of this question.

part a

What does each row in the data frame correspond to?

part b

Identify each variable in the data frame according to the Taxonomy of Data.

part c

Consider how you might take the data of `Sepal.Length` and transform it into a new, *categorical* variable. What would the categories be? Would the variable be nominal or ordinal?

Question 2

Above is an excerpt of the `Cars93` data set, which is included in the MASS library. Here's a quick way to get a look at it.

	Manufacturer	Type	Min.Price	Max.Price
66	Nissan	Van	16.7	21.5
57	Mazda	Sporty	32.5	32.5
79	Saturn	Small	9.2	12.9
75	Pontiac	Sporty	14.0	21.4
41	Honda	Sporty	17.0	22.7
85	Toyota	Subcompact	14.2	22.6
71	Oldsmobile	Large	17	21.9
19	Chevrolet	Sporty	34.6	41.5
3	Audi	Compact	25.9	32.3
38	Ford	Large	20.1	21.7

```
library(MASS)
Cars93 <- Cars93
```

The data includes 93 cars that were sold in 1993. However, the excerpt has been tampered with. Some entries have been fabricated. We want to find which entries have been fabricated , *without using the brute force method of searching through the data set.* This is because: what if our data is much more than 93 observations?

part a

Which entry in the `Type` column has been fabricated and why?

part b

Which entry in the `MPG.city` column has been fabricated and why?

Question 3

One of the nice things that comes with attending a prestigious school (Berkeley) like you guys are is getting a free dataset in R! The `UCBAdmissions` dataset is a set of *contigency tables* which give us admissions statstics in 1973 for six of Berkeley's top departments. Here is the contigency table for "Dept. A."

	Reported Gender	
Admit	Male	Female
Admitted	512	89
Rejected	313	19

part a

What are the variables in this data set? Identify their types in the Taxonomy of Data.

part b

What percentage of students were admitted to Department A?

part c

What percentage of students were reported as female?

part d

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

part e

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

Question 4

To assess yourself properly for this question, make sure you knit your document.

One of the probability distributions we will study later on is known as the Normal distribution. It is ubiquitous. To generate 100 observations coming from the Normal distribution, we can run the following:

```
rnorm(n = 100)
```

Go ahead and do this and assign the result to an object. Keep your code for this around.

---

part a

What type of object do we get out?

part b

Calculate the sum of this vector.

Now generate another 100 observations coming from the Normal distribution and assign this to a different vector.

part c

Calculate the sum of this new vector.

part d

Write whether the sum of the first vector is larger or smaller than that of the second.

part e

How do your know your answer to **part d** is correct?

part f

Paste the code you wrote from before **part a** and write an R comment above the line of code which ensured your correctness (You do not have to evaluate the code).

Question 5

Teaching evaluations are often seen and used as a tool to measure teacher performance. They might also influence decisions regarding promotions. However, a study published in 2016 by Berkeley faculty titled *Student evaluations of teaching(mostly) do not measure teaching effectiveness* turns this notion on its head. The authors argue that students' evaluations are biased by their own performance in the class and more notably, the gender of the instructor.

part a

If we wanted to use data to answer the question: *"Does the gender of the instructor affect student evaluations of teaching?"*, what might be the unit of observation? What variables would we need to collect data on for our unit of observation?

The authors performed part of their results on a data set (click here) containing student evaluations of TAs (one identifying female, one identifying as male) for an online course in the United States. There were four sections of the course, so two per TA.

Since the course was online, the students never saw what the TAs actually looked like. Therefore, the female TA taught one of her sections as herself and the other as her partner TA; it was vice versa for the male TA. Student evaluation scores were received in all four sections; they were given in different aspects of teaching and then summed to give a total overall score.

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 and `ta_gender` as the gender that they presented themselves as to the class.

Also, the following landing page for the data (click here) may prove useful to you in a moment.

---

part b

What does each row in the data frame correspond to? What are the dimensions of the data frame?

part c

What is the variable type of `student_gender` according to the Taxonomy of Data?

The scores in this data set, including the `overall` score, range from 1 to 5.

part d

If you were concerned with visualizing the distribution of `overall` SET scores (with a bar chart, perhaps) what variable type in the Taxonomy of Data would suit `overall` best?

part e

If you were concerned with comparing the average (mean) of `overall` SET scores when split by students' identified gender, what variable type in the Taxonomy of Data would suit `overall` best?

---

part f

Note that the `age` column seems to be given by birth year. Create a new age variable that gives the actual age (in years) of the student at the time when the data was published.