STA130H1F – Fall 2021

Module 5 Problem Set

L. Bolton & S. Caetano – Sample Solutions

Instructions

How do I hand in these problems for the 11:59 a.m. ET, October 21 deadline?

Your complete .Rmd file that you create for this problem set AND the resulting .pdf (i.e., the one you 'Knit to PDF' from your .Rmd file) must be uploaded into a Quercus assignment (link: https://q.utoronto.ca/courses/235890/assignments/715113) by 11:59 a.m. ET, on October 21. Late problem sets or problems submitted another way (e.g., by email) are *not* accepted.

Problem set grading

There are two parts to your problem set. One is largely R-based with short written answers and the other is more focused on communication We recommend you use a word processing software like Microsoft Word to check for grammar errors in your written work. Note: there can be issues copying from Word to R Markdown so it may be easier to write in this file first and then copy the text to Word. Then you can make any changes flagged in Word directly in this file.

Part 1

[Set up]

- Load the `tidyverse` package, and
- Load and save the two dataset you need for this problem set,
  - `ps5_sample_data.csv` (call it `orig_sample`), and
  - `ps5_census_data.csv` (call it `census`).

```
# message = FALSE prevents the info tidyverse gives when loaded from being printed

# Load the tidyverse package

# Load and save ps5_sample_data.csv (call it orig_sample)

# Load and save ps5_census_data.csv (call it census)
```

[Question 1 — Teaching world]

Dataset 1 (census) is Suppose that **Dataset 1 (census)** is a complete census (survey of the entire population) of people aged 18 and over in Representaville, USA.

(a) Produce (i) a summary table (include: min, q1, median, q3, max, mean, sd, n) and (ii) a relevant visualization (with an appropriate title) of the numeric variable for household income (`hhld_income_num`) in the census dataset.

Comment the shape, centre and spread of this distribution (1–2 sentences) and breifly describe in words what you learn from the summary table (1–2 sentences).

(b) If you had to pick one summary statistic from the table above to descrive this data, which would it be and why?

(c) Filter to just people in household that earn less than $50,000 per year and re-create the historgram from part a) (with an appopriate title and number of bins/binwidth). Briefly describe this plot AND suggest whether or not a box plot would be another appropriate graph for this filtered data. Explain why or why not.

ANSWER After restricting to just households earning less than $50,000 per year, we see a bimodal pattern with a mode around $9,000 and another around $42,000. Due to the bimodal nature of this data, a box plot wouldn't be very appropriate as it would hide this feature. #### ANSWER

(d i) Select a random sample of size n=20 (without replacement) from the census data and produce an appropriate table and visualization (same as in (a), but for the new data). Set the seed as the last *three* digits of your student ID number.

```
# set your seed so your values don't change each time!

# get your sample of 20

# make your plots and summary table
# should be able to reuse code from a) with small tweaks
```

(d ii) Now, select a random sample of size n=100 (without replacement) from the census data and again produce an appropriate table and visualization. Set the seed as the last *three* digits of your student ID number.

```
# set your seed so your values don't change each time!

# get your sample of 10

# make your plots and summary table
# should be able to reuse code from a) / d) with small tweaks
```

(d iii) How do the distributions in d(i) and d(ii) compare to the distribution in (a)? Which one of d(i) and d(ii) resembles the distribution in (a) more? Why is this so?

(e ii) How do these two estimated sampling distributions (i.e., the one for the sample medians when the sample size is 20 versus the one for sample medians when the sample size is 100) compare?

(f) Explain how and why the distributions you estimated in part (e) are different from the distributions you estimated in part (d) above.

[Question 2 — Real world]

It is actually very hard to get a census of a whole town (why many countries only run their census every 10 years and why they cost so much!), so let's suppose we only have the sample of 1000 (`sample`). That is still a pretty big random sample!

(a) Briefly describe the relationship between the two datasets, `census` and `sample`. The list of vocab words in part 2 might help. Aim to write 1–3 sentences.

(b) Simulate 1000 bootstrap samples and calculate the difference between the proportion of people in households earning $50,000+ and those earning less than $50,000 who lost all their saving over the course of the pandemic (`lost_all_savings_q8`). Set the seed as the last *three* digits of your student number.

(c) Calculate a 95% confidence interval for the difference between the proportion of people in households earning $50,000+ and those earning less than $50,000 who lost all their saving over the course of the pandemic (`lost_all_savings_q8`) based on the bootstrap sampling distribution you generated in (b). Write 1–2 sentences correctly interpreting this interval.

(d) If we want to be *more* confident about capturing the true difference between the proportion of people in households earning $50,000+ and those earning less than $50,000 who lost all their saving over the course of the pandemic (`lost_all_savings_q8`), should we use a *wider* confidence level or a *narrower* confidence level? Explain your answer.

(f) Suppose we performed a hypothesis test with:

$H_0 : p_{\text{lost all given 50k+}} - p_{\text{lost all given <50k}} = 0$ vs $H_A : p_{\text{lost all given 50k+}} - p_{\text{lost all given <50k}} \neq 0$

Based on your confidence interval, do you think you'd have evidence against your null hypothesis? Explain your answer.

Part 2

You are once again chatting on the phone to your friend. Your friend enjoyed your previous conversation about data visualization so much that they asked you if you had learned anything new in your STA130 course. You decided to tell them about the fancy new technique you just learned: bootstrapping! Be sure to include **at least 2 vocabulary words** from this module and explain them in simple terms for a lay audience.

Other things to consider:

- Try to not spend more than 20 minutes on the prompt.
- Aim for more than 200 but less than 500 words.
- Use full sentences.
- Grammar is not the main focus of this assessment, but it is important that you communicate in a clear and professional manner (i.e., no slang or emojis should appear).

Vocabulary

- Parameter
- Statistic
- Population
- Sample
- Sampling distribution
- Random sampling
- Resampling
- Bootstrap
- Percentile (quantile)
- Confidence interval
- Confidence level
- Testing
- Estimation
- Representative