



Assignment Two

Data 602 - Assignment Two

Due October 1, 2022 @ 11:59pm

Attempt all problems below. You will be required to complete this assignment in R Notebook or R Markdown, and submit a .pdf file into Gradescope.

Ensure you justify all computation and data visualizations with accompanying code.

1. Refer to Question 11 from Assignment 1:

- Compute the probability that another random sample of the same size will produce a sample mean that is at least the same value as the value of \bar{X} you observed in Question 11 of Assignment 1.
- Observe the value of the sample standard deviation S (which you computed in Exercise 11 of Assignment 1): Compute the probability that another random sample (again, of the same size) will yield a sample standard deviation that is between 0.5 hour and 1 hour.

2. A 2012 poll carried out by Ipsos Reid in found that “42% of Canadians, *who live outside of Quebec*, believe Quebec will separate from Canada at some point in the future.” A pollster wishes to see if this sentiment is still present for Canadians who live outside of the province of Quebec.

- The pollster has determined that they are going to randomly sample $n = 1426$ *Canadian residents who are not residents of Quebec* in an attempt to estimate p - the proportion of all Canadians who live outside of Quebec (“Rest of Canada”) - who believe that Quebec will separate from Canada within the next 10 years. Describe the distribution of \hat{p} , the proportion of $n = 1426$ randomly chosen Canadians who live outside of Quebec who believe Quebec will separate from Canada within the next 10 years. (Ensure that your description provides a (i) distribution shape (ii) a balancing point and (iii) a measure of spread.)
- A recent poll¹ of $n = 1426$ Canadians *who are not residents of Quebec* was taken. Of these, 541 thought that Quebec will separate from Canada in the next 10 years (356 indicated “might happen”; 128 responded “likely to happen”; 57 indicated “definitely will happen”). The sample proportion is computed to be $\hat{p} = \frac{541}{1426} = 0.3794$. Under the condition of the 2012 poll result, how likely is it for another random sample of $n = 1426$ Canadians (who reside outside of Quebec) to produce a sample proportion that is at most as **0.3794**?
- Consider the steps and associated R Code required to generate a distribution of the sample proportion, \hat{p} , when sampling $n = 1426$ Canadians who live outside of Quebec, then determining the

proportion who believe Quebec will separate from Canada within the next 10 years. Carry out a simulation where you simulate 1000 random samples of $n = 1426$. Create, then run your code to determine the proportion of your \hat{p} s that are less than or equal to 0.33794. Provide this proportion.

3. Billy purchases one 6-49 lottery ticket every week and keeps track of the number of “matches” he has on each of his tickets. To be clear, a “match” will occur when a number on his ticket matches a number that appears in the winning combination. A random variable X that keeps track of the number of matching numbers Billy experiences per week has the probability distribution function with a mean and standard deviation of

$$P(X = x) = \frac{\binom{6}{x} \binom{43}{6-x}}{\binom{49}{6}} \quad x = 0, 1, 2, 3, 4, 5, 6.$$

$$E(X) = \mu_X = \frac{36}{49} = 0.7347$$

$$SD(X) = \sigma_X = 0.75998 \approx 0.76$$

Billy claims that in a year (52 weeks), on average, he manages to have at least one matching number on his 6-49 ticket. What do you think about Billy’s claim? Provide a brief commentary about Billy’s claim using your current knowledge of statistics and probability theory.

4. A common measure of toxicity for any pollutant is the concentration of the pollutant that will kill half of the test species in a given amount of time (usually about 96 hours for the fish species). This measurement is called the LC50, which refers to the lethal concentration killing 50% of the test species).

The Environmental Protection Agency has collected data on LC50 measurements for certain chemicals likely to be found in freshwater and lakes. For a certain species of fish, the LC50 measurements (in parts per million) for DDT in 12 experiments to determine the LC50 “dose” are

16, 5, 21, 19, 10, 5, 8, 2, 7, 2, 4, 9

- Use R studio to create the bootstrap distribution of the sample mean \bar{X} . Use 2000 resamples in your work.
- From your result in (a), find a 95% bootstrap confidence interval for μ , the mean LC50 measurement for DDT. Interpret the meaning of your interval in the *context of these data*.
- Compute the 95% confidence interval for μ using the t -version of confidence interval. Ensure you appropriately present your finding/result.
- Compare your results in parts (b) and (c). If you were to report one of these confidence intervals, which would you report? Explain your answer.
- The confidence interval you computed in part (c) is valid provided a certain condition holds. Use `ggplot()` to create a graph that is used to check this condition. From your plot, can you infer that this condition is satisfied? Explain.

5. Ipsos Reid reported² in a 2018 survey conducted on “Baby-Boomer” Canadians (Canadians aged 55 or older) homeowners and found that of $n = 1866$ who have either downsized their home or plan to downsize their home, 571 indicates they either downsized or plan to downsize to take the equity out of their home to live comfortably in retirement.

live comfortably in retirement.

- a. Compute a 95% confidence interval for p , the proportion of all Canadians aged 55 years or older homeowners who have either downsized or plan to downsize to take equity out of their home to live comfortably in retirement.
- b. Similar to your work in Question 4(b), create the distribution of the bootstrap statistic \hat{p} .
- c. From your result in (b), compute the 95% bootstrap confidence interval for p .
- d. Compare your results in (a) and (c). which interval should you report? Report the interval and interpret its meaning on the context of these data.

6. Does one's educational level influence their opinion about vaccinations? A recent Angus Reid³ survey was taken. Each person sampled was asked to respond to the statement "The science around vaccinations isn't clear."

Respondents either "strongly agree", "moderately agree", "moderately disagree", or "strongly disagree". The sample was partitioned by level of education.

There were $n = 670$ respondents who's highest level of education was high school or less, of which 348 "disagreed" (moderately disagree or strongly disagree). There were also $n = 376$ who's highest level of education was at least an undergraduate university education. Of these, 274 disagreed.

- a. Consider the population consisting of all persons, who's highest level of education was high school or less and the bootstrap statistic \hat{p}_{HS} . Using 1000 iterations/replications, create a bootstrap distribution of \hat{p}_{HS} . Display your distribution.
- b. Now consider a *different* population that consists of all persons who's highest level of education was at least an undergraduate degree. Repeat part (a), creating a bootstrap distribution for \hat{p}_{Uni} . (Again, display your distribution).
- c. You wish to estimate $p_{Uni} - p_{HS}$, the difference between the proportion of all university-educated Canadians who disagree that the science of vaccinations isn't clear and the proportion of all Canadians who's highest level of completed education is high school who believe the same. You wish to have 95% confidence in your result. Think about the code you created to generate the bootstrap distributions on parts (a) and (b). Modify the code to you created in parts (a) and (b) to create a distribution of the bootstrap statistic $\hat{p}_{HS} - \hat{p}_{Uni}$.
- d. Consider your finding in part (c). Compute the 95% bootstrap confidence interval for $p_{HS} - p_{Uni}$. From your result, does the proportion of persons with at most a high school education who disagree the science around vaccinations isn't clear greater than the similar proportion of persons with at least an undergraduate university degree? Write a paragraph that supports your answer.

7. Refer to the data encountered in Question 4 of this assignment.

- a. Create a bootstrap distribution of the sample median \widetilde{X} , using the same number of replications as you did in Question 4. From this find a 99% confidence interval for the population median, $\widetilde{\mu}$. Interpret your finding in the context of these data.
- b. Compute the 95% bootstrap confidence interval for the population standard deviation, σ . In addition, interpret the meaning of your interval in the context of these data.

8. The most recent poll⁴ taken about the voting preferences of Albertans found that of $n = 858$ randomly chosen, decided Alberta voters. Each was posted with the following question:

“And if a provincial election were held tomorrow here in Alberta, which party’s candidate would you yourself be most likely to support?”

The results?

- 378 responded “UCP” (United Conservative Party)
- 352 responded “NDP” (New Democratic Party)
- 43 responded the “Independence Party”
- 34 responded the “Alberta Party”
- 17 responded the “Liberal Party”
- 34 responded “some other party”

a. Compute the 95% confidence interval for p_{NDP} , the proportion of all Alberta residents (aged 18 years



Reflect in ePortfolio



Download



Print



Activity Details

You have viewed this topic

Last Visited Sep 29, 2022 3:07 AM