

Homework 7 - MATH 141

Due Date: Monday 11/15/2021, 11:59 PM

Instructions:

- Please provide complete answers/solutions for each question/problem.
- **If it involves mathematical computations, please provide your reasoning and/or detailed solutions.**
- There are two ways you can write your answers, a: by handwriting (either physically or digitally), or b: by typing on a template document with file type options, Word or RMarkdown, which can be downloaded from the [course website](#).
- If you had handwritten your answers/solutions on a physical paper, make sure to label it properly and please scan your document using a scanner app for convenience. Suggestions: (1) [“Tiny Scanner” for Android](#) or (2) [“Scanner App” for iOS](#).
- If a problem asks you to show your R code, R outputs, or R plots, please provide them as additional pages into your current homework pdf while labeling them properly. This means that, **if you have handwritten your homework solutions and saved it as pdf, you would need to merge the separate pdf which contains your R code, R outputs, or R plots. Note that all of the problems that require R does not require you to show your R code - unless the problem specifically says so.**
- If you have questions or concerns, please feel free to ask the instructor.
- **Please save your work as one pdf file, don't put your name in any part of the document, and submit it to the Gradescope page for this course. Your document upload will correspond to your name automatically in Gradescope.**

I. Inference for a Single Proportion

Legalization of Marijuana - A *Frequentist* approach.

The exercise problems shown below was taken and slightly modified from your textbook [OpenIntro: Introduction to Modern Statistics Section 16.4](#).

Consider the research study described below.

The 2018 General Social Survey asked a random sample of 1,563 US adults: “Do you think the use of marijuana should be made legal, or not?” 60% of the respondents said it should be made legal. [NORC 2018](#) Consider a scenario where, in order to become legal, more than 55% of voters must approve.

1. Null and Alternative Hypothesis

- What is the null and alternative hypotheses for evaluating whether these data provide convincing evidence that, if voted on, marijuana would be legalized in the US?
- Describe the population parameter and the sample statistic in this study. Write out the point estimate and the null value.

2. Simulation Method

- Perform a parametric bootstrapping procedure with 1000 bootstrapped samples and display the resulting distribution.
- Determine the p-value using the bootstrapped simulations and interpret it in the context of the simulations.
- Determine the 95% confidence interval and the standard error using the percentile method. Does the interval contain the null value? What does this mean?

3. Theoretical Method

- Check the conditions (independence and success-failure) conditions and calculate the standard error of the sample proportion using the standard error formula for a single proportion.
- Compute the test statistic Z and the p-value. Interpret the p-value in this context. What should you decide given a significance value of 0.05?
- Using the standard error formula for one proportion, calculate the 95% confidence interval for the population proportion. Interpret this confidence interval. A news piece on this survey’s findings states, “Majority of US adults think marijuana should be legalized.” Based on your confidence interval, is this statement justified?

4. Simulation vs Theoretical

- Do the theoretical method and parametric bootstrap give similar standard errors?
- In this setting (to test whether the true underlying population proportion is greater than 0.55), would there be a strong reason to choose the mathematical model over the parametric bootstrap (or vice versa)?
- A critic points out that this 95% confidence interval is only accurate if the statistic follows a normal distribution, or if the normal model is a good approximation. Is this true for these data? Explain.

II. Inference for Comparing Two Proportions

Disaggregating Asian American Tobacco Use - A *Frequentist* approach.

The exercise problems shown below was taken and slightly modified from your textbook [OpenIntro: Introduction to Modern Statistics Section 17.5](#).

Consider the research study described below.

Understanding cultural differences in tobacco use across different demographic groups can lead to improved health care education and treatment. A recent study disaggregated tobacco use across Asian American ethnic groups including Asian-Indian ($n = 4,373$), Chinese ($n = 4,736$), and Filipino ($n = 4,912$), in comparison to non-Hispanic Whites ($n = 275,025$). The number of current smokers in each group was reported as Asian-Indian ($n = 223$), Chinese ($n = 279$), Filipino ($n = 609$), and non-Hispanic Whites ($n = 50,880$). Based on a study on the degree to which smoking practices differ across ethnic groups, a confidence interval for the difference in current smoking status for Filipino versus Chinese Americans is desired. [Rao et al. 2021](#)

1. Simulation Method The R code below creates the necessary data to perform the bootstrapping procedure to answer the following questions.

```
asian_smoke <- tibble(
  ethnic = c(rep("Filipino", 4912), rep("Chinese", 4736)),
  outcome = c(
    rep("smoke", 609), rep("healthy", 4303),
    rep("smoke", 279), rep("healthy", 4457)
  )
)
```

- Perform a bootstrapping procedure of the difference in sample proportions of current smokers (Filipino Americans minus Chinese Americans) in 1,000 bootstrap repetitions as above. Estimate the standard error of the difference in sample proportions.
- Using the standard error from the bootstrap distribution, find a 95% bootstrap standard error confidence interval for the true difference in proportion of current smokers (Filipino Americans minus Chinese Americans) in the population. Interpret the interval in the context of the problem.
- Using the entire bootstrap distribution, find a 95% bootstrap percentile confidence interval for the true difference in proportion of current smokers (Filipino Americans minus Chinese Americans) in the population. Interpret the interval in the context of the problem.

2. Theoretical Method

- Are the conditions (independence and success-failure) met for the sampling distribution for the difference in proportions to be normal? What is the point-estimate and the null value?
- Compute the standard error using the formula for the standard error of the difference in two proportions.
- Determine the 95% confidence interval between two proportions and interpret this interval in the context of the study. Does the interval contain the null value? What does this mean?

3. Simulation vs Theoretical

- Do the theoretical method and bootstrap simulations give similar standard errors?
- Explain the advantages of using the theoretical method versus the bootstrapping method?
- In what situations can we use the bootstrapping method if the theoretical method does not work?