Homework 8

Stats 20 Lec 1

Winter 2022

General Guidelines

Please use R Markdown for your submission. Include the following files:

- Your .Rmd file.
- The compiled/knitted HTML document.

Name your .Rmd file with the convention `123456789_stats20_hw0.Rmd`, where `123456789` is replaced with your UID and `hw0` is updated to the actual homework number. Include your first and last name and UID in your exam as well. When you knit to HTML, the HTML file will inherit the same naming convention.

The knitted document should be clear, well-formatted, and contain all relevant R code, output, and explanations. R code style should follow the Tidyverse style guide: https://style.tidyverse.org/.

Any and all course material, including these homework questions, may not be posted online or shared with anyone at any time without explicit written permission by the instructor (Michael Tsiang), even after the quarter is over. Failure to comply is a breach of academic integrity.

Note: All questions on this homework should be done using only functions or syntax discussed in Chapters 1–11 of the lecture notes. No credit will be given for use of outside functions.

Basic Questions

Collaboration on basic questions must adhere to Level 0 collaboration described in the Stats 20 Collaboration Policy.

The following information is used in Questions 1, 2, and 3.

Consider the dataset found at: http://www.isi-stats.com/isi/data/chap3/CollegeMidwest.txt

The data contains two variables gathered from the registrar at a small midwestern college on all students at the college in spring 2011.

The variables are:

- `OnCampus`: Whether or not a student lives on campus (`Y` or `N`)
- `CumGpa`: The student's cumulative GPA.

Note: Since this is data on *all* students at the college, we will treat the students observed in this data to be the population.

Question 1

The objective of this question is to show how R can be applied in the context of simulating a sampling distribution.

(a)

Set the seed to 9999, and simulate the sampling distribution of the difference in mean cumulative GPA between the students who live off campus and the students who live on campus. Simulate the difference in sample means from 1000 random samples of size 30.

Hint 1: For consistency, compute $\bar{x}_{\text{on}} - \bar{x}_{\text{off}}$, where $\bar{x}_{\text{on}}$ and $\bar{x}_{\text{off}}$ are the respective sample means for students who live on and off campus.

Hint 2: The total sample size for each repetition should be 30, not 60. Do not assume that you already know whether the students live off campus or on campus prior to sampling.

(b)

Plot a histogram of the sampling distribution of differences in sample means from part (a). Add vertical lines that show the differences in sample means that are 2 standard errors away from the mean.

(c)

Compute the mean and standard deviation of the simulated distribution of differences in sample means. Use these values to superimpose a normal curve over the histogram.

(d)

Suppose we observe a random sample of size 30 with an observed difference in mean cumulative GPA between off campus and on campus students to be 0.48. Based on your simulated (approximate) sampling distribution in part (a), what is the (approximate) probability of observing a difference in sample means greater than 0.48?

Hint 1: Part (d) does not rely on the Central Limit Theorem.

Hint 2: You may ignore/remove `NA`s if necessary.

Question 2

The objective of this question is to introduce the `t.test()` function, show how the course material can be used to understand objects of a new object class, and show how R can be applied in the context of hypothesis testing.

Suppose we are interested in using a random sample of 30 students to decide if the mean cumulative GPA of the population of students at the College of the Midwest is different from 3.5 or not. The null and alternative hypotheses are given by

$$\begin{aligned} H_0 &: \quad \mu = 3.5 \\ H_a &: \quad \mu \neq 3.5 \end{aligned}$$

The **`t.test()`** function performs one and two sample t-tests on vectors of numeric data. The basic syntax for `t.test()` is `t.test(x ,y, alternative, mu, conf.level)`.

- The `t.test()` function inputs a vector `x` of values from your sample and conducts a one-sample t-test for the mean. If a second vector in the argument `y` is included, `t.test()` will conduct a two-sample t-test for a difference in means.
- The `alternative` argument inputs a character value of `"two.sided"`, `"greater"`, or `"less"`, depending on the alternative hypothesis we are considering. By default, `t.test()` will conduct a two-sided hypothesis (i.e., `alternative = "two.sided"`).
- The `mu` argument inputs a numeric value that specifies the value of the mean parameter μ *under the null hypothesis*. The default value is `mu = 0`, i.e., the default null hypothesis is $\mu = 0$.
- In addition to a hypothesis test, the `t.test()` function also outputs a confidence interval, with confidence level set by the `conf.level` argument. By default, the confidence level is set to `conf.level = 0.95`.

(a)

Set the seed to 30 and draw a random sample of size 30 from the `CollegeMidwest.txt` data.

(b)

For the random sample in (a), compute the observed t-statistic $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$, where $\bar{x}$ is the sample mean, s is the sample standard deviation, and n is the sample size. How would you interpret this value?

(c)

Use the `t.test()` function to conduct a one-sample t-test to decide if the true mean cumulative GPA of all the students at the College of the Midwest is different 3.5. Use a significance level of $\alpha = 0.05$.

(d)

What is the mode and class of the output of `t.test()` from (c)? Use this information to extract the 95% confidence interval vector from the `t.test()` output object. Is 3.5 inside this interval? What does this say about whether the true mean cumulative GPA is 3.5 or not?

Note: The t-test (and `t.test()`) relies on the normal approximation to the sampling distribution of the sample mean. When conducting a t-test, it is assumed that the conditions for the Central Limit Theorem are satisfied.

Intermediate Questions

Collaboration on intermediate questions must adhere to Level 1 collaboration described in the Stats 20 Collaboration Policy.

Question 3

The objective of this question is to give further practice with random simulation and basic graphics, and show how R can be used to construct and visualize confidence intervals.

The **confidence level** refers to the long-run proportion of random samples (of a fixed size) whose confidence intervals contain the true population parameter. We want to illustrate this by simulation.

(a)

Suppose conducting a survey of students from the College of the Midwest consists of the following steps:

(1) Select a random sample of 30 students.

(2) Compute the mean cumulative GPA for the 30 students in the sample.

(3) Construct a 95% confidence interval for the population mean cumulative GPA.

Set the seed to 24601 and repeat steps 1, 2, and 3 a total of 10,000 times. For each random sample, calculate $\bar{x}$ and construct a 95% confidence interval.

Hint: You can use the `t.test()` function from the previous question to construct the 95% confidence interval.

(b)

Use the full data to compute the true mean cumulative GPA for the population of all students at the College of the Midwest. Find the proportion of the 10,000 confidence intervals that contain the true population mean. Is this proportion consistent with what you expected?

(c)

Create a plot of the first 100 confidence intervals. Be sure that the plot satisifies the following criteria:

- The limits of the axes should be large enough to contain the lengths of all of the confidence intervals.
- Represent each sample mean by a point and each corresponding confidence interval by a line segment through the point.
- Color the points and intervals to correspond to whether the interval was successful at capturing the true population mean. In other words, use one color for the intervals that contain the true mean, and use a different color for the intervals that do not contain the true mean.
- Add a straight line that shows the true population mean.
- Add a legend that explains the color coding in the plot.

Hint: Use the `segments()` function. You do not need a `for()` loop to create this plot.

Question 4

The objective of this question is to give further practice with random simulation and show how to simulate sampling from a null distribution.

Lecturer Dr. Mike has a Campuswire forum for his Stats 20 class where students can upvote helpful replies to questions students make about course content. Dr. Mike tries to make helpful replies, but he suspects that students prefer to upvote replies made by other students over replies made by the teaching team (instructor, TAs, LAs).

Dr. Mike hypothesizes that the mean number of upvotes for student replies is higher than the mean number of upvotes for instructor replies. To investigate his suspicions, he takes a random sample of replies from Campuswire and notes the number of upvotes for each reply and whether the reply was made by a "student" or an "instructor" (i.e., a member of the teaching team).

In this scenario, the null and alternative hypotheses are given by

$$\begin{aligned} H_0: &\quad \mu_{\text{student}} - \mu_{\text{instructor}} = 0 \\ H_a: &\quad \mu_{\text{student}} - \mu_{\text{instructor}} > 0 \end{aligned}$$

Since the sample is too small to invoke the Central Limit Theorem, we will use a simulation-based approach to simulate from the null distribution of the difference in sample means.

The data is contained in the `cw_points.txt` file posted on Bruin Learn.

(a)

Create side-by-side boxplots that show the distributions of upvotes split by the type of reply (instructor or student).

(b)

Compute the observed difference in the mean number of upvotes between student and instructor replies.

Hint: For consistency, compute $\bar{x}_{\text{student}} - \bar{x}_{\text{instructor}}$, where $\bar{x}_{\text{student}}$ and $\bar{x}_{\text{instructor}}$ are the respective sample means for student replies and instructor replies.

(c)

To simulate sampling from the null distribution of the difference in sample means, we can simulate random assignment by reassigning the type (student or instructor) for each upvote in the data. That is, we follow the steps below:

(1) Shuffle the order of the `type` values and pair them with the `upvotes` values in their original order. This simulates observing a different sample of replies assuming that there is no difference in the mean number of upvotes between student and instructor replies.

(2) For the shuffled data, compute the difference in the mean number of upvotes between student and instructor replies.

Set the seed to 2021 and repeat steps (1) and (2) 1000 times. Store the 1000 simulated differences in sample mean number of upvotes between student and instructor replies in a vector. This vector represents the simulated null distribution.

(d)

Visualize the simulated null distribution of the difference in sample means with a histogram. Add a vertical line to show the observed difference in the mean number of upvotes between student and instructor replies from part (b).

(e)

Compute the proportion of simulated differences in sample means that are larger than the observed difference in sample means from part (b). This is the simulation-based p-value. Based on this p-value, does Dr. Mike have strong evidence that students prefer to upvote student replies instead of instructor replies?