ST 411/511 Homework 2

Pavan Sai Nallagoni

Summer 2022

Instructions

This assignment is due by 11:59 PM, July 26th on Canvas via Gradescope. **You should submit your assignment as a PDF which you can compile using the provide .Rmd (R Markdown) template.**

> Note: Create a PDF by either compiling a PDF directly (via LaTeX) or from a Word Doc. Do not submit a PDF of the HTML output as they're cumbersome and difficult to read.

Include your code in your solutions and indicate where the solutions for individual problems are located when uploading into Gradescope (Failure to do so will result in point deductions). You should also write complete, grammatically correct sentences for your solutions.

Once you've completed the assignment, and before you submit it to Gradescope, you should read the document to ensure that (1) The computed values show up in the document, (2) the document "looks nice" (i.e. doesn't have extraneous code/outputs and includes *just* the essentials), and (3) ensure the document is "easy" to read.

Goals:

1. Learn how to conduct Z tests for testing claims about unknown population means.
2. Practice writing conclusions for hypothesis tests that are both statistically rigorous but also convey information to non-statistical audiences.
3. Learn how to compute and interpret confidence intervals for unknown population means.
4. Practice determining whether a study is an observational study or an experiment.
5. Based on the design of a study, determine the largest population for which the outcome of the study applies.
6. Focus on understanding some of the *nuance* of statistical hypothesis testing. Help students understand the definition of a p-value, understand the underlying uncertainty in hypothesis testing, and differences between the Z and t distributions.
7. Practice conducting one sample t-tests by hand in order to gain a sense of *intuition* as to how they're conducted.

Question 1 (12 points)

A random sample of $n = 500$ books is selected from a library and the number of words in the title of each book is recorded. The sample mean number of words in the title is 6.2 words. The population variance is 40 words-squared.

(a) (2 points) Compute the z-statistic for testing the null hypothesis $H_0 : \mu = 7$.

z-statistic can be obtained using: Z <- ($\bar{X}$ - μ)/sqrt(σ^2/size)

$\bar{X} = 6.2$ $\mu = 7$ $\sigma^2 = 40$ $n = 500$

```
Xbar <- 6.2
mu <- 7
var <- 40
n <- 500

Z <- (Xbar - mu)/sqrt(var/n)
Z
```

```
## [1] -2.828427
```

(b) (3 points) Perform a level $\alpha = 0.1$ test of $H_0 : \mu = 7$ vs. the one-sided lesser alternative $H_A : \mu < 7$ by comparing the *computed test statistic* (from part (a)) to the correct *critical value*. Be sure to include a *complete* conclusion for your test which states (1) whether you reject or fail to reject the null hypothesis, (2) the reasoning behind why you reject/fail to reject, and (3) what the conclusion means in terms of the context of the question.

Solution: Given $\alpha = 0.1$ $H_0 : \mu = 7$ $H_A : \mu < 7$ $Z(\mu o) = -2.828427$ Therefore, we need to calculate critical value $Z(\alpha)$ using $qnorm(\alpha)$ function.

```
qnorm(0.1)
```

```
## [1] -1.281552
```

Conclusion: We *reject the null hypothesis* $H_0 : \mu = 7$ since the value of *computed test statistic* is less than *critical value* i.e., $Z(\mu o) < Z(\alpha) => (-2.828427 < -1.281552)$ which is for one-sided lesser.

Therefore, we can infer that the true population mean is not equal to 7 and is most likely to be somewhere less than 7 ($H_A : \mu < 7$).

(c) (2 points) What is the one-sided lesser p-value for the statistic you computed in part (a)?

This can be achieved by using $pnorm(z)$ function where $z = -2.828427$ (Z is computed in part (a)).

```
pnorm(Z)
```

```
## [1] 0.002338867
```

(d) (2 points) What is the two-sided p-value for the statistic you computed in part (a)?

p-value for a two-sided test: this can be achieved by using $p = 2 * (1 - pnorm(abs(z))$

```
P <- 2 * (1 - pnorm(abs(Z)))
P
```

```
## [1] 0.004677735
```

(e) (2 points) Construct a 95% confidence interval for the population mean number of words per title. Hint: recall that a 95% confidence interval is formed by the sample mean $\pm 1.96\times$ standard deviation of the sampling distribution. Write a sentence which communicates the bounds of the confidence interval.

From (a), Point estimate: $\bar{X} = 6.2$ Variance : $\sigma^2 = 40$ Sample size : $n = 500$

```
confidenceInterval <- 0.95
sd <- sqrt(var/n)
alpha <- (1 - confidenceInterval)
zValue <- qnorm(1 - alpha/2)

lowerBound <- Xbar - zValue * sd
upperBound <- Xbar + zValue * sd

print(paste0(
    "The 95% confidence interval is ",
    sprintf("[ %f, %f ]", lowerBound, upperBound)))
```

```
## [1] "The 95% confidence interval is [ 5.645638, 6.754362 ]"
```

```
# REF: To print the interval:
# https://stackoverflow.com/questions/47745846/how-to-print-a-confidence-interval-into-a-sentence-in-r-
```

The lower bound of the confidence interval is 5.645638 and the upper bound of the confidence interval is 6.754362.

From the computed confidence interval, we can infer that we're 95 confident that the true population mean is somewhere in between the values 5.645638 and 6.754362 with a point estimate of 6.2.

(f) (1 point) Based on your confidence interval from part (e), would a level $\alpha = 0.05$ two-sided hypothesis test reject or fail to reject the null hypothesis that the population mean is 6.5 words per title? How do you know? Answer this question without conducting the two-sided test with $alpha = 0.05$**.**

We fail to reject the null hypothesis that the population mean is 6.5 *words*. (From e) We're 95 confident that the true population mean is somewhere in between the values 5.645638 and 6.754362.

However, the value of 6.5 lies within this range, and since there is a chance that the true population mean can be equal to 6.5, we fail to reject the null hypothesis.

Question 2 (10 points)

Consider the `rivers` data set in `R`, which is a vector of the lengths (in miles) of 144 "major" rivers in North America, as compiled by the US Geological Survey.

```
data(rivers)
```

(a) (1 point) What is the length of the longest "major" river in North America? Hint: you can find the maximum of a vector using the `max` function.

```
length <- max(rivers)
length
```

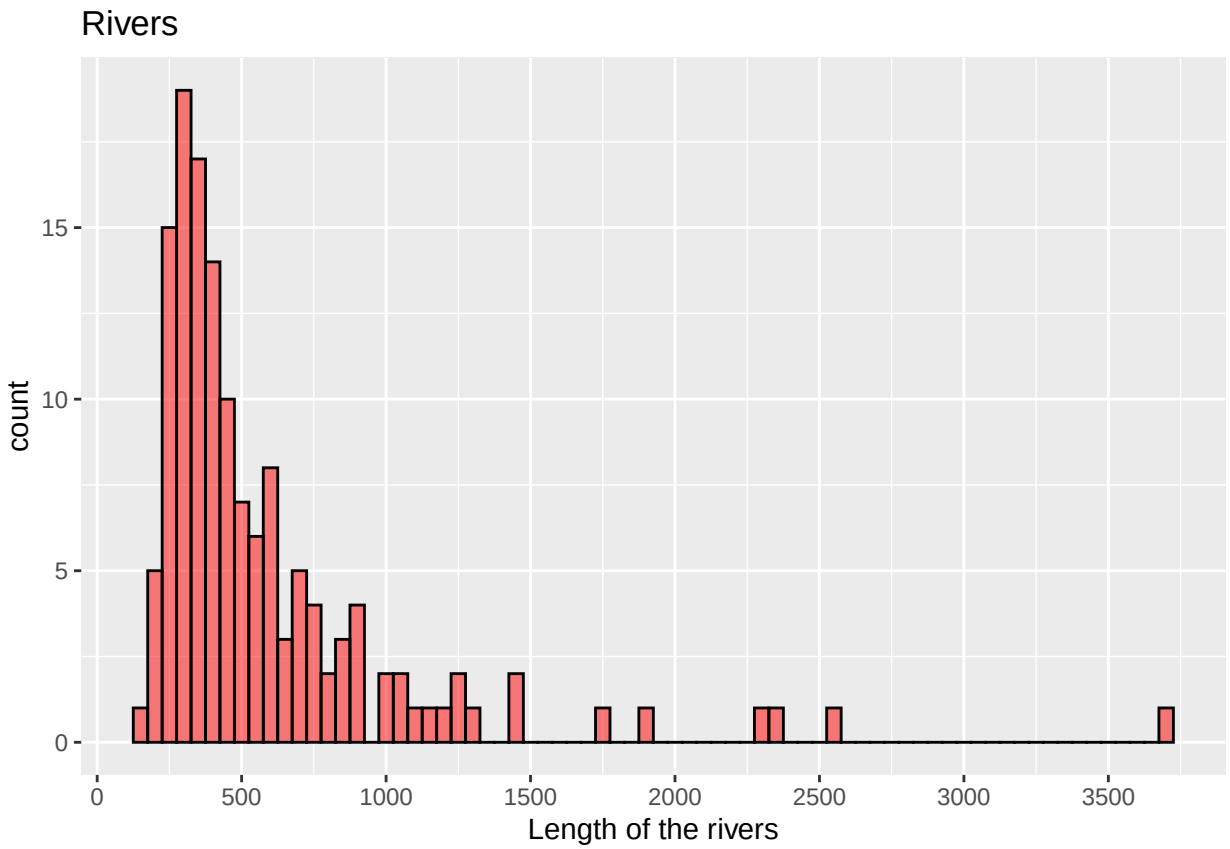
```
## [1] 3710
```

The length of the longest "major" river in North America is 3710 *miles*.

(b) (2 points) Create a *population* histogram of the lengths of the major rivers. Describe the center, shape, and spread of the distribution based solely on the plotted distribution. Note: to use ggplot, the data have to be formatted as a data frame. I have given you the line of code that does this.

```
riversdf <- data.frame(rivers)

ggplot(riversdf, aes(x = rivers)) +
  geom_histogram( binwidth = 50, color = "black", alpha = 0.5, fill = "red") +
  ggtitle("Rivers") + xlab("Length of the rivers") +
  scale_x_continuous(breaks = seq(0, 10000, by=500))
```



From the histogram, we can assume that the *center* could be somewhere around 600 (which could be the mean of the population). The histogram tells us that the *shape* is nearly a right-skewed distribution as it is denser at the start and gradually decreases as we go right. The population is *spread* from left (near to center) to extreme right, where it is denser on the left (near to center) and less towards the right.

(c) (1 point) Select a random sample of $n = 30$ rivers, using `set.seed(411511)` to make sure you draw the same random sample each time. What is the sample mean? Note: To use `set.seed(411511)`, you need to include this command *before* you draw your random sample.

```
set.seed(411511)
samp <- sample(riversdf$rivers, size=30, replace=FALSE)
new_Xbar <- mean(samp)
new_Xbar
```

```
## [1] 661.9
```

The mean of the random sample with size $n = 30$ is 661.9.

(d) (2 points) Compute the test statistic for a z-test of $H_0 : \mu = 600$ versus $H_A : \mu \neq 600$.

```
# Given
new_Mean <- 600
```

```
new_Variance <- var(riversdf$rivers)
new_Size <- 30
new_SD <- sqrt(new_Variance / new_Size)

new_Z <- (new_Xbar - new_Mean) / new_SD
new_Z
```

```
## [1] 0.6864958
```

(e) (2 points) Find the p-value corresponding to your test statistic from part (d). Recall that you are using a two-sided alternative hypothesis.

```
new_Pvalue <- 2 * (1 - pnorm(abs(new_Z)))
new_Pvalue
```

```
## [1] 0.4924005
```

(f) (2 points) What do you conclude from this hypothesis test at the 0.05 significance level? State your conclusion in a few short sentences. Be sure to include a *complete* conclusion for your test which states (1) whether you reject or fail to reject the null hypothesis, (2) the reasoning behind why you reject/fail to reject, and (3) what the conclusion means in terms of the context of the question.

We fail to reject the null hypothesis since the p-value (0.4924005) is greater than the significance level of $\alpha = 0.05$. Hence, we can conclude that the population mean is equal to 600.

Question 3 (3 points)

Researchers are curious about how soil type affects plant growth. To study this, they obtain 100 seeds of a particular plant species from a local seed collector. They randomly choose 50 seeds and plant each in a separate pot filled with soil type A. The remaining 50 seeds are each planted in a separate plot filled with soil type B. The plants receive the same care, and at the end of 3 months the height of each plant is measured.

(a) (1 point) Is this an example of a randomized experiment or an observational study? Justify your answer.

This is an example of a randomized experiment since the scientists were conducting an experiment by randomly assigning the participants to the subjects.

Firstly, they obtained 100 seeds of a particular plant species (participants) and then randomly assigned 50 each to soil type A and type B (subjects). Finally, they've provided same care to the plants and measured their heights after 3 months. Hence, this is a randomized experiment.

(b) (2 points) What is the largest population for which an inference can be made based on the design of this study? Justify your answer.

Based on the design of this study, the largest population for which an inference can be made is only the plant species that is selected in the experiment. It can't be generalized to other plant species since the experiment is done only on one particular plant species and they all received the same care.

However, different plant species need different types of care. Treating them with the same care would fetch us very insignificant results and therefore, the study design is not applicable.

Question 4 (6 points)

Answer whether each statement is True or False, and explain your reasoning.

(a) (2 points) A p-value tells you the probability that the null hypothesis is true.

False. Because, p-value is the probability under the null hypothesis of observing a result at least as extreme as the statistic you observed. It means p-value only tells us how likely is our observed statistic is/have occurred under the null hypothesis and not the probability of the occurrence of null hypothesis.

(b) (2 points) It is possible for a hypothesis test procedure to reject the null hypothesis even when the null hypothesis is true.

True. It is possible for a hypothesis test procedure to reject the null hypothesis even when the null hypothesis is true. Usually, this type of error is known as *Type I* error or $false - positive$ where the null hypothesis is incorrectly rejected.

(c) (2 points) Consider the null hypothesis $H_0 : \mu = \mu_0$ versus a one-sided greater alternative $H_A : \mu > \mu_0$. For a fixed significance level α the critical value $z_{1-\alpha}$ will be greater than the critical value $t_{(4)1-\alpha}$ (i.e., the critical value for a t-distribution with 4 degrees of freedom).

False. For a fixed significance level α the critical value $z_{1-\alpha}$ will not be greater than the critical value $t_{(4)1-\alpha}$. It is because the t distribution usually has wider tails (larger area for the critical region) compared to the (z) Normal distribution. And, with 4 degrees of freedom, the t distribution deviates from the Normal distribution and has a fatter tail.

Hence, the critical value $t_{(4)1-\alpha}$ is greater than the $z_{1-\alpha}$.

(Also, s^2 is always greater than σ^2, such that the Standard error tends to be smaller for the t.)

Question 5 (4 points)

A random sample of $n = 10$ OSU students is obtained, and the college GPA of each is recorded. The GPAs of the 10 students in the sample are provided in the vector `gpa`.

```
gpa <- c(3.1, 3.7, 4.0, 2.7, 2.5, 3.4, 3.5, 3.0, 1.9, 3.4)
```

(4 points) Test the null hypothesis $H_0 : \mu = 3.0$ versus the one-sided greater alternative $H_A : \mu > 3.0$ at significance level $\alpha = 0.05$. Write a *complete* conclusion stating the outcome of the test, the reason why you chose that conclusion, and what this conclusion means in the context of the question.

Note: Perform these calculations "by hand" (i.e. do not use the `t.test()` function or other built-in equivalents) using either mathematical notation or by writing code to compute the necessary values. If you write code to compute the values, make sure to output the value of the test statistic, the critical value and/or the p-value.

```
gpa_sample_Mean <- mean(gpa)
gpa_sample_Mean
```

```
## [1] 3.12
```

```
gpa_sample_Variance <- var(gpa)
gpa_sample_Variance
```

```
## [1] 0.3862222
```

```
gpa_sample_Size <- 10
gpa_Mean <- 3
```

Rather than computing z we'll use t-statistic since we are replacing σ^2 with an estimate s^2.

```
gpa_t <- (gpa_sample_Mean - gpa_Mean) / sqrt(gpa_sample_Variance / gpa_sample_Size)
gpa_t
```

```
## [1] 0.6106082
```

```
gpa_df <- gpa_sample_Size - 1
gpa_pValue <- 1 - pt(gpa_t, gpa_df)
gpa_pValue
```

```
## [1] 0.2782812
```

Conclusion:

We fail to reject the null hypothesis $H_0 : \mu = 3.0$ since the p-value (0.2782812) is greater than the significance level of $\alpha = 0.05$. We have evidence which indicates that GPAs of the students is equal to 3, on average.