ST 411/511 Homework 2

E. Alex Soderquist

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 $H_0 : \mu = 7$.

```
Zmu0 <- (6.2-7)/(sqrt(40/500))
Zmu0
```

```
## [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.

```
qnorm(0.1)
```

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

We reject the null hypothesis in favor of the alternative hypothesis because our rejection region is any z-score less than -1.28 - which our z-score associated with the mean of 6.2 is, i.e., given our null hypothesis is true, there is a less than 10% chance that we would obtain a sample mean of 6.2 or less (more extreme in this case). This means that we can conclude the true population mean of words in the title of all books in the library is less than 7 words.

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

```
pnorm(Zmu0)
```

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

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

```
2*(1-pnorm(abs(Zmu0)))
```

```
## [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.

‘

```
6.2+1.96*sqrt(40/500)
```

```
## [1] 6.754372
```

```
6.2-1.96*sqrt(40/500)
```

```
## [1] 5.645628
```

Our 95% confidence interval is (5.65,6.75) meaning we are 95% confident that the average number of words in titles of all books at the library is between 5.65 words and 6.75 words.

(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 $alpha = 0.05$.

In such a hypothesis test we would fail to reject the null hypothesis, because 6.5 words per title is well within our confidence interval. So we can't disregard the fact that 6.5 words per title may be the population mean.

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.

```
max(rivers)
```

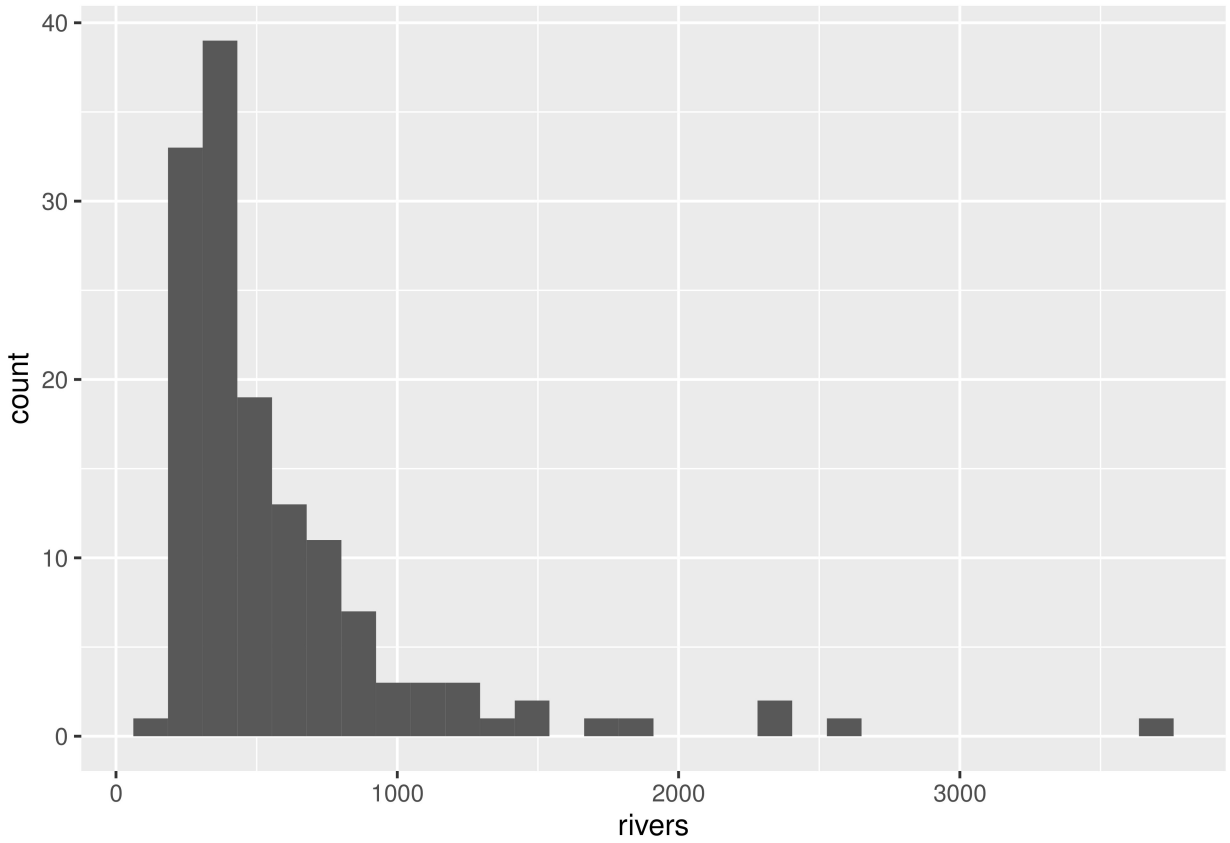
```
## [1] 3710
```

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)
# Don't delete the line of code above this line.

ggplot(data = riversdf, aes(x=rivers))+geom_histogram()
```



The data is largely right skewed, large spread with a median around 500 miles.

(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)
samp1 <- sample(rivers, size=30, replace=FALSE)
mean(samp1)
```

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

The mean of the 'random' sample of 30 rivers is 661.9 miles.

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

```
va <- var(rivers)
Zmu01 <- (661.9-600)/(sqrt(va/30))
Zmu01
```

```
## [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.

```
2*(1-pnorm(Zmu01))
```

```
## [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 hypothesis test at the 0.05 significance level. Our p-value tells us there is a 49% chance that we would find a sample mean as or more extreme given the null hypothesis is true, which is much greater than our 5% significance level. Therefore, we cannot refute the possibility that the average length of all 'large' rivers in North America is 600 miles.

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 a randomized experiment because the seeds are each chosen randomly, and there is some control exerted over how the individual plants are being grown in which soil type.

(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.

The largest population they can make claims about is every plant of the particular species. They are only experimenting with the one species, and other types of plants may react better or worse in certain manners, e.g., they may depend more or less on the type of soil for optimal growth.

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! The p-value tells you the probability for which you would observe a value of the test statistic as or more extreme as the one you just computed, given the null hypothesis is true.

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

True, given a small sample size, or an 'abnormal' random sample. Consider the 5% of constructed confidence intervals which do not contain the null hypothesis value in a 95% confidence interval test!

(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! Because the t-distribution has larger tails than the normal distribution (which z-statistics are approximated with), there is more area under the curve nearer the ends of the tails in the t-distribution. Because there is more area with less length, it takes less 'length' in the distribution to achieve the same signifiance level, so the critical value for a t-distribution will be closer to the end of the tail than the z critical value. So the z critical value will actually be less than the critical value based on the t-distribution.

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.

```
tmu <- (mean(gpa)-3.0)/(sqrt(var(gpa)/10))
tmu
```

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

```
1-pt(tmu,df=9)
```

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

We fail to reject the null hypothesis that the mean GPA of students at OSU is 3.0, meaning, we cannot discount the possibility that the true mean GPA of all OSU students is 3.0. Because our t-statistic is 27.8% likely to be 'found' with this sample given our null hypothesis is true, at the 0.05 significance level we fail to reject the null hypothesis.