ST 411/511 Homework 2

Due on January 22

Aashish Adhikari

Winter 2020

Instructions

This assignment is due by 11:59 PM, January 22, 2020 on Canvas via Gradescope. **You should submit your assignment as a PDF which you can compile (should you choose – recommended) using the provide .Rmd (R Markdown) template.** If you opt to not use R Markdown, please format your solutions in a similar manner as provided in this document. Include your code in your solutions and indicate where the solutions for individual problems are located when uploading into Gradescope. You should also use complete, gramatically correct sentences for your solutions.

Problems (25 points total)

Question 1

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^2.

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

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

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

The z statistic is -2.8284. ### (b) (3 points) Perform a level $\alpha = 0.1$ test of $H_0 : \mu = 7$ vs. the one-sided lesser alternative $H_A : \mu < 7$.

```
qnorm(0.1)
```

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

Since z(miu) < z-alpha, we reject the null hypothesis.

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

```
pnorm((6.2-7)/sqrt(40/500))
```

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

The one-sided lesser p-value is 0.0023388.

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

```
2 * (1-pnorm(abs(-2.8284)))
```

```
## [1] 0.004678131
```

0.004678131 is the two-sided p value for this test statistic.

(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.

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

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

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

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

The confidence interval is between 5.645628 and 6.754372.

(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?

Since 6.5 lies within an interval, it should fail to reject the hypothesis.

Question 2

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.

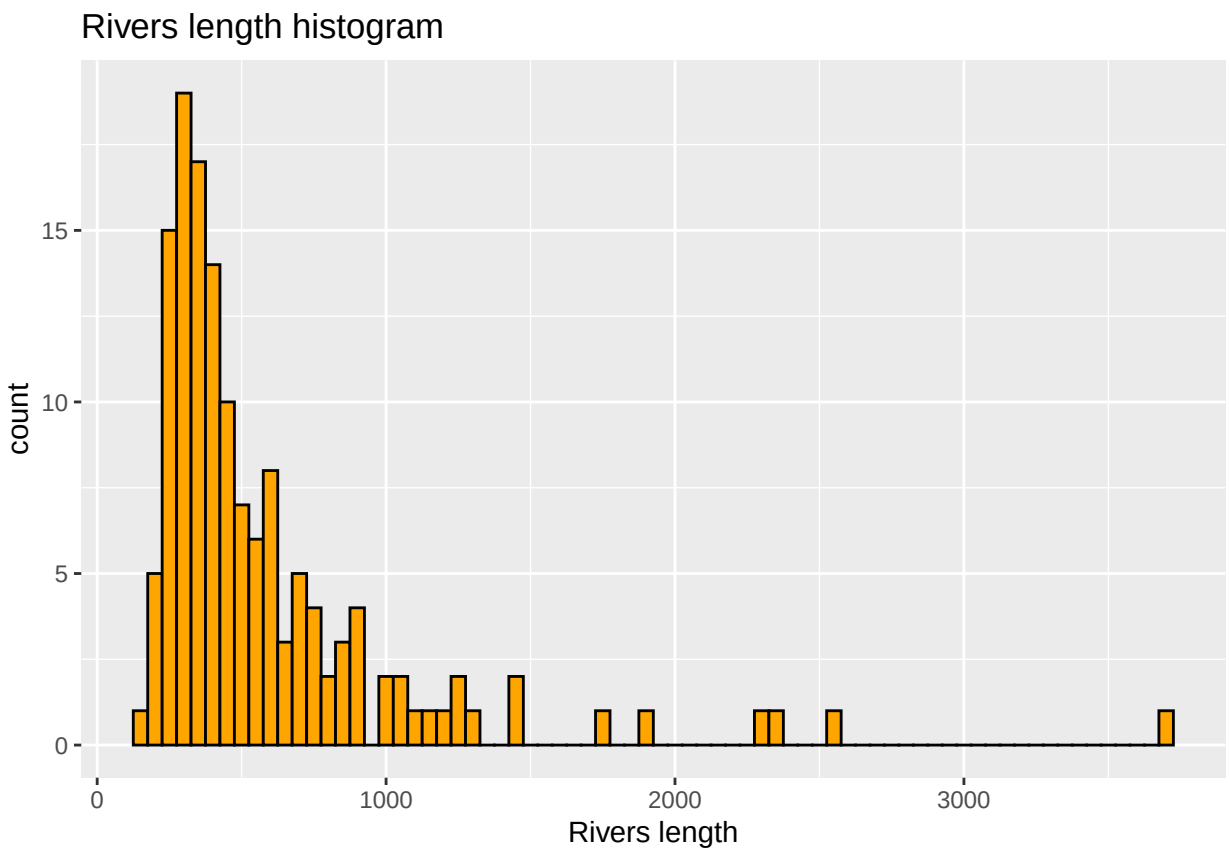
```r
max(rivers)
```

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

The length of the longest river is 3170.

(b) (2 points) Create a population histogram of the lengths of the rivers. Describe the center, shape, and spread of the distribution (You can describe the center and spread based solely on the plotted distribution. Just ensure the values are reasonable). 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.

```r
riversdf <- data.frame(rivers)
#riversdf
ggplot(riversdf, aes(x=rivers))+ggtitle("Rivers length histogram")+geom_histogram(binwidth = 50, color=
```




```r
#hist(riversdf$rivers)
```

The center of the distribution is low giving an idea that many rivers have a low length. The shape is half-bell curve. Or, we can also call it right-skewed normal distribution. Few extreme values are spread farther from the mean but most of the values.

(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?

```
set.seed(411511)
sample_taken <- sample(rivers, size=30)

sample_mean <- mean(sample_taken, )
sample_mean
```

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

The sample mean is 661.9.

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

```
z <- (sample_mean-600)/sqrt(var(rivers)/30)
z
```

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

```
#qnorm(1-(0.05/2))
```

The test statistic is 0.68649.

(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(abs(z)))
```

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

The p value required is 0.4924005.

(f) (2 points) What do you conclude from this hypothesis test? State your conclusion in a few short sentences.

We fail to reject the null hypothesis 600 at this significance level 0.05 given that our p-value is greater than 0.05. There is no strong evidence that the population mean is greater than 600. Thus, we cannot rule out the null hypothesis.

Question 3

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?

It is an example of a randomized experiment.

(b) (2 points) To what group can inference be made?

The inference can be made on this particular plant species at the local seed collector.