Lab 6 (100 points + 10 points BONUS) - One Sample t Confidence Interval and Hypothesis Test

Objectives: Confidence interval and hypothesis tests for one-sample

A. (10 points) Online Prelab

B (90 points) In this lab, we are concerned with the Average Test Score (TestScore) (Data Set: Clean US Data) In high schools, they once said that a grade of 75% is “average.” A high schooler, Antonio, wants to test whether the population mean of the Average Test Score is 75%. Since the maximum grade on this test is 2400, the hypothesis is whether the population mean of Average Test Score is $0.75 \times 2400 = 1800$. Note that the average test score in each county is considered a random variable, so it has a population mean. We are inferring about a population average for a variable that is itself an average.

1. (10 points) Code. There should not be a separate command for the confidence interval; though there will be two commands for hypothesis tests. See the tutorial for an explanation of what is meant by this.

Solution:

```
####################################
# Lab 6
#####################################
library(ggplot2)
# Read in data
# using the interface
# Import Dataset --> From CSV --> browse to the file
#    --> set delimiter to tab --> Change name to USData_cleaned_spring
#    --> Import
#
```

Note: If you are using method 2 (via R command, as discussed in Lab 1) then after importing the dataset run the following code:

```
class(USData_cleaned)
```

If the output is

```
[1] "tbl_df"     "tbl"        "data.frame"
```

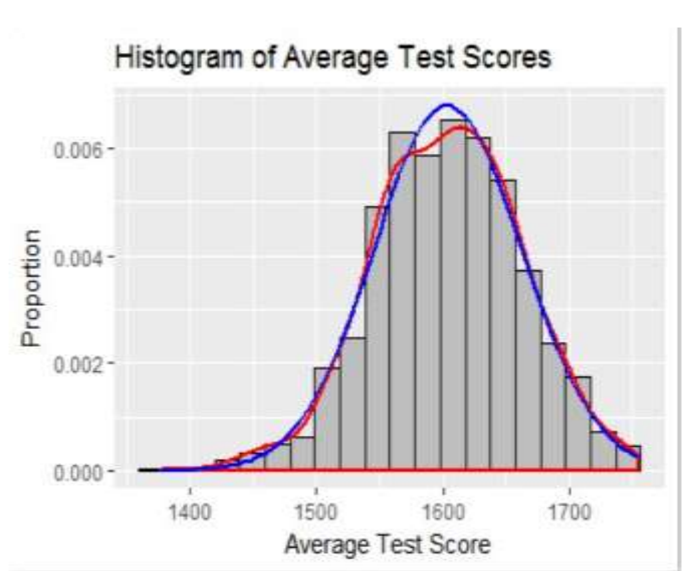
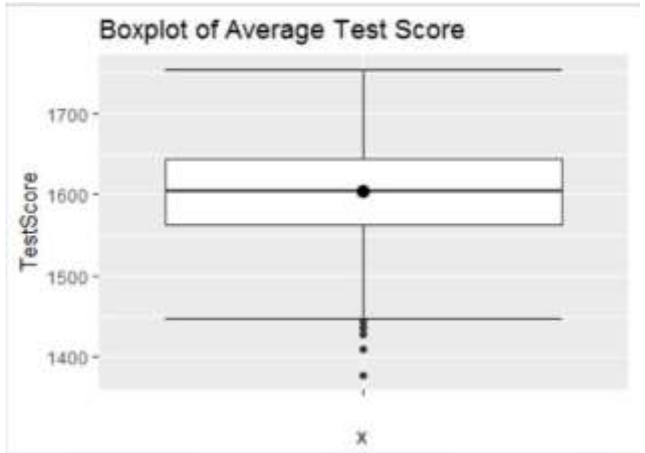
(meaning that it has `tbl_df`, `tbl` and `data.frame` as classes), then use the command

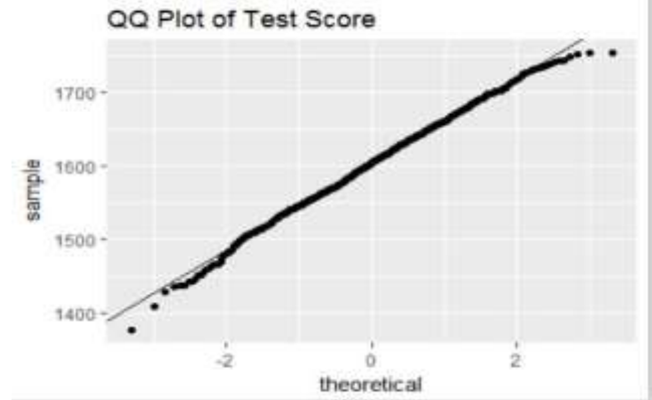
```
USData_cleaned <- as.data.frame(USData_cleaned)
```

```
# 2. Histogram
#
xbar <- mean(USData_cleaned$TestScore)
s <- sd(USData_cleaned$TestScore)
windows()
ggplot(USData_cleaned, aes(x=TestScore)) +
  geom_histogram(aes(y = ..density..), bins = 20,
                 fill = "grey", col = "black") +
  geom_density(col = "red", lwd = 1) +
  stat_function(fun = dnorm, args = list(mean = xbar, sd = s),
                col="blue", lwd = 1) +
  ggtitle("Histogram of Average Test Scores") +
  xlab("Average Test Score") +
  ylab("Proportion")
#
# 2, Boxplot
#
windows()
ggplot(USData_cleaned, aes(x = "", y = TestScore)) +
  stat_boxplot(geom = "errorbar") +
  geom_boxplot() +
  ggtitle("Boxplot of Average Test Score") +
  stat_summary(fun.y = mean, col = "black", geom = "point", size = 3)
#
# 2. QQPlot
#
windows()
ggplot(USData_cleaned, aes(sample = TestScore)) +
  stat_qq() +
  geom_abline(slope = s, intercept = xbar) +
  ggtitle("QQ Plot of Average Test Score")
#
# 4. Mean and SD
#    SE (may be done via computer or via hand)
#
xbar # These were calculated when we made the histogram
s
s/sqrt(nrow(USData_cleaned)) # SE, optional
#
# 5./7. Confidence Interval and Hypothesis Test (vs. 1800)
#
t.test(USData_cleaned$TestScore, conf.level = 0.95, mu = 1800,
       alternative = "two.sided")
#
# 8. Hypothesis test (vs. 1608)
#)
t.test(USData_cleaned$TestScore, conf.level = 0.95, mu = 1608,
       alternative = "two.sided")
#
# 10. Bonus
#
t.test(USData_cleaned$TestScore, conf.level = 0.95, mu = 1800,
       alternative = "greater")
```

2. (10 points) Create a histogram, boxplot, and a Normal probability plot using these data. Briefly describe each plot and what it indicates about the data. After you have described each graph, provide a general description of the data including information about the shape of the distribution and any unusual aspects of the data, including outliers.

Solution:

	The histogram is unimodal and symmetric. The estimated density curve closely follows that of the theoretical normal density. This suggests that the distribution of Average Test Score is approximately normal.
	The boxplot reveals a generally symmetric shape. It also indicates that there is a potential low outlier; the Test Score less than 1400. However, it is not so dramatic that we have to question the validity of a normal model for the data or t-procedures.

	The points on the QQ Plot mostly follow the line, with a minor concave pattern exhibited towards the high range of Quantile. This indicates that the distribution resembles a normal distribution, but has a weak skew.

The distribution is well approximated by a normal distribution. While there are small deviations in the form of minor skewness and a potential outlier, these are not too dramatic with the large sample size and it is appropriate to deem this data approximately normal.

3. (6 points) Is it appropriate to analyze these data using the t-procedures? Explain your response by stating what the assumptions are and then use the graphs to justify your answer. Be sure to include all of the assumptions even those that can not be confirmed via graphs so have to be assumed.

Solution:

The first assumption is that the data come from a simple random sample. This can be assumed true for this lab. The second is that the data come from a normal distribution or the conditions are in place for us to take advantage of CLT. In our case, the sample size is large and the distribution is unimodal with minor skewness. CLT is appropriate in this case, and the distribution of the test statistic will be very well approximated by a t-distribution in spite of the minor skewness.

Note: a condition where CLT may be inappropriate would be the existence of several extreme outliers or extreme skewness with a smaller sample size.

4. (5 points) Find the sample mean, sample standard deviation, and the standard error of the mean (standard deviation of the estimator) for "TestScore." What does this tell you about the data? Please write at least one sentence concerning how these numeric values describe the distribution. You may calculate the standard error by hand or via computer software. If you perform the calculations by hand, please show your work.

Solution:

```
> xbar # These were calculated when we made the histogram
[1] 1603.538
> s
[1] 58.74789
> s/sqrt(nrow(USData cleaned spring)) #SE, optional
[1] 1.772928
```

The sample mean is $\bar{x} = 1603.538$, the sample standard deviation is $s = 58.74789$, and the standard error of the mean is

$$SE(\bar{X}) = \frac{s}{\sqrt{n}} = \frac{58.74789}{\sqrt{1098}} = 1.772928$$

This means that the data is centered at 1603.538 with a relatively small spread. The standard deviation of the sample mean is much smaller than the sample standard deviation so the sampling distribution has a much narrower spread than the population does.

5. (10 points) Find the 95% confidence interval for the mean “TestScore.” Please interpret your result.

Solution:

```
        One Sample t-test

data:  USData cleaned spring$TestScore
t = -110.81, df = 1097, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 1800
95 percent confidence interval:
 1600.059 1607.017
sample estimates:
mean of x 
 1603.538 
```

The 95% confidence interval is (1600.059,1607.017).
We are 95% confident that the population mean of Average Test Score is covered by the interval given by 1600.059 and 1607.017.

6. (5 points) Before performing the hypothesis test (or looking at the output for the hypothesis test), would you reject or fail to reject the claim that the mean of Average Test Score is 1800 at a 5% significance level? Please explain your answer. Hint: Use the results of question 5. You will receive **0 points** if you refer to the results of the hypothesis test.

Solution:

I would reject the claim that the mean of Average Test Score is 1800 at the 5% significance level, since the 95% confidence interval does not contain the value 1800.

Optional comments:
In fact, almost all the data points are less than 1800, so this is result is not a surprise. Since we appropriately specified a hypothesis in advance of seeing the data, it is acceptable that our hypothesis was dramatically inconsistent with the data. There is nothing statistically wrong with testing such a hypothesis, it simply means the high schooler’s guess failed to accurately reflect the reality of “TestScore”.

7. (12 points) Do these data provide evidence that the average “TestScore” is different from 1800? Carry out a hypothesis test using the four-step procedure, with a significance level of 5%. Please provide the relevant output required for the steps and explicitly include all four steps in your answer. No calculations are required because the necessary information is obtained from the software output.

Solution:

```
        One Sample t-test

data:  USData cleaned spring$TestScore
t = -110.81, df = 1097, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 1800
95 percent confidence interval:
 1600.059 1607.017
sample estimates:
mean of x
 1603.538
```

Step 1: Define Parameters

Let μ denote the population mean of Average Test Score

Step 2: State Hypothesis

$H_0: \mu = 1800$

$H_A: \mu \neq 1800$

Step 3: Test Statistic, Degrees of Freedom, P-value

$t_{ts} = -110.81$

$df = 1097$

$\text{p-value} < 2.2 \times 10^{-16}$

Step 4: Conclusion

Since $2.2 \times 10^{-16} < 0.05$, we reject the null hypothesis.

The data shows strong support (p = 2.2×10^{-16}) for the claim that the population mean of Average Test Score is not equal to 1800.

8. (12 points) Another high school student, Bhudevi, thinks that the percentage should be closer to $^2/_3$'s (67%). That would mean that we are testing to see if the Average Test Score to 0.67 × 2400 = 1608. Do these data provide evidence that the average “TestScore” is different from 1608? Carry out a hypothesis test using the four-step procedure, with a significance level of 5%. Please provide the relevant output required for the steps and explicitly include all four steps in your answer. No calculations are required because the necessary information is obtained from the software output.

Solution:

```
        One Sample t-test

data:  USData_cleaned_spring$TestScore
t = -2.5168, df = 1097, p-value = 0.01199
alternative hypothesis: true mean is not equal to 1608
95 percent confidence interval:
 1600.059 1607.017
sample estimates:
mean of x 
 1603.538 
```

Step 1: Define Parameters
Let μ denote the population mean of Average Test Score

Step 2: State Hypothesis
$H_0: \mu = 1608$
$H_A: \mu \neq 1608$

Step 3: Test Statistic, Degrees of Freedom, P-value
t_{ts} = -2.5168
$df = 1097$
$\text{p-value} = 0.01199$

Step 4: Conclusion
Since $0.01199 < 0.05$, we reject the null hypothesis.
The data might show support (p = 0.01199) for the claim that the population mean of Average Test Score is not equal to 1608.

9. (20 points) What would you tell the high school students concerning the population mean of Average Test Score ("TestScore")? Remember that you performed two separate hypothesis tests so that you will have to provide an answer to each student, Antonio (hypothesis test with μ_0 = 1800) and Bhudevi (hypothesis test with μ_0 = 1608). Make sure that you do not use technical terms. However, you do need to justify your answers using the results of your inference and additional information. New information (not already presented in the report) is required for full credit. This information may be additional calculations or from personal knowledge. In your discussion, please include the answers to the following questions:

a) Is it appropriate to perform the hypothesis test?
b) Is there an effect? If there is an effect, how big a difference is there?
c) In practical terms, what would you tell the student concerning their hypotheses?
d) Can your conclusion be generalized to the states that were not included in the data set?

The answers to questions a) and d) only need to be provided once. Please explain all of your answers and write them in complete English sentences. Full credit for b) - d) will not be given for answers without explanations.

Solution:

I am writing this to answer each of the questions to make it easier to see. It is acceptable if you write this in paragraph form.

a)The assumptions required to perform the calculation are met. We are assuming that the data is representative of the population and that the shape of the variables are appropriate.

To Antonio:

b) The assumption of a mean of 1800 is far outside of the range of 1600.059 to 1607.017. The difference of roughly 1800 – 1607 = 193 points is a very large effect.

c) Practically, 193 is a very large difference. Therefore, I would say that the mean of the Average Test Score is definitely not 1800, but much lower.

To Bhudevi:

b) The assumption of a mean of 1608 is very close to the range of 1600.059 to 1607.017. In fact, the difference is only about 1608 – 1607 = 1 point.

c) I would say that this difference is practically not important when applying to a college. Therefore, I would say that the mean of the Average Test Score could be 1608. Note that even though it is not practically important, the inference stated that it was statistically significant.

d) Without analyzing which states are missing from the data set, I would assume that the missing states are spread out over the whole country; therefore, the states that are included would be similar to the ones that are missing. Therefore, the answer is yes, we can generalize this conclusion to the states that are not explicitly included.

10. (10 points) BONUS How are the results for calculating the lower confidence bound different from calculating the lower limit of the confidence interval?

a) (5 points) Code to create a lower 95% confidence bound for the mean of Average Test Score. Even if the code is in Part 1, please repeat it here.

Solution:

```
# 10. Bonus
#
t.test(USData_cleaned_spring$TestScore, conf.level = 0.95, mu = 1800,
alternative = "greater")
```

b) (5 points) Report the confidence bound (provide the appropriate output) and the output for the confidence interval (copied from Part 5). How are the results of the two calculations different? Specifically, which quantities in the formulae are different and which is larger, the lower confidence bound or the lower limit of the confidence interval?

Solution:

Lower Bound	Confidence Interval
One Sample t-test data: USData_cleaned_spring$TestScore t = -110.81, df = 1097, p-value = 1 alternative hypothesis: true mean is greater than 1800 95 percent confidence interval: 1600.619 Inf sample estimates: mean of x 1603.538	One Sample t-test data: USData_cleaned_spring$TestScore t = -110.81, df = 1097, p-value < 2.2e-16 alternative hypothesis: true mean is not equal to 1800 95 percent confidence interval: 1600.059 1607.017 sample estimates: mean of x 1603.538
The lower confidence bound is 1600.619 In other words, the "interval" is (1600.619, ∞)	The 95% confidence interval is (1600.059, 1607.017).

The lower bound is one-sided while the confidence interval is two-sided. In addition, the lower bound is 1600.619 when the lower part of the range of the confidence interval is 1600.059.

The formulae for the lower confidence bound and two-sided confidence intervals are given below:

One-sided bound (lower)

$$\bar{x} - t_{\alpha,n-1} \times \frac{s}{\sqrt{n}}$$

Two-sided interval.

$$\bar{x} \pm t_{\frac{\alpha}{2},n-1} \times \frac{s}{\sqrt{n}}$$

The difference between the two formulae is the t critical value. The two-sided interval uses $t_{\frac{\alpha}{2},n-1}$ whereas the bound uses $t_{\alpha,n-1}$, and $t_{\frac{\alpha}{2},n-1} > t_{\alpha,n-1}$. In our case,

$$t_{\frac{\alpha}{2},n-1} = t_{0.025,1097} = 1.962 > 1.646 = t_{0.05,1097} = t_{\alpha,n-1}$$

The remaining components of the formulae are the same between the two cases. This causes the lower bound to be higher than the lower limit of the two-sided interval.