Jordan Mayer
STAT 350
Lab 06
March 22, 2018

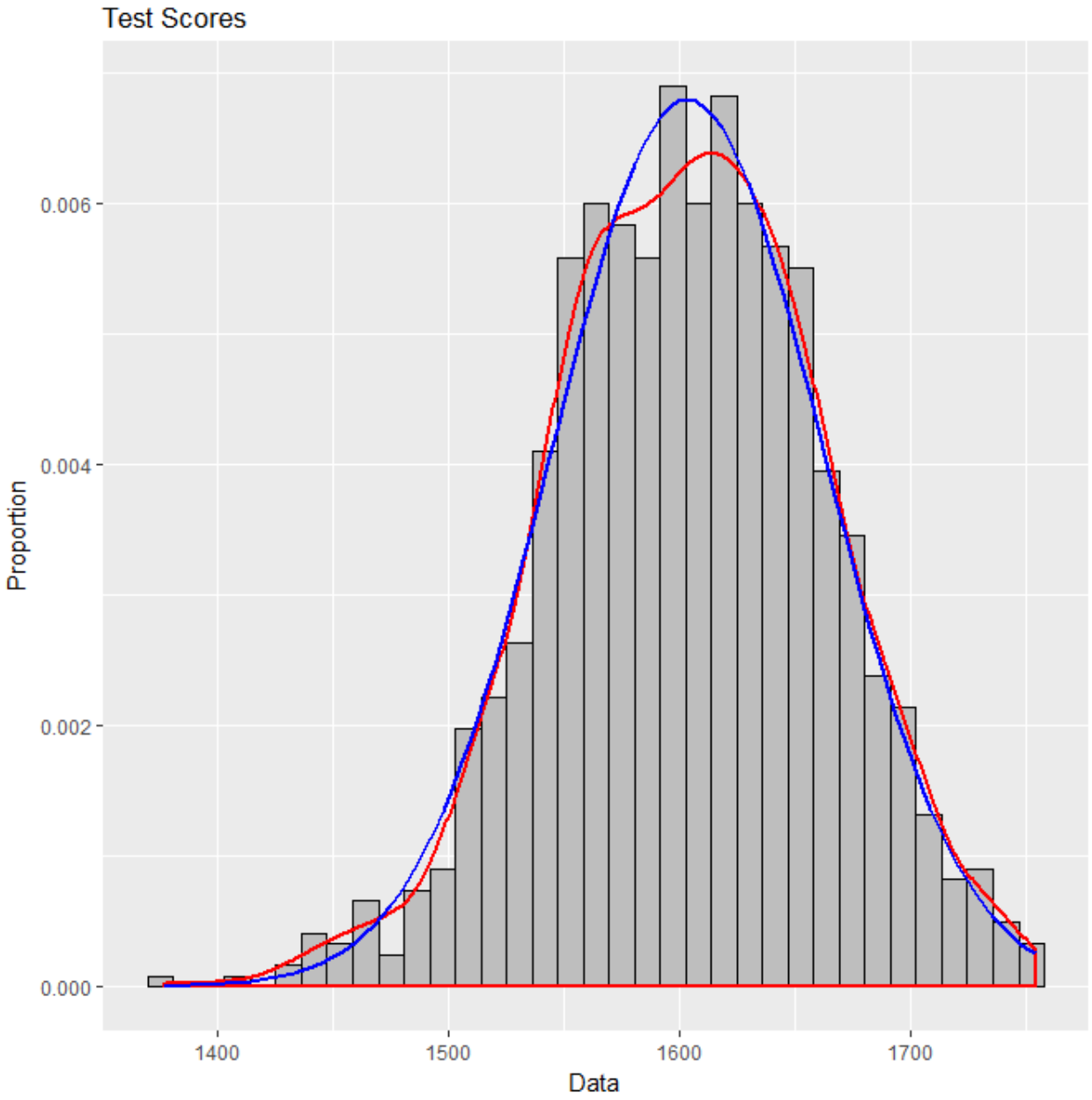
Part B.

1. Code

```
#####
# Jordan Mayer
# STAT 350
# Lab 06
# March 22, 2018
#####

# set working directory to STAT 350
setwd("C:/Users/jordan/Google Drive/Courses Spring 2018/STAT 350/STAT 350 Labs/Lab 06")
# set up ggplot2 for plotting
library(ggplot2)
# close open figures
graphics.off()
# get US Data
USData <- read.table("US_Data.txt", header=TRUE, sep="\t")
USData_clean <- USData[complete.cases(USData),]
attach(USData_clean)

# create histogram
windows()
ggplot(data.frame(TestScore=TestScore), aes(TestScore))+
  geom_histogram(aes(y=..density..),
                 bins=sqrt(length(TestScore))+2,
                 fill="grey",col="black")+
  geom_density(col="red",lwd=1)+
  stat_function(fun=dnorm,args=list(mean=mean(TestScore),
                                    sd=sd(TestScore)),
                col="blue",lwd=1)+
  ggtitle("Test Scores")+
  xlab("Data")+
  ylab("Proportion")
# create boxplot
windows()
ggplot(data.frame(TestScore=TestScore), aes(x="", y=TestScore))+
  stat_boxplot(geom="errorbar")+
  geom_boxplot()+
  ggtitle("Test Scores")+
  stat_summary(fun.y=mean,col="black",geom="point",size=3)
# create normal probability plot
windows()
ggplot(data.frame(TestScore=TestScore),aes(sample=TestScore))+
  stat_qq()+
  geom_abline(slope=sd(TestScore),intercept=mean(TestScore))+
  ggtitle("Test Scores")+
  xlab("Theoretical")+
  ylab("Sample")

# display sample mean, sample standard deviation, and sample
# standard error of the mean
sample_mean<-mean(TestScore)  # sample mean
sample_stdev<-sd(TestScore)      # sample standard deviation
sample_sem<-sd(TestScore)/sqrt(length(TestScore))  # sample standard
                                          # error of mean
sample_mean
sample_stdev
sample_sem

# find 95% confidence interval for mean TestScore
t.test(TestScore,conf.level=0.95,mu=1800,alternative="two.sided")
t.test(TestScore,conf.level=0.95,mu=1608,alternative="two.sided")
t.test(TestScore,conf.level=0.95,mu=1800,alternative="greater")
```

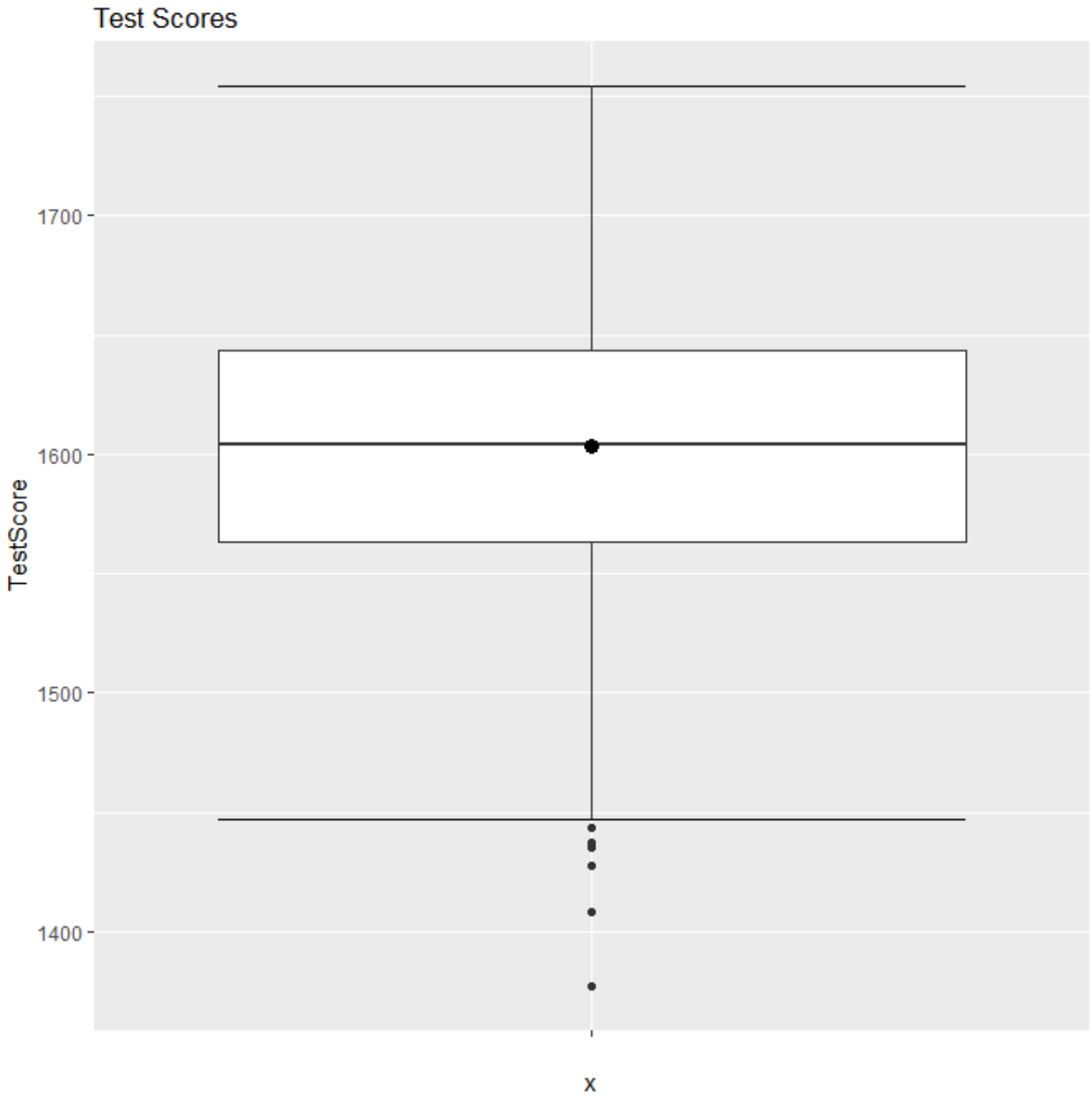
Jordan Mayer
STAT 350
Lab 06
March 22, 2018

2. Histogram, Boxplot, and Normal Probability Plot



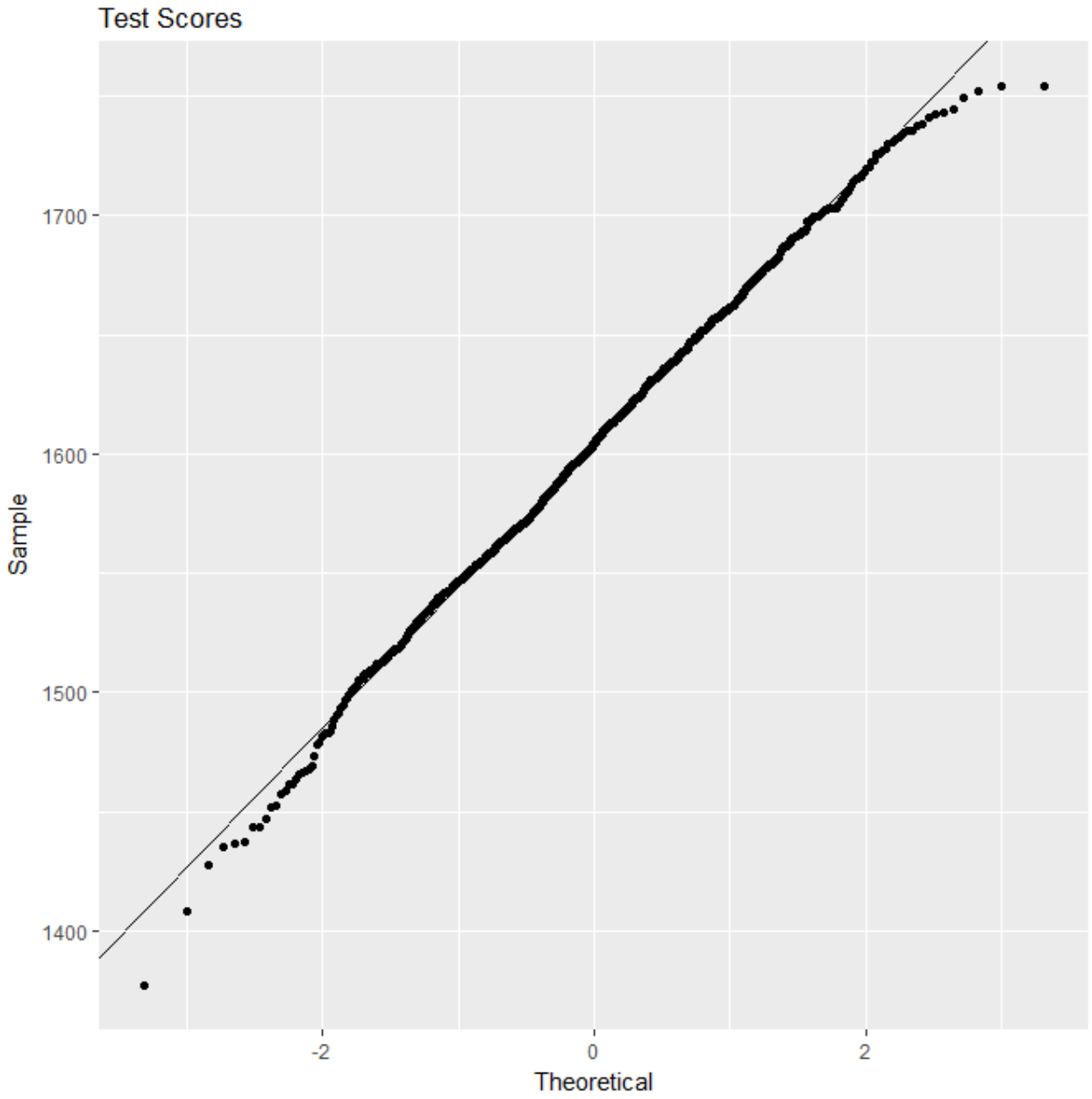
This histogram shows a roughly normal distribution with a slight left skew due to some outliers on the left.

Jordan Mayer
STAT 350
Lab 06
March 22, 2018



This boxplot, as well, shows a roughly normal distribution, with outliers on the negative end of the distribution (the bottom, in this figure).

Jordan Mayer
STAT 350
Lab 06
March 22, 2018



The normal probability plot also shows a nearly normal distribution, with some deviation at the very ends, and more at the negative end of the distribution – in agreement with the histogram and the boxplot.

Jordan Mayer
STAT 350
Lab 06
March 22, 2018

3. To appropriately analyze data using t-procedures, two criteria must be met:

First, the population distribution must be assumed to be normal or the sample size large.

Second, the sample standard deviation must be known, but the population standard deviation unknown.

Here, we assume the population distribution is normal and we do not know the population standard deviation – thus t-procedures are appropriate.

4.

```
> sample_mean
[1] 1603.538
> sample_stdev
[1] 58.74789
> sample_sem
[1] 1.772928
```

From these values, we learn that the sample mean is significantly lower than the population mean supposed by the null hypothesis (1800). We also learn that the standard deviation is significantly lower than the difference between these two means (~200). From the very small sample standard error of the mean, we can infer that our sample mean is fairly consistent – and, thus, may be fairly close to the population mean.

5.

```
alternative hypothesis: true mean is not equal to 1800
95 percent confidence interval:
 1600.059 1607.017
```

This result tells us that the interval from 1600.059 to 1607.017 has a 95% chance of capturing the population mean.

Jordan Mayer
STAT 350
Lab 06
March 22, 2018

6. Based purely on the results of question 5, I would reject the claim that the mean of Average Test Score is 1800 at a 5% significance level. This is because 100-5 = 95 and the 95% confidence interval found in question 5 does not contain 1800.

7.

1. Parameters of interest: μ, population mean of Average Test Score in the US Dataset.
2. Hypotheses:
   H_0: $\mu = 1800$
   H_a: $\mu \neq 1800$
3. Test statistic and P-value:

```
data:  TestScore
t = -110.81, df = 1097, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 1800
```

4. Decision:
   $p << \alpha = 0.05$ → reject H_0.
   The data does not give support (P-value = 2.2e-16) to the claim that the overall average test score in the US is equal to 1800.

8.

1. Parameters of interest: μ, population mean of Average Test Score in the US Dataset.
2. Hypotheses:
   H_0: $\mu = 1608$
   H_a: $\mu \neq 1608$
3. Test statistic and P-value:

```
data:  TestScore
t = -2.5168, df = 1097, p-value = 0.01199
alternative hypothesis: true mean is not equal to 1608
```

4. Decision:
   $p < \alpha = 0.05$ → reject H_0.
   The data does not give support (P-value = 0.01199) to the claim that the overall average test score in the US is equal to 1608.

9. a) It is indeed appropriate to perform the hypothesis test. We have a null hypothesis ($\mu = \mu_0$), we have an alternative hypothesis ($\mu \neq \mu_0$), and we have a large data set such that we can assume the population distribution is normal.

In layman's terms: we have a claim about what the average score is; we have an alternative claim (that the average test score is not that); and we have all the test scores, which are spread out in a way that lets us perform a good hypothesis test.

b) To Antonio: there is a great effect if we claim the average test score is equal to 1800, compared to the test scores we have available to us. The average of the average test scores we know is around 1603, and these scores do not vary a whole lot.

Jordan Mayer
STAT 350
Lab 06
March 22, 2018

To Bhudevi: there is a minor effect if we claim the average test score is equal to 1608, compared to the test scores available to us. The average of the average test scores we know is around 1603, and these scores do not vary a whole lot.

c) To Antonio: after crunching some numbers, we know that the interval from roughly 1600 to roughly 1607 has a 95% chance of capturing the true overall average test score. Clearly, this is very far from 1800. Now, there is always the possibility that the average test scores we analyzed were just flukes and don't really reflect most average test scores in this respect – but our hypothesis test tells us that the chances of this are less than one *billionth* of a percent. Thus, we have some pretty strong evidence suggesting that the overall average test score is *not* 1800.

To Bhudevi: after crunching some numbers, we know that the interval from roughly 1600 to roughly 1607 has a 95% chance of capturing the true overall average test score. So your hypothesis of 1608 is just outside of this range. Now, there is always the possibility that the average test scores we analyzed were just flukes and don't totally reflect most average test scores in this respect – but our hypothesis test tells us that the chances of this are around 1.1%. Because this is below 5%, we have some evidence to suggest that the overall average test score is *not* 1608.

d) This conclusion cannot be generalized to the states that were not included in the data set. There could be a lot of things that vary from state to state, especially in regards to education systems, so we should not assume the trends we see here will hold in other states as well.

10. a)

```
t.test(TestScore,conf.level=0.95,mu=1800,alternative="greater")
```

b)

```
alternative hypothesis: true mean is not equal to 1800
95 percent confidence interval:
 1600.059 1607.017


alternative hypothesis: true mean is greater than 1800
95 percent confidence interval:
 1600.619      Inf
```

These results are quite close, but differ by 0.56, with the lower confidence bound being slightly larger. Formulaically speaking, the confidence interval uses $t_{\alpha/2}$ as the test variable, while the confidence bound uses t_{α}.