# Sai Lakshman Rao Duggirala – M15848995

#### Wine Project Part B

**library**(tidyverse)

wine<-read.csv(file="D:/Applied Stat methods/winequality-red.csv", header=T, sep = ";")

Question 1

1:Suppose the population mean of the variable “density” is μ , do the following-inferences:  
a. Provide an estimate of μ based on the sample

mean\_density <- mean(density)

mean\_density

## [1] 0.9967467

The estimate of μ based on the sample is calculated by using the mean. The value is 0.9967467.

b. Use the Central Limit Theorem (CLT) to quantify the variability of your estimate

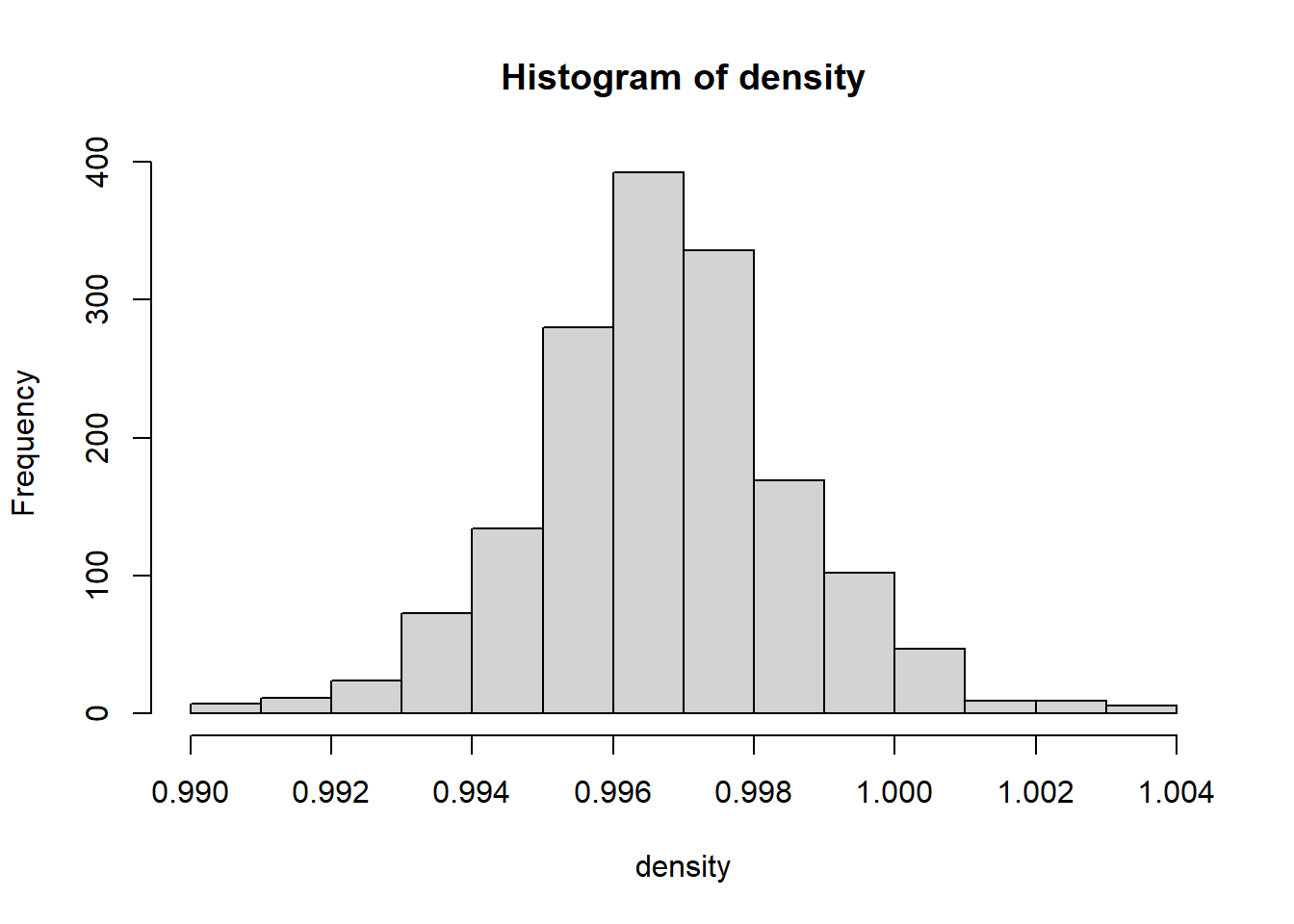
sd\_density <- sd(density)/sqrt(length(density))

sd\_density

## [1] 4.71981e-05

hist(density)

The histogram representation here shows the distribution of density before considering the central limit theorem. We further investigate with the use of Central Limit Theorem and we will plot a new histogram which can identify whether the distribution is normal or not.



From this we can observe that the mean value is very close to 1 and the standard deviations is near 0. With the use of Central Limit Theorem, we can showcase that the distribution of the density is normal distribution using histogram.

1. Use the CLT to give a 95% confidence interval for μ.

c(mean\_density-2\*sd\_density, mean\_density+2\*sd\_density)

## [1] 0.9966523 0.9968411

By using the formula to identify the confidence interval for 95% for Central Limit Theorem, we can observe the interval lies in between 0.9966523 and 0.9968411

1. Use the bootstrap method to do parts b and c, and compare the results with those obtained from the CLT. State your findings.

density\_mu.set <- NULL

**for**(k **in** 1:3500) {

bootstrap <- sample(density, size=1599, replace=T)

mu\_density <- mean(bootstrap)

density\_mu.set[k] <- mu\_density

}

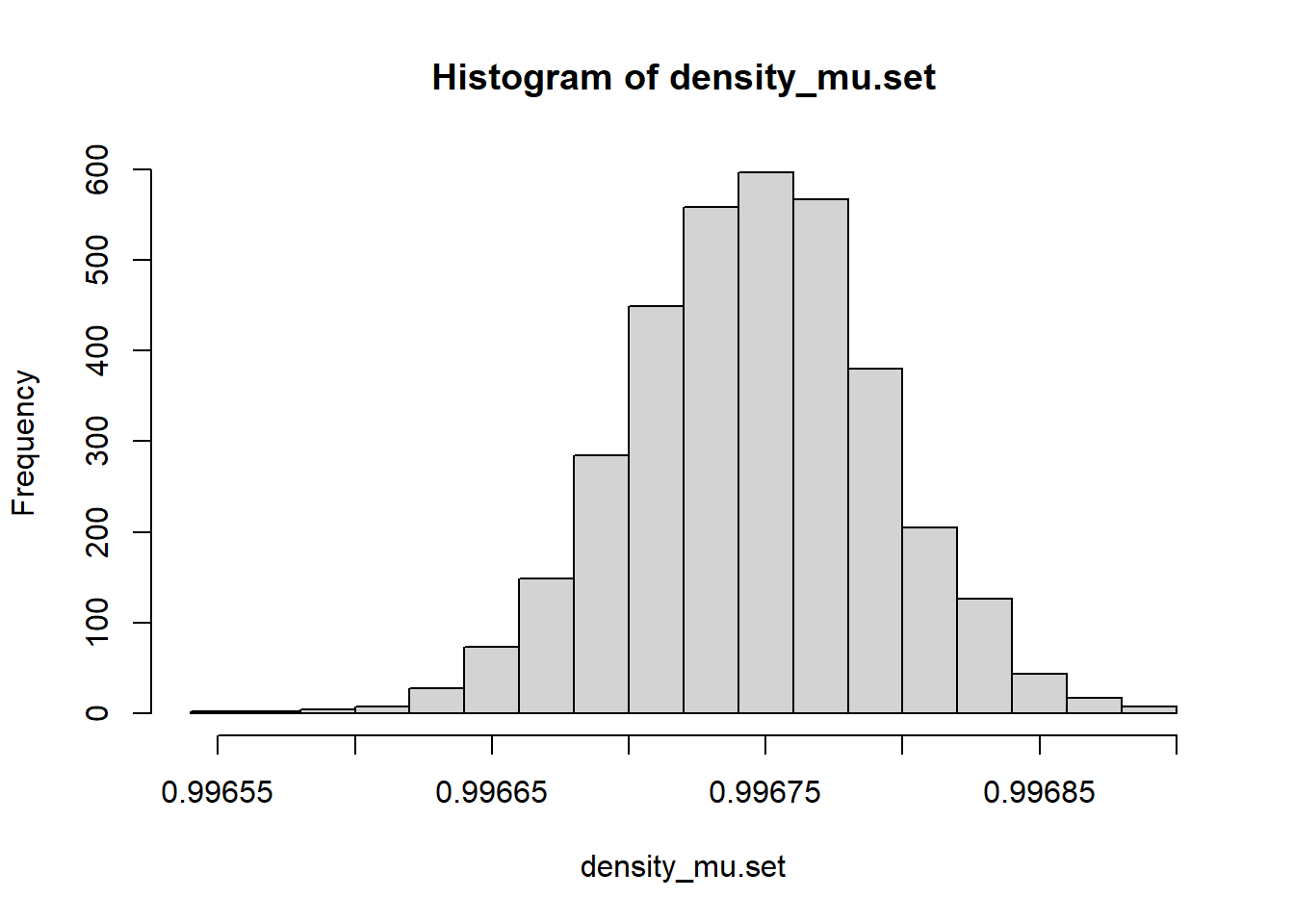
mean(density\_mu.set)

[1] 0.9967457

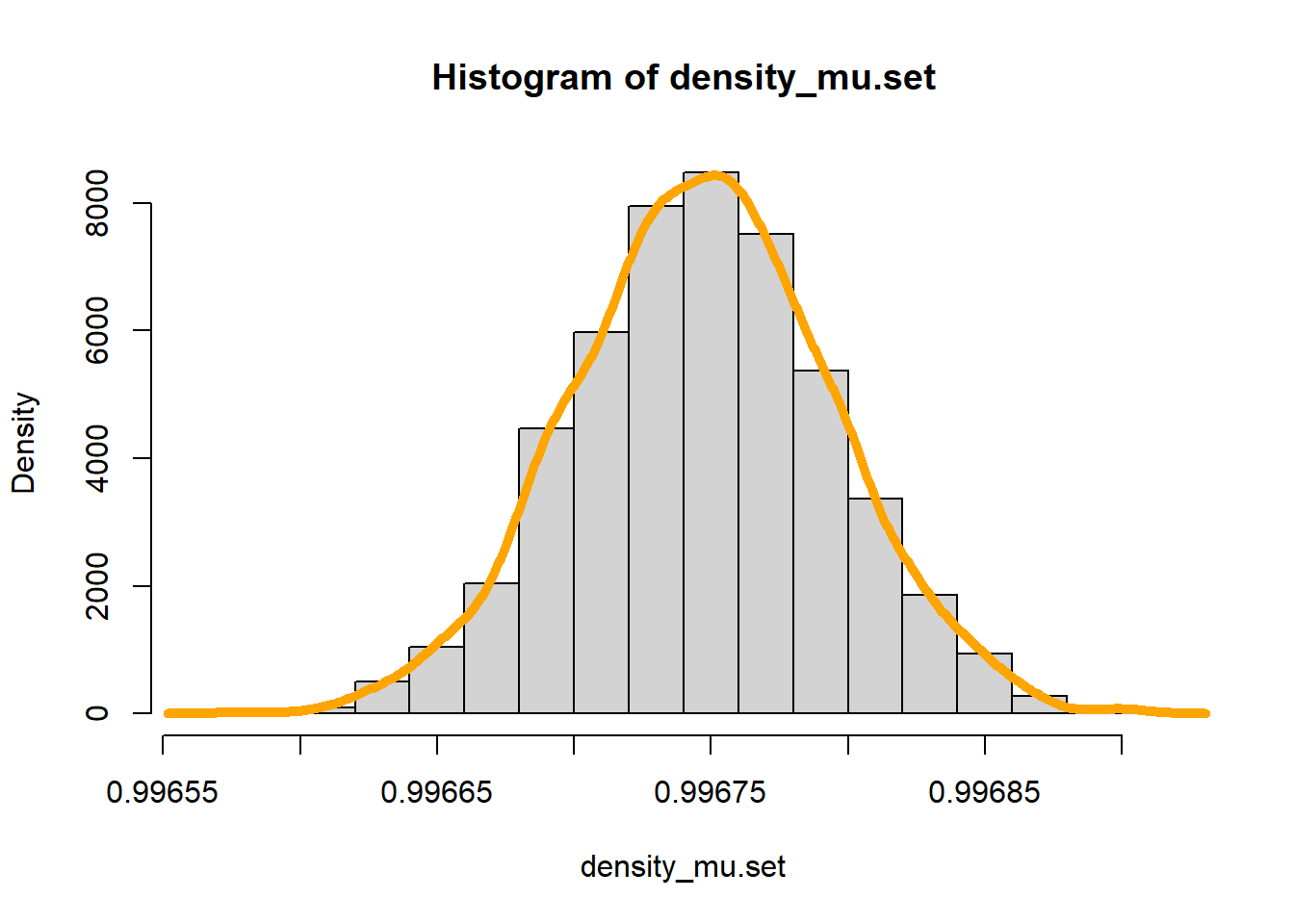
sd(density\_mu.set)

## [1] 4.694278e-05

hist(density\_mu.set)



lines(density(density\_mu.set), lwd=5, col='orange')



*### Bootstrap Confidence Interval*

boot\_ci <- quantile(density\_mu.set, probs = c(0.025,0.975))

boot\_ci

## 2.5% 97.5%

## 0.9966571 0.9968373

From the above set of results, we can say that the bootstrap method also gives similar results to central limit theorem. The sample may vary slightly each time the loop is run but the results would be close to central limit theorem. We can also identity that the loop is a normal distribution.

1. Suppose the population mean of the variable “residual sugar” is μ , answer the following questions.
2. Provide an estimate of μ based on the sample

resid\_sugar<- mean(residual.sugar)

resid\_sugar

## [1] 2.538806

The estimate of μ based on the sample is calculated by using the mean. The value is 2.538806.

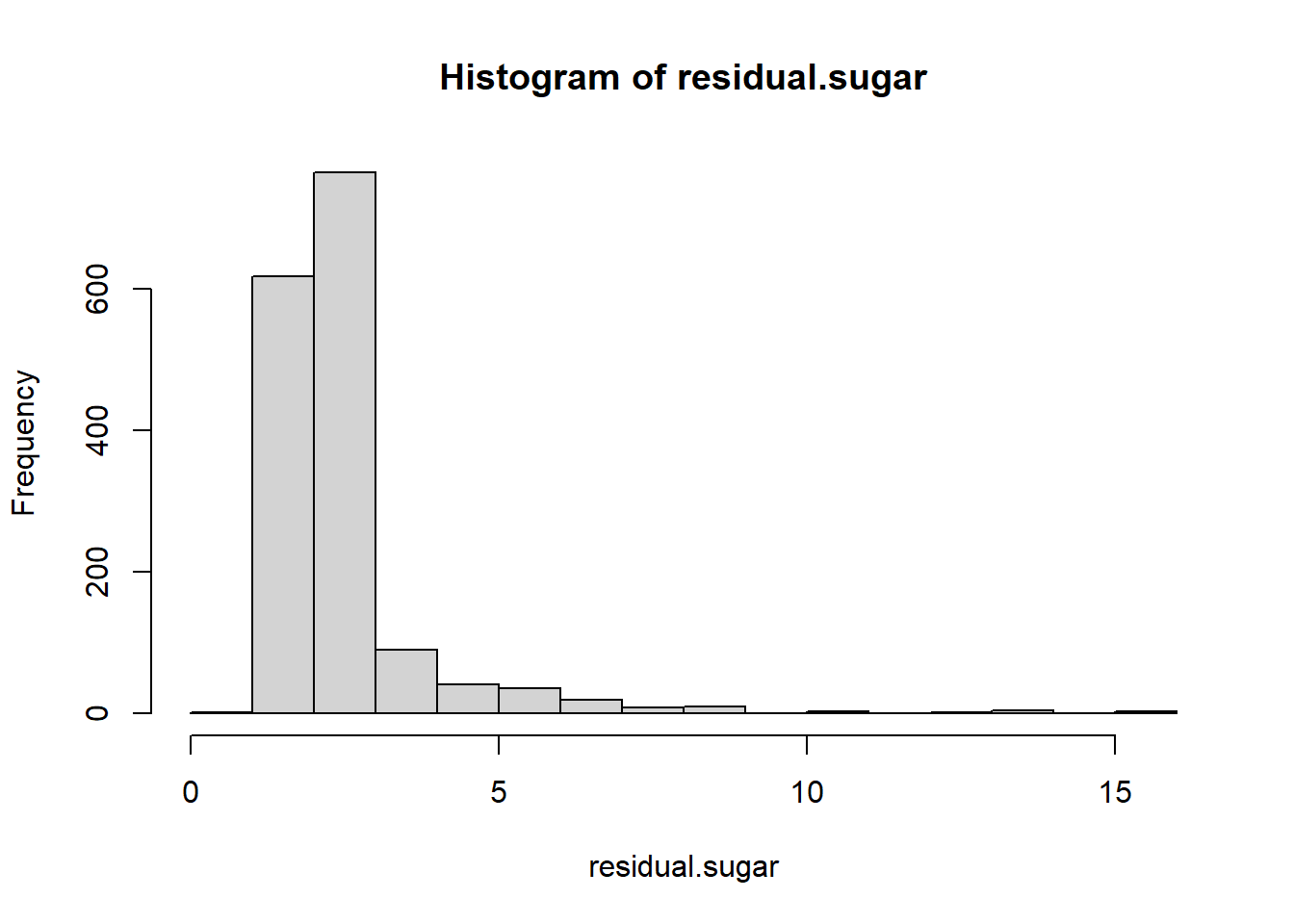
1. Noting that the sample distribution of “residual sugar” is highly skewed, can we use the CLT to quantify the variability of your estimate? Can we use the CLT to give a 95% confidence interval for μ? If yes, please give your solution. If no, explain why.

sd\_resid\_sugar<- sd(residual.sugar)/sqrt(length(residual.sugar))

sd\_resid\_sugar

## [1] 0.03525922

hist(residual.sugar, freq = FALSE, ylim= c()))



We can see from the histogram for residual sugar, that the distribution is not symmetric. We further investigate it with use of Central Limit Theorem and we will plot a new histogram to identify whether the distribution is normal or not.

c(resid\_sugar-2\*sd\_resid\_sugar, resid\_sugar+2\*sd\_resid\_sugar)

## [1] 2.468287 2.609324

By using the formula to identify the confidence interval for 95% for Central Limit Theorem, we can observe the interval lies in between 2.468287 and 2.609324. With this we can say that the Central Limit Theorem can be used for residual sugar.

C. Use the bootstrap method to do part b. Is the bootstrap confidence interval symmetric?

mu\_sugar.set <- NULL

**for**(k **in** 1:3500) {

sugar.bootstrap <- sample(residual.sugar, size=1599, replace=T)

mu\_sugar <- mean(sugar.bootstrap)

mu\_sugar.set[k] <- mu\_sugar

}

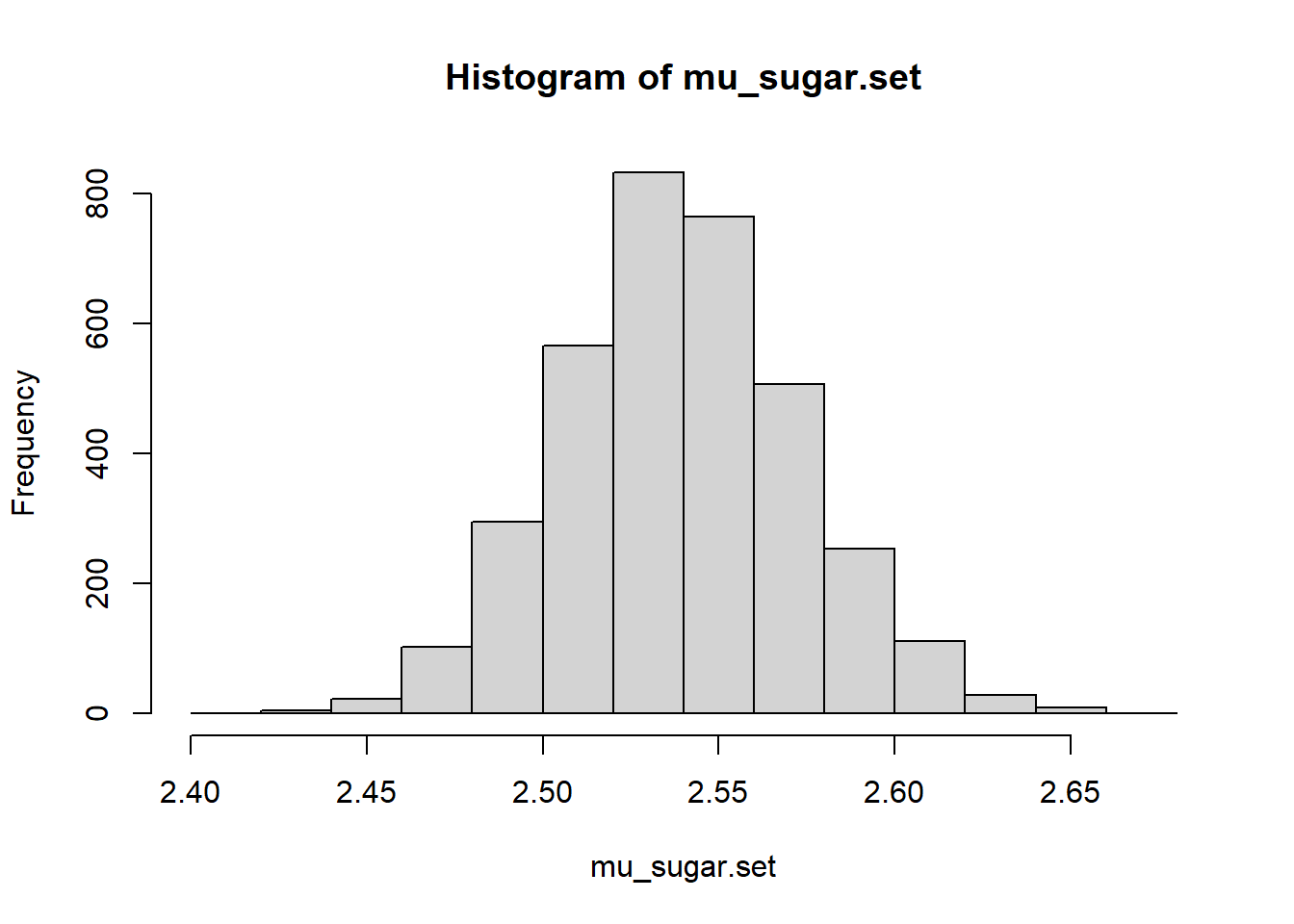
mean(mu\_sugar.set)

## [1] 2.539393

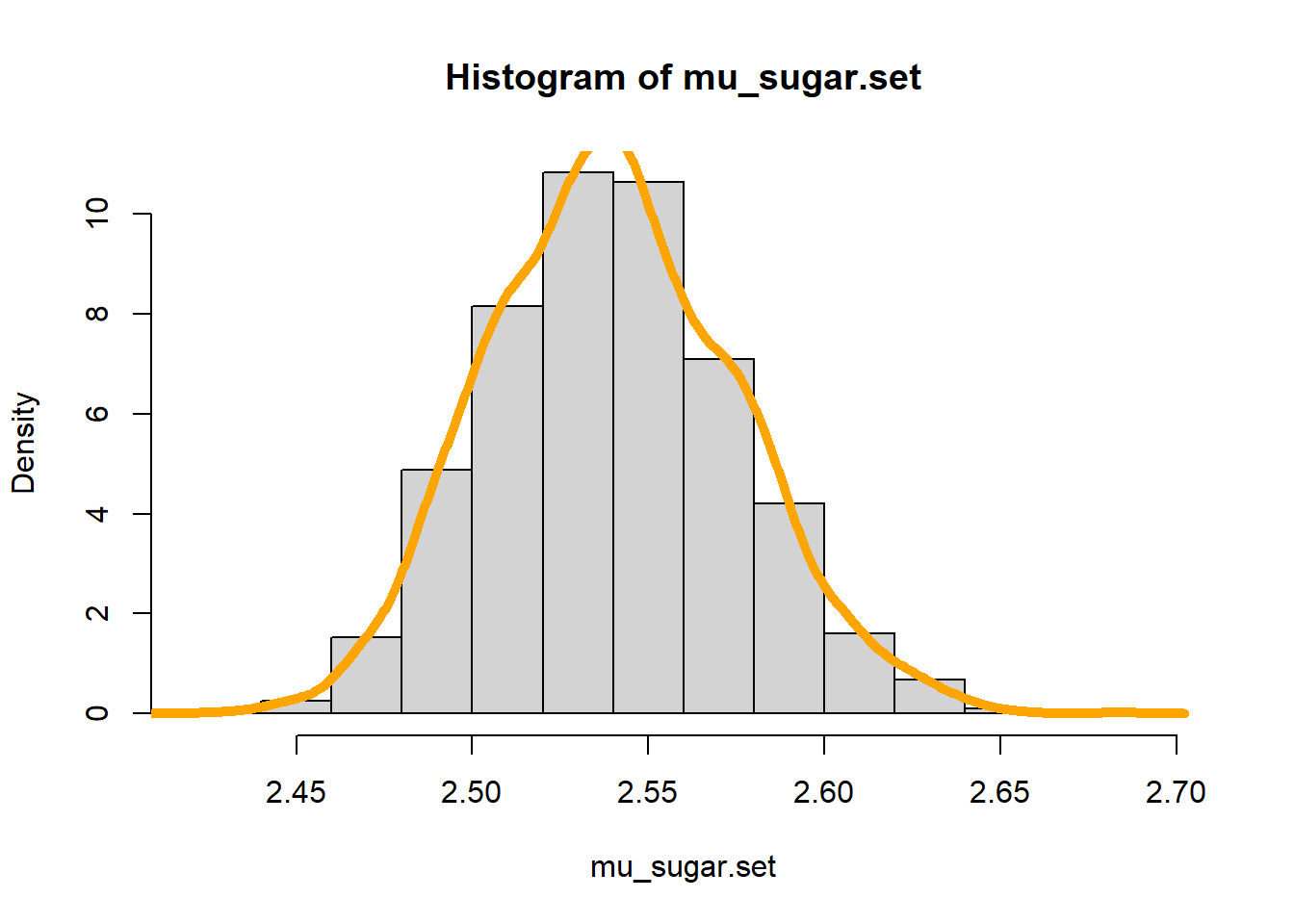
sd(mu\_sugar.set)

## [1] 0.03526239

hist(mu\_sugar.set, freq = FALSE, ylim= c())



lines(density(mu\_sugar.set), lwd=5, col='orange')



*### Bootstrap Confidence Interval*

boot\_ci <- quantile(resid\_sugar, probs = c(0.025,0.975))

boot\_ci

## 2.5% 97.5%

## 2.538806 2.538806

With this confidence interval formula for Bootstrap, we can identify that the values of mean are identical. We can use bootstrap method as well as we can clearly state that the distribution is normal and looks symmetric in nature from the histogram.

1. We classify those wines as “excellent” if their rating is at least 7. Suppose the population proportion of excellent wines is p. Do the following:
2. Use the CLT to derive a 95% confidence interval for p;

excellent\_wine<- as.numeric(wine$quality> 6)

mean\_excellent<-mean(excellent\_wine)

mean\_excellent

## [1] 0.1357098

variance\_excellent<- sd(excellent\_wine/sqrt(excellent\_wine))

variance\_excellent

## [1] 0.00856736

c(mean\_excellent-2\*variance\_excellent, mean\_excellent+2\*variance\_excellent)

## [1] 0.1185751 0.1528445

The confidence interval using Central Limit Theorem lies between 0.1185751 and 0.1528445.

1. Use the bootstrap method to derive a 95% confidence interval for p;

mu\_quality.set <- NULL

**for**(k **in** 1:3500) {

quality.bootstrap <- sample(excellent\_wine, size=1599, replace=T)

mu\_quality <- mean(quality.bootstrap)

mu\_quality.set[k] <- mu\_quality

}

mean(mu\_quality.set)

## [1] 0.1359182

*### Bootstrap Confidence Interval*

quantile(mu\_quality.set, probs = c(0.025,0.975))

## 2.5% 97.5%

## 0.1194497 0.1525954

With this confidence interval formula for Bootstrap, we can identify that the values of mean are identical.

1. Compare the two intervals. Is there any difference worth our attention?

With the comparison from above results, we can clearly identify that the intervals are similar and there is no specific investigation needed further.