Final Project Report

**Submitted to:**

Kevin Donges

**Created by:**

Harshil Patel

***Statistics 3201***

***The Ohio State University***

***Columbus, OH***

***3rd December 2018***

**Table of Contents**

Part 1……………………………………………………………………...3

*Overview……………………………………………………..................*3

*Results……………………………………………………......................*3

*Conclusion……………………………………………………...............*5

Part 2……………………………………………………………………...6

*Overview……………………………………………………..................*6

*Results……………………………………………………......................*6

*Conclusion……………………………………………………...............*8

Part 3……………………………………………………………………...9

*Overview……………………………………………………..................*9

*Results……………………………………………………......................*10

*Conclusion……………………………………………………...............*11

Appendices……………………………………………………………….13

*Appendix A: Program Codes …………………………………………*13

Contact……………………………………………………………………22

**Part 1**

***Overview:***

The average time spent studying for an exam in hours was modeled with three different probability distributions of the time spend studying for an individual. In this simulation using R and Studio, 1,000 random samples of size 25 (individuals) were generated from the distributions below. The program operating code is shown in *Appendix A.1A* through *Appendix A.1C*.

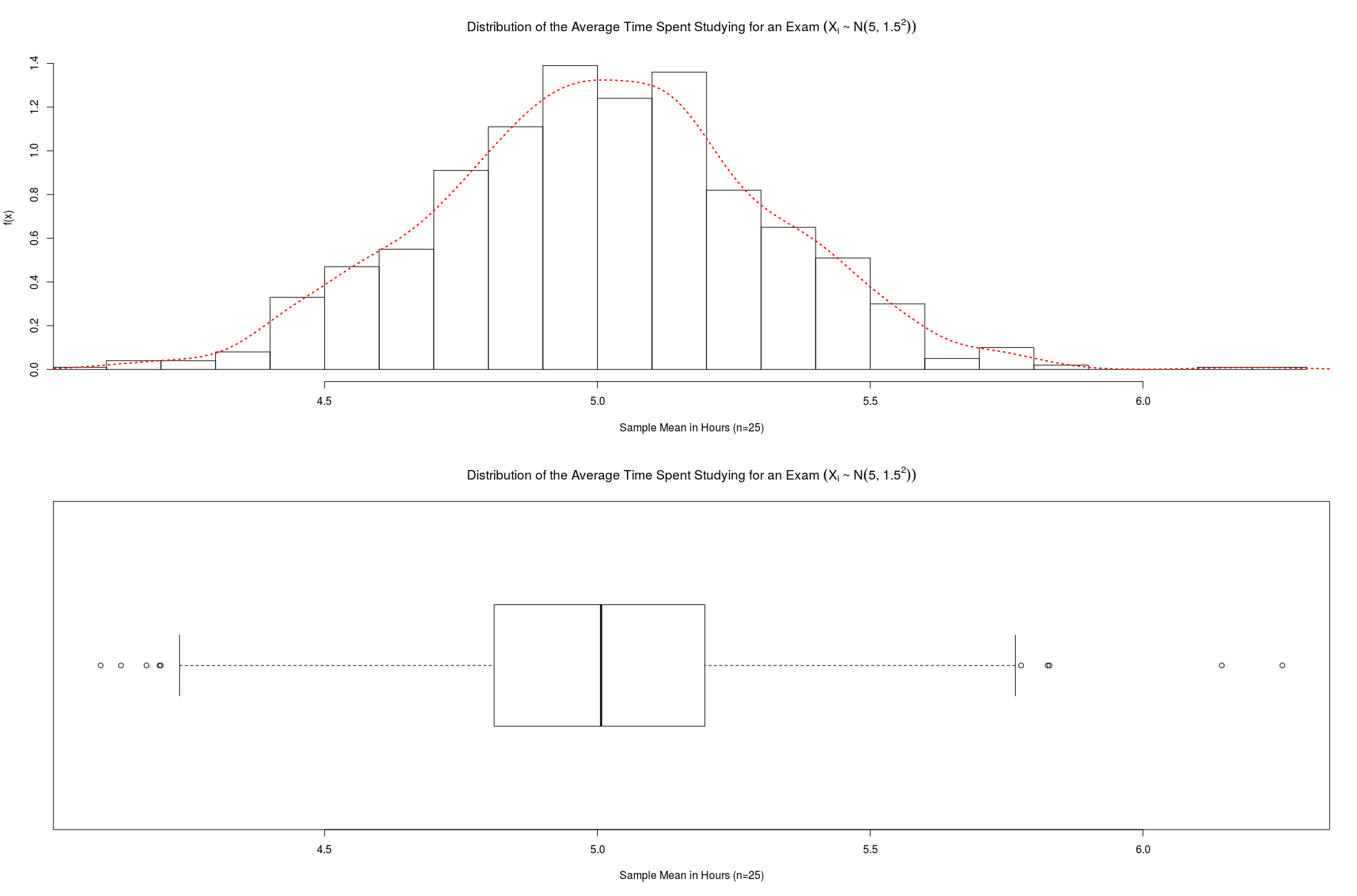
Additionally, the empirical distribution of the 1000 sample means were plotted, and the estimated mean of the sample mean, and estimated standard deviation of the sample mean were determined for each distribution below, along with the theoretical values of the former & latter.

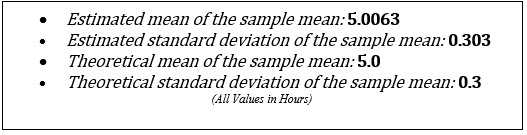
The three distribution of an individual’s time spent studying for an exam in hours are:

1. Normally distributed, with mean 5 hours and standard deviation of 1.5 hours.
2. Uniformly distributed with minimum of 0 hours to maximum of 10 hours.
3. Gamma distribution with parameters:.

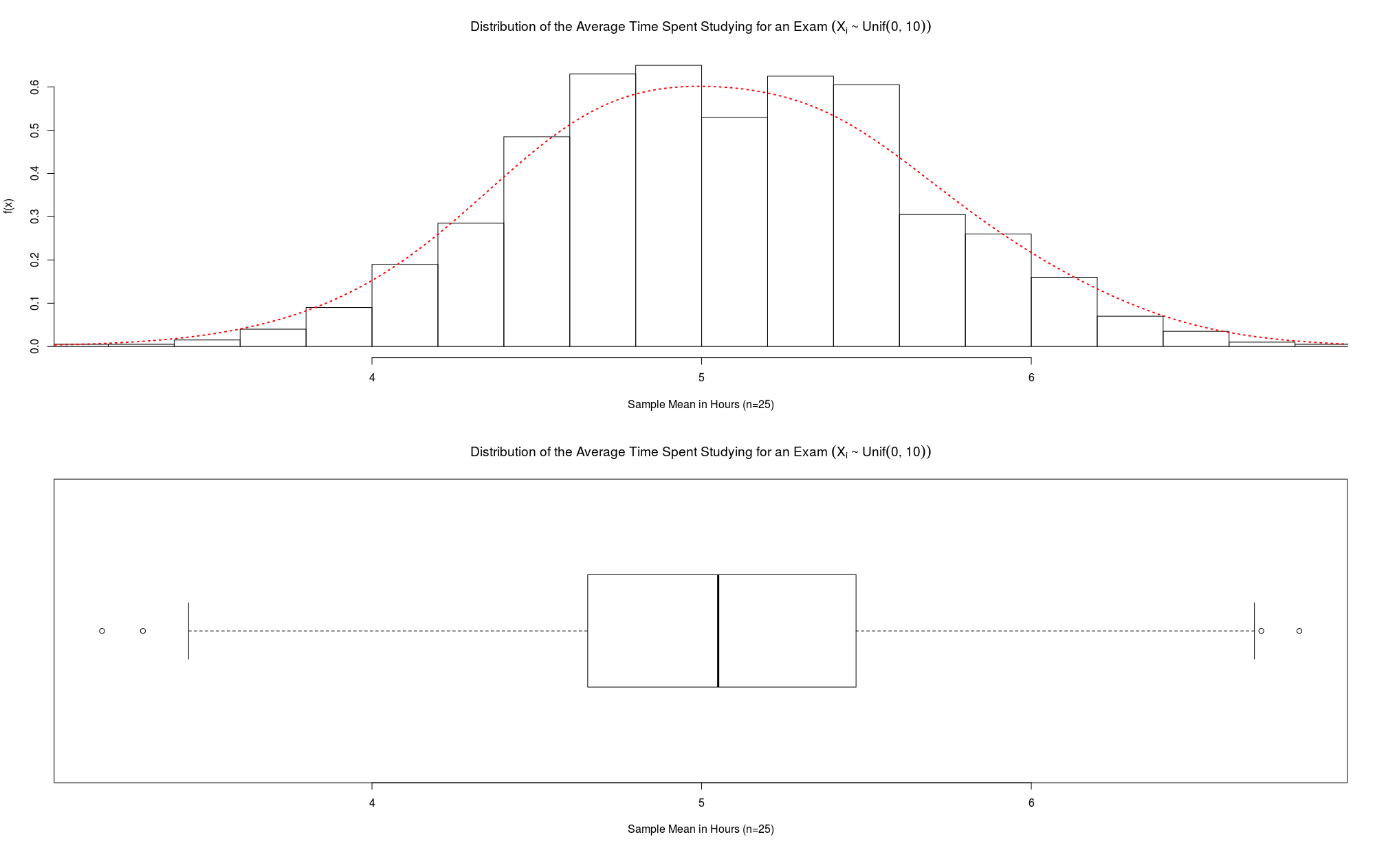
***Results:***

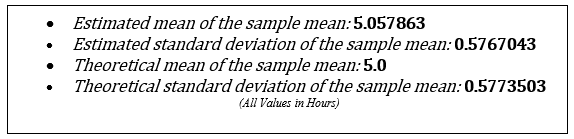
A

******

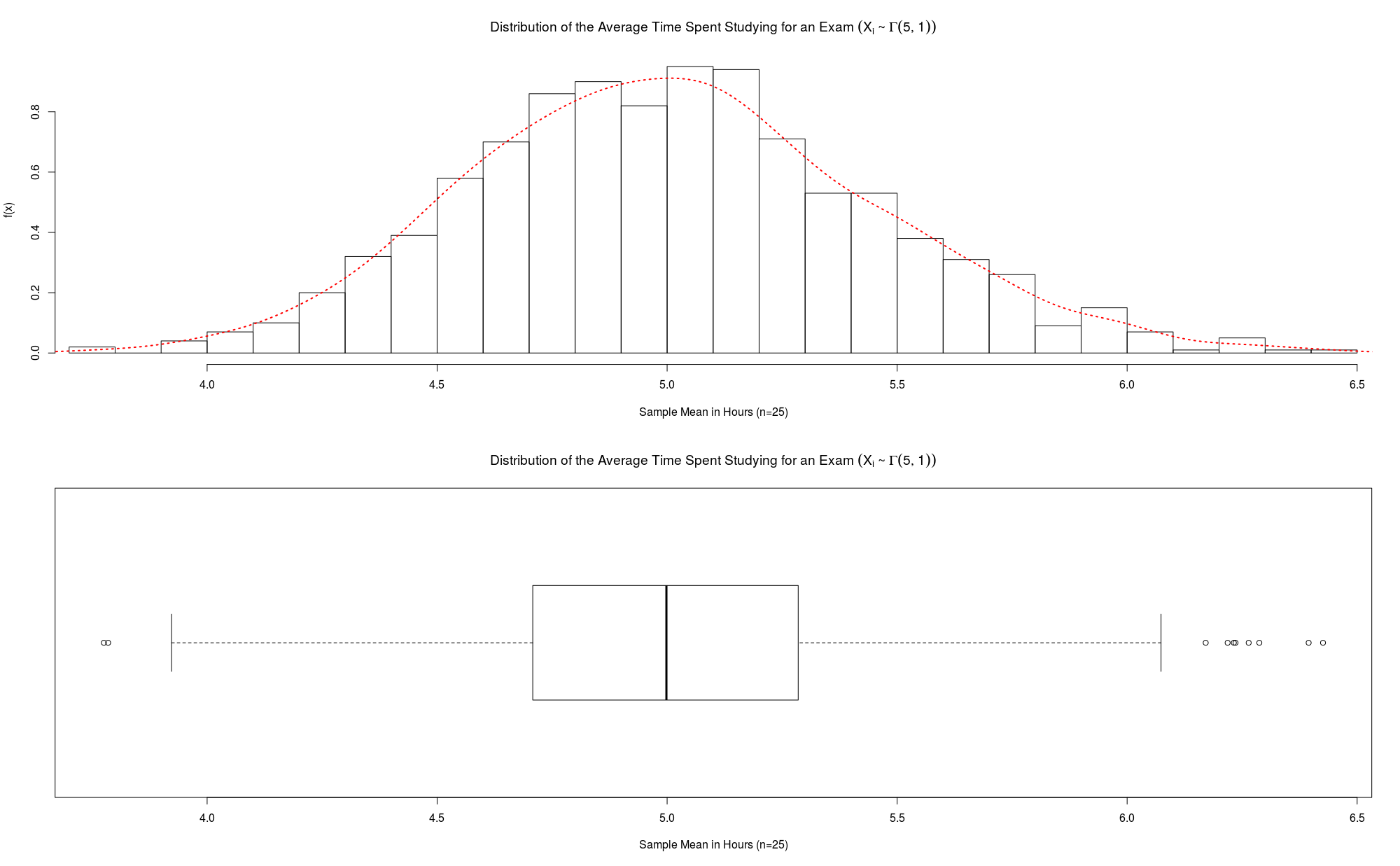


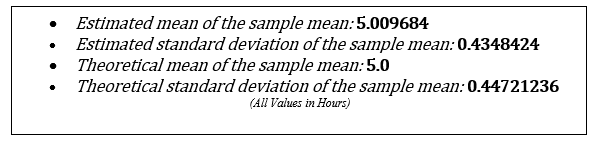
B

******



C

******



***Conclusion:***

The time spend studying for an individual in hours can be assumed to be independent and identically distributed. When regarding at the characteristics of the 1,000 random samples of size 25 (individuals) taken from the three distributions, it can be concluded that the estimated mean and variance of the sample means closely approximates the theoretical mean and variance for all the distributions of an individual’s time spent studying for an exam in hours. But they cannot be exactly the same value due to sampling error. Sampling error arises from the differences in the observations of the sample from the whole population and whole population of individuals studying for exam taken.

Furthermore, the sampling distribution of the sample mean for all three original distributions of an individual’s time spent studying for an exam in hours closely approximates a normal distribution with its corresponding theoretical mean and standard deviation as its parameters. The red density curve in all three graph visualize a normal distribution with some inconsistencies but overall these density curves would be very similar to normal probability curves with respect to each parts (A,B &C) thermotical and estimated mean. In conclusion, these results are expected due to the Central Limit Theorem and each individual time to be independent and identically distributed. The sample mean distribution of a large enough number of samples, the result would approximate a normal distribution with its expected value closely matching the theoretical mean from the original distribution as the mean of the sample means.

**Part 2**

***Overview:***

The average time spent studying for an exam in hours was modeled using bootstrapping with three different probability distributions of the time spend studying for an individual. In this simulation using R and Studio, 1 random samples of size 25 (individuals) were generated from the distributions below. Then 1,000 resamples of size 25 from the original size 25 sample with replacement were taken, this method is known as bootstrapping. The program operating code is shown in *Appendix A.2A* through *Appendix A.2C*.

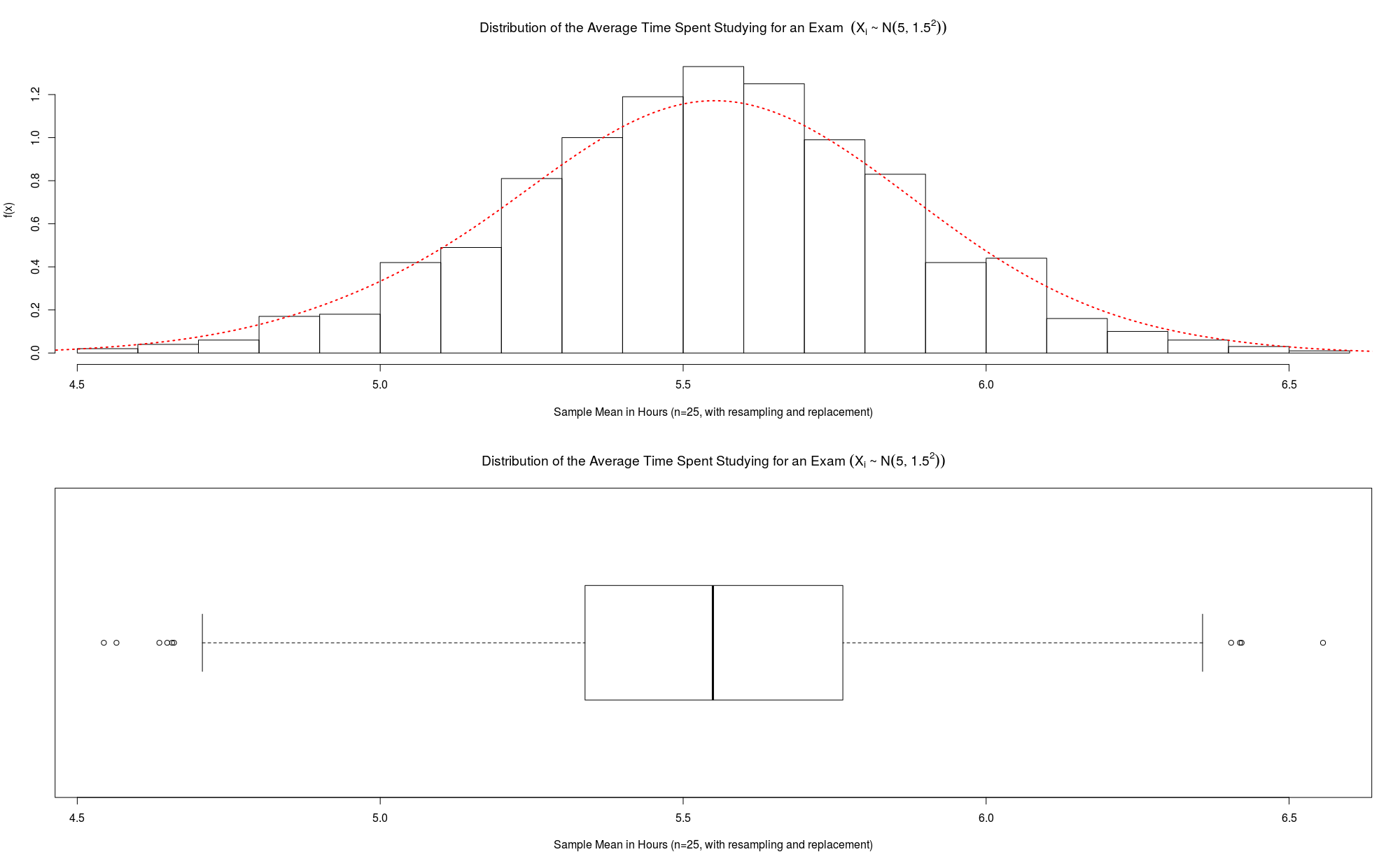
Additionally, the empirical distribution of the 1000 resample’s sample means were plotted, and the estimated mean of the sample mean, and estimated standard deviation of the sample mean were determined for each distribution below, along with the theoretical values of the former & latter.

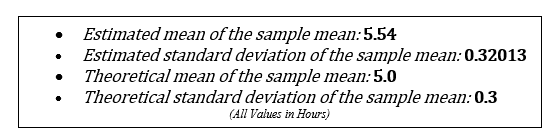
The three distribution of an individual’s time spent studying for an exam in hours are:

1. Normally distributed, with mean 5 hours and standard deviation of 1.5 hours.
2. Uniformly distributed with minimum of 0 hours to maximum of 10 hours.
3. Gamma distribution with parameters:.

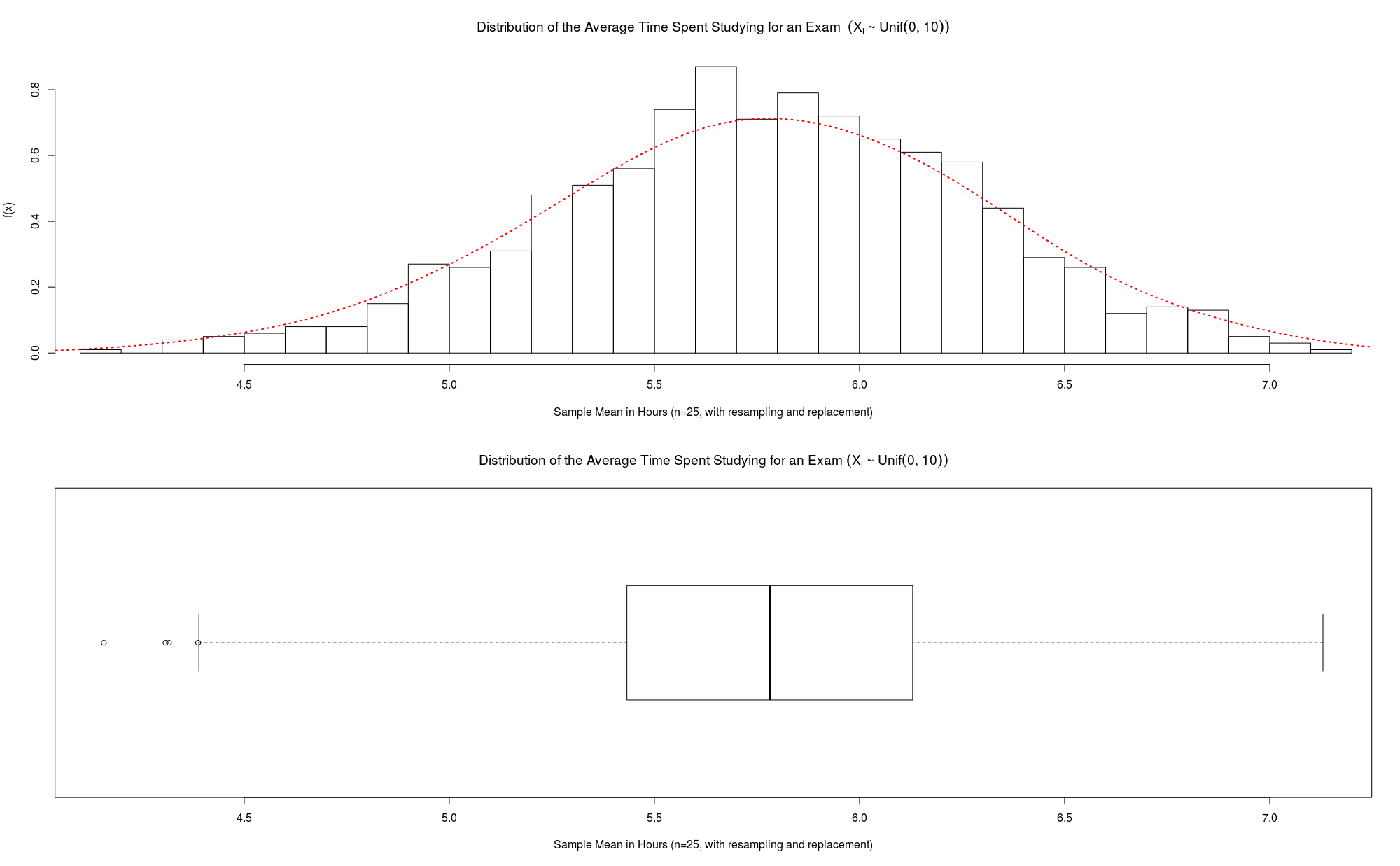
***Results:***

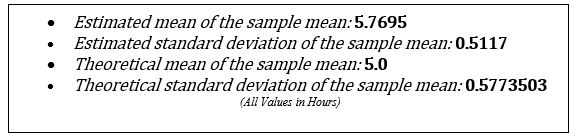
A



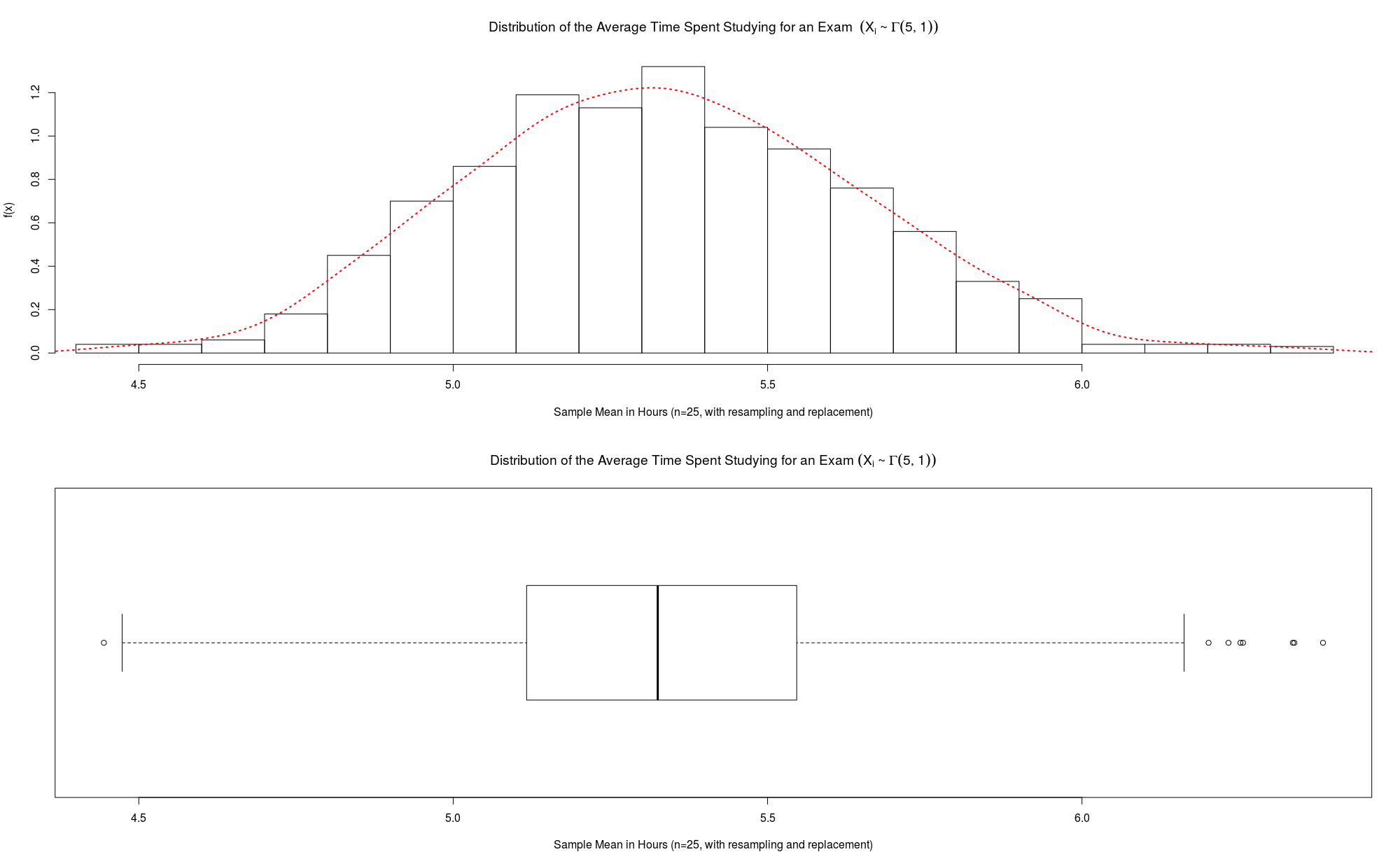


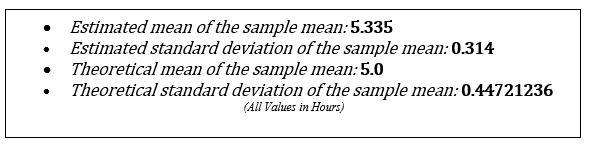
B

******



C

******



***Conclusion:***

Bootstrapping is a method that estimates the sampling distribution by taking multiple (1000) samples with replacement from a single random sample with a certain distribution or data set. These repeated samples are called resamples. Each resample is the same size as the original sample of 25. The original sample represents the population from which it was drawn which in this simulation had the distributions presented in A, B&C. Therefore, the resamples from this original sample represent the distribution of the sample by taking many samples from the population. The bootstrap distribution of a statistic or in this case, the sample mean of the resamples, based on the resamples, represents the sampling distribution of that statistic with sampling error.

The time spend studying for an individual in hours can be assumed to be independent and identically distributed. For A, the density curve of the sample mean or average time spending for an exam for the 1000 resamples is very similar to a normal probability curve. This is expected as the original sample that was used to resample derived from a normal distribution with the theoretical parameters. This would allow the 1000 resamples to have a similar normal distribution with the estimated mean and standard deviation of the sample means to be its parameter. But for A, the estimated mean of the sample means has shifted .54 from its theoretical value, although the standard deviation of the sample mean was approximate to the theoretical standard deviation. Additionally, comparing to Part 1-A, both have an approximate normal distribution for the sampling distribution for the sample mean from the simulation, but in Part 2-A, there is shift in the mean of the sample mean as described earlier, but the standard deviations are very similar. This shift in the sample mean of the resamples mean is most probable due to sampling error caused by the obtaining the original example and the error caused in resampling as the values for the resample have a tighter constraint on what they can be, causing changes in the statistic mentioned.

For B, the density curve of the sample mean or average time spending for an exam for the 1000 resamples is very similar to a normal probability curve with the estimated mean and standard deviation of the resample means. This is expected due the Central Limit Theorem, which allows the distribution of the sample means convey much closer to a normal distribution even though the original sample derived from a uniform distribution. This resemblance increases as the number of resamples increases. With 1,000 resamples, the distribution of the sample mean of the time spend studying for an exam is approximately normal. But the estimated mean of resamples is very different the theoretical sample mean expected by a shift of 0.77, although the standard deviation of the sample mean was almost similar to the theoretical standard deviation. Additionally, comparing to Part 1-B, both have an approximate normal distribution for the sampling distribution for the sample mean from the simulation, but in Part 2-A, there is shift in the mean of the sample mean as described earlier, but the standard deviations are very similar. This shift in the sample mean of the resamples mean is most probable due to sampling error caused by the obtaining the original example and the error caused in resampling as the values for the resample have a tighter constraint on what they can be, causing changes in the statistic mentioned.

Lastly in C, as before the distribution and density curve of the sample mean or average time spending for an exam given the original sample having a gamma distribution and 1000 resamples, is very similar to a normal distribution. This arises again due the Central Limit Theorem, which allows the distribution of the sample means convey much closer to a normal distribution even though the original sample derived from a gamma distribution. This resemblance increases as the number of resamples increases. With 1,000 resamples, the distribution of the sample mean of the time spend studying for an exam is approximately normal. Once again, the estimated mean of resamples is very different the theoretical sample mean expected by a shift of 0.335. Comparing it to Part 1-A, both of the distributions of the sample mean are similar as they have very comparable standard deviations, but the mean of sample mean is shifted as described earlier. This is most probable due to sampling error caused by the obtaining the original example and the error caused in resampling as the values for the resample have a tighter constraint on what they can be, causing changes in the statistic mentioned.

**Part 3**

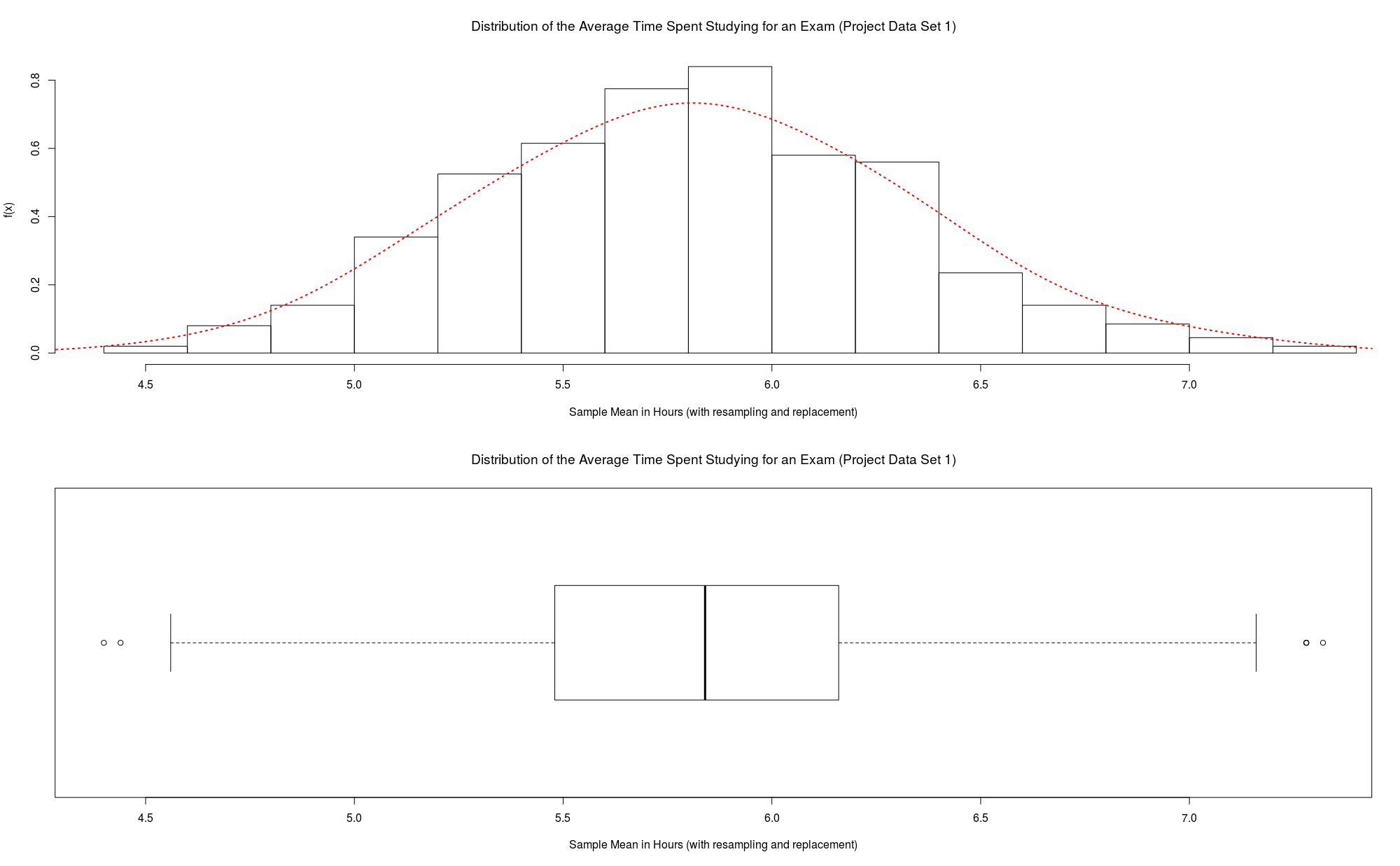
***Overview:***

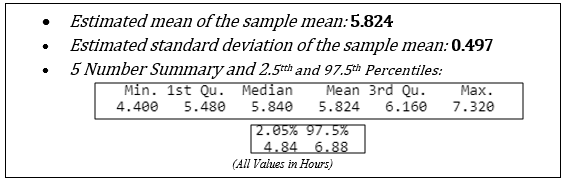
The time spent studying for an exam in hours was observed for 25 individuals and another study was conducted for 250 individuals. Project Data 1.csv and Project Data 2.csv display the collected data. For the first dataset of 25 individuals, 1,000 resamples of size 25 with replacement were taken from the data. For the second dataset of 250 individuals, 1,000 resamples of size 250 with replacement were taken from the data. This method of resampling is known as bootstrapping. The program operating code is shown in *Appendix A.3A* through *Appendix A.3B*.

Additionally, the empirical distribution of the 1000 resample’s sample means for both datasets were plotted, and the estimated mean of the sample mean, and estimated standard deviation of the sample mean were determined for each distribution below, along with the estimated 5 number summary, 2.5th and 97.5th percentiles.

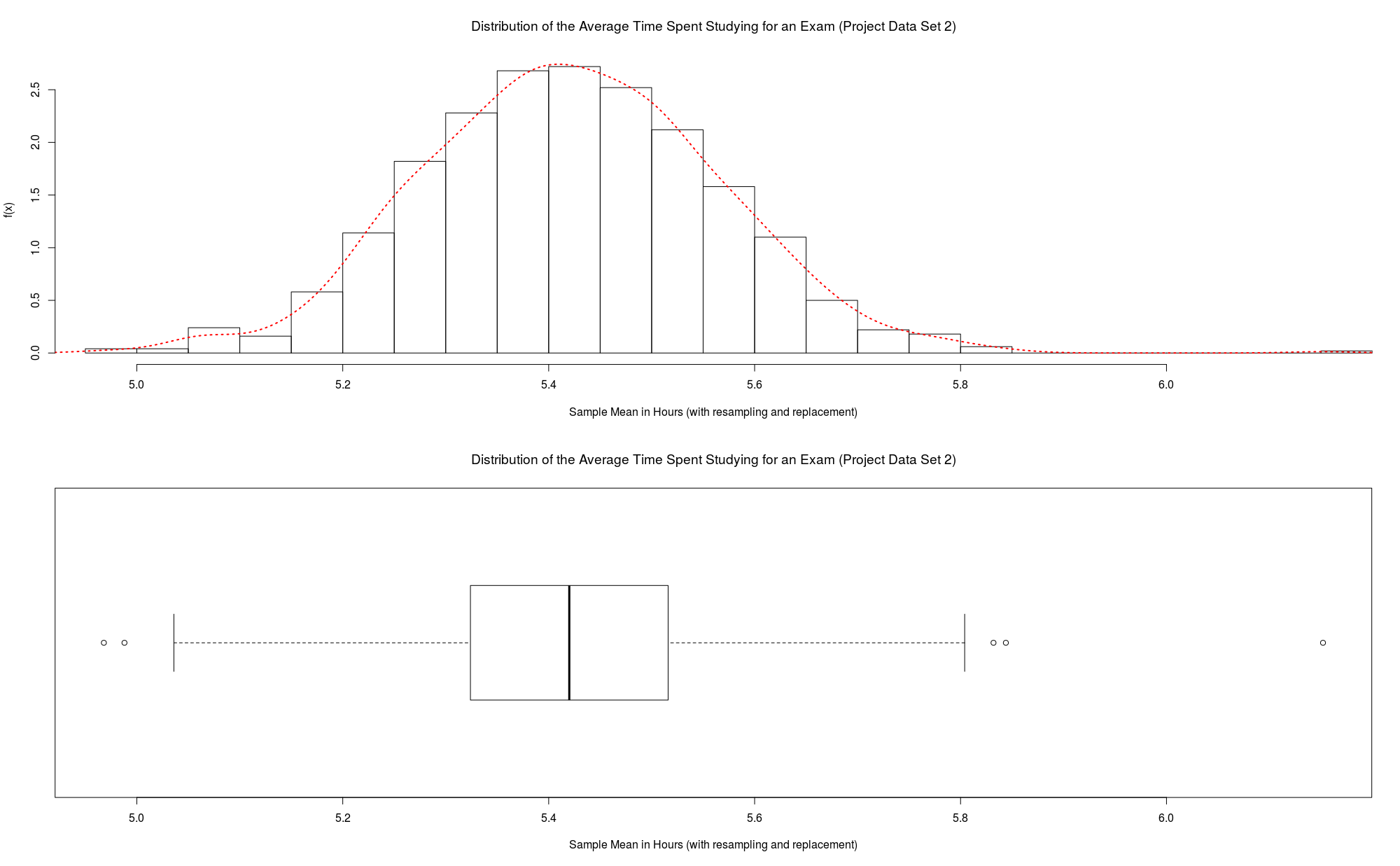
***Results:***

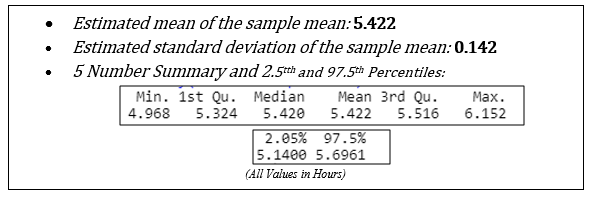
A: Dataset 1 of 25 Individuals





B: Dataset 2 of 250 Individuals





***Conclusion:***

The bootstrap method used in this simulation was a resampling technique used to estimate the sample mean of the time spend studying for an exam on the population by sampling a original observed dataset (size 25 and size 250) with replacement of size 25 and size 250 respectively. The bootstrap resamples are the same size as the original dataset. As a result, some samples will be represented multiple times in the bootstrap sample while others will not be selected at all.

For the original dataset of size 25, the sample mean was evaluated to be 5.84. The distribution of the sample mean, or the average time spend studying for an exam using the 1000 resamples and bootstrapping was determined in A. The estimated mean for the resamples was determined to be 5.824 which very approximate to the original sample’s mean. This distribution also approximates a normal distribution as shown by the histogram & density curve with the estimated standard deviation of the sample mean shown in A. Furthermore, for the original dataset of size 250, the sample mean was evaluated to be 5.248. The distribution of the sample mean using the bootstrapping method was plotted in B. The estimated mean of a sample mean was calculated to be 5.422. This is a shift of 0.174. The distribution is also very approximate to a normal distribution with its estimated mean and standard deviation as it’s parameters.

Even though for both the datasets with different sample sizes, the distribution of the sample means or average time spend studying for an exam using was approximately normal. The standard deviation differed as the dataset with size 250 was much smaller, causing a tighter and less spread distribution compared to the dataset with size 25. In terms of the mean of the sample mean distribution and the original sample mean, the dataset with size 25 was much bigger than the other dataset with size 250. These difference in the mean and standard deviation of the sample mean distribution is most likely caused by the circumstance that when resampling from data with replacement as done in this simulation, then as the sample size increases, the distribution of the two statistics mentioned converge to the population’s (all taking an exam) true distribution and parameters. Therefore, in the size 250 dataset, the standard deviation of the sample mean is much smaller than the size 25 dataset as there is less variability in the sample mean when there is a larger sample size in estimating the population mean as the sampling error gets averaged out. Additionally, the mean of the sample mean of the size 250 dataset varies from the mean of the sample mean of the size 25 dataset as the former dataset because it most likely captures the true population mean or time spend studying for an exam among individuals compared to the latter as there is a difference in sample size.

**Appendices**

***Appendix A: Program Codes:***

A.1A

1. ###################
2. # FINAL PROJECT: PART 1A - STAT 3201(DONGES)
3. # HARSHIL PATEL, 11/23/2018
4. ###################
6. #set working directory
7. setwd("C:/Users/offic/OneDrive - The Ohio State University/data analytics/R programming Files/project")
9. set.seed(35) #this ensures you'll get the same random process every time you run the code
11. #################################
12. # Part 1A:
13. #################################
15. #Model the time spent studying for an exam. Take 1,000 random samples of size 25 from the distributions below.
16. #Plot the empirical distribution of the sample mean, estimate the mean of the sample mean, and estimate the
17. #standard deviation of the sample mean. Compare the results to the theoretical results.
19. #The distribution of one student's time spent studying for an exam is
20. #normally distrubuted, with mean 5 hours and standard deviation of 1.5 hours.
22. mean<-5  #The parameter of individual time spend studying dist.
23. stdeviation<-1.5  #The parameter of individual time spend studying dist.
24. sampleSize<-25
25. numberOfSamples<-1000
26. vectorOfSampleMeans<-vector()
27. samples<-vector()
28. # use a loop to generate and record 1000 random sample means with x[i]~N(5,1.5^2) from samples of size 25.
29. **for** (i **in** 1:numberOfSamples) {
30. samples<- rnorm(sampleSize,mean = mean,sd = stdeviation)
31. vectorOfSampleMeans[i]<-mean(samples)
32. }
33. #Determine the estimated mean & st. deviation of sample mean.
34. avgOfSampleMean<-mean(vectorOfSampleMeans)
35. sdtOfSampleMean<-sd(vectorOfSampleMeans)
37. #Determine the theoretical mean & st. deviation of sample mean.
38. theoreticalMeanOfSampleMean<-mean
39. theoreticalSDTOfSampleMean<-(stdeviation)/(sqrt(sampleSize))
41. #Data Visualization, create a histrogram & boxplot of sample mean generated from previous operation.
42. par(mfrow=c(2,1))
43. hist(vectorOfSampleMeans,
44. main = expression("Distribution of the Average Time Spent Studying for an Exam"~ (X[i]~"~"~N(5,1.5^{2}))),
45. ylab='f(x)',
46. xlab = "Sample Mean in Hours (n=25)",
47. breaks = 25,
48. xlim = c(min(vectorOfSampleMeans),max(vectorOfSampleMeans)),
49. freq = FALSE)
50. boxplot(vectorOfSampleMeans,
51. main = expression("Distribution of the Average Time Spent Studying for an Exam"~ (X[i]~"~"~N(5,1.5^{2}))),
52. xlab = "Sample Mean in Hours (n=25)",
53. horizontal = TRUE)

A.1B

1. ###################
2. # FINAL PROJECT: PART 1B - STAT 3201(DONGES)
3. # HARSHIL PATEL, 11/23/2018
4. ###################
6. #set working directory
7. setwd("C:/Users/offic/OneDrive - The Ohio State University/data analytics/R programming Files/project")
9. set.seed(35) #this ensures you'll get the same random process every time you run the code
11. #################################
12. # Part 1B:
13. #################################
15. #Model the time spent studying for an exam. Take 1,000 random samples of size 25 from the distributions below.
16. #Plot the empirical distribution of the sample mean, estimate the mean of the sample mean, and estimate the
17. #standard deviation of the sample mean. Compare the results to the theoretical results.
19. #The distribution of one student's time spent studying for an exam is
20. #uniformally distrubuted, with time spend can be between 0 to 10 hours.
22. alpha<-0 #min amount student could study
23. beta<-10 #max amount student could study
24. sampleSize<-25
25. numberOfSamples<-1000
26. vectorOfSampleMeans<-vector()
27. samples<-vector()
28. # use a loop to generate and record 1000 random sample means with x[i]~Unif(0,10) from samples of size 25.
29. **for** (i **in** 1:numberOfSamples) {
30. samples<- runif(sampleSize, min = alpha, max = beta)
31. vectorOfSampleMeans[i]<-mean(samples)
32. }
33. #Determine the estimated mean & st. deviation of sample mean.
34. avgOfSampleMean<-mean(vectorOfSampleMeans)
35. sdtOfSampleMean<-sd(vectorOfSampleMeans)
37. #Determine the theoretical mean & st. deviation of sample mean.
38. theoreticalMeanOfSampleMean<-(beta+alpha)/2
39. theoreticalSDTOfSampleMean<-sqrt((((beta-alpha)^2)/12)/(sampleSize))
41. #Data Visualization, create a histrogram & boxplot of sample mean generated from previous operation.
42. par(mfrow=c(2,1))
43. hist(vectorOfSampleMeans,
44. main = expression("Distribution of the Average Time Spent Studying for an Exam"~ (X[i]~"~"~Unif(0,10))),
45. ylab='f(x)',
46. xlab = "Sample Mean in Hours (n=25)",
47. breaks = 25,
48. xlim = c(min(vectorOfSampleMeans),max(vectorOfSampleMeans)),
49. freq = FALSE)
50. boxplot(vectorOfSampleMeans,
51. main = expression("Distribution of the Average Time Spent Studying for an Exam"~ (X[i]~"~"~Unif(0,10))),
52. xlab = "Sample Mean in Hours (n=25)",
53. horizontal = TRUE)

A.1C

1. ###################
2. # FINAL PROJECT: PART 1C - STAT 3201(DONGES)
3. # HARSHIL PATEL, 11/23/2018
4. ###################
6. #set working directory
7. setwd("C:/Users/offic/OneDrive - The Ohio State University/data analytics/R programming Files/project")
9. set.seed(35) #this ensures you'll get the same random process every time you run the code
11. #################################
12. # Part 1C:
13. #################################
15. #Model the time spent studying for an exam. Take 1,000 random samples of size 25 from the distributions below.
16. #Plot the empirical distribution of the sample mean, estimate the mean of the sample mean, and estimate the
17. #standard deviation of the sample mean. Compare the results to the theoretical results.
19. #The distribution of a student's time spent studying for an exam is
20. #distrubuted with a gamma distribution, with parameters: alpha=5, beta=1.
22. alpha<-5
23. beta<-1
24. sampleSize<-25
25. numberOfSamples<-1000
26. vectorOfSampleMeans<-vector()
27. samples<-vector()
28. # use a loop to generate and record 1000 random sample means with x[i]~Gamma(5,1) from samples of size 25.
29. **for** (i **in** 1:numberOfSamples) {
30. samples<- rgamma(sampleSize,alpha,scale = beta)
31. vectorOfSampleMeans[i]<-mean(samples)
32. }
33. #Determine the estimated mean & st. deviation of sample mean.
34. avgOfSampleMean<-mean(vectorOfSampleMeans)
35. sdtOfSampleMean<-sd(vectorOfSampleMeans)

38. #Determine the theoretical mean & st. deviation of sample mean.
39. theoreticalMeanOfSampleMean<-(beta\*alpha)
40. theoreticalSDTOfSampleMean<-sqrt((alpha\*(beta^2))/(sampleSize))
42. #Data Visualization, create a histrogram & boxplot of sample mean generated from previous operation.
43. par(mfrow=c(2,1))
44. hist(vectorOfSampleMeans,
45. main = expression("Distribution of the Average Time Spent Studying for an Exam"~ (X[i]~"~"~Gamma(5,1))),
46. ylab='f(x)',
47. xlab = "Sample Mean in Hours (n=25)",
48. breaks = 25,
49. xlim = c(min(vectorOfSampleMeans),max(vectorOfSampleMeans)),
50. freq = FALSE)
51. boxplot(vectorOfSampleMeans,
52. main = expression("Distribution of the Average Time Spent Studying for an Exam"~ (X[i]~"~"~Gamma(5,1))),
53. xlab = "Sample Mean in Hours (n=25)",
54. horizontal = TRUE)

A.2A

1. ###################
2. # FINAL PROJECT: PART 2A - STAT 3201(DONGES)
3. # HARSHIL PATEL, 11/23/2018
4. ###################
6. #set working directory
7. setwd("C:/Users/offic/OneDrive - The Ohio State University/data analytics/R programming Files/project")
9. set.seed(35) #this ensures you'll get the same random process every time you run the code
11. #################################
12. # Part 2A:
13. #################################
15. #Model the time spent studying for an exam. Take one random sample of size 25 from the distribution below. Then, take 1,000 resamples (i.e.,
16. #sample with replacement) of size 25 from the sample. Plot the empirical distribution of the sample mean, estimate the mean of the sample mean,
17. #and estimate the standard deviation of the sample mean.
19. #the distribution of a student's time spent studying for an exam is
20. #distrubuted with a normal distribution, with parameters: mean=5, SD=1.5
22. mean<-5
23. stdeviation<-1.5
24. sampleSize<-25
25. originalSample<-rnorm(sampleSize,mean = mean,sd = stdeviation) #generate one random sample of size 25 with x[i]~N(5,1.5^2)
26. numberOfResamples<-1000
27. vectorOfSampleMeans<-vector()
28. samples<-vector()
29. # use a loop to generate and record 1000 random sample means from original sample of size 25.
30. **for** (i **in** 1:numberOfResamples) {
31. samples<- sample(originalSample,25, replace = TRUE)
32. vectorOfSampleMeans[i]<-mean(samples)}
34. #Determine the estimated mean & st. deviation of sample mean.
35. avgOfSampleMean<-mean(vectorOfSampleMeans)
36. sdtOfSampleMean<-sd(vectorOfSampleMeans)
38. #Determine the theoretical mean & st. deviation of sample mean.
39. theoreticalMeanOfSampleMean<-mean
40. theoreticalSDTOfSampleMean<-(stdeviation)/(sqrt(sampleSize))
42. #Data Visualization, create a histrogram & boxplot of sample mean generated from previous operation.
43. par(mfrow=c(2,1))
44. hist(vectorOfSampleMeans,
45. main = expression("Distribution of the Average Time Spent Studying for an Exam "~ (X[i]~"~"~N(5,1.5^{2}))),
46. ylab='f(x)',
47. xlab = "Sample Mean in Hours (n=25, with resampling and replacement)",
48. breaks = 25,
49. xlim = c(min(vectorOfSampleMeans),max(vectorOfSampleMeans)),
50. freq = FALSE)
51. boxplot(vectorOfSampleMeans,
52. main = expression("Distribution of the Average Time Spent Studying for an Exam"~ (X[i]~"~"~N(5,1.5^{2}))),
53. xlab = "Sample Mean in Hours (n=25, with resampling and replacement)",
54. horizontal = TRUE)

A.2B

1. ###################
2. # FINAL PROJECT: PART 2B - STAT 3201(DONGES)
3. # HARSHIL PATEL, 11/23/2018
4. ###################
6. #set working directory
7. setwd("C:/Users/offic/OneDrive - The Ohio State University/data analytics/R programming Files/project")
9. set.seed(35) #this ensures you'll get the same random process every time you run the code
11. #################################
12. # Part 2B:
13. #################################
15. #Model the time spent studying for an exam. Take one random sample of size 25 from the distribution below. Then, take 1,000 resamples (i.e.,
16. #sample with replacement) of size 25 from the sample. Plot the empirical distribution of the sample mean, estimate the mean of the sample mean,
17. #and estimate the standard deviation of the sample mean.
19. #the distribution of a student's time spent studying for an exam is
20. #distrubuted with a uniform distribution, with parameters: min=0, max=10
22. alpha<-0 #min amount student could study
23. beta<-10 #max amount student could study
24. sampleSize<-25
25. originalSample<-runif(sampleSize, min = alpha, max = beta) #generate one random sample of size 25 with x[i]~Unif(0,10)
26. numberOfResamples<-1000
27. vectorOfSampleMeans<-vector()
28. samples<-vector()
30. # use a loop to generate and record 1000 random sample means with from orginal sample of size 25.
31. **for** (i **in** 1:numberOfResamples) {
32. samples<- sample(originalSample,25, replace = TRUE)
33. vectorOfSampleMeans[i]<-mean(samples)}
35. #Determine the estimated mean & st. deviation of sample mean.
36. avgOfSampleMean<-mean(vectorOfSampleMeans)
37. sdtOfSampleMean<-sd(vectorOfSampleMeans)
39. #Determine the theoretical mean & st. deviation of sample mean.
40. theoreticalMeanOfSampleMean<-(beta+alpha)/2
41. theoreticalSDTOfSampleMean<-sqrt((((beta-alpha)^2)/12)/(sampleSize))
43. #Data Visualization, create a histrogram & boxplot of sample mean generated from previous operation.
44. par(mfrow=c(2,1))
45. hist(vectorOfSampleMeans,
46. main = expression("Distribution of the Average Time Spent Studying for an Exam "~ (X[i]~"~"~Unif(0,10))),
47. ylab='f(x)',
48. xlab = "Sample Mean in Hours (n=25, with resampling and replacement)",
49. breaks = 25,
50. xlim = c(min(vectorOfSampleMeans),max(vectorOfSampleMeans)),
51. freq = FALSE)
52. boxplot(vectorOfSampleMeans,
53. main = expression("Distribution of the Average Time Spent Studying for an Exam"~ (X[i]~"~"~Unif(0,10))),
54. xlab = "Sample Mean in Hours (n=25, with resampling and replacement)",
55. horizontal = TRUE)

A.2C

1. ######################################
2. # FINAL PROJECT: PART 2C - STAT 3201(DONGES)
3. # HARSHIL PATEL, 11/23/2018
4. ######################################
6. #set working directory
7. setwd("C:/Users/offic/OneDrive - The Ohio State University/data analytics/R programming Files/project")
9. set.seed(35) #this ensures you'll get the same random process every time you run the code
11. #################################
12. # Part 2C:
13. #################################
15. #Model the time spent studying for an exam. Take one random sample of size 25 from the distribution below. Then, take 1,000 resamples (i.e.,
16. #sample with replacement) of size 25 from the sample. Plot the empirical distribution of the sample mean, estimate the mean of the sample mean,
17. #and estimate the standard deviation of the sample mean.
19. #the distribution of a student's time spent studying for an exam is
20. #distrubuted with a gamma distribution, with parameters: alpha=5, beta=1
22. alpha<-5
23. beta<-1
24. sampleSize<-25
25. originalSample<-rgamma(sampleSize,alpha,scale = beta) #generate one random sample of size 25 with x[i]~Gamma(5,1)
26. numberOfResamples<-1000
27. vectorOfSampleMeans<-vector()
28. samples<-vector()
30. # use a loop to generate and record 1000 random sample means with from orginal sample of size 25.
31. **for** (i **in** 1:numberOfResamples) {
32. samples<-sample(originalSample,25, replace = TRUE)
33. vectorOfSampleMeans[i]<-mean(samples)}
35. #Determine the estimated mean & st. deviation of sample mean.
36. avgOfSampleMean<-mean(vectorOfSampleMeans)
37. sdtOfSampleMean<-sd(vectorOfSampleMeans)
39. #Determine the theoretical mean & st. deviation of sample mean.
40. theoreticalMeanOfSampleMean<-(beta\*alpha)
41. theoreticalSDTOfSampleMean<-sqrt((alpha\*(beta^2))/(sampleSize))
43. #Data Visualization, create a histrogram & boxplot of sample mean generated from previous operation.
44. par(mfrow=c(2,1))
45. hist(vectorOfSampleMeans,
46. main = expression("Distribution of the Average Time Spent Studying for an Exam "~ (X[i]~"~"~Gamma(5,1))),
47. ylab='f(x)',
48. xlab = "Sample Mean in Hours (n=25, with resampling and replacement)",
49. breaks = 25,
50. xlim = c(min(vectorOfSampleMeans),max(vectorOfSampleMeans)),
51. freq = FALSE)
52. boxplot(vectorOfSampleMeans,
53. main = expression("Distribution of the Average Time Spent Studying for an Exam"~ (X[i]~"~"~Gamma(5,1))),
54. xlab = "Sample Mean in Hours (n=25, with resampling and replacement)",
55. horizontal = TRUE)

A.3A

1. #########################################################
2. # FINAL PROJECT: PART 3.Data Set 1 - STAT 3201(DONGES)
3. # HARSHIL PATEL, 11/24/2018
4. #########################################################
6. #set working directory
7. setwd("C:/Users/offic/OneDrive - The Ohio State University/data analytics/R programming Files/project/part3")
8. set.seed(35)
10. #################################
11. # Part 3 Data Set 1:
12. #################################
14. #Project Data Set 1.csv gives 25 observations of the time spent studying for an exam of students. Use this data to take 1,000 resamples (i.e.,
15. #sample with replacement) of size 25 from the observed data. Plot the empirical distribution of the sample mean, estimate the mean of the sample mean,
16. #and estimate the standard deviation of the sample mean. Also ind the estimated 5 number summary and the estimated 2.5th and 97.5th
17. #percentiles. Compare and contrast your results for the two sample sizes.
19. #EXTRACT DATA FROM DATA SET FILE(.CSV)
20. dataset1<-read.csv("Project Data 1.csv", header = TRUE)
21. studyTimeValues<- dataset1$x #vector of 25 observations of a student's time spent studying for exam.
23. numberOfResamples<-1000
24. vectorOfSampleMeans<-vector()
25. samples<-vector()
27. # use a loop to generate and record 1000 random sample means from observed data using resampling technique from part two .
28. **for** (i **in** 1:numberOfResamples) {
29. samples<- sample(studyTimeValues,25, replace = TRUE)
30. vectorOfSampleMeans[i]<-mean(samples)}
32. #Determine the estimated mean & st. deviation of sample mean, and estimated 5 number summary.
33. avgOfSampleMean<-mean(vectorOfSampleMeans)
34. sdtOfSampleMean<-sd(vectorOfSampleMeans)
35. summary(vectorOfSampleMeans)
37. #Determine estimated estimated 2.5th and 97.5th percentiles
38. quantile(vectorOfSampleMeans, c(.0205, .975))
40. #Data Visualization, create a histrogram & boxplot of sample mean generated from previous operation.
41. par(mfrow=c(2,1))
42. hist(vectorOfSampleMeans,
43. main = expression("Distribution of the Average Time Spent Studying for an Exam (Project Data Set 1)"),
44. ylab='f(x)',
45. xlab = "Sample Mean in Hours (with resampling and replacement)",
46. breaks = 20,
47. xlim = c(min(vectorOfSampleMeans),max(vectorOfSampleMeans)),
48. freq = FALSE)
49. boxplot(vectorOfSampleMeans,
50. main = expression("Distribution of the Average Time Spent Studying for an Exam (Project Data Set 1)"),
51. xlab = "Sample Mean in Hours (with resampling and replacement)",
52. horizontal = TRUE)

A.3B

1. #########################################################
2. # FINAL PROJECT: PART 3.Data Set 2 - STAT 3201(DONGES)
3. # HARSHIL PATEL, 11/24/2018
4. #########################################################
6. #set working directory
7. setwd("C:/Users/offic/OneDrive - The Ohio State University/data analytics/R programming Files/project/part3")
8. set.seed(35)
10. #################################
11. # Part 3 Data Set 2:
12. #################################
14. #Project Data Set 2.csv gives 250 observations of the time spent studying for an exam of students. Use this data to take 1,000 resamples (i.e.,
15. #sample with replacement) of size 250 from the observed data. Plot the empirical distribution of the sample mean, estimate the mean of the sample mean,
16. #and estimate the standard deviation of the sample mean. Also ind the estimated 5 number summary and the estimated 2.5th and 97.5th
17. #percentiles. Compare and contrast your results for the two sample sizes from Project Data Set 1.csv.
19. #EXTRACT DATA FROM DATA SET FILE(.CSV)
20. dataset1<-read.csv("Project Data 2.csv", header = TRUE)
21. studyTimeValues<- dataset1$x #vector of 250 observations of a student's time spent studying for exam.

24. numberOfResamples<-1000
25. vectorOfSampleMeans<-vector()
26. samples<-vector()
28. # use a loop to generate and record 1000 random sample means from observed data using resampling technique from part two .
29. **for** (i **in** 1:numberOfResamples) {
30. samples<- sample(studyTimeValues,250, replace = TRUE)
31. vectorOfSampleMeans[i]<-mean(samples)}
33. #Determine the estimated mean & st. deviation of sample mean, and estimated 5 number summary.
34. avgOfSampleMean<-mean(vectorOfSampleMeans)
35. sdtOfSampleMean<-sd(vectorOfSampleMeans)
36. summary(vectorOfSampleMeans)
38. #Determine estimated estimated 2.5th and 97.5th percentiles
39. quantile(vectorOfSampleMeans, c(.0205, .975))
41. #Data Visualization, create a histrogram & boxplot of sample mean generated from previous operation.
42. par(mfrow=c(2,1))
43. hist(vectorOfSampleMeans,
44. main = expression("Distribution of the Average Time Spent Studying for an Exam (Project Data Set 2)"),
45. ylab='f(x)',
46. xlab = "Sample Mean in Hours (with resampling and replacement)",
47. breaks = 25,
48. xlim = c(min(vectorOfSampleMeans),max(vectorOfSampleMeans)),
49. freq = FALSE)
50. boxplot(vectorOfSampleMeans,
51. main = expression("Distribution of the Average Time Spent Studying for an Exam (Project Data Set 2)"),
52. xlab = "Sample Mean in Hours (with resampling and replacement)",
53. horizontal = TRUE)

**Author Contact Information:**

**Harshil Patel**

Student at Ohio State University

[patel.3001@osu.edu](mailto:patel.3001@osu.edu)

513-349-6426