Statistical Inference: Peer Assessment 1, Question 1

Introduction

This is the project for the statistical inference class. In it, I will use simulation to explore inference and do some simple inferential data analysis. The project consists of two parts:

1. Simulation exercises.
2. Basic inferential data analysis.

I will create a report to answer each of the questions. I use knitr to create the reports and convert to a pdf. Each pdf report will be no more than 2 pages with 3 pages of supporting appendix material if needed (code, figures, etcetera).

Report Number 1: Simulation exercises

The exponential distribution can be simulated in R with `rexp(n, lambda)` where `lambda` is the rate parameter. The mean of exponential distribution is `1/lambda` and the standard deviation is also also `1/lambda`.

I'll set `lambda = 0.2` for all of the simulations. In this simulation, I will investigate the distribution of averages of 40 `exponential(0.2)`s. Note that I will need to do a thousand or so simulated averages of 40 exponentials, and will generate those as follows:

```
set.seed(18532)
sample.size <- 40
num.trials <- 1000
lambda <- 0.2
data <- matrix(data=NA,nrow=0,ncol=sample.size)
for (i in 1:num.trials) {
    data <- rbind(data, rexp(sample.size, lambda))
}
```

Now, I will illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 `exponential(0.2)`s.

1. Show where the distribution is centered at and compare it to the theoretical center of the distribution.

As stated in the introductory paragraph to this section, the theoretical mean of the distribution is `1/lambda`, which, for a `lambda` of **0.2**, equates to **5**.

First, we'll calculate the mean and median of these sample means to see how closely it lines up with the theoretical value:

```
sample.means <- c(1:num.trials)
for (j in 1:num.trials) {
    sample.means[j] <- mean(data[j,])
}
mean.of.sample.means <- mean(sample.means)
median.of.sample.means <- median(sample.means)
```

So, as we can see, compared to the theoretical mean of **5**, the sample mean of **5.0095247** and the median **4.9604287** line up quite nicely.

2. Show how variable it is and compare it to the theoretical variance of the distribution.

As stated in the introductory paragraph to this section, the theoretical standard deviation of the distribution is also `1/lambda`, which, for a `lambda` of **0.2**, equates to **5**. The variance is the square of the standard deviation, which is **25**.

To get the observed variance, we'll calculate the mean and median of these sample means to see how closely it lines up with the theoretical value:

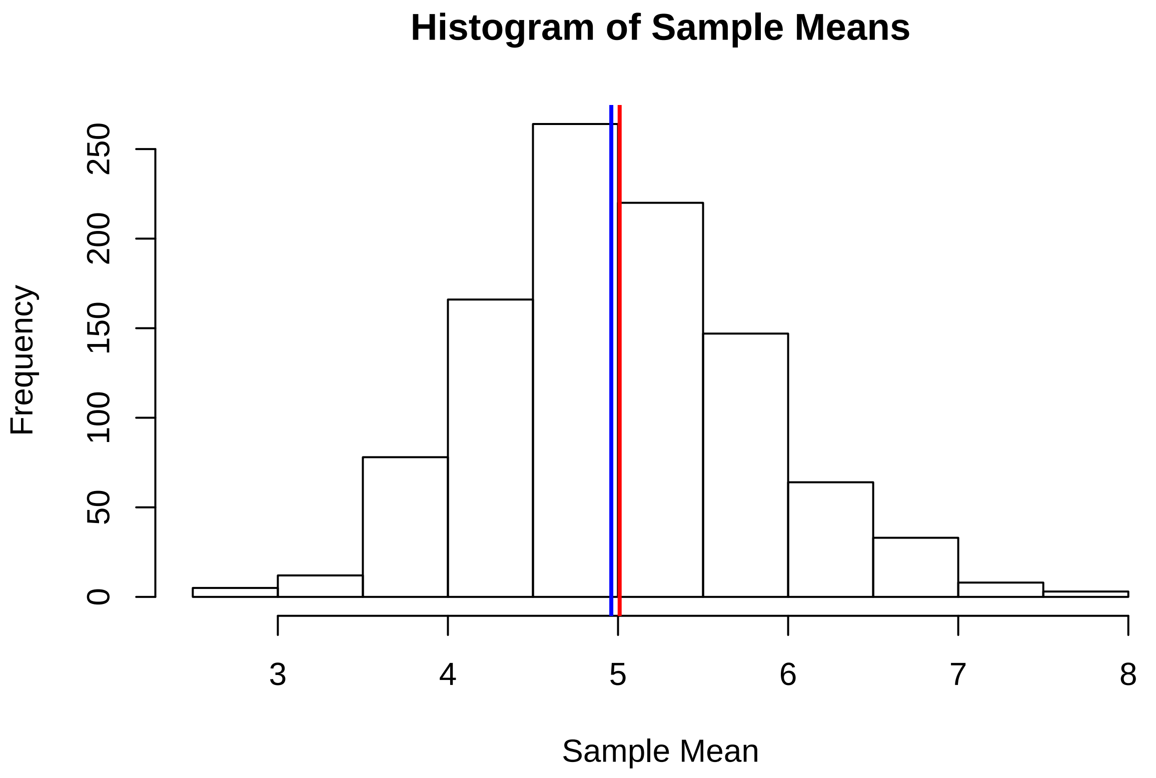
```r
all.data <- as.vector(data)
variance <- var(all.data)
unbiased.variance <- variance * num.trials/(num.trials-1)
```

So, as we can see, compared to the theoretical variance of **25**, the sample variance of **25.0412814** lines up quite nicely. As expected, the unbiased variance **4.9604287** lines up even better.

3. Show that the distribution is approximately normal.

To get a bit more insight into the distribution of observed sample means, we can also plot a histogram of sample means along with the mean value (in red) and median value (in blue) to have some idea of how the data lays out:

```r
hist(sample.means, xlab="Sample Mean", main="Histogram of Sample Means")
abline(v = mean.of.sample.means, col = "red", lwd = 2)
abline(v = median.of.sample.means, col = "blue", lwd = 2)
```



4. Evaluate the coverage of the confidence interval for 1/lambda: X¯±1.96Sn.

```r
levels <- mean.of.sample.means + c(-1,1) * 1.96 * sqrt(var(sample.means))
number.of.values.below <- length(sample.means[sample.means<levels[1]])
number.of.values.above <- length(sample.means[sample.means>levels[2]])
coverage <- 1 - (number.of.values.below + number.of.values.above) / length(sample.means)
```

The coverage evaluated to **94.8**%, which is what we'd expect based on the theoretical value of 95% from the central limit theorem.