Statistical Inference Project Part 1

William L.

Saturday, October 25, 2014

This analysis involves an exploration of the Exponential Distribution. A random sampling of 40 observations will be taken from the Exponential Distribution with a rate parameter (lambda) of 0.2 and then the mean of these observations will be calculated. This process will be repeated 1,000 times in order to obtain a better understanding of the mean of 40 exponentials with lambda 0.2:

Load libraries:

```
library("ggplot2")
```

Set parameters for rexp function:

```
n <- 40
lambda <- 0.2
```

Set the theoretical mean and standard deviation for the Exponential Distribution with lambda 0.2:

```
mean <- 1/lambda
sd <- 1/lambda
```

The theoretical mean is: **5**

The theoretical standard deviation is: **5**

Take a sample of 40 observation from the Exponential Distribution and store the mean, variance, and standard deviation of the sample for future analysis. Repeat the process across 1,000 simulations.

```
distavg <- vector(mode="numeric", length=0)
distvar <- vector(mode="numeric", length=0)
distsd <- vector(mode="numeric", length=0)
for (i in 1:1000) {
    sample <- rexp(n,lambda)
    distavg[[i]] <- mean(sample)
    distvar[[i]] <- var(sample)
    distsd[[i]] <- sd(sample)
}
```

Convert the sample vectors into a data frame for additional analysis and plotting:

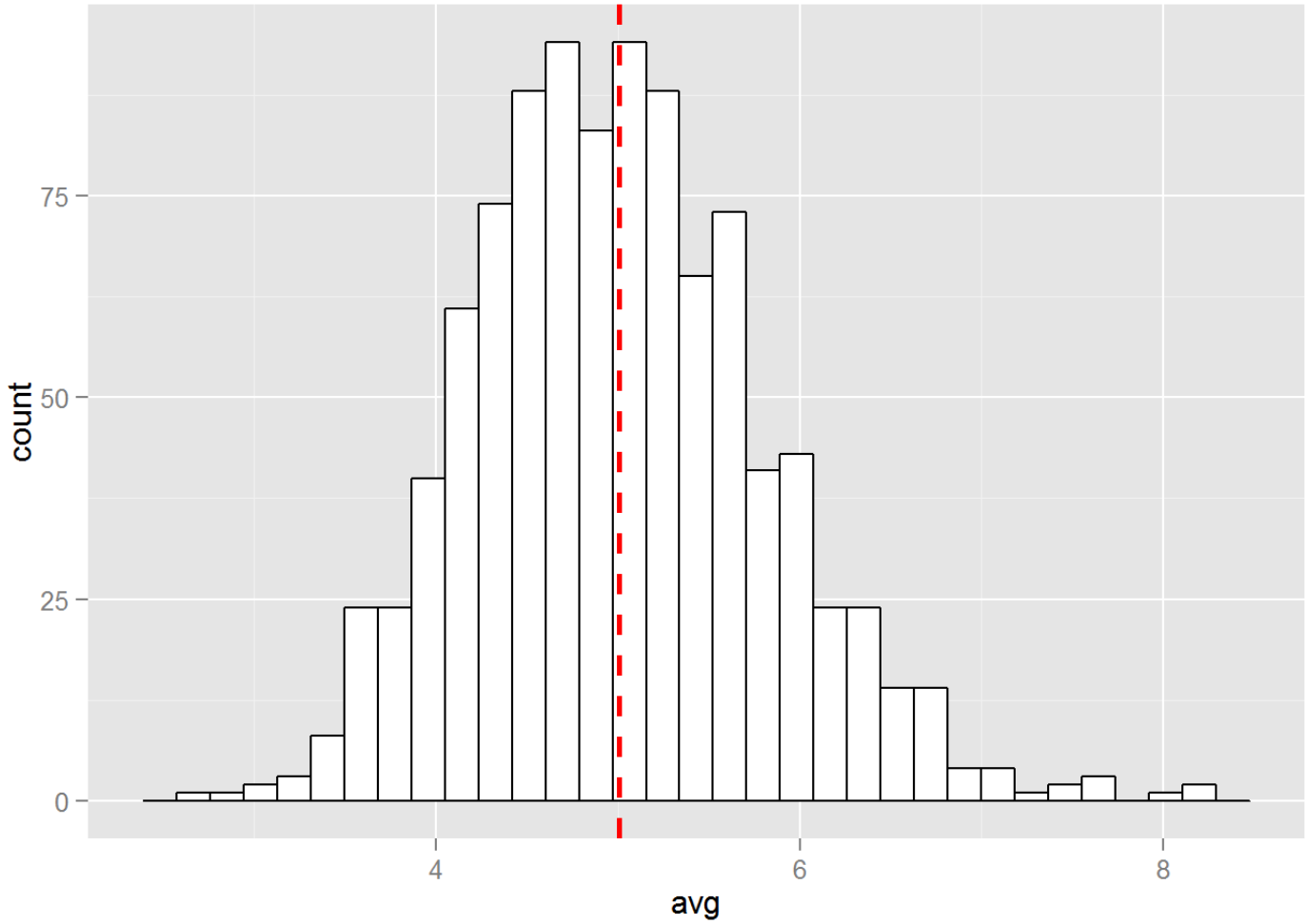
```
df <- data.frame(avg=distavg, var=distvar, sd=distsd)
```

Question 1

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

The below plot shows a histogram of the averages from the Exponential Distribution with a dotted red line indicating the mean or center of the distribution. The mean is located at **5.0095** and this is very close to the theoretical mean of **5**. This is expected as it should be centered around the population mean after 1,000 simulations.

```
ggplot(df, aes(x=avg)) +
    geom_histogram(binwidth=diff(range(df$avg))/30, colour="black", fill="white") +
    geom_vline(aes(xintercept=mean(avg)), color="red", linetype="dashed", size=1)
```



Question 2

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

The theoretical variance of the distribution would be equal to the populations standard deviation squared divided by the number of observations:

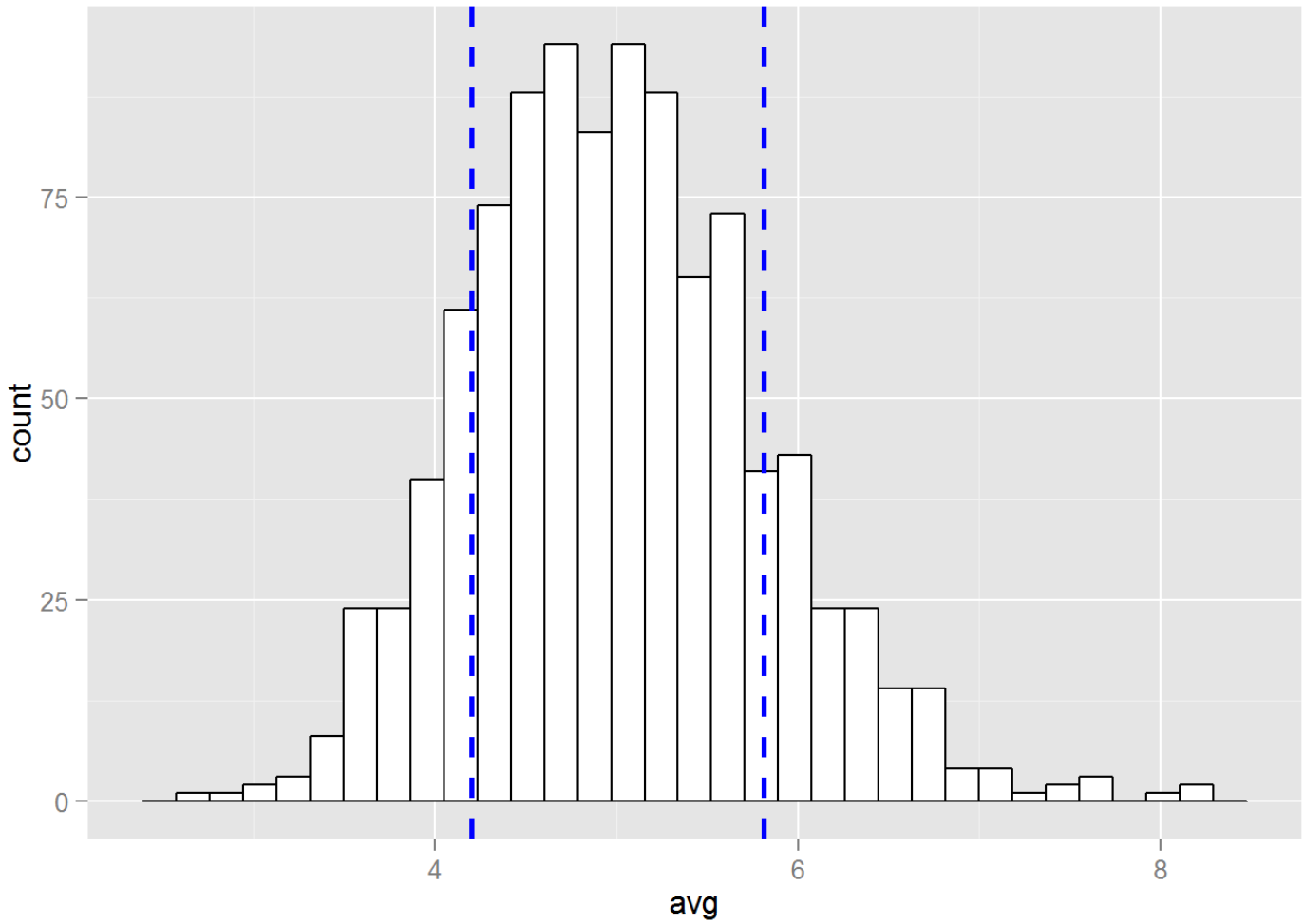
```
tvar <- sd^2/n
```

The simulations variance is equal to:

```
svar <- var(distavg)
```

The theoretical variance is **0.625** and the simulations variance is **0.6469** which is an absolute difference of **0.0219**. The simulation variance is close to the expected variance and the below plot with standard deviations indicated helps to demonstrate the variability of the samples:

```
ggplot(df, aes(x=avg)) +
    geom_histogram(binwidth=diff(range(df$avg))/30, colour="black", fill="white") +
    geom_vline(aes(xintercept=mean(avg)+sd(avg)), color="blue", linetype="dashed", size=1) +
    geom_vline(aes(xintercept=mean(avg)-sd(avg)), color="blue", linetype="dashed", size=1)
```

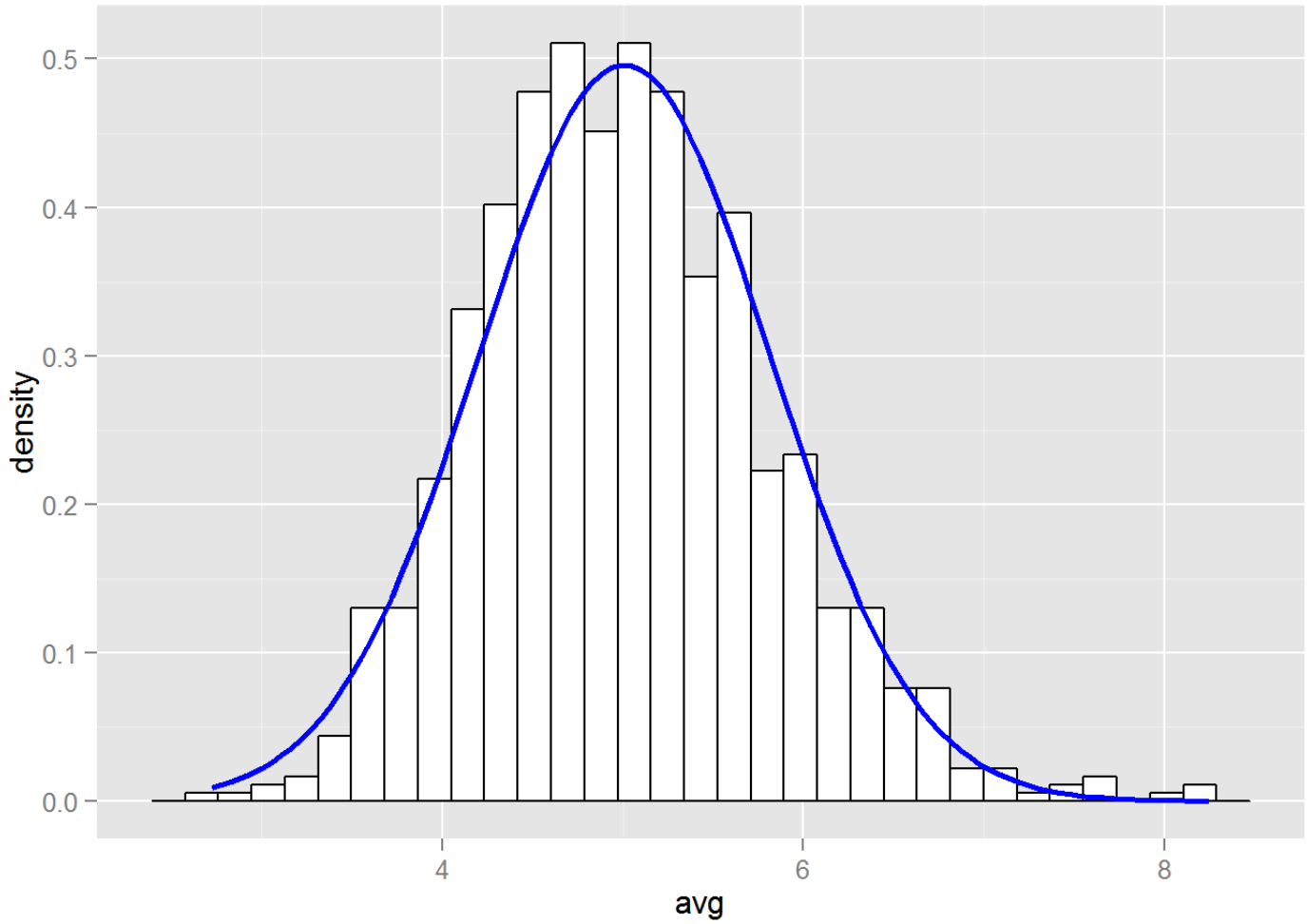


Question 3

Show that the distribution is approximately normal

The simulated distribution can be seen to be approximately normal by viewing a density plot of the distribution with a normal curve highlighted:

```
ggplot(df, aes(x=avg)) +
geom_histogram(aes(y = ..density..),binwidth=diff(range(df$avg))/30, color="black", fill="white") +
stat_function(fun = dnorm, color="blue", size=1, args = list(mean=mean(df$avg), sd=sd(df$avg)))
```



Question 4

Evaluate the coverage of the confidence interval for 1/lambda

The lower and upper bounds of the required confidence interval is calculated using the sample mean plus/minus 1.96 times the standard error (standard deviation over the square root of the number of observations):

```
low <- mean(df$avg) - (1.96 * (sd(df$avg)/sqrt(n)))
up <- mean(df$avg) + (1.96 * (sd(df$avg)/sqrt(n)))
```

The lower bound of the confidence interval is **4.7603** and the upper bound is **5.2588**. The coverage of this confidence interval can be calculated using these limits by determining how many samples fall within this range:

```
coverage <- nrow(df[df$avg <= up & df$avg >= low,])/nrow(df)*100
```

The calculated confidence interval includes **4.7603%** of the simulated values. A visual representation of this confidence interval is below:

```
ggplot(df, aes(x=avg)) +
    geom_histogram(binwidth=diff(range(df$avg))/30, colour="black", fill="white") +
    geom_vline(aes(xintercept=low), color="blue", linetype="dashed", size=1) +
    geom_vline(aes(xintercept=up), color="blue", linetype="dashed", size=1)
```