Peer-graded Assignment: Statistical Inference Course Project Part 1

Kan Chuen LAM

January 1, 2018

Overview

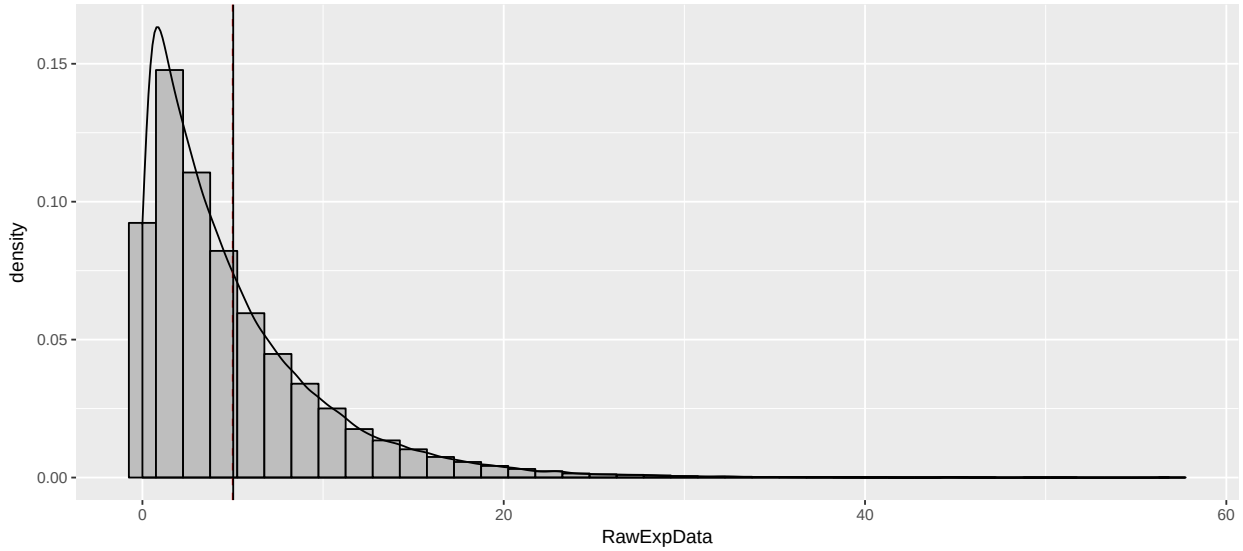
This is a data analysis to investigate the exponential distribution and compare it to the Central Limit Theorem. For this analysis, the lambda is set as 0.2 for all of the simulations. This investigation will compare the distribution of the means of 40 exponentials random over 1000 simulations.

Initialize variables and calculate the sample means & sample variances.

```
set.seed(4)
nsim <- 1000
nsamp <- 40
lambda <- 0.2
rexpdat <- rexp(nsim * nsamp, rate = lambda)
rexpdatDF <- data.frame(RawExpData = rexpdat)
mat <- matrix(rexpdat, nrow = nsim)
samp.mean <- data.frame(SampMeans = rowMeans(mat))
samp.var <- data.frame(SampVar = apply(mat, 1, var))
```

First let's look at the distribution of the original "rexp()" generated data, 40000 data.

```
ggplot(data=rexpdatDF,aes(x=RawExpData)) +
        geom_histogram(aes(y =..density..), fill="grey",binwidth=1.5,colour="black") +
        geom_density(colour="black") +
        geom_vline(xintercept=1/lambda,col="red",linetype="dashed",show.legend=T) +
        geom_vline(xintercept=mean(rexpdatDF$RawExpData),col="black",show.legend=T)
```



It is far the shape of normal distribution. The given theoretical mean is $\mu = \frac{1}{\lambda}$

```
1/lambda
```

```
## [1] 5
```

Whereas the mean of the 40000 data is

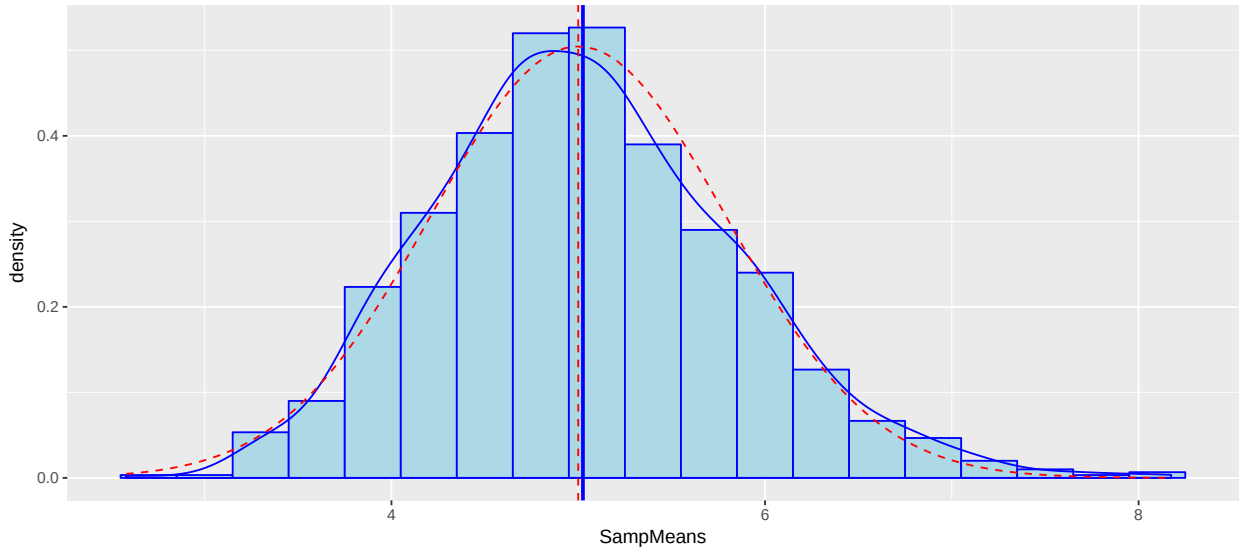
```
mean(rexpdatDF$RawExpData)
```

```
## [1] 5.025289
```

Sample Mean versus Theoretical Mean

Now plot the distribution of the sample mean of the 1000 samples(each with 40 data)

```
ggplot(data=samp.mean,aes(x=SampMeans)) +
        geom_histogram(aes(y=..density..),fill="lightblue",binwidth=0.3,colour="blue") +
        geom_density(colour="blue") +
        geom_vline(xintercept=mean(samp.mean$SampMeans),col="blue",size=1) +
        stat_function(fun=dnorm,args=list(mean=1/lambda,sd=1/lambda/sqrt(nsamp)),
                      colour="red",linetype="dashed") +
        geom_vline(xintercept=1/lambda,col="red",linetype="dashed",show.legend=T)
```



The distribution centers at the mean (blue vertical line) :

```
mean(samp.mean$SampMeans)
```

```
## [1] 5.025289
```

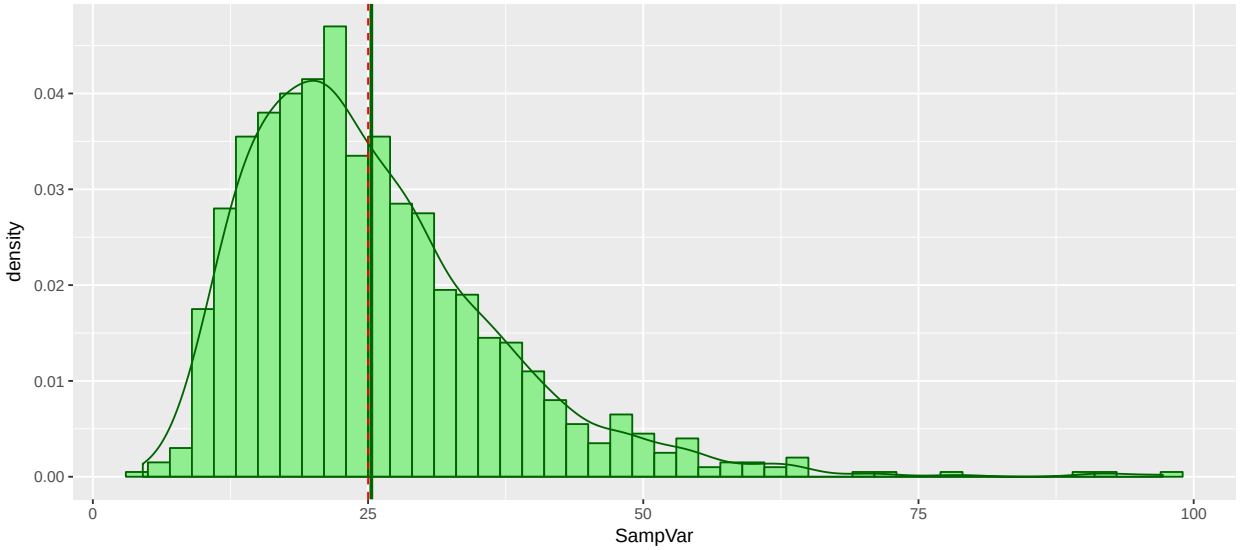
Compare to the theoretical normal distribution (red dashed line) with $\mu = \frac{1}{\lambda} = 5$, and variance $\sigma^2/n = 1/(\lambda^2 n) = 1/(0.04 \times 40) = 0.625$

Sample Variance versus Theoretical Variance

Now let's plot the distribution of the sample variance of the 1000 samples(each with 40 data)

```
ggplot(data=samp.var,aes(x=SampVar)) +
        geom_histogram(aes(y=..density..),fill="lightgreen",binwidth=2,colour="darkgreen") +
        geom_density(colour="darkgreen") +
```

```
        geom_vline(xintercept=mean(samp.var$SampVar),col="darkgreen",size=1) +
        geom_vline(xintercept=1/lambda^2,col="red",linetype="dashed")
```



Note the mean of the unbiased sample variance distribution equals 25.2965456, which is very close to the theoretical variance(red dashed line), $\sigma^2 = \frac{1}{\lambda^2} = 25$