Investigation of Distributions

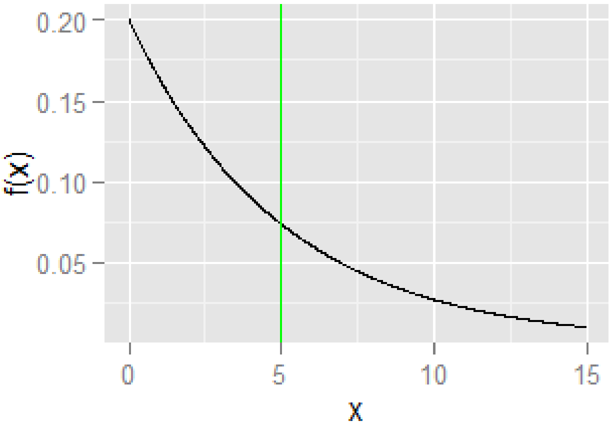
Author: Jeffrey Thatcher

Overview

The goal of this report is to use the central limit theorem to show that an estimation of the mean of a probability density function, specifically the exponential density function (EDF) in this case, is a good approximation of the calculated theoretical mean of the density function.

In the following experiment, we will estimate the mean of the EDF by taking 1000 random samples of the EDF in 40 different experiments. The average of these 40 different experiments will be calculated and it is expected that this value is very close to the theoretical mean of the EDF.

lambda is 0.2 for all the following experiments and the exponential density function has the following appearance:



Simulations

The sample mean will be simulated by taking 1000 random samples of the exponential distribution in 40 separate experiments. From each experiment we will calculate the sample mean (represents the population mean) and then average these 40 sample means to estimate the theoretical mean.

To begin we generate a matrix with 40 columns that each contain 1000 random samples of the EDF.

```
DF <- as.data.frame(matrix(1,1000,40))
DFexp <- sapply(DF,rexp, rate=0.2)
```

Next we calculate the mean and variance for each column and store these for the next sections.

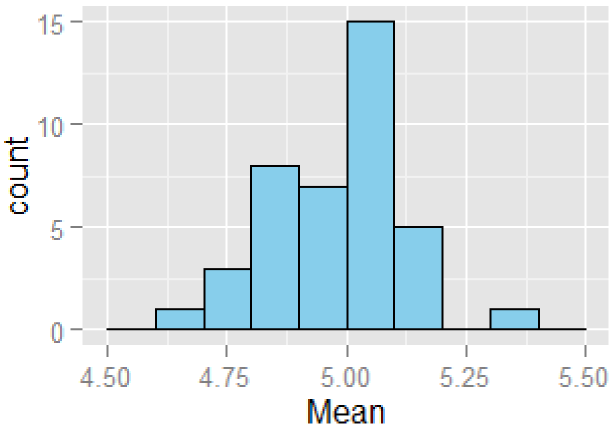
```
#Mean of each experiment
DFmean <- as.data.frame(cbind(seq(1,40,1),colMeans(DFexp)))
colnames(DFmean) <- c("Experiment","Mean")

#Variance of each experiemnt
DFvar <- apply(DFexp,2,var)
DFvar <- as.data.frame(cbind(seq(1,40,1), DFvar))
colnames(DFvar) <- c("Experiment","Variance")
```

Sample Mean versus Theoretical Mean

The EDF function is f(x) = lambda{e}^{-lambda*x} for x >= 0. Lambda is the rate parameter. The theoretical mean of the EDF distribution is 1/lambda and in this case is 1/0.2 or 5.

From our 40 experiments where we randomly sampled the EDF 1000 times, we calculated the mean. These 40 means are represented as a histogram in the following plot:



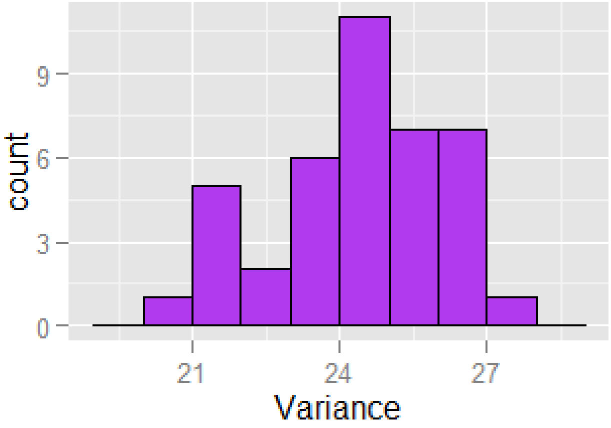
If we calculate the average of these means we obtain the value: 4.969789

The value 4.969789 is very close to the theoretical mean 5. This is not surprising and shows that this experiment did agree with the central limit theorem.

Sample Variance versus Theoretical Variance

We expect the same tendency of the experimentally calculated mean to be true for the experimentally calculated variance of the EDF. The theoretical Variance of the EDF can be calculated as the square of the standard deviation, and the standard deviation of the EDF is calculated as 1/lambda. Therefore, the variance for the EDF is 5^2 or 25.

From our 40 experiments where we randomly sampled the EDF 1000 times, we calculated the variance for these experiments individually and present them in the following histogram:



If we calculate the average of these variances we obtain the value: 24.4068993

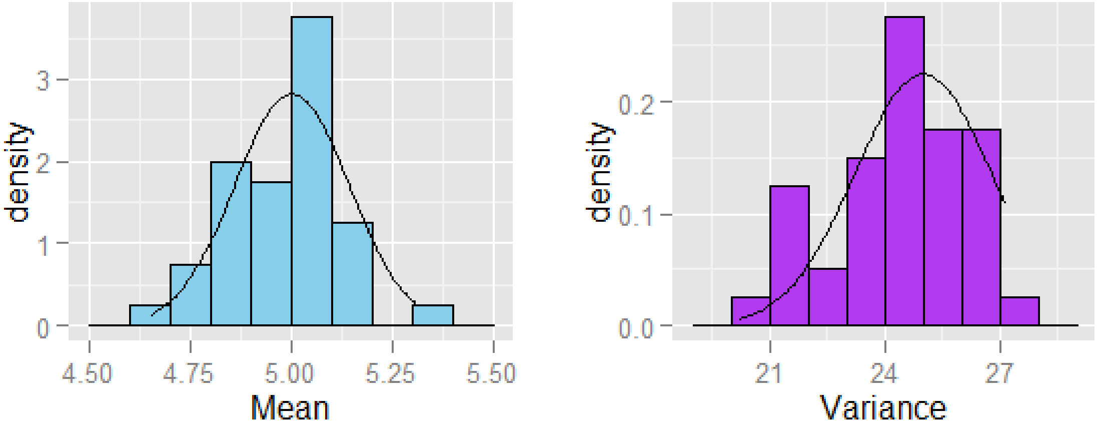
The value 24.4068993 is very close to the theoretical variance 25. This is not surprising and shows that this experiment did agree with the central limit theorem.

Distribution

One can tell the distribution of Means gathered from our 40 experiments is approximately normal by an informal approach of plotting a normal distribution over the histogram. In this case the y-axis must be normalized to 1 prior to mapping the normal distribution to this figure. The same principle applied to the distribution of variances.

From the following figure, we can tell that the two histograms approximate a normal distributions.

```
## Loading required package: grid
```



Appendix: R-code

```
library(ggplot2)
library(stats)
source("multiplot.R")

set.seed(777)
lambda = 0.2

#Plot of the exponential density function with line at the mean
x = seq(0,15,0.01)
p1 = qplot(x, dexp(x,lambda), geom="line", ylab="f(x)") +
        geom_vline(xintercept = 5, color="green")
p1

DF <- as.data.frame(matrix(1,1000,40))
DFexp <- sapply(DF,rexp, rate=0.2)

#Mean of each experiment
DFmean <- as.data.frame(cbind(seq(1,40,1),colMeans(DFexp)))
colnames(DFmean) <- c("Experiment","Mean")

#Variance of each experiemnt
DFvar <- apply(DFexp,2,var)
DFvar <- as.data.frame(cbind(seq(1,40,1), DFvar))
colnames(DFvar) <- c("Experiment","Variance")

ggplot(DFmean, aes(x=Mean)) +
        geom_histogram(fill = "skyblue", color = "black", size = 0.2,
binwidth=0.1)

ggplot(DFvar, aes(x=Variance)) +
        geom_histogram(fill = "darkorchid2", color = "black", size = 0.2,
binwidth=1)

Mstdev <- sd(DFmean[,2])
Vstdev <- sd(DFvar[,2])

#Comparison to normal distribution
p1 <- ggplot(DFmean, aes(x=Mean)) +
        geom_histogram(aes(y=..density..),fill = "skyblue", color = "black",
                       size = 0.2, binwidth=0.1) +
        stat_function(fun = dnorm, args = list(mean = 5, sd = Mstdev))

p2 <- ggplot(DFvar, aes(x=Variance)) +
        geom_histogram(aes(y=..density..),fill = "darkorchid2", color =
"black",
                       size = 0.2, binwidth=1) +
        stat_function(fun = dnorm, args = list(mean = 25, sd = Vstdev))
```

```r
multiplot(p1, p2, cols = 2)
```