Statistical Inference Project Pt.1: The Central Limit Theorem

Valentin Goverdovsky

24 December 2015

Overview

This report explores the central limit theorem (CLT) which states that the distribution of the sample means follows a normal distribution even if the samples are drawn from a population with not normal distribution. The mean value of the sample mean distribution is equal to that of the population mean and its standard deviation is $\frac{\sigma}{\sqrt{n}}$, where n is the sample size and σ^2 is the population variance. These ideas are illustrated with simulations using draws from a population with exponential distribution.

Simulations

```
#Load the simulation parameters and necessary libraries
library(ggplot2)
set.seed(555)
lambda=0.2
n=40
B=1000
```

First we draw 1000 numbers from the exponential distribution with $\lambda = 0.2$.

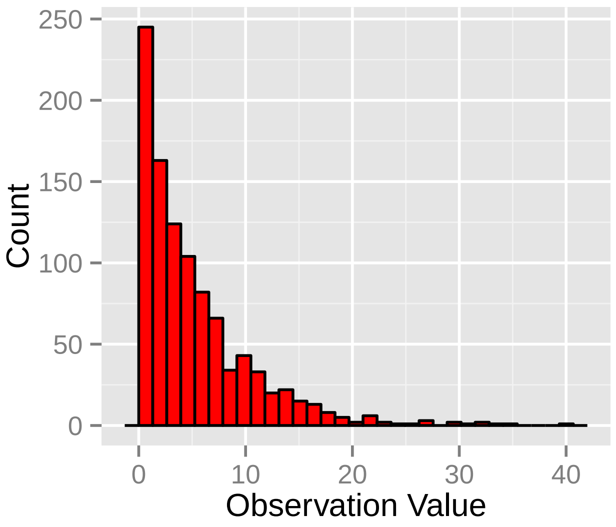
```
expDraw <- data.frame(expData=rexp(B,lambda),id='exp')
```

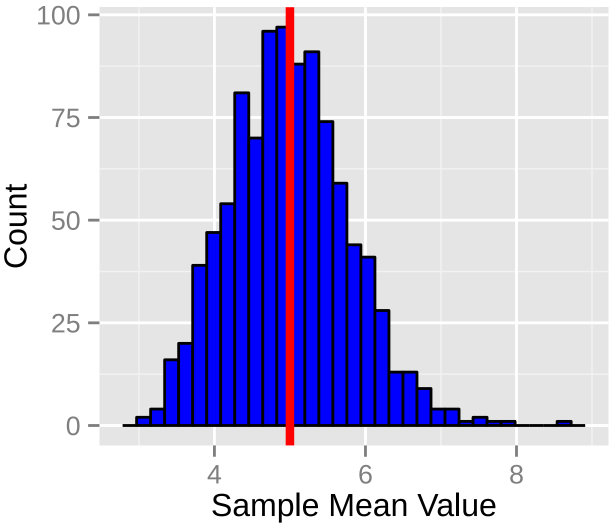
Now we draw 1000 samples of size 40 from the same distribution and find mean of each sample.

```
meanDraw <- data.frame(meanData=apply(matrix(rexp(B*n,lambda),B,n),1,mean),id='mean')
```

Now plot the histograms for draws from the exponential distribution as well as for sample means.

```
e <- ggplot(expDraw, aes(x=expData)) +
     geom_histogram(colour='black', fill='red',show_guide=FALSE) +
     labs(x = "Observation Value", y = "Count")
m <- ggplot(meanDraw, aes(x=meanData)) +
     geom_histogram(colour='black', fill='blue',show_guide=FALSE) +
     geom_vline(aes(xintercept=1/lambda), colour='red',lwd=1.5) +
     labs(x = "Sample Mean Value", y = "Count")
e;m
```





It is clear from these plots that although the samples follow the exponential distribution (left plot), the distribution of the sample means is close to normal (right plot).

Sample Mean versus Theoretical Mean

At this point we can clearly see from the sample mean plot that its mean of 4.99 as indicated by the red vertical line is very close to $\frac{1}{\lambda} = 5$ - the population mean, just as postulated by the CLT.

Sample Variance versus Theoretical Variance

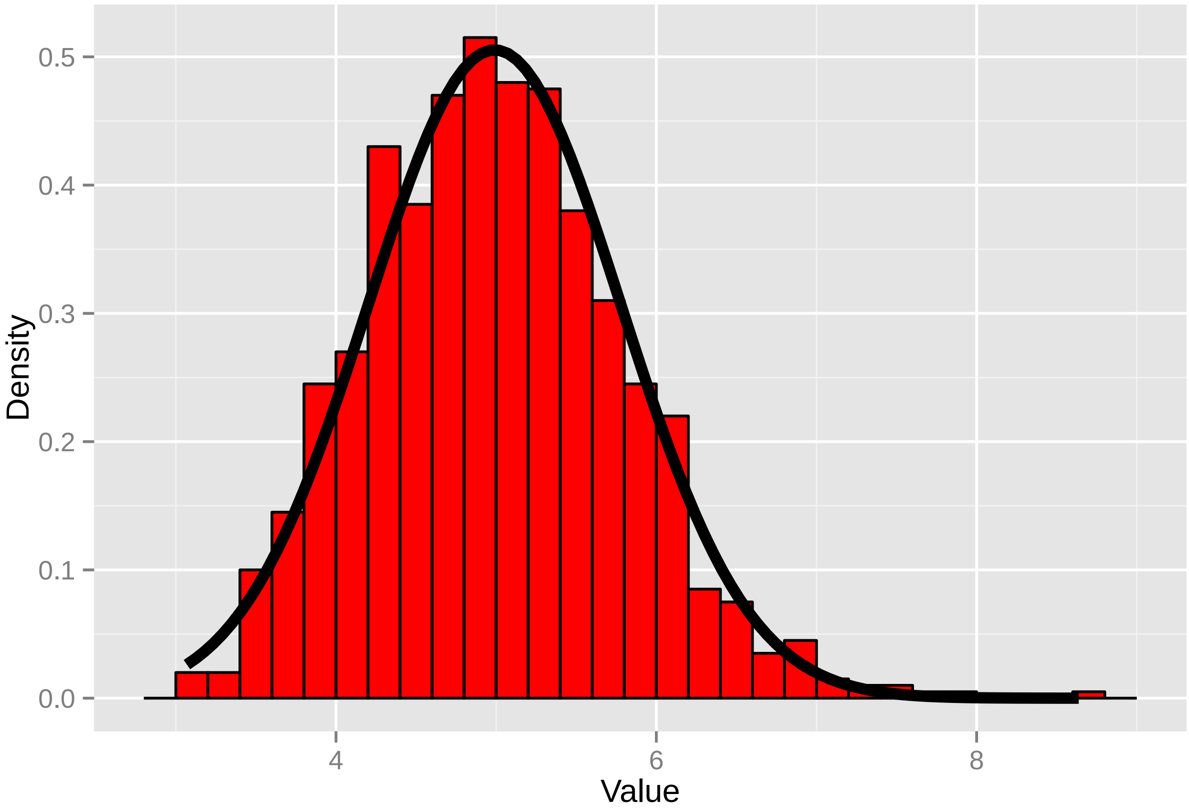
In addition to predicting the mean value of the distribution of the sample means, central limit theorem states that the variance of the distribution of the sample means is $\frac{\sigma^2}{n}$, where n is the sample size and σ is the standard deviation of the population, which in this case is equal to $\frac{1}{\lambda}$.

To verify this aspect of the CLT we can calculate the variance of 1000 sample means we simulated earlier and it is equal to 0.623, while the theoretical value using CLT is $\frac{1}{n\lambda^2} = \frac{1}{40 \times 0.2^2} = 0.625$. Both of these once again are very close to each other.

Distribution

There are mathematical methods and tools which allow one to formally assess the so called non-Gaussianity of the data (e.g. kurtosis is used in the Independent Component Analysis algorithm). For the purposes of this class we can appeal to a somewhat informal but illustrative method of comparing the histogram of the sample means that we obtained earlier and the probability density function of the normal distribution with mean and standard deviation chosen appropriately. An overlay plot of these is shown below, where we rescaled the y-axis of the sample mean histogram to display proportion rather than count values.

```r
m <- ggplot(meanDraw, aes(x=meanData)) +
     geom_histogram(binwidth=.2, aes(y = ..density..),
                    colour='black', fill='red', show_guide=FALSE) +
     stat_function(fun=dnorm,size=2,arg = list(mean = xBar, sd = sqrt(sSQ))) +
     labs(x = "Value", y = "Density")
m
```



This plot clearly illustrates that the distribution of the sample means very closely follows the normal distribution with mean 4.99 and variance 0.623.