A Simulation Exercise of Statistical Inference

Jonas Wiorek

Sunday, April 26, 2015

The Central Limit Theorem (CLT) is one of the most important theorems in statistics. The CLT states that the distribution of averages of independent and identical distributed (iid) variables becomes that of a standard normal as the sample size, n, increases.

In this report we will investigate the exponential distribution and compare it with the CLT. We take a population of exponential distribution with the rate lambda. The mean of the exponential distribution is 1/lambda and the standard deviation is also 1/lambda.

```
lambda <- 0.2
mu <- 1/lambda
sigma <-1/lambda
```

With lambda = 0.2, the mean of the exponential distribution is 5 and the standard deviation is 5.

```
n <- 40
nosim <- 1000
```

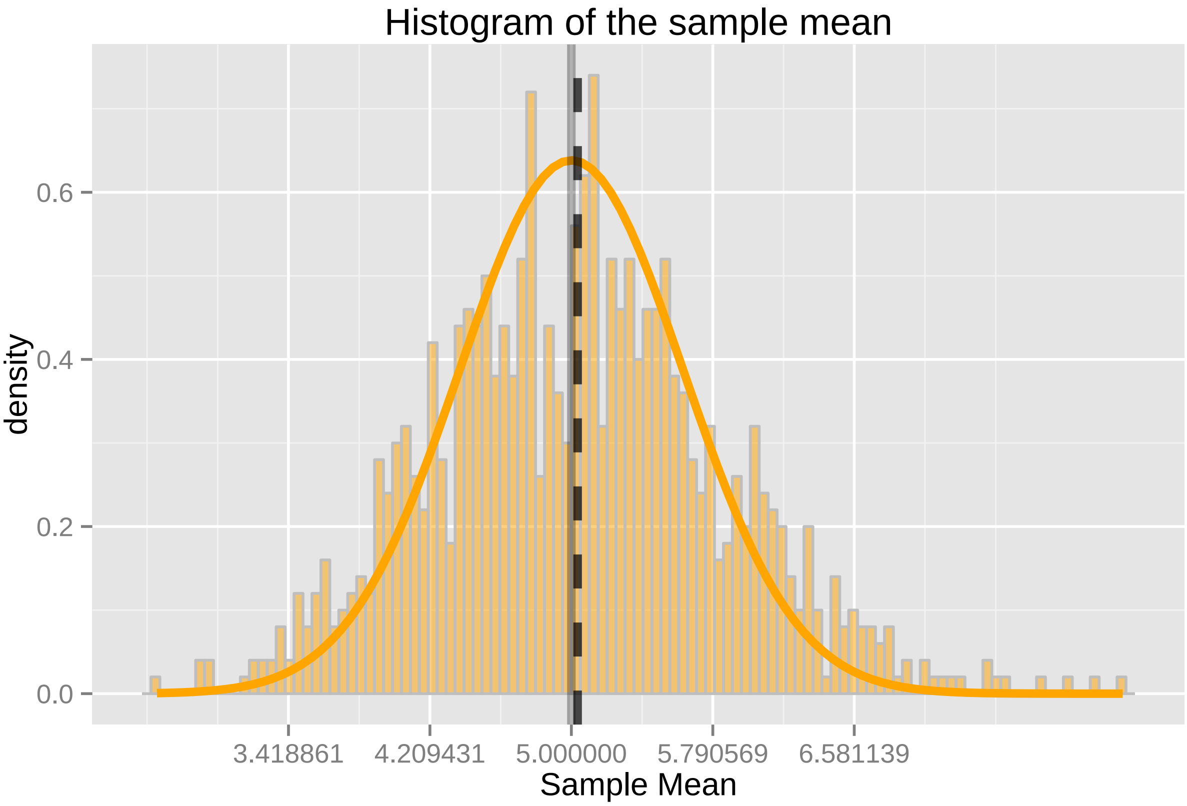
We will investigate the distribution of averages of 40 exponentials. We will do 1000 simulations.

The theory says that n samples of the exponential distribution has a mean asymptotically equal to lambda with increasing n, and that the mean of n samples of this exponetial distribution has a standard deviation asymptotically equal to lambda/sqrt(n).

Sample Mean

The histogram of the sample means from the simulations of the exponential distribution is plotted below. The expected value of the sample mean, xhat, is marked with a black dashed line and the population mean, mu, is marked with a solid grey line.

```
meansamples <- apply(matrix(rexp(nosim*n, lambda),nosim),1,mean)
xhat <- mean(meansamples)
df <- data.frame(meansamples)
ggplot(df, aes(x=meansamples)) +
        geom_histogram(aes(y = ..density..), binwidth=0.05, color='gray', fill = 'orange', alpha = 0.5)
        stat_function(fun = dnorm, geom='line',args=list(mean=mu,sd=sigma^2/n),col='orange',size=1.5) +
        geom_vline(xintercept = mu, color='black', alpha=0.3,linetype = 'solid', size=1.5) +
        geom_vline(xintercept = xhat, color='black',alpha=0.7,linetype = 'dashed', size=1.5) +
        scale_x_continuous(breaks=c(mu-2*sigma/sqrt(n),mu-sigma/sqrt(n),mu,mu+sigma/sqrt(n),mu+2*sigma/
        xlab("Sample Mean") +
        ggtitle('Histogram of the sample mean')
```



From the histogram above we could observe that the distribution of the sample means of the exponential distribution becomes that of a normal distribution. That is what is stated by the CLT. A normal distribution with mean = mu and standard deviation = sigma/sqrt(n) is overlayed to the histogram to illustrate this. The expected value of the sample mean, xhat, is equal to the population mean. xhat = 5.035. Thus, close to the populatin mean = 5. It would get even closer with a larger n according to the CLT.

```
# The variance of the sample mean is sigma^2/n
sdMean <- sd(apply(matrix(rexp(nosim*n, lambda),nosim),1,mean))
# The logical estimate of the variance of the sample mean is S^2/n,
# where S^2 is the sample variance
stddevsamples <- apply(matrix(rexp(nosim*n, lambda),nosim),1,sd)
S <- mean(stddevsamples)
```

Further, the variance of the sample mean is sigma^2/n, i.e 0.625. The variance of the sample mean for the simulations is 0.595. The logical estimate of the variance of the sample mean is S^2/n, where S^2 is the sample variance. The sample variance over n for the simulations is 0.592.

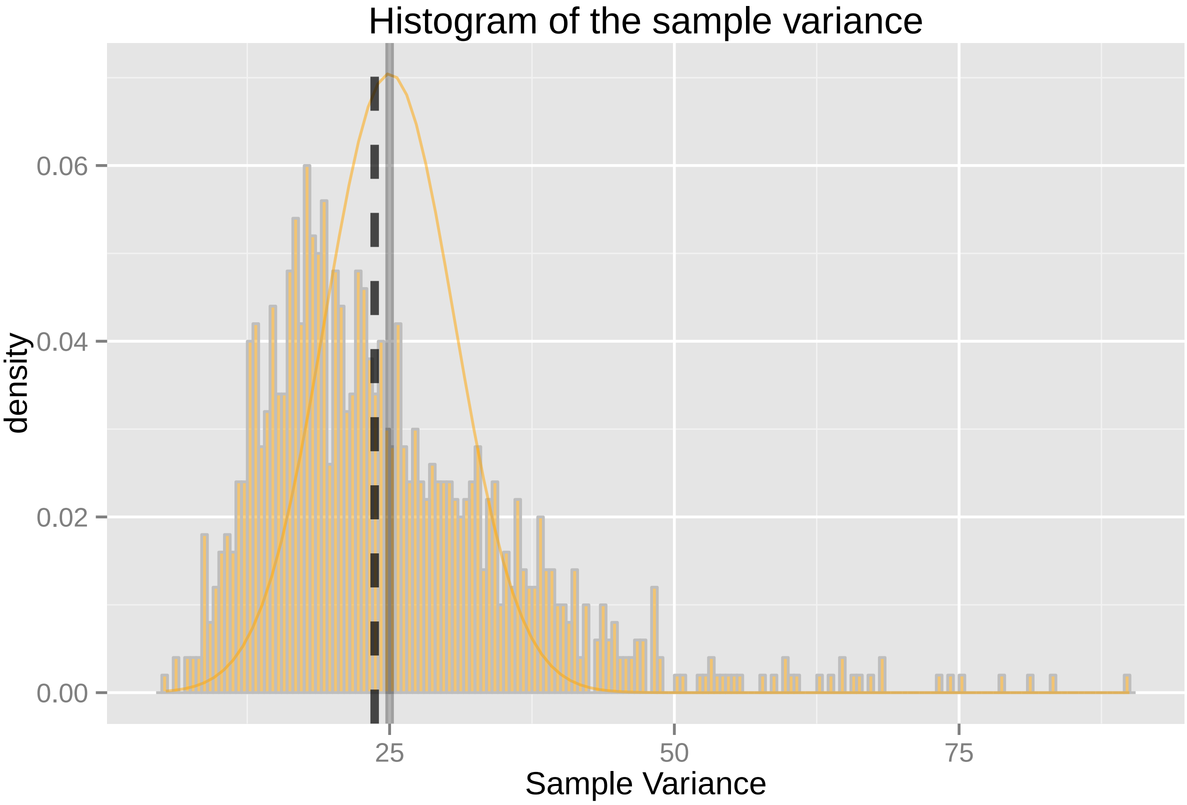
Sample Variance

The sample variance, S^2, estimates the population variance sigma^2. The sample variance expected value is equal to the population variance, i.e. the distribution of the sample variance is centered around sigma^2.

The histogram of the sample variance of the simulations is plotted below. The expected value of the sample varaince, S^2, is marked with a black dashed line and the population variance, sigma^2, is marked with a solid grey line.

```
variancesamples <- apply(matrix(rexp(nosim*n, lambda),nosim),1,var)

 df <- data.frame(variancesamples)
 ggplot(df, aes(x=variancesamples)) +
         geom_histogram(aes(y = ..density..), binwidth=0.5, color='gray', fill = 'orange', alpha = 0.5)
         stat_function(fun = dnorm, geom='line', args=list(mean=sigma^2,sd=sqrt(2*sigma^4/(n-1))),col='
         geom_vline(xintercept = sigma^2, color='black', alpha=0.3,linetype = 'solid', size=1.5) +
         geom_vline(xintercept = S^2, color='black',alpha=0.7,linetype = 'dashed', size=1.5) +
#          scale_x_continuous(breaks=c(sigma^2-10,sigma^2,sigma^2+10) +
        xlab("Sample Variance") +
        ggtitle('Histogram of the sample variance')
```



```
# The distribution of the sample variance gets more centered around
# the populatin variance sigma^2 with more data
sdvs <- sd(variancesamples)
```

From the plot above we could observe that the distribution of the sample variance of the exponential distribution becomes that of a normal distribution with large n. The expected value of the sample variance, S^2, is equal to the population variance. S^2 = 23.686. i.e. close to the populatin variance = 25.

Finally, the distribution of the sample variance gets more centered around the populatin variance sigma^2 with increasing n. In the simulation with n equals to 40 the standard deviation of the sample variance is about 11.611.