Statistical Inference Project: Part 1

jmtaysom

05/08/2015

Executive summary

In this report we will explore the Central Limit Therom and how it applys to multiple draws of a sample from a population. Specifically we will look at how 40 means of draws of sample of size 40 random exponential numbers with a lambda of 0.2 compares to 1000 random normal observations. The random normal numbers will have a mean and standard deviation equivilant to the theoretical mean and standard deviation that we would expect for the exponential numbers. The mean and standard deviation of the population of eponential numbers are both equal to $\frac{1}{\lambda}$ or in ths case equal to 5. The theoretical standard deviation of the sample will be equal to $\frac{\sigma}{\sqrt{40}}$ on in this case 0.7905694. The actuall mean of the 1000 means of 40 exponential numbers is 4.9896776 and the standard deviation is 0.7862304 show that the observed values are representitive of the population according to the Central Limit Therom.

Introduction

In this experiment we will be comparing the distribution of the of 1000 means of samples of 40 exponential numbers to their expected distribution through the use of the central limit therom. The poultaion of exponential numbers will have a mean and standard deviation of $\frac{1}{\lambda}$ or $\frac{1}{0.2} = 5$. The standard deviation of the sample will be equal to $\frac{\sigma}{\sqrt{40}}$ on in this case 0.7905694.

Calculations

Comparing to expected values

Upon sampling the exponential numbers we can start comparing them to the expected values. The expected mean of the samples is 5 while the observed mean is 4.9896776 as seen in Figure 1 in the appendix. The expected variance is 0.625 and the observed variance is 0.6181582. However with the sample size being 40 the variance is much smaller than the variance of the population. The variance of the population would be $(\frac{1}{\lambda})^2$ or 25 which is greater than the observed variance as seen in figure 2 in the appendix. With the observed mean and variance so close to the expected mean and standard deviation we can see that the central limit therom governs this dataset.

Comparing to normal data

While the distibution of exponential numbers is not normal the distribution of the mean of the samples should be normal. This can be tested by comparing the distibution of the samples to the distribution of the same number of random normal numbers that have a mean and standard deviation equal to that of the samples of exponential numbers. The result as seen in figure 3 show that distibution of the means of the samples of exponential numbers are distributed normally.

Appendix

Figures

Figure 1

Histogram of the sampled expoential numbers with the expected mean and actual mean

```
xdf <- data.frame(expn)
x <- ggplot(xdf,aes(x=expn)) + geom_histogram(aes(fill = ..count..), binwidth=.25) + xlim(2,8) +ylim(0,
x + geom_vline(xintercept = mean(expn), color='red') + geom_vline(xintercept = 1/lambda, color='black')
```

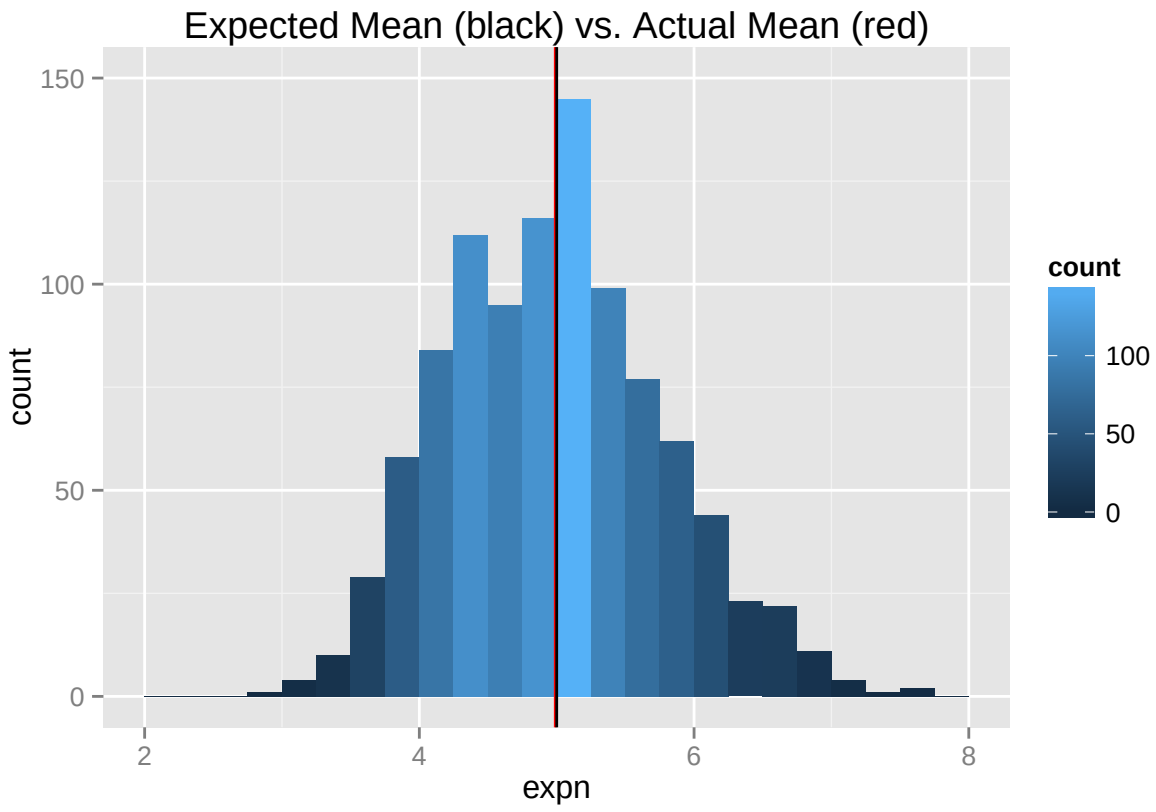


Figure 2

The variance of the sample is smaller than the variance of the poplutation

```
norm <-rnorm(1000,mean=5, sd = 5)
ndf <- data.frame(norm)
ndf$distribution <- 'population'
names(ndf)[names(ndf)=="norm"] <- "value"
xdf$distribution <- 'sample'
names(xdf)[names(xdf)=="expn"] <- "value"
combined <- rbind(xdf,ndf)

ggplot(combined, aes(x=value, colour=distribution)) + geom_density() +ggtitle('Sample Variance vs. Popul
```

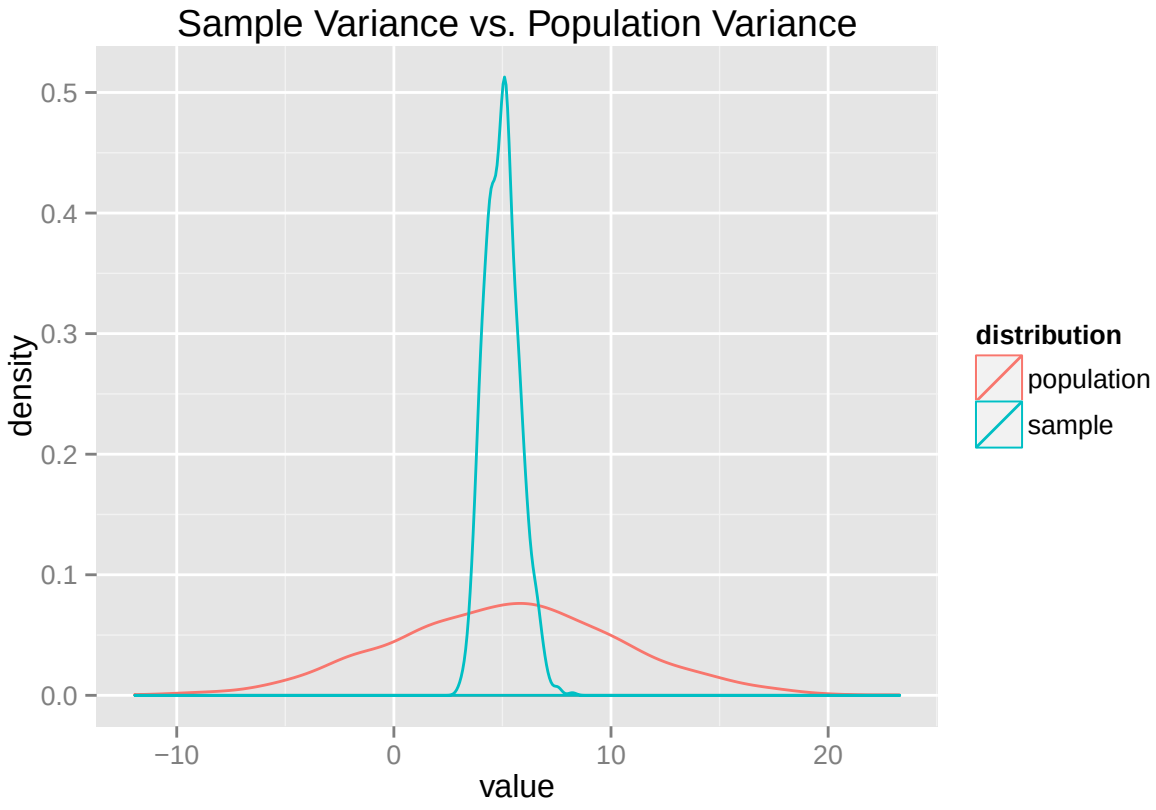
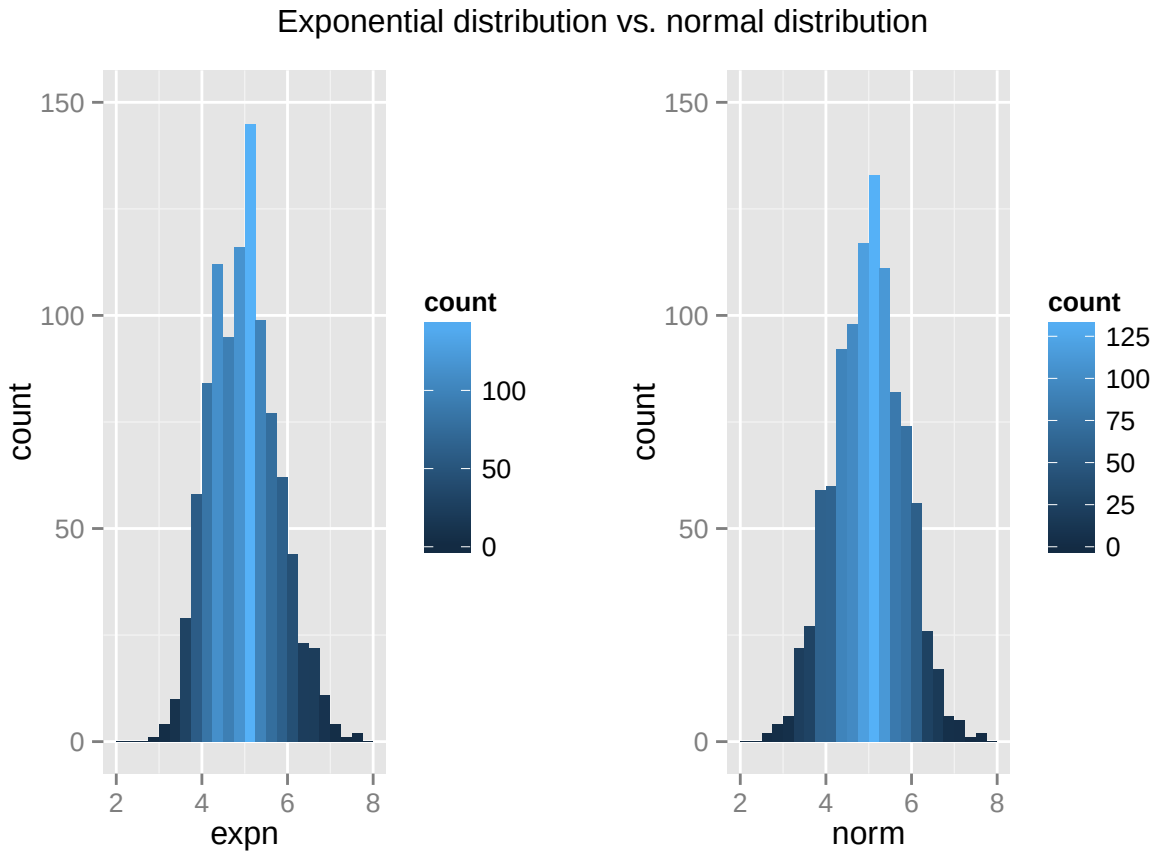


Figure 3

Comparing the distribution of the samples to a normal distribution

```
norm <-rnorm(1000,mean=5, sd = 5/sqrt(40))
ndf <- data.frame(norm)
n <- ggplot(ndf, aes(x=norm)) + geom_histogram(aes(fill = ..count..), binwidth=.25) + xlim(2,8) +ylim(0

grid.arrange(x,n, ncol=2, main='Exponential distribution vs. normal distribution')
```



Loading Data

Here we set the seed to verify the reproducibility of the results. Set the variables according to the experiment. And create two samples; one for 1000 of the mean of 40 exponential numbers, and one for 1000 of normal numbers that have a mean and sd equal to the exponential numbers.

```
set.seed(0)
lambda <- 0.2
observations <- 40
trials <- 1000

expn = NULL
for (i in 1:trials) expn = c(expn, mean(rexp(observations, lambda)))

exsigma <- (1/lambda)/sqrt(observations)
exmean <- (1/lambda)
acsigma <- sd(expn)
acmean <- mean(expn)

norm <- rnorm(trials,mean=1/lambda, sd = (1/lambda)/sqrt(observations))
```