Stats Inference Project

Michael Pearson

August 13, 2016

Overview

I will compare the theoretical and simulated means and standard deviations of a large set of samples of the random exponential distribution.

Simulations

I ran a simulation where exponential random sets of 40 results - with a mean of 5 and a standard deviation of - were generated 1000 times. I found the mean and standard deviation of each of the 100 sets of 40 results and compared the simulated results of the 1000 trials with the theoretical, expected results (expected mean of 5 and standard deviation of 5). The code generates a matrix with each row being 40 exponential random numbers of theoretical mean 5 and standard deviation 5 (1/lambda, with lambda as 0.2), then calculates the mean and standard deviation of each row of the matrix.

```
set.seed(500)
library(matrixStats)
```

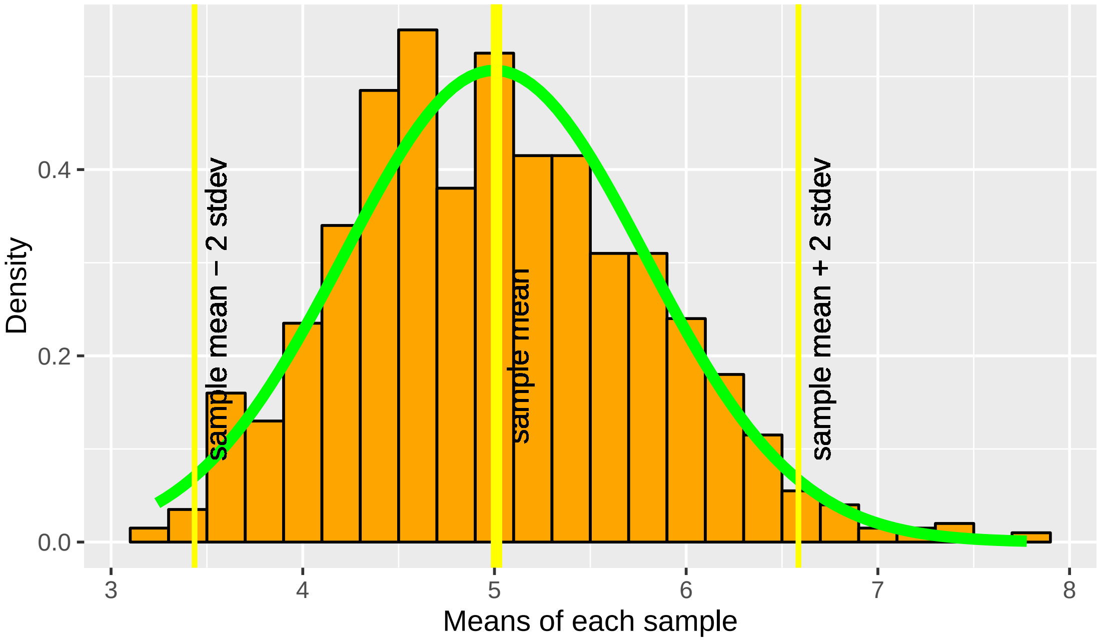
```
## matrixStats v0.50.2 (2016-04-24) successfully loaded. See ?matrixStats for help.
```

```
## create a matrix of 1000 rows and 40 columns - with each row generated by a random exponential distril
trippy <- matrix(data = 1, nrow = 1000, ncol = 40)
for (i in 1:1000){
  trippy[i,] <- rexp(40, 0.2)
}
## get the means of each row of exponential randoms
meanz <- rowMeans(trippy)
## get sd and var of each row
sdz <- rowSds(trippy)
varz <- rowVars(trippy)
## find the mean, variance and sd of the means of each row
meanit <- mean(meanz)
varmeans <- var(meanz)
sdomeans <- sd(meanz)
expectedsd <- 1/(0.2*sqrt(40))
## now for the mean of the standard deviation and its standard deviation
meansd <- mean(sdz)
sdzsd <- sd(sdz)
vart <- var(varz)
```

```
library(ggplot2)
valu <- t.test(meanz, mu = 5)$p.value
mns <- data.frame(meanz)
par(mfrow = c(1,1) ,mar = c(1,4,1,1))
g <- ggplot(mns, aes(x = meanz, fill = NULL)) + geom_histogram(aes(y=..density..), binwidth = 0.2, col =
g <- g +  stat_function(fun=dnorm, args=list(mean = 5, sd = sd(meanz)), colour = "green", size = 2)
```

```r
g<- g + geom_vline(xintercept = meanit, colour = "yellow", size = 2, show.legend = TRUE)
g<- g + geom_vline(xintercept = meanit-2*sd(meanz), colour = "yellow", size = 1, show.legend = TRUE)
g<- g + geom_vline(xintercept = meanit+2*sd(meanz), colour = "yellow", size = 1, show.legend = TRUE)
g<- g + geom_text(aes(x = meanit, label = "\nsample mean", y= 0.2), angle = 90) +  xlab("Means of each s
g<- g + geom_text(aes(x = meanit-2*sd(meanz), label = "\nsample mean - 2 stdev", y= 0.25), angle = 90)
g<- g + geom_text(aes(x = meanit+2*sd(meanz), label = "\nsample mean + 2 stdev", y= 0.25), angle = 90)
g <- g + theme(plot.margin = unit(c(1, 1, 1, 1), "cm"))
 g
```

Figure 1: Distribution of means of
40 random exponentials from 1000 simulations



Sample Mean versus Theoretical Mean:

The mean of the 1000 simulations was 5.0105617 as compared to the theoretical expected value of 5.

The standard error of the means was 0.7874779. We expected this to be 5 divided by the square root of 40: 0.7905694. The difference between the inferred standard error and the theoretical is -0.0030915. So this was a good confirmation of our idea that the simulated values would be close to the theoretical ones.

The figure above has a normal distribution (in green) overlaid with the simulated results, to help given an idea of how well the data fits the theoretical expectations.

Testing the hypothesis that the mean is 5

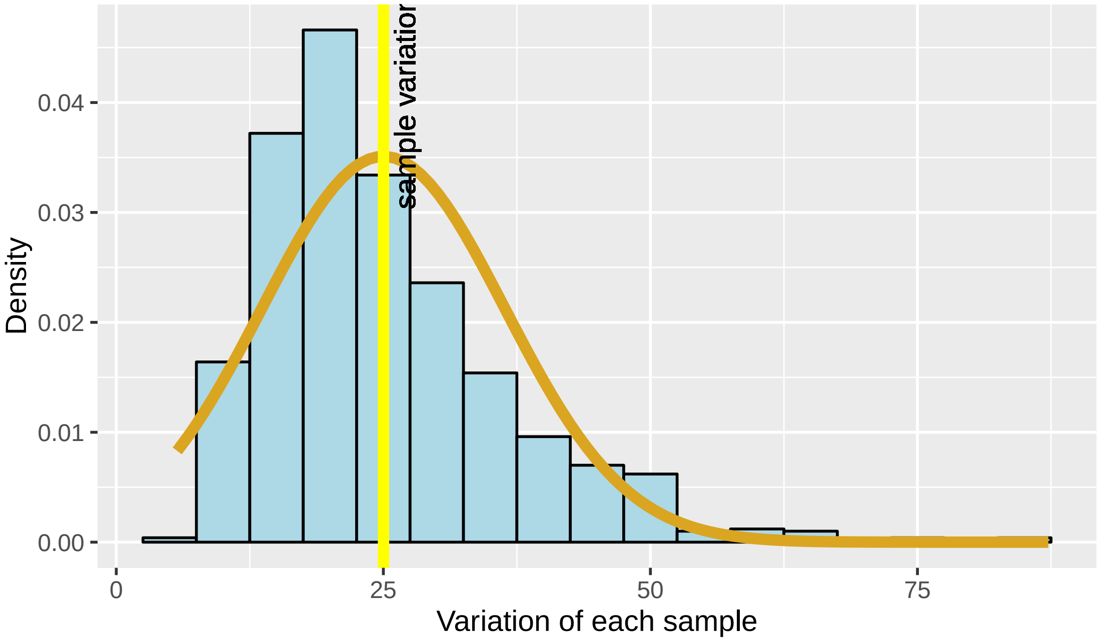
We can test the hypothesis that the mean of the simulated means is 5 by using a t test. The p-value we get for the null hypothesis (mu = 5) is 0.6715655 , which is a very convincing value to not reject the null hypothesis.

Sample Variance versus Theoretical Variance:

For the sample variance, the mean of the simulated variation was 25.0162204, very close to the theoretical value of 25 (since variation is the standard deviation squared).

```
vns <- data.frame(varz)
par(mfrow = c(1,1) ,mar = c(1,4,1,1))
v <- ggplot(vns, aes(x = varz, fill=NULL)) + geom_histogram(aes(y=..density..), binwidth = 5, col = "bla
v <- v +  stat_function(fun=dnorm, args=list(mean = 25, sd = sd(varz)), colour = "goldenrod", size = 2)
v <- v + geom_vline(xintercept = mean(varz), colour = "yellow", size = 2, show.legend = TRUE)
v <- v + geom_text(aes(x = mean(varz), label = "\nsample variation", y= 0.04), angle = 90) +  xlab("Var
v <- v + theme(plot.margin = unit(c(1, 1, 1, 1), "cm"))
 v
```

Figure 2: Distribution of Variations of
40 random exponentials from 1000 simulations



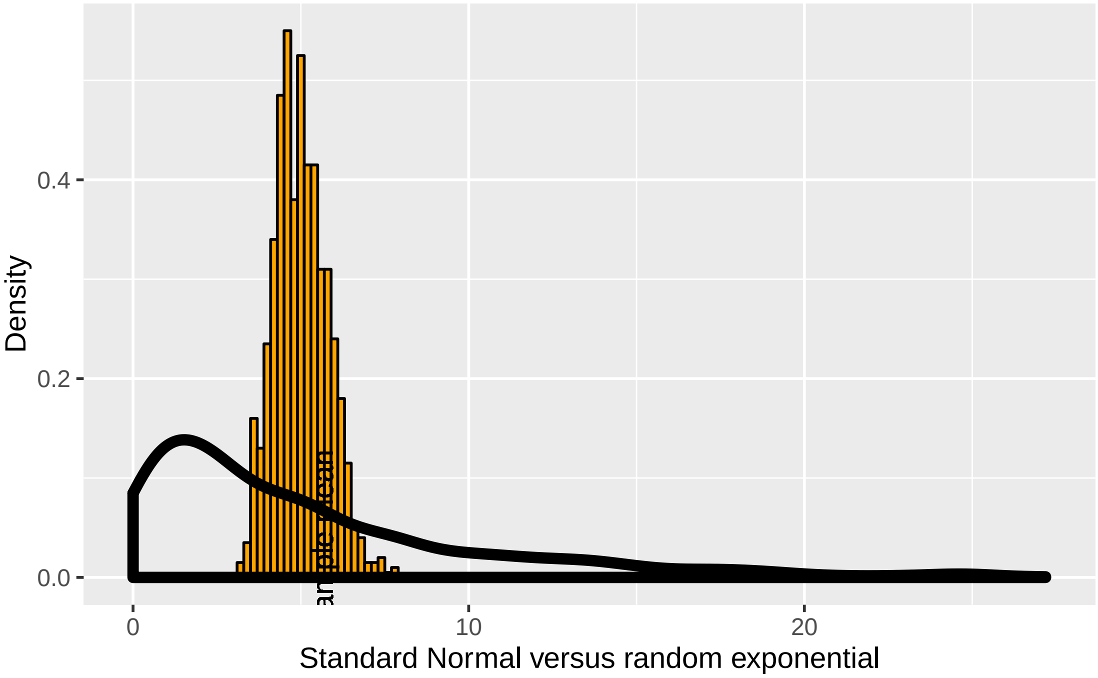
The goldenrod line is a theoretical normal density with mean of 25 and standard deviation of 11.3661229, the measured standard deviation of the samples.

Testing the hypothesis that the variance is 25

We can test the hypothesis that the mean of the simulated variances is 25 by using a t test. The p-value we get for the null hypothesis (mu = 25) is 0.964014 , which is a very convincing value to not reject the null hypothesis.

```
par(mfrow = c(1,1) ,mar = c(1,4,1,1))
h <- ggplot(mns, aes(x = meanz, fill=NULL)) + geom_histogram(aes(y=..density..), binwidth = 0.2, col =
h <- h + geom_density(aes(x=rexp(1000, 0.2)), size = 2, colour = "black")
h <- h + geom_text(aes(x = mean(meanz), label = "\nsample mean", y= 0.04), angle = 90) +  xlab("Standar
h <- h + theme(plot.margin = unit(c(1, 1, 1, 1), "cm"))
h
```

Figure 3: Comparison of Random Normal of Means to Random Exponential Distril



Distribution

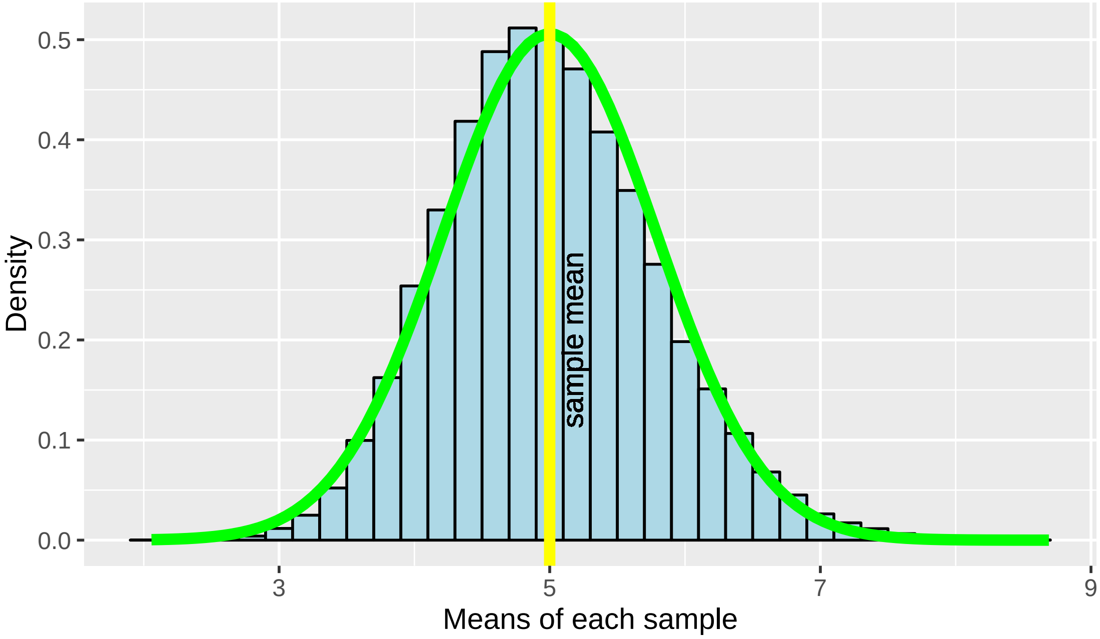
How can we show that the distribution of means is normal? If we compare a random exponential distribution to the distribution of the means, we can see that they are radically different. This alone does not demonstrate that the distribution of means is normal. If we greatly increase the number of simulations we can see how the distribution of means gets closer to a normal distribution. This is a figure that compares the two distributions.

The figure below shows what a simulation of 100,000 random exponential functions would look like.

```
trappy <- matrix(data = 1, nrow = 100000, ncol = 40)
for (i in 1:100000){
  trappy[i,] <- rexp(40, 0.2)
}
meanza <- rowMeans(trappy)
## get sd and var of each row
sdza <- rowSds(trappy)
meanita <- mean(meanza)
sdomeansa <- sd(meanza)
meansda <- mean(sdza)
sdzsda <- sd(sdza)
mnsa <- data.frame(meanza)
c <- ggplot(mnsa, aes(x = meanza, fill = NULL)) + geom_histogram(aes(y=..density..), binwidth = 0.2, co
c <- c +  stat_function(fun=dnorm, args=list(mean = meanita, sd = sdomeansa), colour = "green", size = 
c <- c + geom_vline(xintercept = meanita, colour = "yellow", size = 2, show.legend = TRUE)

c <- c + geom_text(aes(x = meanit, label = "\nsample mean", y= 0.2), angle = 90) +  xlab("Means of each
c<- c + theme(plot.margin = unit(c(1, 1, 1, 1), "cm"))
c
```

Figure 4: Distribution of means of
40 random exponentials from a hundred thousand simulations



Conclusion

We can see that the distribution of means of a random exponential will approach a normal distribution with the same mean as the exponential, and the same variation. This is one way to demonstrate the central limit theorem.