Statistical Inference Project (Coursera)

Jason Murray

2017-01-17

Overview

In this project we will be exploring the exponential distribution using R. Looking at how the theoretical mean and variance compare to a sample mean and variance. We will also be looking at how the values of the exponential function are distributed compared to how the values of the means of samples of exponential function are distributed as a demonstration of the central limit theorem in action.

Simulations

Required Libraries

```
library(ggplot2)
library(dplyr)
```

First let's set a random seed and generate a 1000 means of sample size 40 from the exponential distribution. We will be using $\lambda = .2$ throughout.

```
set.seed(42)

# Generate a 1000 means of random values form the exponential distribution
re1000m = NULL
for (i in 1 : 1000) re1000m = c(re1000m, mean(rexp(n= 40, rate = .2)))
```

Sample Mean vs Theoretical Mean

The theoretical mean of the exponetial distribution is equal to $1/\lambda$. So with $\lambda = .2$ the theoretical mean would be equal to

$1/\lambda = 1/.2 = 5$

If we now calculate the mean for our 1000 means we get.

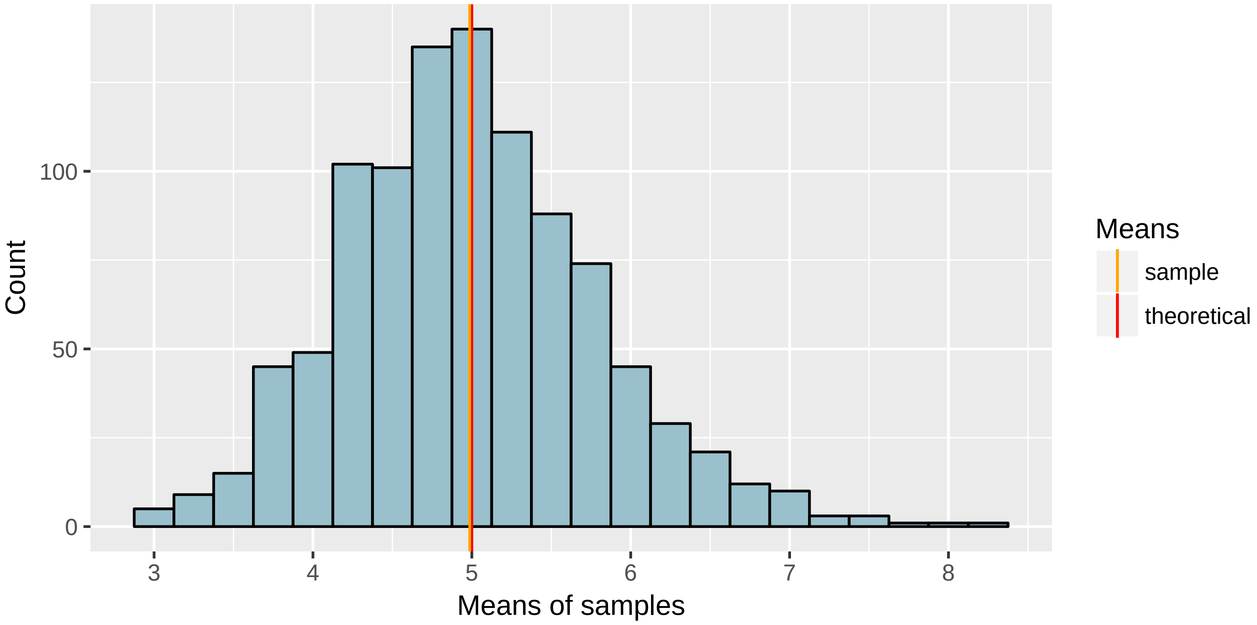
```
mre1000m <- mean(re1000m)
mre1000m
```

```
## [1] 4.986508
```

Which is very close to the theoretical value of 5.

Let's graph the distribution of our sample means and so we can see how close.

```
ggplot() +
    aes(re1000m) +
    geom_histogram(binwidth = .25, fill = "light blue 3", color = "black") +
    geom_vline(aes(xintercept = 5, color = "theoretical")) +
    geom_vline(aes(xintercept = mre1000m, color = "sample")) +
    scale_color_manual(name = "Means", values = c(theoretical = "red", sample = "orange")) +
    labs(x = "Means of samples", y = "Count")
```



Sample Variance vs Theoretical Variance

The theoretical variance for the distribution of samples means is σ^2/N where sigma is equal to $1/\lambda$. We already know $1/\lambda = 5$ so plugging in we get: $5^2/N = 25/40 = .625$

```
var(re1000m)
```

```
## [1] 0.6344405
```

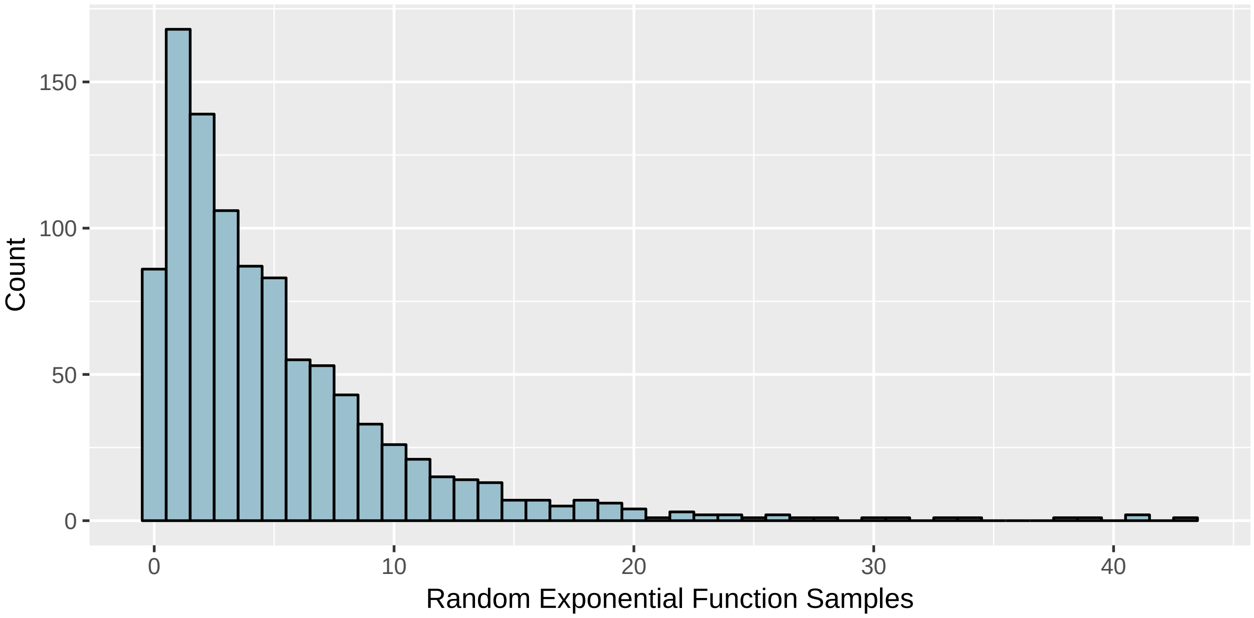
Which is again very close to the theoretical variance of .625.

Distribution

We have so far been looking at the distribution of sample means but what does the distribution of the sample exponential function variables actually look like?

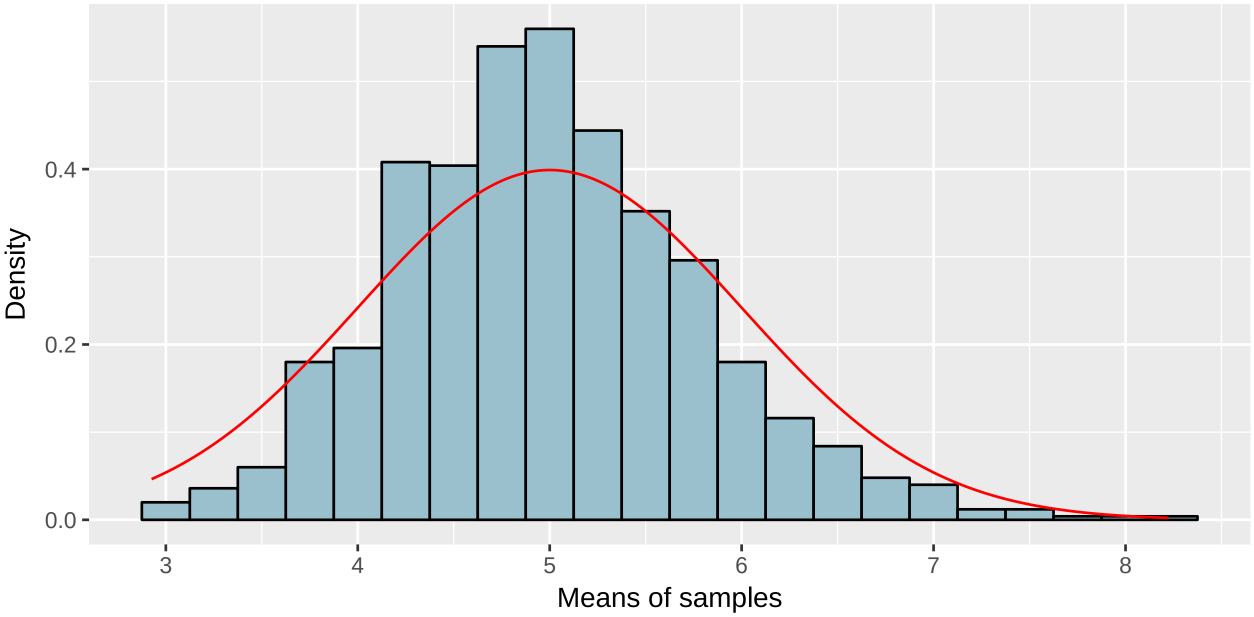
Let's sample a 1000 random variables from the exponential function and see how they are distributed

```
re1000 <- rexp(1000, rate = .2)
ggplot() +
    aes(re1000) +
    geom_histogram(binwidth = 1, fill = "light blue 3", color = "black") +
    labs(x = "Random Exponential Function Samples", y = "Count")
```



So why does the distribution of 1000 means of 40 samples look different than 1000 sample values?

```
ggplot() +
    aes(re1000m) +
    geom_histogram(aes(y=..density..),binwidth = .25, fill = "light blue 3", color = "black") +
    stat_function(fun = dnorm, color = "red", n = 1000, args = list(mean = 5)) +
    labs(x = "Means of samples", y = "Density")
```



This is because of the central limit theorem. Regardless of how the original data is distributed the distribution of the means of samples approaches a normal distribution centered at the mean of the population as the sample size get's larger. I changed the y axis to density above and overlayed a normal curve to show that the shape of the distribution of means is approaching normal.