Statistical Inference Course Project

Shayaan Ahmed Farooqi

5/20/2020

```
library(ggplot2)
set.seed(1)
```

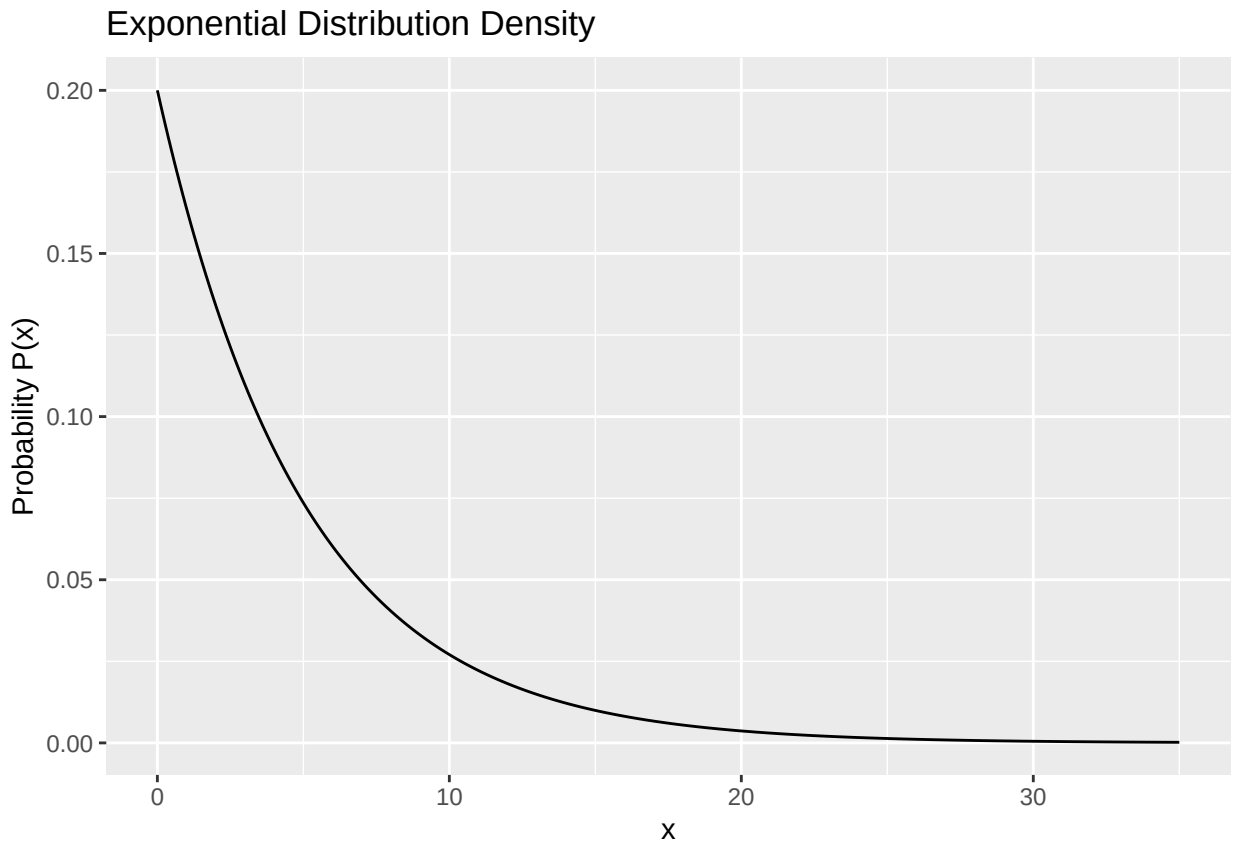
Overview

We are going to simulate samples from the exponential distribution and use it to verify the central limit theorem by comparing the mean and variance of the sampling distribution to the actual population mean and variance.

The exponential Distribution:

This is what the exponential distribution looks like with theoretical mean and standard deviation $= \frac{1}{\lambda} = \frac{1}{0.2} = 5$:

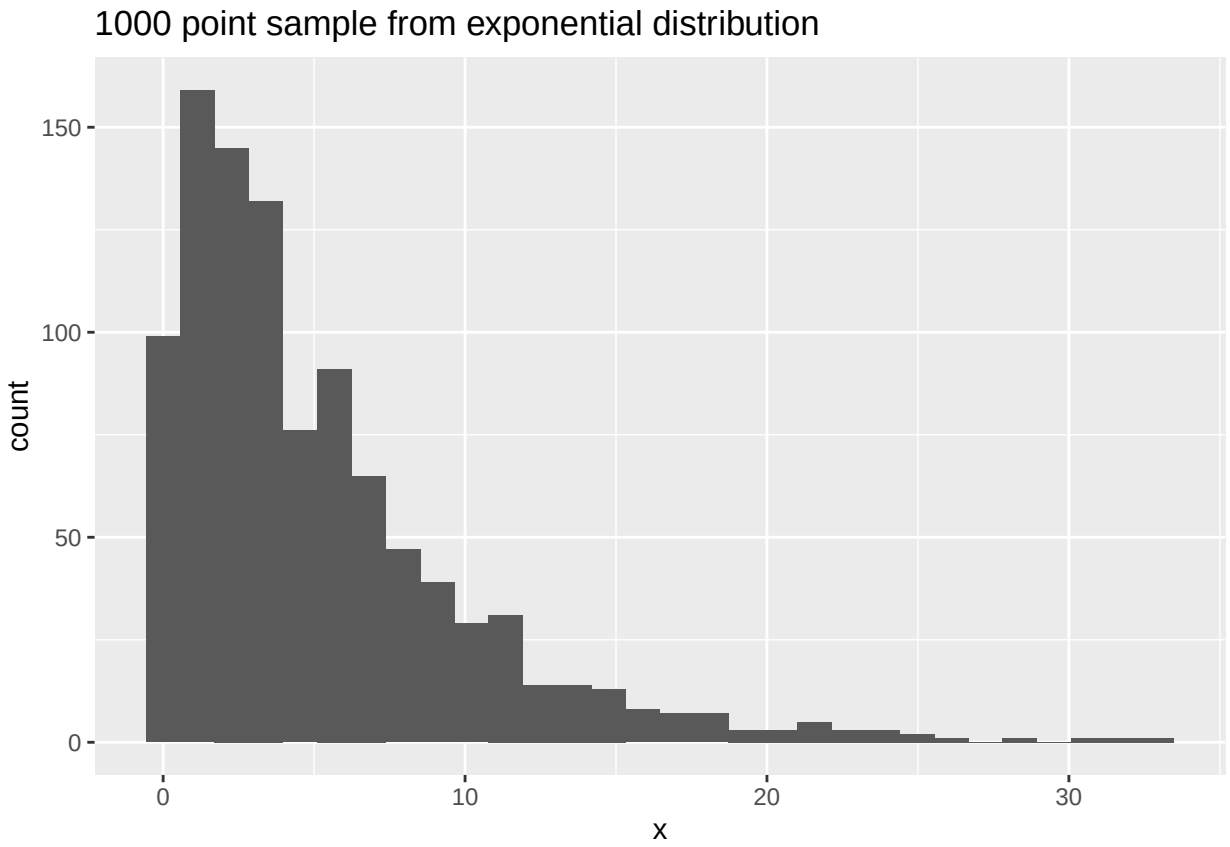
```
ggplot() +
    aes(y=dexp(seq(0,35,length.out = 1000), 0.2), x=seq(0,35,length.out = 1000)) +
    geom_line() +
    xlab("x") +
    ylab("Probability P(x)") +
    ggtitle("Exponential Distribution Density")
```



Lets take a sample of 1000 points from this exponential distribution and see what its histogram looks like:

```
expdata = rexp(1000, 0.2)
ggplot() +
    aes(x=expdata) +
    geom_histogram() +
    xlab("x") +
    ggtitle("1000 point sample from exponential distribution")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



As we expected, it follows the shape of the exponential distribution.

Now instead of taking just one sample of 1000 points, we will take 1000 samples of 40 points each from the exponential distribution. Then we will calculate the mean of all 1000 samples and then plot the histogram of these 1000 sample means. First lets see what the mean and variance turn out to be for these 1000 sample means:

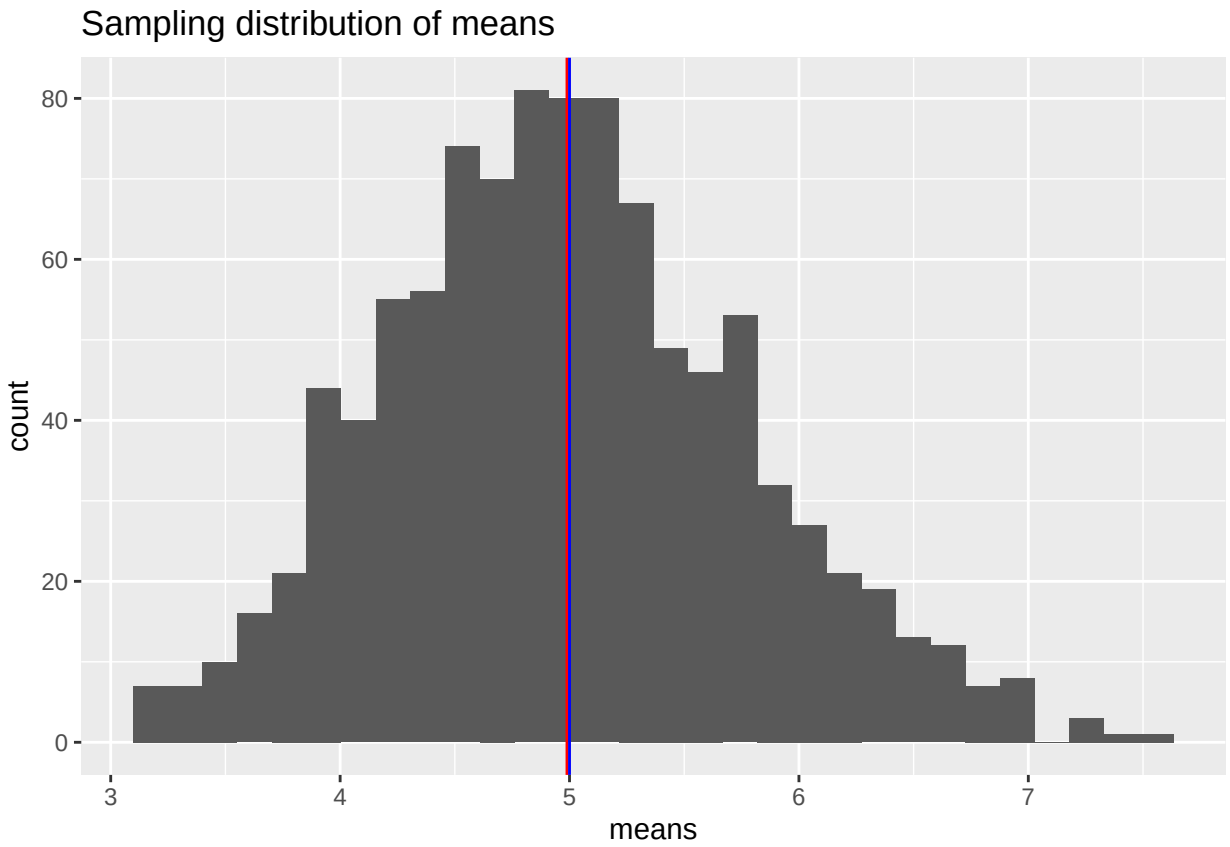
```
lambda = 0.2
means = NULL
for (i in 1 : 1000) means = c(means , mean(rexp(40, lambda)))
sample_mean <- mean(means)
sample_sd <- sd(means)
```

The mean of these 1000 sample means turned out to be 4.988882 which is approximately equal to the theoretical mean given by $\frac{1}{\lambda} = \frac{1}{0.2} = 5$. The variance turned out to be 0.6119066 which, according to the central limit theorem is related to the theoretical variance as $sd^2 = \frac{\sigma^2}{(N)}$ which is $\frac{25}{40} = 0.625$

Now lets see what the distribution of these sample means looks like:

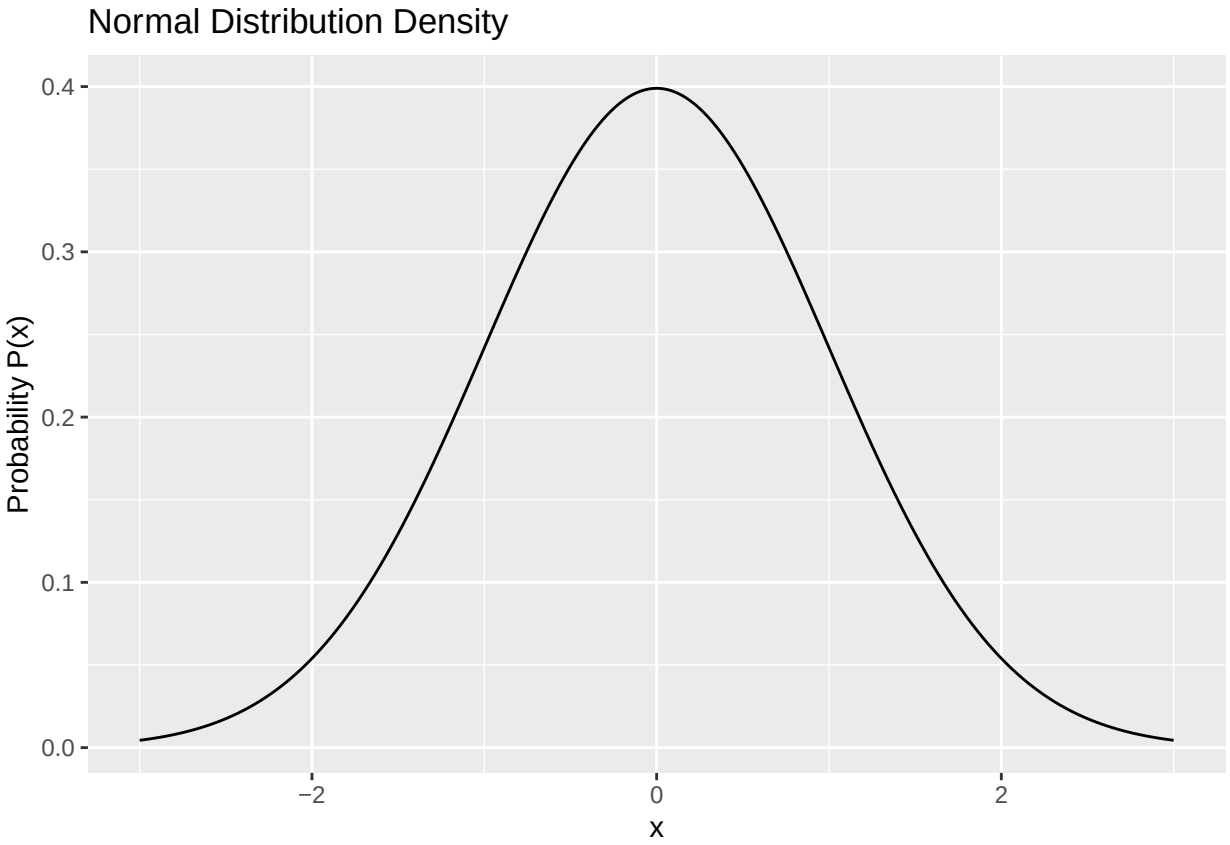
```
ggplot() +
    aes(x=means) +
    geom_histogram() +
    geom_vline(xintercept=mean(means), color='red') +
    geom_vline(xintercept=5, color='blue') +
    ggtitle("Sampling distribution of means")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We see that the sampling distribution of the means looks like its following a gaussian or normal distribution which we can compare with the standard normal figure below. We can also see the vertical blue line representing the true mean and the red line representing the mean of these means and it is closely approximating the true value of the exponential distribution $\mu = 5$. The thing to note here is that earlier on we saw the histogram for a SINGLE sample with 1000 points and we saw that it had an exponential distribution. The figure we see here is the histogram for the means of a THOUSAND samples with 40 points and we see that it follows a normal distribution.

```
ggplot() +
    aes(y=dnorm(seq(-3,3,length.out = 1000)), x=seq(-3,3,length.out = 1000)) +
    geom_line() +
    xlab("x") +
    ylab("Probability P(x)") +
    ggtitle("Normal Distribution Density")
```


Normal Distribution Density
Probability P(x)
0.4
0.3
0.2
0.1
0.0
−2
0
2
x