Demonstrate Central Limit Theorem using Exponential Distribution

Abhilash Itharaju

Overview

This is a simulation experiment to see if Exponential Distribution obeys **Central Limit Theorem**. Performing thousands of simulations on a sample size of 40 - Sample mean, sample variance are measured and compared against known population mean, variance. The sample mean and variance are also checked to see if they fit a normal distrubition.

Goals

Using simulation we try to verify following

1. Distribution of sample mean is centered around population mean (Part 1)
2. Variance of sample mean is equal to variance of population mean/sample size (Part 1)
3. Distribution of sample variance is centered around population variance (Part 2)
4. Distribution of sample mean is normal even though population distribution is exponential (Part 3)

Simulation

First step of the project is to obtain simulated data. We simulate one thousand samples each of size 40 from a exponential distribution with a known lambda of **0.2**. Population mean of such a exponential distribution is 1/lambda (5) and population variance is also 1/lambda (5). For sake of reproducibility, it is important to set seed.

```
nsim <- 1000
sampleSize <- 40
lambda <- 0.2

set.seed(17)

simResults <- matrix(rexp(sampleSize*nsim, lambda), sampleSize, nsim)
dim(simResults)
```

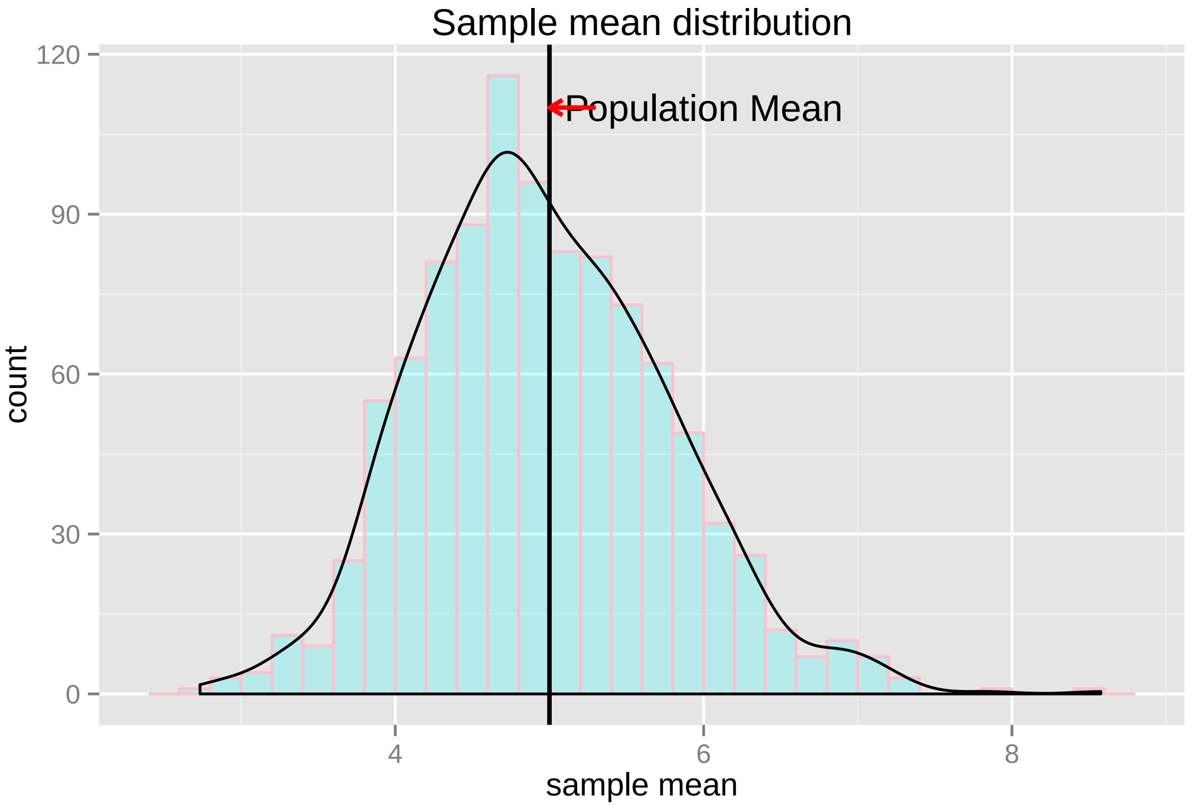
```
## [1]   40 1000
```

Variable *simResults* has the result of simulation. Each column represents a sample. Hence there are 40 rows and 1000 columns.

Sample mean Vs Population mean (Part 1)

We know that the population mean is 1/lambda = 5. For each of the 1000 simulations, sample mean is calculated and a histogram for sample mean is overlayed with the population mean.

```
sampleMeans <- apply(simResults, 2, mean)
library(ggplot2)
library(grid)
plot <- ggplot(data=NULL, aes(x=sampleMeans))
plot <- plot + geom_histogram(binwidth = 0.2,
                      fill=I("cyan"), col=I("pink"),
                      alpha=I(.2))
plot <- plot + labs(title = 'Sample mean distribution',
              x = 'sample mean', y = 'count')
plot <- plot + geom_vline(x=5, size = 0.8) +
      annotate("text", label="Population Mean", y = 110, x = 6) +
      annotate('segment', x=5.3, y=110, xend=5, yend=110, size=0.8, col='red',
      arrow=arrow(length=unit(.2, 'cm')))
plot + geom_density(aes(y=..density..*(1000*0.2)))
```



From the graph, it can be observed sample mean is centered around the population mean and follows a distribution that can be approximated to a normal distribution.

```
mean(sampleMeans)
```

```
## [1] 4.961683
```

```
1/lambda
```

```
## [1] 5
```

It can be seen from the simulation that the mean of sample means is very close to the population mean.

```
var(sampleMeans)
```

```
## [1] 0.6408254
```

```
(1/lambda)^2/sampleSize
```

```
## [1] 0.625
```

Also the variance of sample mean is very close to population variance/sample size.

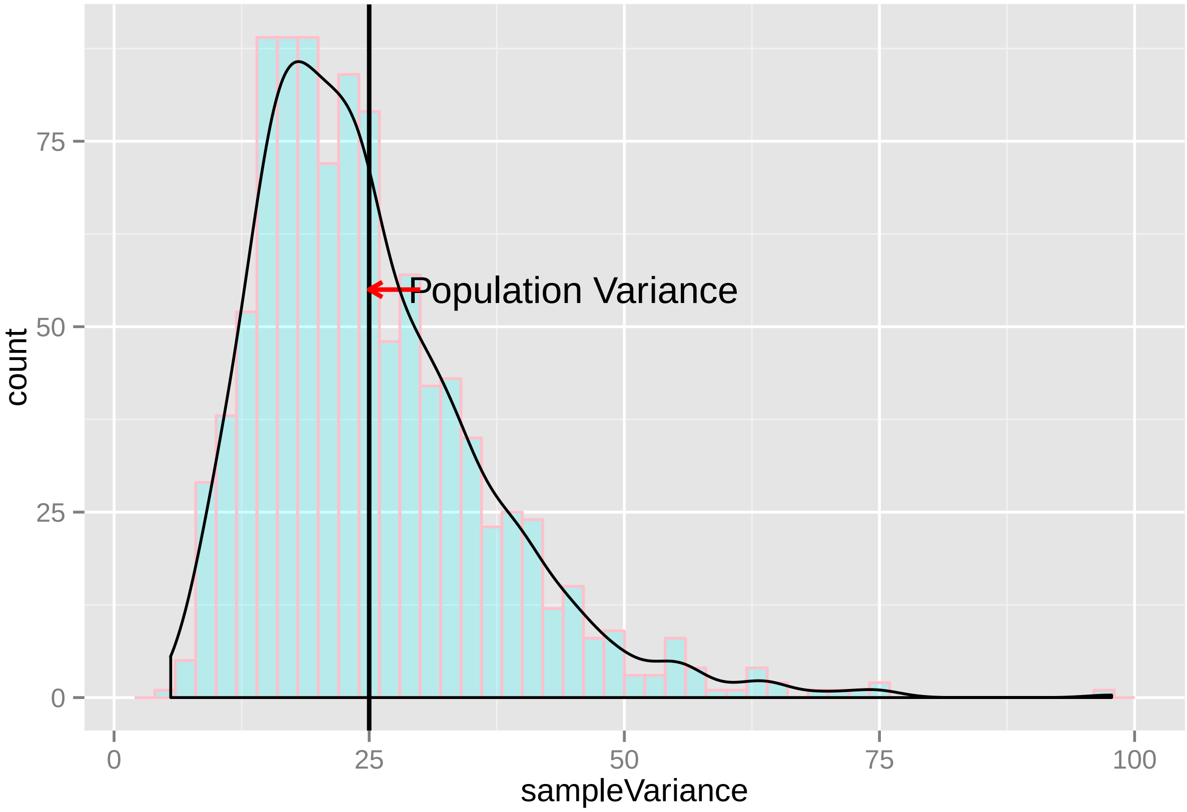
Sample variance Vs Population variance (Part 2)

We know that the population variance is 1/lambda = 5. For each of the 1000 simulations, sample variance is calculated and a histogram for sample variance is overlayed with the population variance.

```
sampleVariance <- apply(simResults, 2, var)
library(ggplot2)
library(grid)
plot <- ggplot(data = NULL, aes(x=sampleVariance))
plot <- plot + geom_histogram(binwidth = 2,  main = "Sample Variance Distribution",
              xlab = "Sample Variance",
              fill=I("cyan"), col=I("pink"),
              alpha=I(.2))

plot <- plot + geom_vline(x=25, size = 0.8) +
      annotate('text', label='Population Variance', y = 55, x = 45) +
      annotate('segment', x=30, y=55, xend=25, yend=55, size=0.8, col='red',
      arrow=arrow(length=unit(.2, 'cm')))

plot + geom_density(aes(y=..density..*(1000*2)))
```



From the graph, it can be observed sample variance is centered around the population variance and follows a distribution that can be approximated to a normal distribution.

```
mean(sampleVariance)
```

```
## [1] 25.04348
```

```
(1/lambda)^2
```

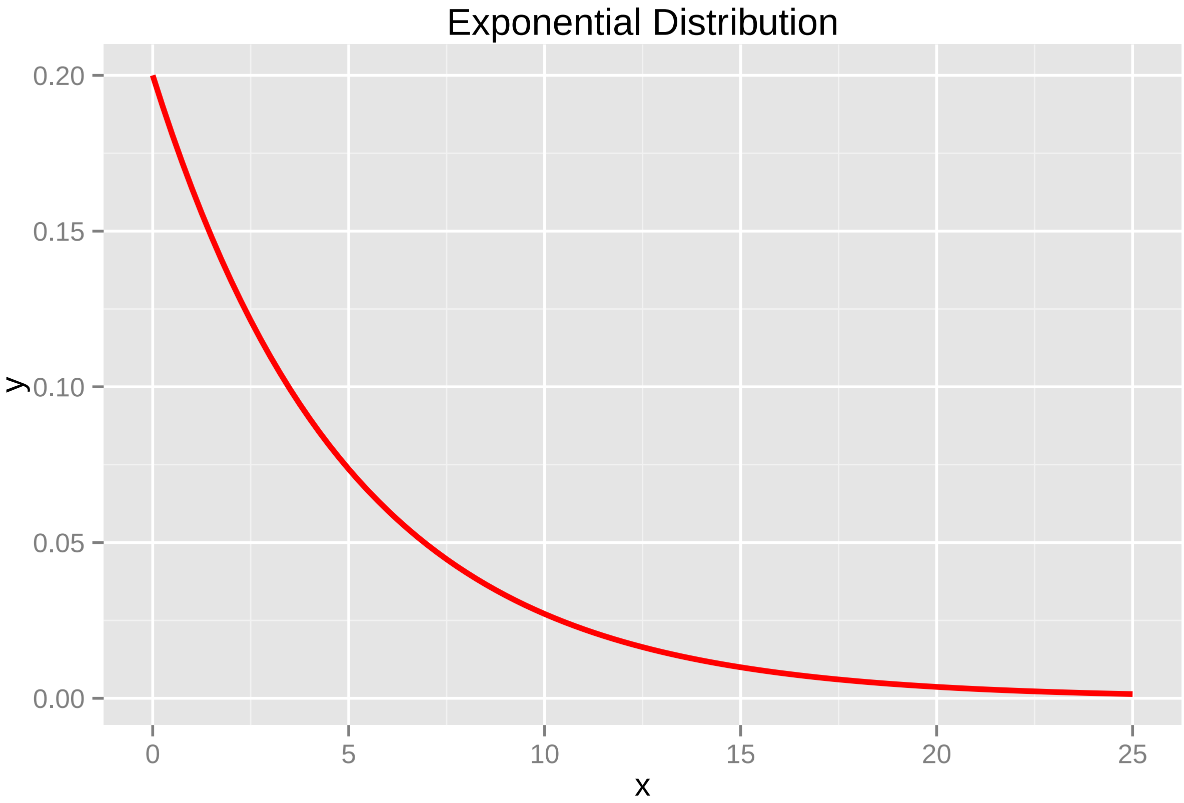
```
## [1] 25
```

It can also be noted that sample variance from simulation is centered very close to population variance.

Population Distribution vs Sample Mean Distribution (Part 3)

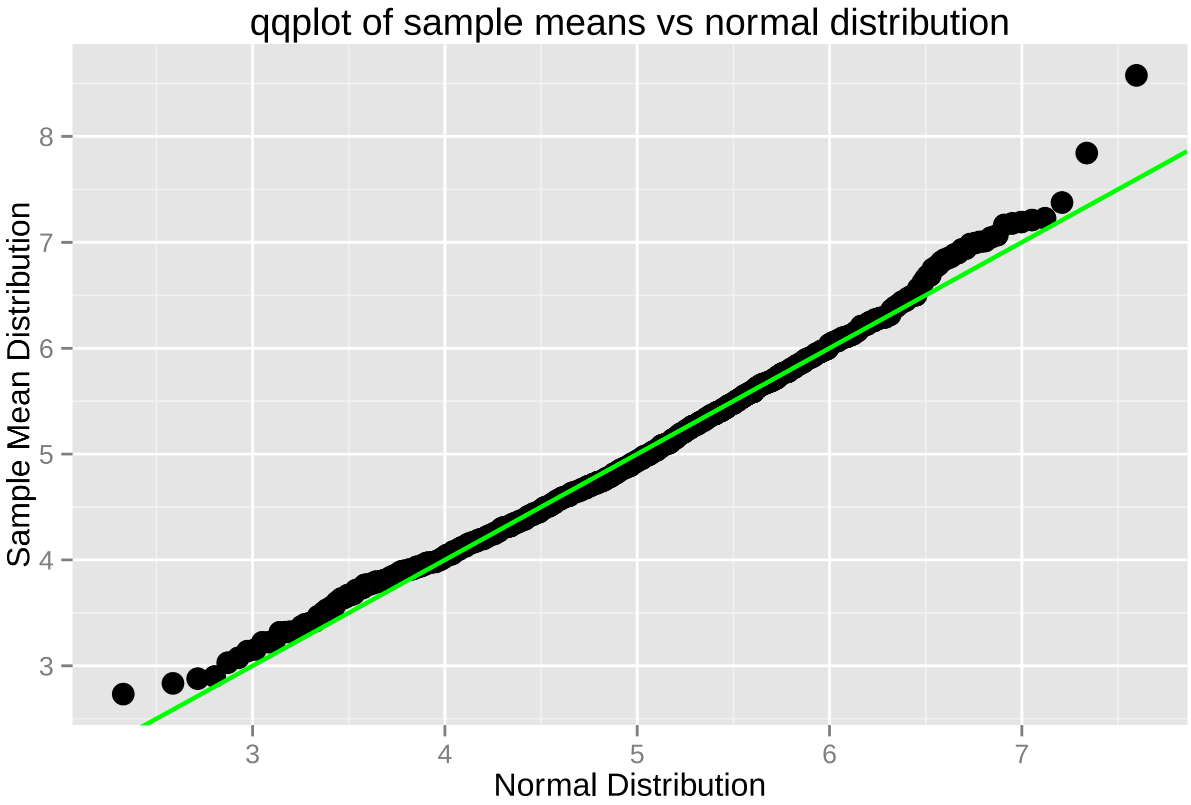
There is no need to use simulated data to visualize population distribution. That would only be an approximation. Instead, we could use the lambda value and plot the true population distribution.

```
library(ggplot2)
x <- seq(0,25,0.05)
plot <- ggplot(data=NULL, aes(x=x))
plot <- plot + stat_function(fun=dexp, args=list(rate=lambda), col = 'red', size = 1)
plot + labs(title='Exponential Distribution')
```



It can be observed that the population distribution is very different from normal distribution.

```
library(ggplot2)
plot <- ggplot(data=NULL, aes(sample = sampleMeans))
plot <- plot + stat_qq(distribution = qnorm, dparams = list(mean=mean(sampleMeans), sd = sd(sampleMeans
plot + geom_abline(slope=1, size = .8, col = 'green') +
  labs(title='qqplot of sample means vs normal distribution', x='Normal Distribution',y='Sample Mean Di
```



In above QQ plot for sample mean vs normal distribution, it can be observed that the sample means are very close to x=y line indicating that the sample mean distribution closely follows a normal distribution.