A Simulation Exercise

7/1/2020

SYNOPSIS

This is the project for statistical inferential class. The project consists of two segments of code chunks-

1. A Simulation Exercise
2. Basic Inferential Data Analysis

The project aims to compare exponential distribution with the central limit theorem and also generates the normal distribution curve for these random exponential values for sample size = 40 and lambda = 0.2.

Simulation

In order to produce reproducible random values , make use of [set.seed] function.

```
set.seed(8)
```

The exponential distribution is simulated in R with rexp(n,lambda) for n=40,lambda = 0.2 and the number of simulation is 1000 times. The 1000 sample means are stored in means. View the first 6 of the means dataset.

```
n <- 40
lambda <- 0.2
nosim <- 1000
set.seed(8)
means = 0
for (i in 1 : 1000) {
  means = c(means, mean(rexp(n,lambda)))
}
head(means)
```

```
## [1] 0.000000 4.442077 4.377342 6.612987 5.233305 5.038121
```

Compare theoretical mean of the exponential distribution with the actual mean

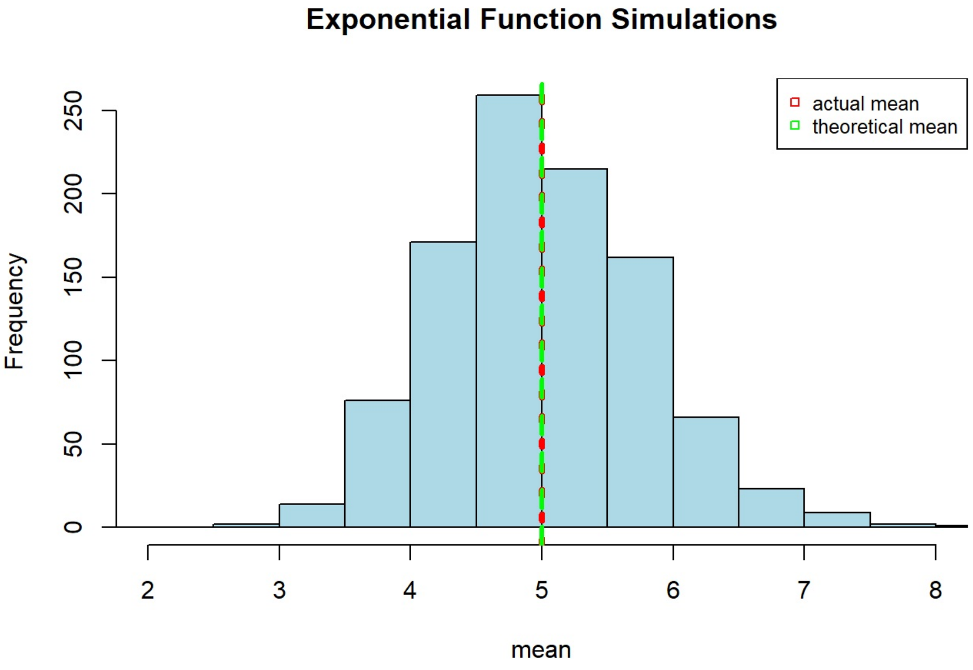
```
## calculate theoretical and actual mean of the distribution
theromean <- round(1/lambda)
theromean
```

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

```
actualMean <- round(mean(means),3)
actualMean
```

```
## [1] 5.001
```

```
## Hence the theoretical mean is 5 and actual mean is 5.001.
## Make a histogram to demonstrate the comparison between actual and theoretical mean
hist(means, xlab = "mean", main = "Exponential Function Simulations", xlim = c(2,8), breaks = 14, col = "lightblue
")
abline(v = actualMean, col = "red", lty = 3, lwd = 4)
abline(v = theromean, col = "green", lty = 2, lwd = 3)
legend("topright", legend = c("actual mean", "theoretical mean"), col = c("red", "green"), pch = 22, cex = 0.8)
```



The green dashed vertical line indicate the theoretical sample mean, which is 1/lambda=5. The red dashed vertical line is the actual sample (size of 40) mean from 1000 samples. The two means are very close.

Compare the actual variance with the theoretical variance of the distribution

```
therovar <- round((1/lambda)^2/n,3)
therovar
```

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

```
actualVar <- round(var(means),3)
actualVar
```

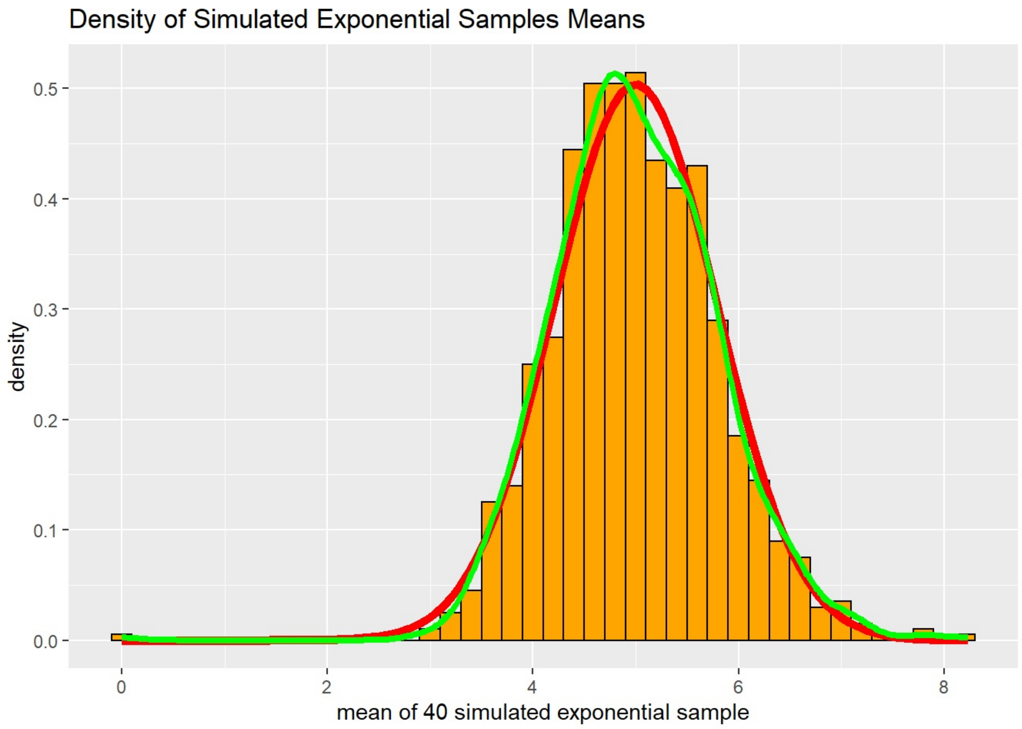
```
## [1] 0.627
```

Hence the theoretical variance is 0.625,actual variance is 0.627.

Show that the distribution is normal

Make a histogram with the density and sample means. Add density curve of the normal distribution and the sample distribution:

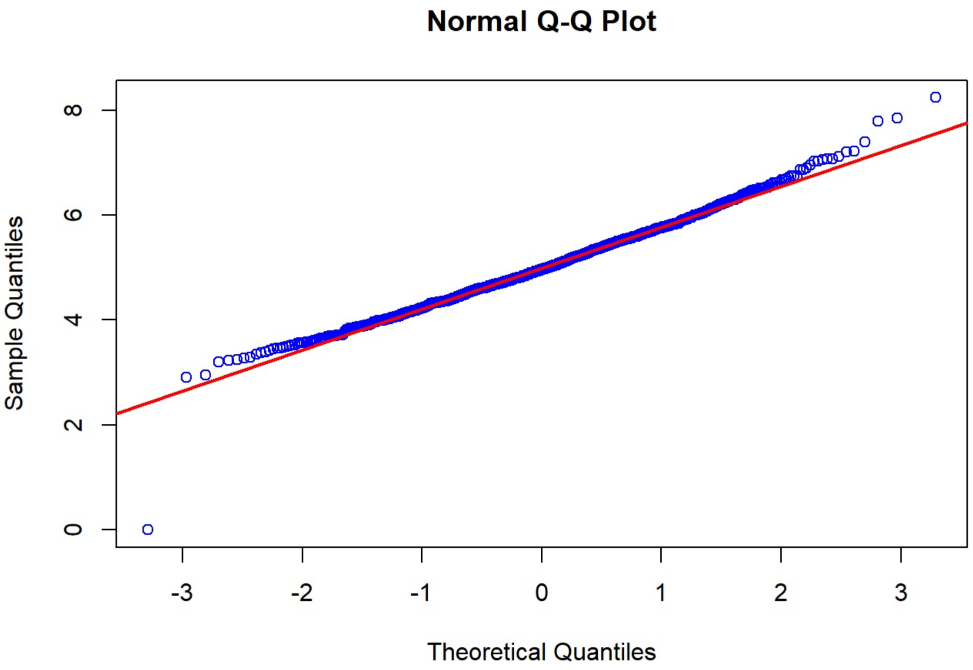
```
library(ggplot2)
sDf <- as.data.frame(means)
ggplot(sDf, aes(x=means)) + geom_histogram(binwidth = .2, color="black", fill="orange" , aes(y=..density..))+
  stat_function(fun=dnorm, args=list(mean=theromean, sd=sd(means)),
                color="red", size =2, geom = "line") +
  stat_density(geom = "line", color = "green", size =1.5)  +
  labs(x="mean of 40 simulated exponential sample", y= "density",
       title="Density of Simulated Exponential Samples Means")
```



The above plot shows the density curve (green curve) is very similar as the normal distribution curve (red curve). It indicates that the distribution is approximately normal.

Q-Q Normal plot to check the normal distribution of data.

```
qqnorm(means, col = "blue")
qqline(means, col = "red", lwd = 2)
```



If the data is normally distributed, the points in the QQ-normal plot lie on a straight diagonal line (red line, illustrated with R code qqline). The deviations from the straight line are minimal. This indicates the sample mean is normal distribution, although the initial sample are not normal distributed.

Confidence Interval for the given random exponential distribution

```
round((actualMean + c(-1,1)*qnorm(.975)*sd(means)/sqrt(n)),3)
```

```
## [1] 4.764 5.256
```

The 95% confidence interval of the sample mean is 4.764 to 5.256, which is very close to the theoretical mean of 5.000. It means there is 95% probability that the sample may contain population mean which lies between these two bounds.