Exponential Distributions: Testing Central Limit Theorem

Janamejaya Chowdhary

Synopsis

For the exponential distribution, using simulations, it is demonstrated that the distribution of sample means and sample variances are in good agreement with the theoretical predictions based on the Central Limit Theorem.

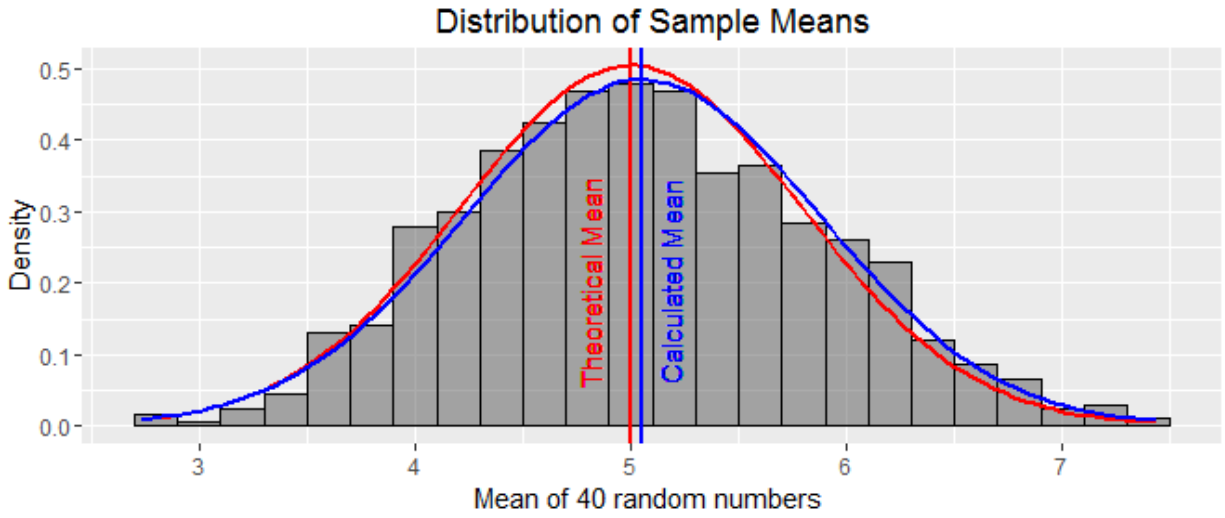
Simulating the Exponential Distribution

The 1000 simulations, each of drawning 40 random numbers drawn from an exponential distributions are set up as follows:

```
set.seed(10252016)     # Set up seed for reproducible sampling from distributions
lambda<-0.2            # Set the value of lambda for the exponential distribution
nrn <- 40              # Set the number of random numbers per simulation
nsim <- 1000           # Set the number of simulations
data <- as.data.frame(matrix(rexp(nsim*nrn, lambda), nrow=nsim, ncol=nrn)) # Generate data
```

Analysis of Sample Mean

The simulated data generated above (data) is analyzed and the distribution of sample means is presented below. R codes are in Appendix 1 and 2.



For this distribution, the actual mean (blue vertical line in Figure) and variance are

```r
print(c(mean_actual, var_actual))
```

```
## [1] 5.0552664 0.6764016
```

and the theoretical mean ($\frac{1}{\lambda}$), red vertical line in Figure) and variance ($\frac{1}{n\lambda^2}$) are

```r
print(c(mean_theo, var_theo))
```

```
## [1] 5.000 0.625
```

The center of the distribution is close to the population mean and the sample variance is close to the population variance.

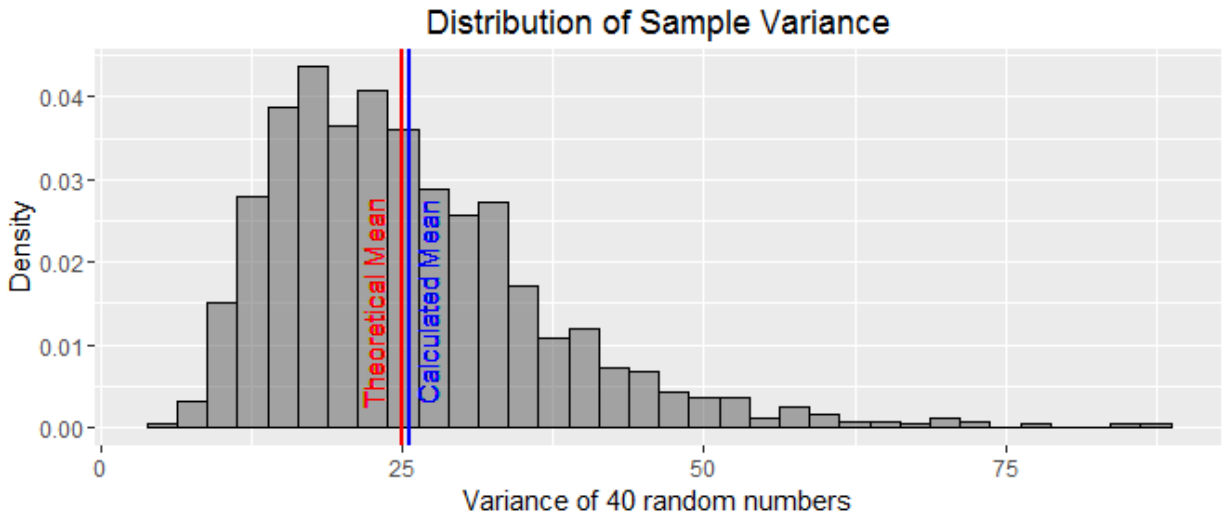
Analysis of Sample Variance

Here the sample variance for each of the 1000 sets of samples is calculated and presented below. R codes are included in Appendices 3 and 4, respectively.

```r
var_list <- as.data.frame(apply(data, 1, var))
colnames(var_list)<-c("SD")

# The theoretical variance is assigned to var_theo
var_theo <- (1/lambda)**2

# The actual variance is calculated as var_actual
var_actual <- mean(var_list$SD)
```



The theoretical (red vertical line in Figure) and actual (blue vertical line in Figure) variance values are

```r
print(c(var_theo, var_actual))
```

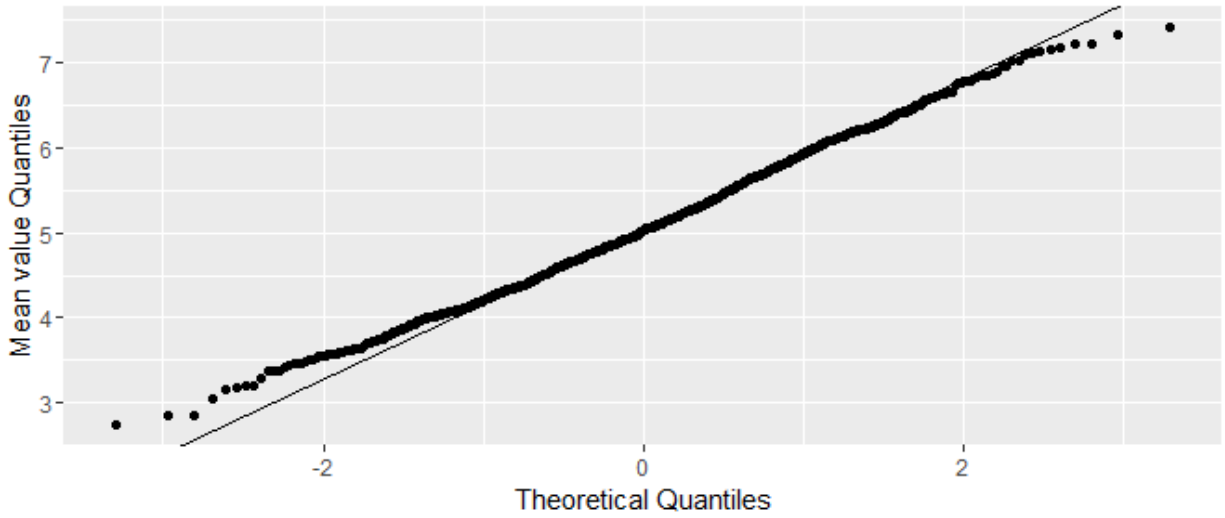
```
## [1] 25.00000 25.52441
```

Clearly, the mean value of variance is in good agreement with the theoretically expected value. However, note that the peak of the distribution appears to be shifted to a value smaller than the mean variance. It would appear that there is more variability in variance due to sampling, i.e., this is a Monte Carlo error.

Central Limit Theorem

Here, the objective is to verify if the distribution of sample means is Normal. One way to verify this is by visual comparison of the distribution of sample means with the distribution for the population. Normal distributions with the theoretical mean and variance (red curve), as well as sample mean and variance(blue curve) are plotted with the distribution of sample means above. The two normal distributions describe the overall shape of the distribution of sample means well.

A second way of testing for normality is based on the Quantile-Quantile plot for the sample mean data. A straight line with slope 1 implies a Gaussian distribution. The QQ-plot is shown below and is mostly linear except at small and large quantiles where sampling errors lead to non-linearity.



A third way is by testing the hypothesis that the sample means are normally distributed ? Since this is a course on Statistical Inference, hypothesis testing will be used here. The objective is to check if the distribution of sample mean values has a normal distribution with mean 1/lambda. So, the hypothesis to be tested is formulated as

H_0: $\mu = \frac{1}{lambda}$ (Null Hypothesis)

H_a: $\mu \neq \frac{1}{lambda}$ (Alternative Hypothesis)

To test this hypothesis, using the actual sample mean and standard deviation, the 95% confidence interval is

```
conf_int <- mean_actual + c(-1,1)*qnorm(0.975)*sqrt(var_actual)/sqrt(nrn)
conf_int
```

```
## [1] 3.489612 6.620921
```

Since the confidence interval (conf_int above) contains $\frac{1}{\lambda} = 5$, we fail to reject the null hypothesis. With probability 0.95, the distribution of sample mean values is Normal in agreement with the Central Limit Theorem.

Appendix

1. R code to calculate the mean values from 1000 simulations

```
mean_list <- as.data.frame(apply(data, 1, mean)) # Mean value of each sample of 40
colnames(mean_list)<-c("Means")
mean_theo <- 1/lambda                # The theoretical mean is assigned to mean_theo
mean_actual <- mean(mean_list$Means) # Actual mean of sample means
var_theo <- (1/lambda^2)/nrn         # The theoretical variance is assigned to var_theo
var_actual <- var(mean_list$Means)   # The actual variance of mean_list
```

2. R code to plot the distribution of sample means

```
library(ggplot2)
g1 <- ggplot(mean_list, aes(x=Means))
g1 <- g1 + geom_histogram(aes(y=..density..),binwidth=0.2, alpha=0.5, col="black")
g1 <- g1 + geom_vline(xintercept = mean_theo, colour="red", size=1, show.legend=TRUE)
g1 <- g1 + geom_text(aes(x=mean_theo-0.15, label="Theoretical Mean", y=0.2), colour="red",
                     angle=90, vjust = 0.1, size=4)
g1 <- g1 + stat_function(fun = dnorm, color="red", size=1.0, args = list(mean = mean_theo,
                                                                         sd = sqrt(var_theo)))
g1 <- g1 + geom_vline(xintercept = mean_actual, colour="blue", size=1, show.legend=TRUE)
g1 <- g1 + geom_text(aes(x=mean_actual+0.17, label="Calculated Mean", y=0.2), colour="blue",
                     angle=90, vjust = 0.1, size=4)
g1 <- g1 + stat_function(fun = dnorm, color="blue", size=1.0, args = list(mean = mean_actual,
                                                                          sd = sqrt(var_actual)))
g1 <- g1 + ylab("Density")
g1 <- g1 + xlab("Mean of 40 random numbers ")
g1 <- g1 + ggtitle("Distribution of Sample Means")
plot(g1)
```

3. R code to plot the distribution of sample variance

```
g2 <- ggplot(var_list, aes(x=SD))
g2 <- g2 + geom_histogram(aes(y=..density..), binwidth=2.5, alpha=0.5, col="black")

g2 <- g2 + geom_vline(xintercept = var_theo, colour="red", size=1, show.legend=TRUE)
g2 <- g2 + geom_text(aes(x=var_theo-1.7, label="Theoretical Mean", y=0.015), colour="red",
                     angle=90, vjust = 0.02, size=4)
g2 <- g2 + geom_vline(xintercept = var_actual, colour="blue", size=1, show.legend=TRUE)
g2 <- g2 + geom_text(aes(x=var_actual+2.3, label="Calculated Mean", y=0.015), colour="blue",
                     angle=90, vjust = 0.02, size=4)
g2 <- g2 + ylab("Density")
g2 <- g2 + xlab("Variance of 40 random numbers")
g2 <- g2 + ggtitle("Distribution of Sample Variance")
plot(g2)
```

4. R code to plot the Quantile-Quantile plot for the distribution of sample means

```
# Find the 1st and 3'rd quantiles
y <- quantile(mean_list$Means, c(0.25, 0.75))
x <- qnorm(c(0.25, 0.75))
slope <- diff(y)/diff(x)
int   <- y[1] - slope * x[1]
g3 <- ggplot(mean_list, aes(sample=Means))
g3 <- g3 + stat_qq()
g3 <- g3 + geom_abline(intercept=int, slope=slope)
g3 <- g3 + xlab("Theoretical Quantiles")
g3 <- g3 + ylab("Mean value Quantiles")
plot(g3)
```