Statistical Inference : Report 1

Mohd Azil Noor

Sunday, 22 November 2015

Statistical Inference - Exponential Distribution

Introduction

Mutliple simulation was done for the exponential distribtution in this report. The results analyzed to visualized the distribution of the averages of the exponential distribution.

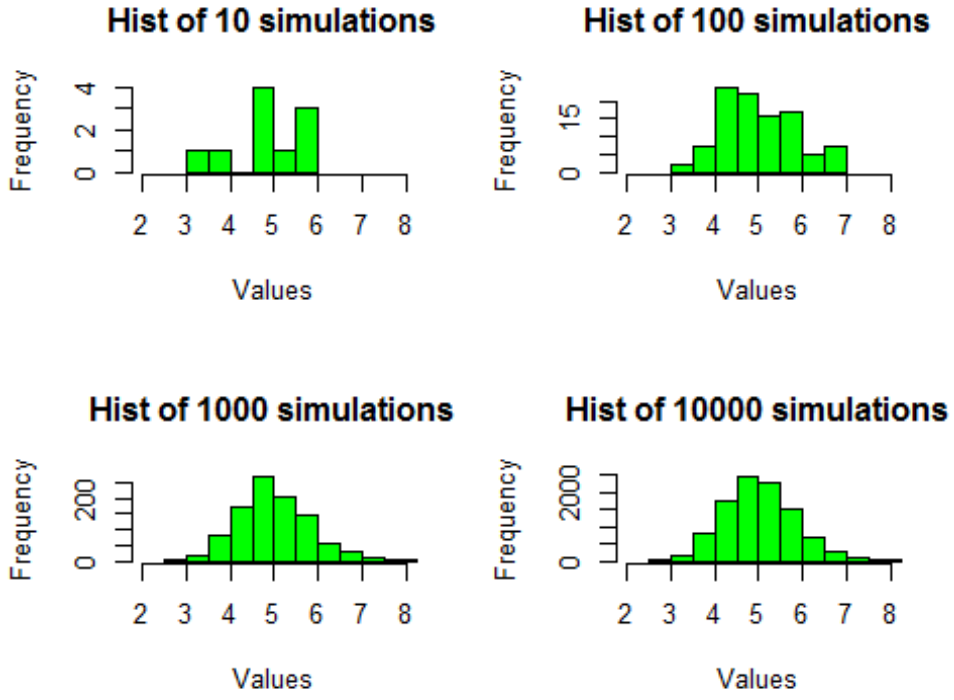
Method

```
library(knitr)
library(plotrix)

opts_chunk$set(eval = TRUE)
opts_chunk$set(fig.height = 4)
```

Each simulation using a fixed parameter of lambda = 0.2 and was used to produce n = 40 random variables via the rexp function in R. The resulting mean and standard deviation of the 40 values were then calculated. The simulations were run 10 times with each mean and standard deviation recorded. These points were plotted on a histogram to show the general distribution.

```
# set the random seed
set.seed(1230)

# set the fixed parameters
lambda <- .2; n <- 40

# prepare the device for a 2x2 plot to show distribution of 10, 100, 1000 and
10000.
par(mfrow = c(2,2))

# generate the random variables for n-values
for (no_sim in c(10, 100, 1000, 10000)){

  # clear the vectors
  mean_values <- NULL; mean_sds <- NULL

  for (i in 1:no_sim){
    # calculate the mean & sd of all the sample means
```

```
    values <- rexp(n, lambda)
    means <- mean(values); sds <- sd(values)
    mean_values  <- c(mean_values, means); mean_sds <- c(mean_sds, sds)
  }

  myhist <- hist(mean_values , freq = TRUE, xlim = c(2, 8),
                 main = paste("Hist of", no_sim, "simulations"), xlab =
"Values", col="green")
}
```



Results

The expected value for an exponential distribution is the inverse of its rate parameter (i.e. $\mathbb{E}[X] = 1/\text{lambda}$) which is equal to 5 in this case. The average value for the n = 10,000 simulation was calculated by taking the mean which worked out to be 5.00.

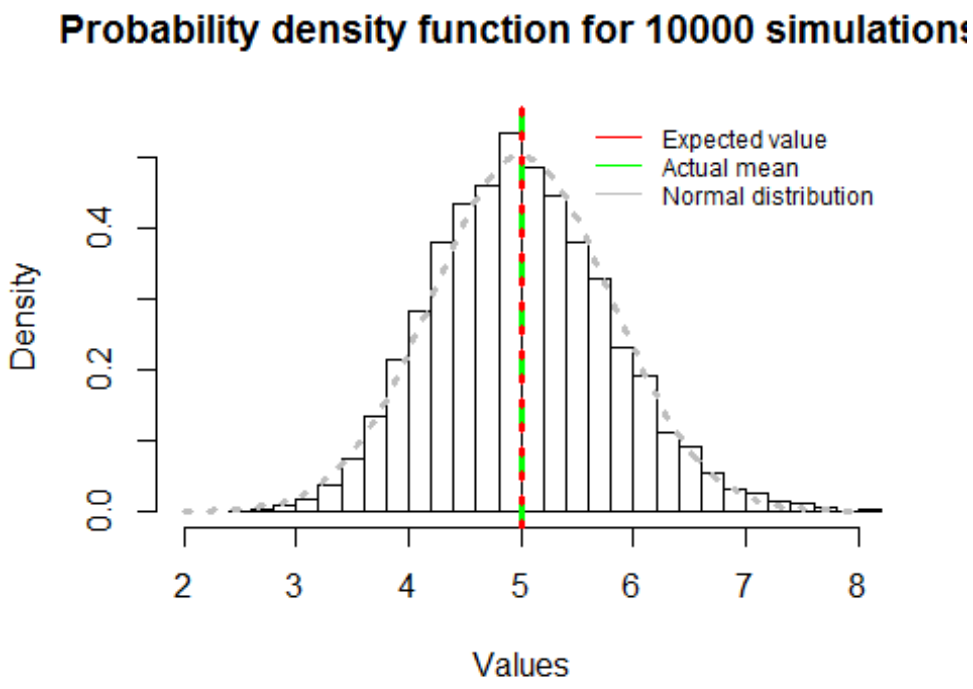
The results concur with the Central Limit Theorem (CLT). The expected value of the sample mean is equal to the population mean it's trying to estimate. The distribution of the sample mean is gausian, centered at 5 and disbured at the center as shown below.

```
mean(mean_values)
```

```
## [1] 5.006232
```

```
# n: 10,000 - histogram of probability density
par(mfrow = c(1,1))
myhist <- hist(mean_values , freq = FALSE, xlim = c(2, 8), ylim = c(0, .55),
```

```
                breaks = 25, main = paste("Probability density function for",
no_sim, "simulations"),
                xlab = "Values")

# calculate the total mean and standard deviation
avg <- mean(mean_values)
s <- sd(mean_values)

# Average value from the data set is plotted.
abline(v = avg , col = "green", lwd = 3, lty = 2)

# Expected valued of exponential distribution is plotted
abline(v = 5, col = "red", lwd = 3, lty = 9)

# plot the theoretical normal distribution for the data set
x <- seq(min(mean_values ), max(mean_values ), length = 100)
y <- dnorm(x, mean = avg, sd = s)
curve(dnorm(x, mean = avg, sd = s),
      col = "gray", lwd = 3, lty = 3, add = TRUE)

legend('topright', c("Expected value", "Actual mean", "Normal distribution"),
       lty=1, col=c('red', 'green', "gray"), bty='n', cex=.75)
```



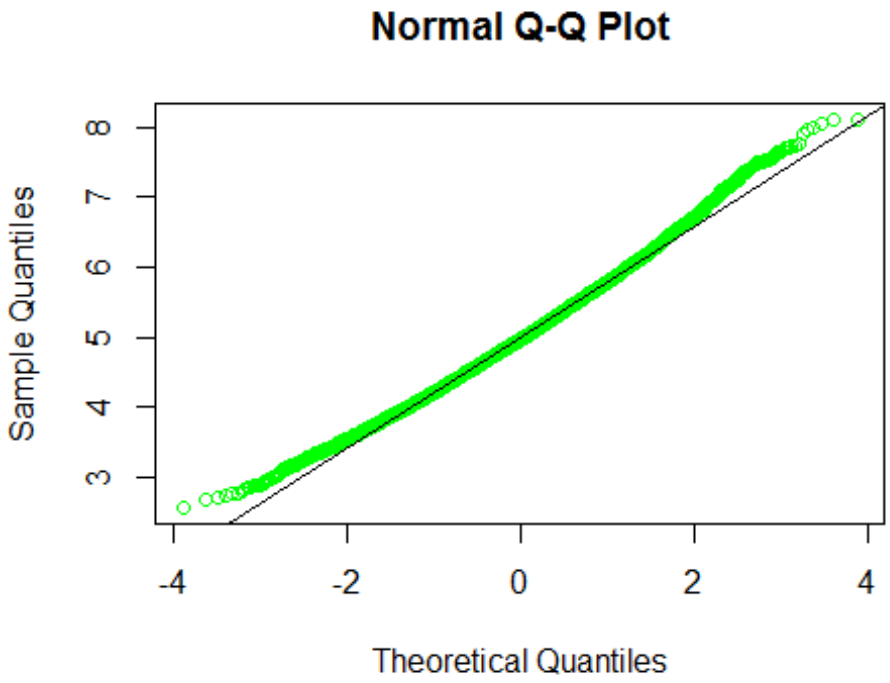
Next, the variance of the sample mean was worked out to be 0.79. This corresponds exactly with the standard error of the mean(i.e. SE = sigma/$\sqrt{n}$) which is equal to 0.79 for the 40 observations.

```
sd(mean_values)
```

```
## [1] 0.7928717
```

A Q-Q plot of the mean values was also plotted below. There is insignificant deviation between the actual quantile values and the theoretical; this indicates aggregated sample distribution is indeed normal.

```
qqnorm(mean_values, col = "green")
qqline(mean_values)
```



The 95% COnfidence Interval (CI) of each simulation was worked out using the interval's own standard deviaton and mean according to the equation $\overline{X} \pm 1.96 sigma/\sqrt{n}$. The coverage was computed as the percent of times the true mean fell within each CI.

```
# construct 95% CI for each simulation
upper <- mean_values +  1.96 * (mean_sds/sqrt(n))
lower <- mean_values -  1.96 * (mean_sds/sqrt(n))
sum(lower < 5 & 5 < upper)/10000 * 100
```

```
## [1] 92.33
```

The simulation for 100 simulations was rerun and the CI for each simulation was plotted for visualization purposes.

```
# rerun for no_sim <- 100
no_sim <- 100
```

```r
mean_values <- NULL; mean_sds <- NULL

for (i in 1:no_sim){
  # calculate the mean & sd of all the sample means
  values <- rexp(n, lambda)
  means <- mean(values); sds <- sd(values)
  mean_values  <- c(mean_values, means); mean_sds <- c(mean_sds, sds)
}
# construct 95% CI for each simulation
upper <- mean_values +  1.96 * (mean_sds/sqrt(n))
lower <- mean_values -  1.96 * (mean_sds/sqrt(n))
sum(lower < 5 & 5 < upper)/10000 * 100
```

```
## [1] 0.95
```

```r
index <- c(1:no_sim)

plot(index, upper, ylim = c(0, 10), type = "n", xlab = "Index", ylab =
"Mean",
     main = "Plot of CI coverage for 100 simulations")

segments(index, upper, index, lower, col = "green", lwd = 3)
#ablineclip(h = 5, col = "red", lwd = 2, lty = 2)
text(-8, 5, expression(paste("", mu, "")), cex = 1.5)
ablineclip(h=5, x1 = -2.5, lty = 2, col="red")
```