Statistical Inference part1

Introduction

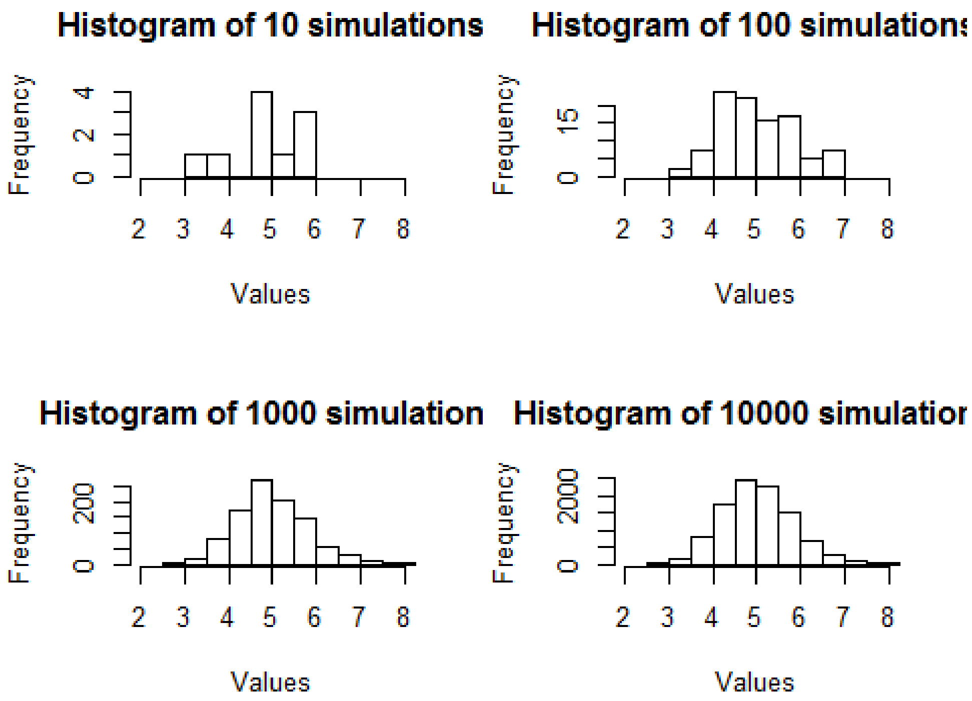
In this report the exponential distribution was simulated multiple times in R. The resulting data was analyzed to investigate the distribution of the averages of the exponential distribution.

Method

```r
library(knitr)
library(plotrix)

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

Each simulation involved a rate parameter $\lambda = 0.2$ used to produce n = 40 random variables via the rexp function in R. The 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. This experiment was then carried out 3 more times such that, in total, results for 10, 100, 1,000, 10,000 simulations were obtained.

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

# set the experiment values
lambda <- .2; n <- 40

# prepare the device for a 2x2 plot
par(mfrow = c(2,2))

# generate the random variables for these 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("Histogram of", no_sim, "simulations"), xlab =
"Values")
}
```

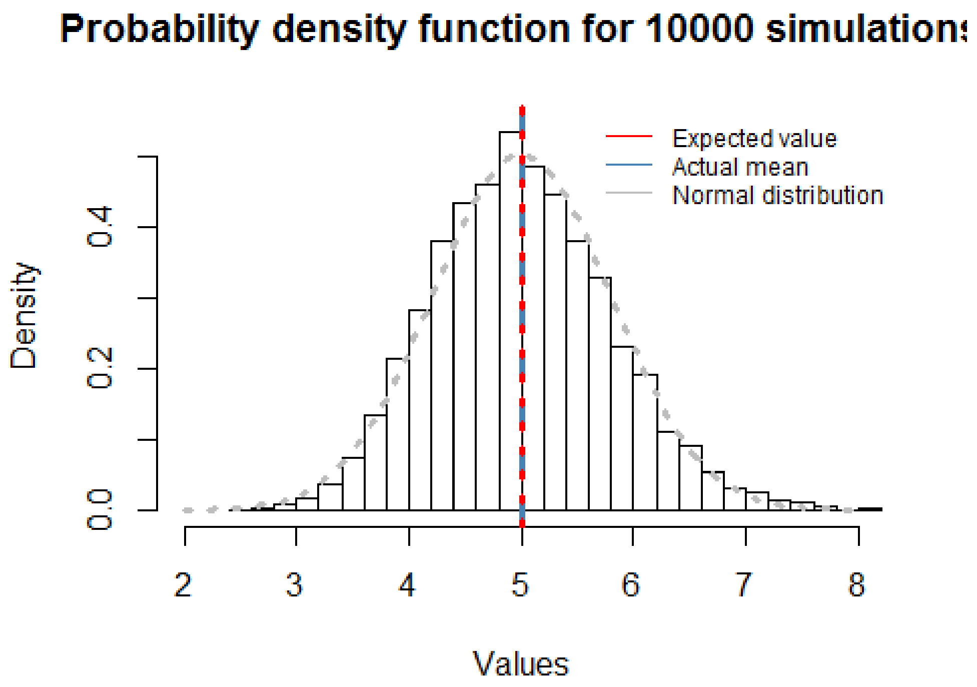


Results

The expected value for an exponential distribution is the inverse of its rate parameter (i.e. $\mathbb{E}[X] = 1/lambda$) 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. As expected, due to the central limit theorem, 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 concentrated at the center as shown below.

```
mean(mean_values)

## [1] 5.006

# 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 of the aggregated samples
avg <- mean(mean_values)
s <- sd(mean_values)
```

```r
# plot the average value from the data set
abline(v = avg , col = "steelblue", lwd = 3, lty = 2)

# plot the expected value of an exponential distribution
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', 'steelblue', 'gray"), bty='n', cex=.75)
```

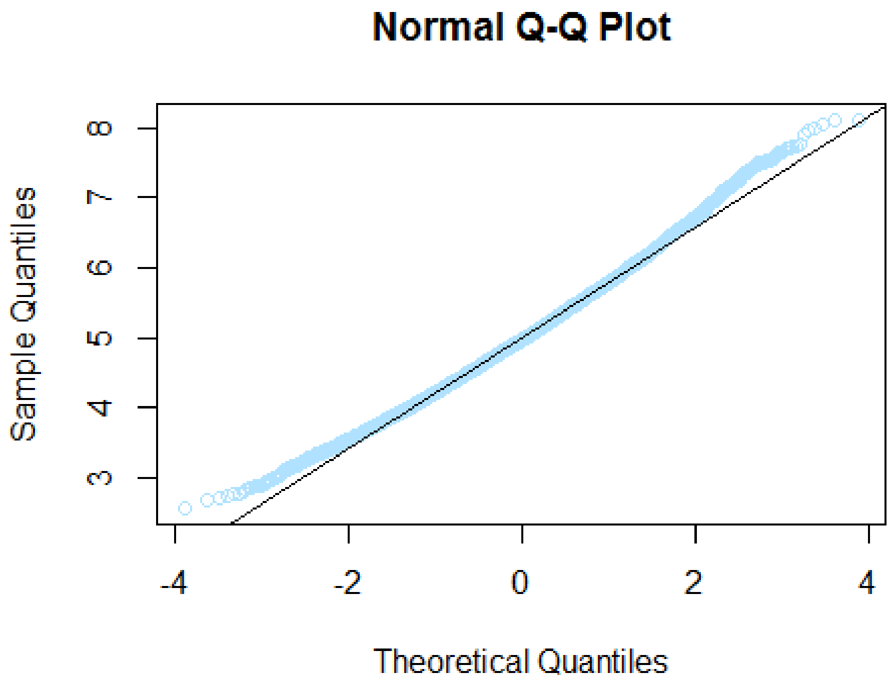


ext, the variance of the sample mean was worked out to be 0.79. This corresponds exactly with the standard error of the mean(i.e. $SE = sigma/\sqrt{n}$) which is equal to 0.79 for the 40 observations.

```r
sd(mean_values)

## [1] 0.7929
```

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

```r
qqnorm(mean_values, col = "lightskyblue1")
qqline(mean_values)
```



The 95% confidence interval 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 interval's confidence interval.

```
 # construct 95% confidence interval 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
```

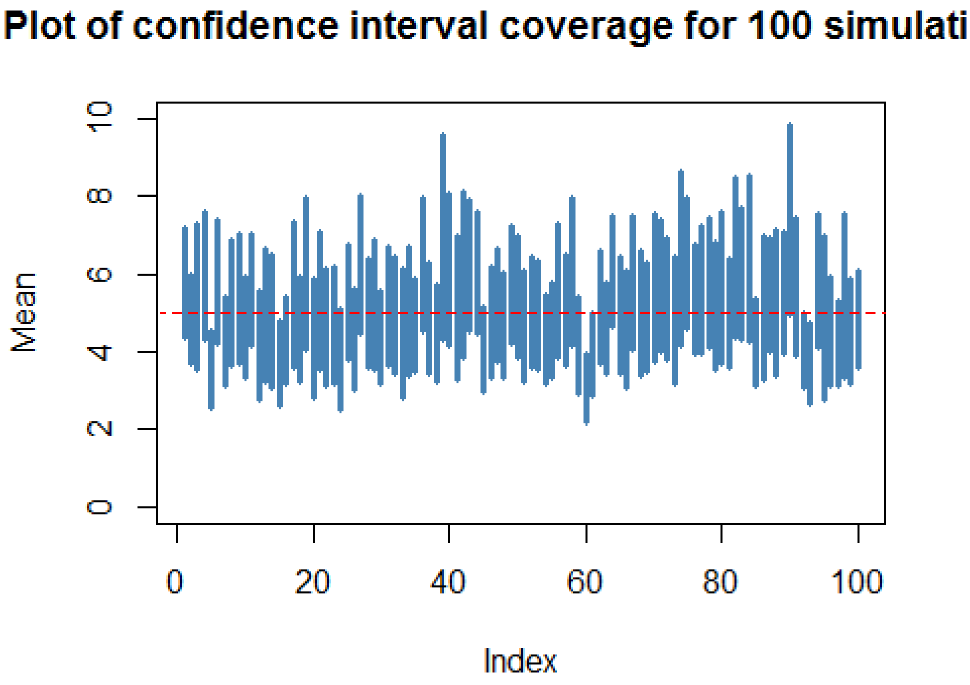
For visualization purposes, the simulation for 100 simulations was rerun and the confidence interval for each simulation was plotted. Here the ~92% coverage is clearly seen.

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

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% confidence interval 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
```

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

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

segments(index, upper, index, lower, col = "steelblue", 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")
```


Plot of confidence interval coverage for 100 simulati
Mean
Index