Statistical Inference-Course Project1 Simulation

English Garden

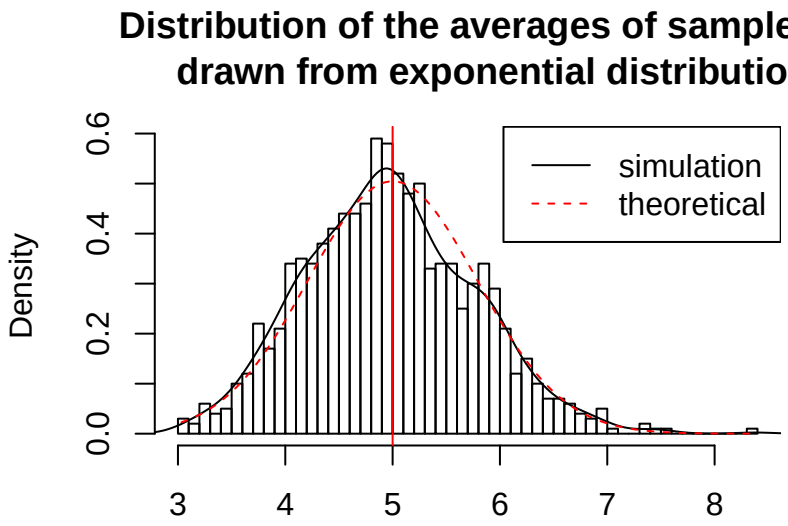
June 19, 2015

The exponential distribution can be simulated in R with `rexp(n, lambda)` where `lambda` λ is the rate parameter. The mean of exponential distribution is $1/\lambda$ and the standard deviation is also $1/\lambda$. In this project, we set $\lambda = 0.2$ for all the simulations. The goal of this project is to investigate the distribution of averages of 40 exponential numbers sampled from exponential distribution with $\lambda = 0.2$, by performing a thousand simulations. So first we obtain a thousand simulated averages of 40 exponentials.

```
set.seed(12345)
lambda <- 0.2
num_sim <- 1000
sample_size <- 40
sim <- matrix(rexp(num_sim*sample_size, rate=lambda), num_sim, sample_size)
row_means <- rowMeans(sim)
```

The distribution of sample means drawn from exponential distribution with $\lambda = 0.2$ is as follows:

```
# plot the histogram of the averages
hist(row_means, breaks=50, prob=TRUE,
     main="Distribution of the averages of samples
     drawn from exponential distribution",
     xlab="")
# density of the averages of samples
lines(density(row_means))
# theoretical center of the distribution
abline(v=1/lambda, col="red")
# theoretical density of the sample means
xfit <- seq(min(row_means), max(row_means), length=100)
yfit <- dnorm(xfit, mean=1/lambda, sd=(1/lambda/sqrt(sample_size)))
lines(xfit, yfit, pch=22, col="red", lty=2)
# add legend
legend('topright', c("simulation", "theoretical"), lty=c(1,2), col=c("black", "red"))
```

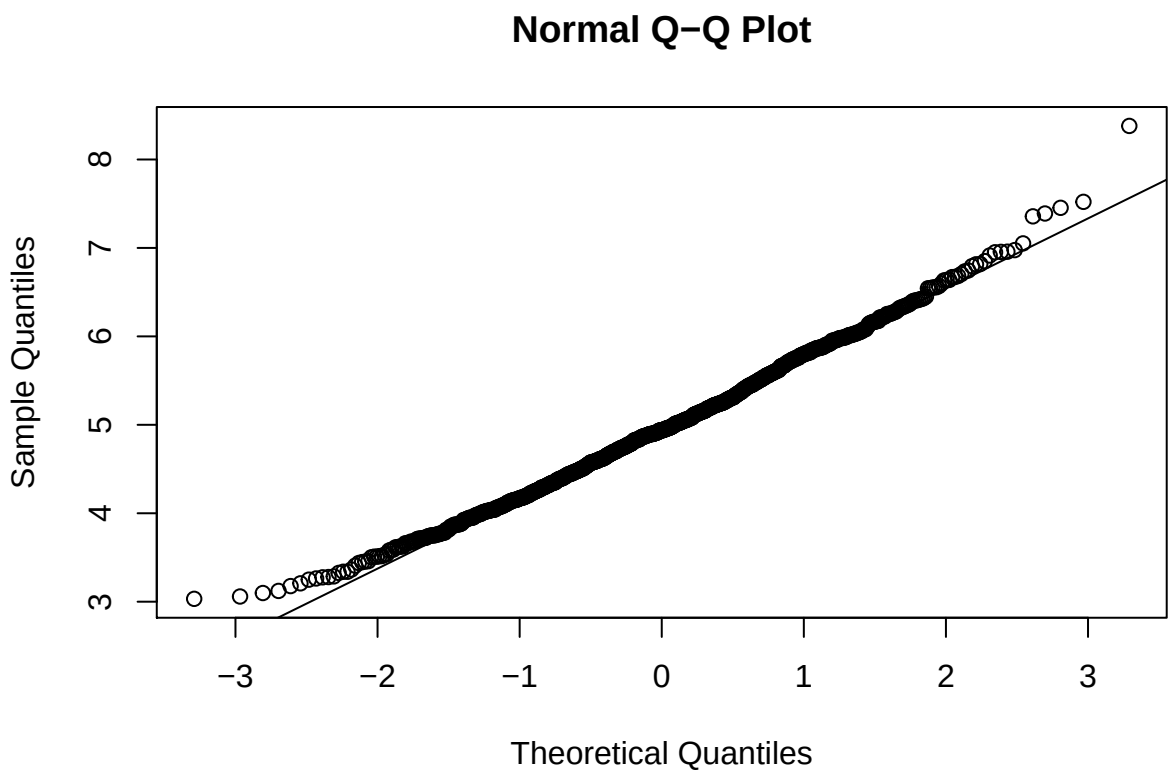


The distribution of sample means is centered at 4.971972. The theoretical center of the distribution is $\lambda^{-1} = 5$. The variance of sample means is 0.6157926. The theoretical variance of the distribution is $\sigma^2/n = 1/(\lambda^2 n) = 1/(0.04 \times 40) = 0.625$. We can see that the simulation mean and variance are close to the theoretical mean and variance.

The central limit theorem states that the averages of a sufficiently large number of samples of indepdent random variables, each with a well-defined expected value and well-defined variance, follow normal distribution regardless of the underlying distribution. The figure plotted above shows the sample density computed using the histogram in black color and the normal density plotted with theoretical mean and variance values in red color.

Q-Q plot is a plot of quantile of the first data set against the quantiles of the second data set. It is a graphical technique for determining if two data sets come from populations with a common distribution. A 45-degree reference line is also plotted. If the two sets of data come from a population with the same distribution, the points should fall approximately along this reference line. The further away the points depart from this reference line, the greater the evidence for the conclusion that the two data sets are from populations with different distributions. Here the q-q plot below suggests the normality of the simulation data set.

```
qqnorm(row_means); qqline(row_means)
```

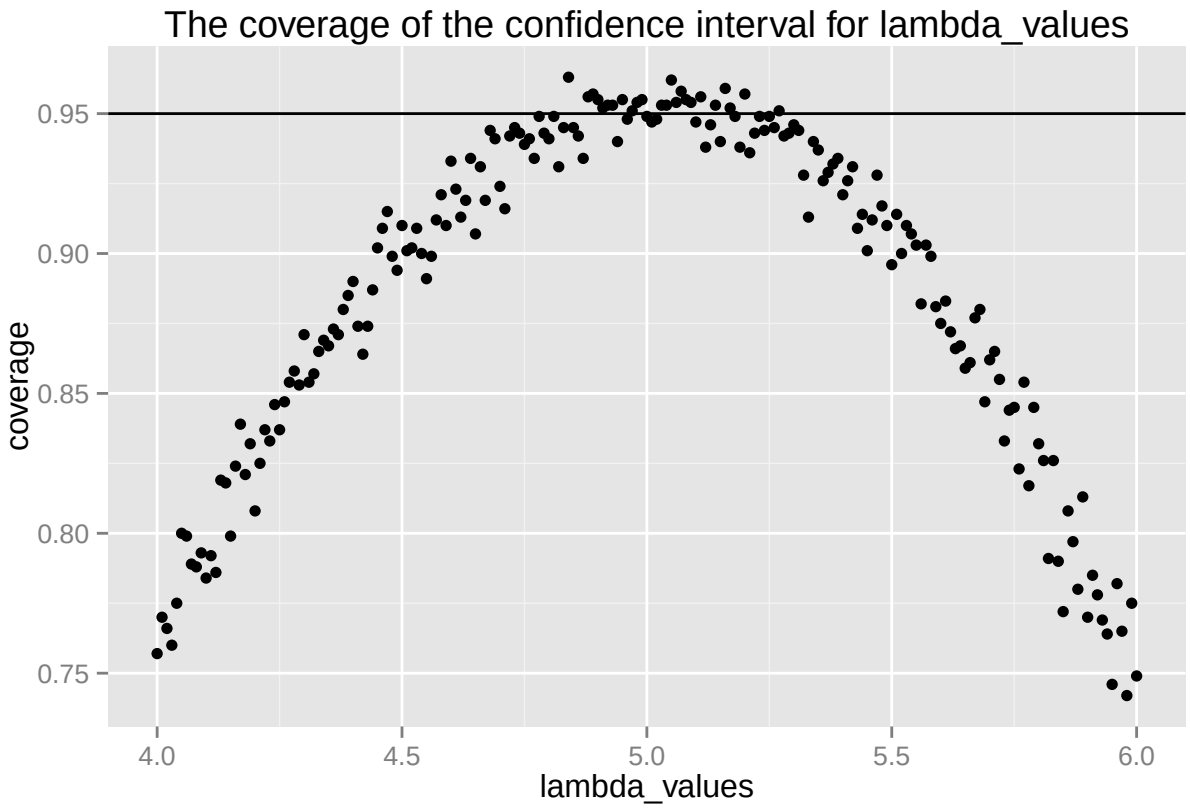


Finally, we evaluate the coverage of the confidence interval for $1/\lambda = \bar{X} \pm 1.96\frac{S}{\sqrt{n}}$

```
lambda_values <- seq(4, 6, by=0.01)
coverage <- sapply(lambda_values, function(x) {
    mu_hats <- rowMeans(matrix(rexp(sample_size*num_sim, rate=0.2),
                               num_sim, sample_size))
    ll <- mu_hats - qnorm(0.975) * sqrt(1/lambda**2/sample_size)
    ul <- mu_hats + qnorm(0.975) * sqrt(1/lambda**2/sample_size)
    mean(ll < x & ul > x)
})
summary(lambda_values)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     4.0     4.5     5.0     5.0     5.5     6.0
```

```
library(ggplot2)
qplot(lambda_values, coverage, main="The coverage of the confidence interval for lambda_values") +
geom_hline(yintercept=0.95)
```



The 95% confidence intervals for the exponential distribution rate parameter λ to be estimated $\hat{\lambda}$ are: $\hat{\lambda}_{low} = \hat{\lambda}(1 - \frac{1.96}{\sqrt{n}})$ and $\hat{\lambda}_{upp} = \hat{\lambda}(1 + \frac{1.96}{\sqrt{n}})$. The figure above shows that for selection of $\hat{\lambda}$ around 5 (the true rate parameter), the average of the sample mean falls within the confidence interval at least 95% of the time.