Statistical Inference course project

Georges Bodiong

Saturday, February 21, 2015

Overview

This is the project for the Coursera Statistical Inference class. We will use simulation to explore inference and do some basic inferential data analysis. There are two parts:

- A simulation exercise
- Basic inferential analysis

Simulations

```
#set lambda to 0.2
lambda = 0.2

# 40 samples
n = 40

# A thousand simulations
sims <- 1000

# Set seed for reproducibility
set.seed(820)

# Simulate
sim_expo <- replicate(sims, rexp(n, lambda))

# Calculate mean of exponentials
means_expo <- apply(sim_expo, 2, mean)

head(means_expo)

## [1] 5.750000 3.808205 4.058154 3.999241 4.312532 4.418246
```

Calculate simulation mean to show where the simulation is centered at and compare it to theoretical center.

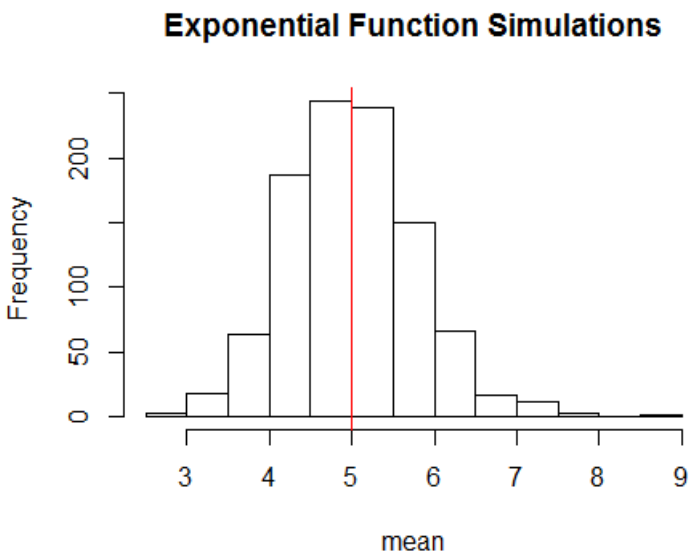
```
dist_mean <- mean(means_expo)
dist_mean

## [1] 4.998812

theo_mean <- 1/lambda
theo_mean
```

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

```
# Let's visualise it...
hist(means_expo, xlab = "mean", main = "Exponential Function Simulations")
abline(v = dist_mean, col = "blue")
abline(v = theo_mean, col = "red")
```



The analytical mean here is 4.9988117 and the theoretical mean 5. The two averages are very close.

Variability

```
# standard deviation of distribution
stand_dev <- sd(means_expo)
stand_dev
```

```
## [1] 0.7909422
```

```
# standard deviation from analytical expression
theo_sd <- (1/lambda)/sqrt(n)
theo_sd
```

```
## [1] 0.7905694
```

```
# variance of distribution
dist_var <- stand_dev^2
dist_var
```

```
## [1] 0.6255895
```

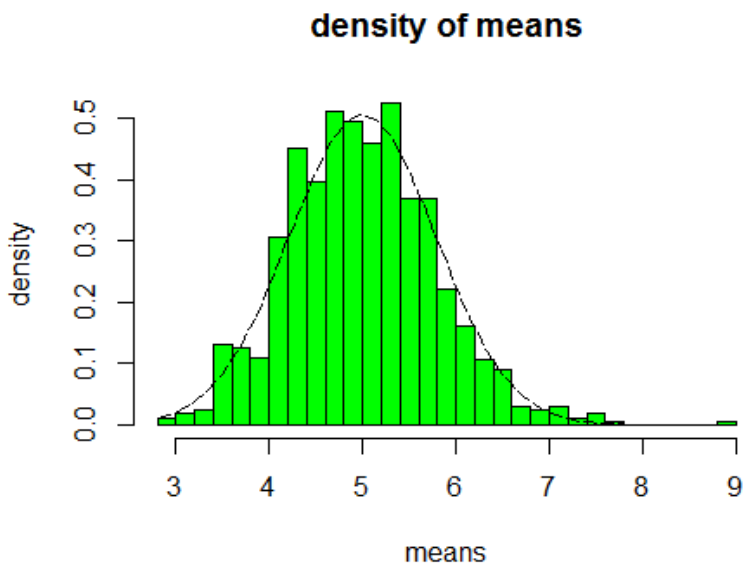
```
#variance from analytical expression
theo_var <- ((1/lambda) * (1/sqrt(n)))^2
theo_var
```

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

The standard deviation of the distribution is 0.7909422 and the theoretical standard deviation is 0.7905694. The theoretical variance is 0.625 while the actual variance of the distribution is 0.6255895.

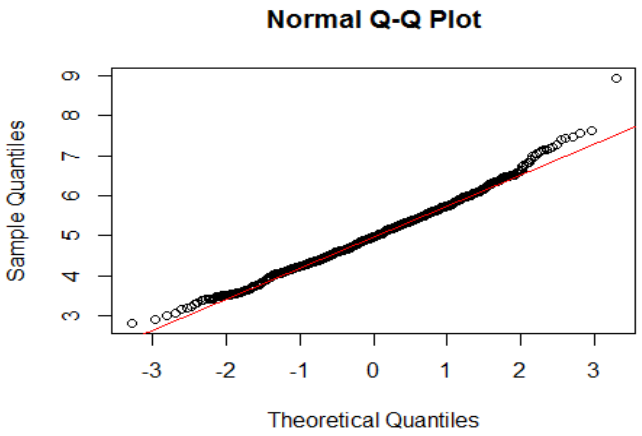
Is the distribution approximately normal?

```
xfit <- seq(min(means_expo), max(means_expo), length=100)
yfit <- dnorm (xfit, mean=1/lambda, sd=(1/lambda/sqrt(n)))
hist(means_expo, breaks=n, prob=T, col="green", xlab="means", main="density of means", ylab="density")
lines(xfit, yfit, pch=22, col="black", lty=5)
```



Let's compare distribution of averages of 40 exponentials to normal distribution

```
qqnorm(means_expo)
qqline(means_expo, col=2)
```



It is obvious that the distribution is close to normal.