Statistical Inference Course Project

Javier Nieto

Friday, January 12, 2015

Part 1: Simulation exercise

Exploring exponential distribution

The exponential distribution can be simulated in R with rexp(n, λ) where λ is the rate parameter. The mean of exponential distribution is $1/\lambda$ and the standard deviation is also $1/\lambda$. In this simulation, we investigate the distribution of averages of 40 numbers sampled from exponential distribution with $\lambda = 0.2$.

Setting up required environment in R

```
# Load libraries
library(ggplot2)
library(knitr)

# Changing locale time to English
Sys.setlocale("LC_TIME", "english")
```

Simulations

Let's do a thousand simulated averages of 40 exponentials:

We generate random numbers from an exponential distribution with $\lambda = 0.20$. For each simulation, we draw 40 samples. We do this 1000 times, taking the average of the values each time.

```
# Lambda value
lambda  <- 0.2

# Number of simulations
num_sim <- 1000

# Sample numbers
n       <- 40

# Simulations
simulaciones <- do.call(rbind, replicate(num_sim, rexp(n, lambda), simplify = F))

# Average simulations
promedios <- data.frame(x = rowMeans(simulaciones))
```

Sample measures versus theoretical measures

To show proximity between sample and theoretical averages we calculate it as following

```r
# Mean of simulation
avg_sim <- mean(promedios$x)

# Mean theoretical
avg_teo <- 1/lambda
```

To show the variability of the simulation results, we solve for both the standard deviaion and variance statistics for both simulated and theoretical data, respectively, and then compare. The code chunk below allows us to do just this.

```r
# Standard deviation of simulation
sd_sim  <- round(sd(promedios$x), 3)

# Variance of simulation
var_sim <- round(sd_sim^2, 3)

# Standard deviation theoretical
sd_teo  <- round((1/lambda)/sqrt(n), 3)

# Variance theoretical
var_teo <- round(sd_teo^2, 3)
```

In the next table we show a summary from computed values previously

```r
# Data frame with simulated and theoretical values
statistics <- data.frame(Mean = c(avg_sim, avg_teo),
                         `Standard deviation` = c(sd_sim, sd_teo),
                         Variance = c(var_sim, var_teo)
                         )

# Rename the row names
row.names(statistics) <- c("Simulated", "Theoretical")

# Show the table
kable(statistics, format = "pandoc", caption = "Sample and theoretical measures", dig=3)
```

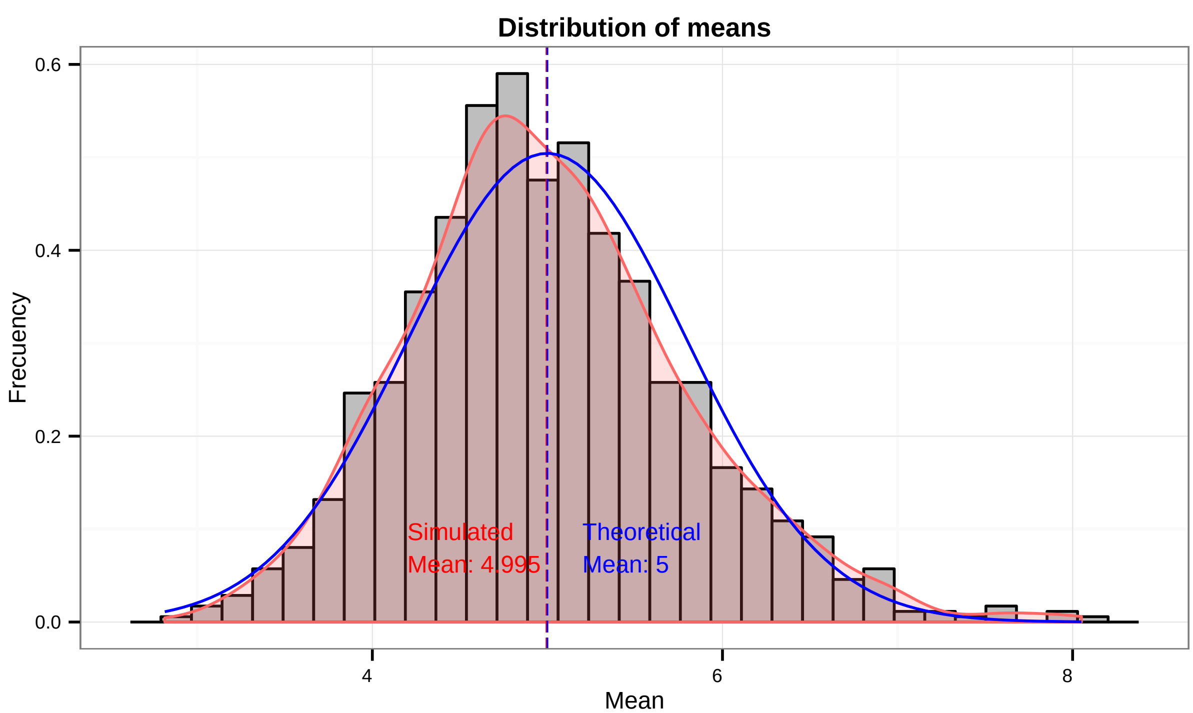
```
Table: Sample and theoretical measures

                 Mean   Standard.deviation   Variance
------------  ------  -------------------  ---------
Simulated      4.995                0.795      0.632
Theoretical    5.000                0.791      0.626
```

The values shown above are very similar as indicated by the central limit theorem.

Data distribution plot

The previous data obtained are shown graphically as follows:

```r
ggplot(promedios, aes(x=x)) +
       geom_histogram(binwidth = diff(range(promedios$x))/30,
                      fill="gray", aes(y=..density..),
                      colour="black") +
       ggtitle("Distribution of means") +
       xlab("Mean") +
       ylab("Frecuency") +
       theme_bw() +
       theme(plot.title   = element_text(lineheight=.8, size = 10, face  = "bold"),
             axis.text.x  = element_text(lineheight=.8, size =  7, hjust = 1),
             axis.text.y  = element_text(lineheight=.8, size =  7),
             axis.title.x = element_text(lineheight=.8, size =  9),
             axis.title.y = element_text(lineheight=.8, size =  9)
             ) +
       geom_density(alpha=.2, fill="#FF6666",color = '#FF6666') +
       geom_vline(xintercept=avg_sim, lwd=.3, col="red", linetype = "longdash") +
       annotate("text", x = 4.2, y = 0.08,
                label = paste("Simulated\nMean:", round(avg_sim,3), " "), cex=3,
                              col="red", hjust=0 ) +
       stat_function(fun = dnorm, arg = list(mean = avg_teo, sd = sd_teo),
                     color="blue") +
       geom_vline(xintercept=avg_teo, lwd=.3, col="blue", linetype = "longdash") +
       annotate("text", x = 5.2, y = 0.08,
                label = paste("Theoretical\nMean:", round(avg_teo,3), " "), cex=3,
                              col="blue", hjust=0 )
```



In the plot we can see that the distribution of averages of 40 exponentials is very close to a normal distribution as indicated by the central limit theorem.