Statistical Inference Project - Part 1

yxy

Saturday, January 17, 2015

Part 1 Simulation Exercise

Overview

The main objective of this project is to investigate the exponential distribution in R and compare it with the Central Limit Theorem.

Simulations

The exponential distribution was simulated in R using **rexp(n, lambda)** where **n** is the number of observations and **lambda** is the exponential rate. A distribution of averages of 40 exponentials with **n** = 1000 and **lambda** (λ) = 0.2 was generated. The R codes to simulate the distribution are as follows:

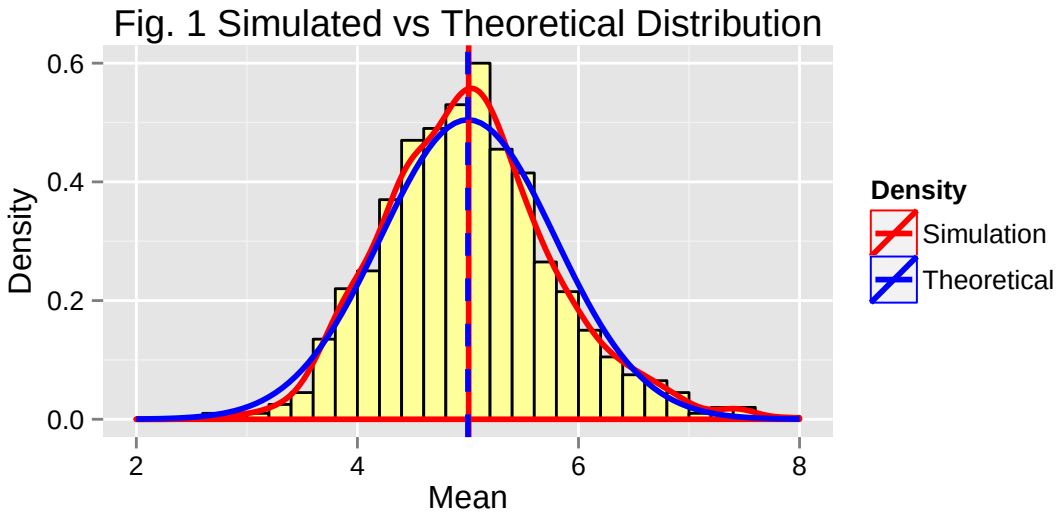
```
set.seed(8)
lambda <- 0.2
nSimulation  <- 1000
nExponential <- 40
# Simulate the exponential distribution
simulation <- matrix(rexp(nSimulation * nExponential, lambda),
                     nSimulation, nExponential)
# Find the distribution (average across the 40 exponentals)
data <- data.frame(mean = apply(simulation, 1, mean))
```

Sample Mean versus Theoretical Mean

The theoretical mean and the standard deviation of the exponential distribution are $\frac{1}{\lambda} = 5.0$ respectively. The following codes generates Fig. 1, which compares the simulated and theoretical distribution.

```
library(ggplot2)
# Theoretical mean
mean_Theoretical <- 1 / lambda
# Standard error of the mean
std_error <- (1 / lambda / sqrt(nExponential))
# Plot the histogram and the theoretical density
ggplot(data, aes(x = mean)) +
     geom_histogram(binwidth = .2, colour = "black", fill = "#FFFF99",
                    aes(y = ..density..)) +
     geom_density(alpha=.05, aes(colour = "line1"), size = 1) +
     stat_function(fun = dnorm, aes(colour = "line2"), size = 1,
                   arg = list(mean = mean_Theoretical, sd = std_error)) +
     geom_vline(aes(xintercept = mean(mean)), color = "red", size = 1) +
     geom_vline(aes(xintercept = mean_Theoretical), color = "blue", size = 1,
                linetype = "dashed") +
     labs(x = expression("Mean"), y = expression("Density"),
```

```
        title = "Fig. 1 Simulated vs Theoretical Distribution") +
    theme(axis.text.y = element_text(colour = "black"),
          axis.text.x = element_text(colour = "black")) +
    scale_colour_manual(values = c("red","blue"), name = "Density",
                        labels = c("Simulation", "Theoretical")) +
    xlim(2, 8)
```



The confidence interval for the theoretical mean when $\alpha = 0.05$ is as follows:

```
mean_Theoretical + c(-1, 1) * qnorm(0.975) * std_error
```

```
## [1] 3.450512 6.549488
```

This shows that the sample mean (5.006) is not significantly different from the theoretical mean (5.000). The comparison between the sample mean and theoretical mean is also presented in Fig. 1.

Sample Variance versus Theoretical Variance

The sample variance (0.5908) is close to the theoretical variance (0.625). Fig. 1 shows that the simulated and theoretical distributions have similar variability.

Distribution

Fig. 1 shows that the simulated distribution is approximately normal. Further, we also look at the difference between a large collection of random exponentials and the distribution of a large collection of averages of 20, 40, and 60 exponentials respectively. Their differences are demonstrated in Fig. 2. The larger the number of exponentials used, the closer the distribution to the normal distribution. This verifies the Central Limit Theorem.

Fig. 2 Simulated vs Theoretical Distribution