Project 1, Part 1 of Statistical Inference

Jin-Keat Lim

Friday, October 24, 2014

Introduction

Before beginning the project, I simply initialized the given project parameters. I am also using set.seed(346) in this case to document reproducibility, and to determine both the mean as well as the standard deviation of each individual rexp(40,lambda) sample.

```
##initializing parameters
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.1.1
```

```
lambda <- 0.2
nosim <- 1000    #of sets to repeat
mean <- 1/lambda
sd <- 1/lambda

##simulation code

set.seed(346)
#taking the mean of rexp(40,lambda) nosim times, or in this case 1000 times
means <- replicate(nosim, mean(rexp(40,lambda)))

#evaluating the std. dev. of rexp(40,lambda) nosim times, or in this case 1000 times
set.seed(346)
sim_sd <- replicate(nosim, sd(rexp(40,lambda)))
```

Part 1 & 2 of the project - calculating simulation vs. theoretical (mean and variance)

```
est_mean <- mean(means)
cat("Simulations distribution refers to the 40 exponentials repeated",nosim,"times", "\n")
cat("\n")
cat("Simulations distribution centered at:", round(est_mean,3),"\n")
cat("Theoretical distribution centered at:", mean, "\n")
cat("Simulations variance:", round(var(means),3),"\n")
cat("Theoretical variance:", ((sd^2)/40),"\n")
```

```
## Simulations distribution refers to the 40 exponentials repeated 1000 times
##
## Simulations distribution centered at: 5.005
## Theoretical distribution centered at: 5
## Simulations variance: 0.644
## Theoretical variance: 0.625
```

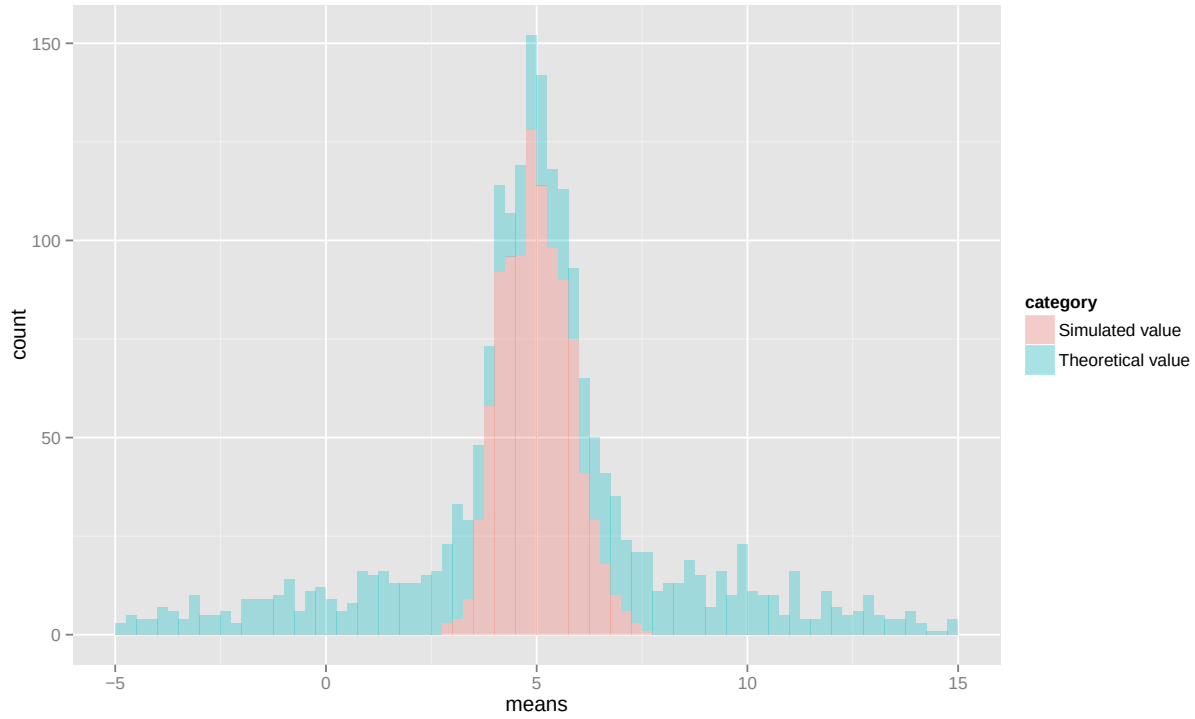
We can see that the simulated mean and variance are very close to the expected theoretical mean and variance.

For **Part 3**, I created a theoretical normal distribution with the same parameters using rnorm, then superimposed it on the simulation distribution

```
#creating a theoretical normal distribution with the same parameters
theo <- data.frame(means = rnorm(nosim, mean=mean, sd=sd))
theo$category <- 'Theoretical value'

#converting the simulated distribution into a dataframe
sim_df <- data.frame(means)
sim_df$category <- 'Simulated value'

#combining the theoretical and simulated distributions into a single dataframe for
#plotting purposes
plot_df <- rbind(theo, sim_df)
```

I then plotted the histograms of both the theoretical and simulated distributions

```
#code to plot and overlay histograms of both theoretical and simulated distributions
plot <- ggplot(plot_df, aes(means, fill = category)) +
        geom_histogram(alpha=0.3, binwidth=.25) +
        xlim(-5,15)

print(plot)
```



We can observe our simulations indeed follow a normal distribution.

For **Part 4** to evaluate the coverage of the confidence interval (hereinafter referred to as "CI"), I took the following approach:

1. Evaluated the CI of each individual rexp(40,lambda) sample
2. Evaluated the % that the overall population mean falls within each of the sample CIs

```
#adding the sd for each 1000 observation to sim_df, and calculating the corresponding CI
#limits
sim_df$sd <- sim_sd
sim_df$ll <- sim_df$means - qnorm(0.975) * (sim_df$sd)/sqrt(40)
sim_df$ul <- sim_df$means + qnorm(0.975) * (sim_df$sd)/sqrt(40)
sim_df$coverage <- (mean > sim_df$ll & mean < sim_df$ul)
coverage_eval <- sum(sim_df$coverage==TRUE) / length(sim_df$coverage)
```

```
cat("Evaluation of the 95% CI coverage:", coverage_eval*100, "%",
    "of observations fall within the 95% CI")
```

```
## Evaluation of the 95% CI coverage: 92.6 % of observations fall within the 95% CI
```

Conclusion

By taking 1000 samples of a 40 exponential simulation, I was able to compare the distribution of the simulations vs. the theoretical distribution and observe they are largely similar, and follow a normal distribution. Based on the simulated data, I was also able to determine a 95% CI with 92.6% coverage.