Statistical Inference - Project 1

Baskaran Viswanathan

September 22, 2015

Project for the "Statistical Inference" course (Coursera, September 2015)

Comparing the simulated mean and variance with the theoretical values

We will run 1000 rounds of simulation of 40 exponentials with $\lambda = 0.2$, using a fixed seed, and comparing the distribution of the simulated mean and variance with the theoretical value of $1/\lambda$:

```r
library(pander)
nsim <- 1000
nvals <- 40
lambda <- 0.2
set.seed(567)
simdata <- t(replicate(nsim, rexp(nvals, lambda)))
df <- data.frame(Mean=c(mean(rowMeans(simdata)), 1/lambda),
                 Variance=c(mean(apply(simdata, 1, var)), 1/lambda^2))
rownames(df) <- c("Simulated", "Theoretical")
pander(df, round=2)
```

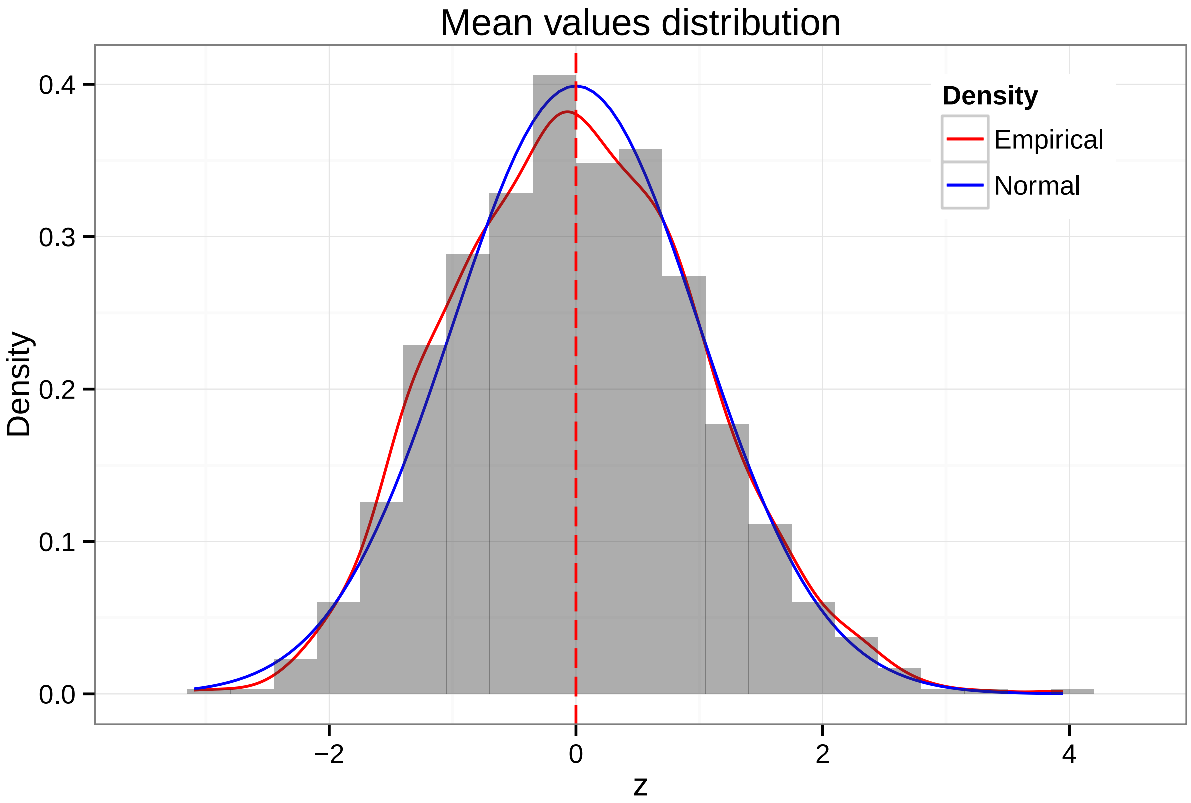
	Mean	Variance
Simulated	4.99	24.78
Theoretical	5	25

The simulated and theoretical values are very close, as expected by the CLT.

Assessing if the simulated values are approximately normal

Also, according to the CLT, the distribution of the simulated means should be approximately normal. To illustrate this we will normalize the vectors and compare it to a $N(0, 1)$ distribution.

```r
library(ggplot2)
meanvals <- rowMeans(simdata)
zmean <- (meanvals - mean(meanvals)) / sd(meanvals)
qplot(zmean, geom = "blank") +
    geom_line(aes(y = ..density.., colour = 'Empirical'), stat = 'density') +
    stat_function(fun = dnorm, aes(colour = 'Normal')) +
    geom_histogram(aes(y = ..density..), alpha = 0.4, binwidth=.35) +
    geom_vline(xintercept=0, colour="red", linetype="longdash") +
    scale_colour_manual(name = 'Density', values = c('red', 'blue')) +
    ylab("Density") + xlab("z") + ggtitle("Mean values distribution") +
    theme_bw() + theme(legend.position = c(0.85, 0.85))
```



Evaluating the coverage of the confidence interval

Theoretically, a 95% confidence interval should contain, if we simulate a big number of them, the mean value for the exponential distribution ($1/\lambda$) 95% of the time.

```
set.seed(567)
lambda <- 0.2
# checks for each simulation if the mean is in the confidence interval
inconfint <- function(lambda) {
                ehats <- rexp(1000, lambda)
                se <- sd(ehats)/sqrt(1000)
                ll <- mean(ehats) - 1.96 * se
                ul <- mean(ehats) + 1.96 * se
                (ll < 1/lambda & ul > 1/lambda)
        }
# estimate the coverage in each round of simulations
coverage <- function(lambda) {
    covvals <- replicate(100, inconfint(lambda))
    mean(covvals)
}
# perform the simulation
simres <- replicate(100, coverage(lambda))
mean(simres)
```

```
## [1] 0.9484
```

As expected, the confidence interval contains the theoretical value 94.84% of the time (close to the expected 95%).