Statistical Inference - Course Project Part 1

Wednesday, August 20, 2014

This is the first part of the project for the Statistical Inference class. It consists of simulating the exponential distribution in R - with the function rexp(n, lambda) and lambda equals to 2 - in order to investigate the distribution of averages of 40 exponential(0.2)s. It is known that both the mean and the standard deviation of exponential distribution are 1/lambda.

Let's start with the initialization of variables:

```
nsim   <- 1000      # number of simulations
n      <- 40        # sample size of 40 as requested
lambda <- 0.2
mu     <- 1/lambda  # mu is the population mean = 5
s      <- mu        # s is the population standard deviation = 5
SE     <- s/sqrt(n) # SE is the theoretical standard error
```

The simulation consists of a loop where the distribution of exponential sample means is built.

```
X  <- NULL # vector of averages of exponential samples
S  <- NULL # vector containing the standard deviation of each exponential sample
sn <- NULL # vector of normalized averages of exponential samples
for(i in 1:nsim){
    exp_i <- rexp(n,lambda)      # exponential sample
    xi    <- mean(exp_i)         # exponential sample mean
    X     <- c(X,xi)
    S     <- c(S,sd(exp_i))
    sn    <- c(sn, (xi-mu)/SE)
}
```

Based on the above simulation, it is possible to:

1. Show where the distribution is centered at and compare it to the theoretical center of the distribution.

   The sample means of a collection of iid observations constitutes a new distribution. According to the Law of Large Numbers the average of this new distribution limits to what it's estimating, i.e. the population mean (theoretical center). We define an estimator as consistent if it converges to what you want to estimate. The simulation results show this trend below.

```
## [1] "sample-mean's estimate = 5.0088 || population mean = 5"
```

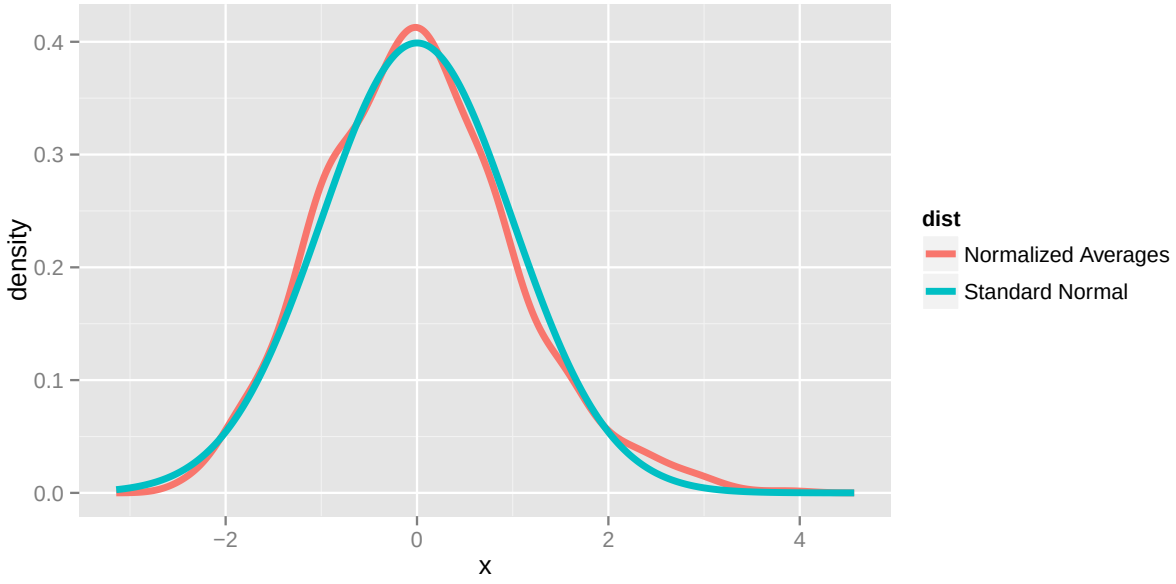
2. Show how variable it is and compare it to the theoretical variance of the distribution.

   Different samples drawn from the same population would in general have different values of the sample mean. So, the distribution of sampled means has its own mean and variance. The standard deviation (square root of variance) of those sample means is knwon as the standard error of the mean, i.e., of using the sample mean as a method of estimating the population mean. The simulation results show that the distribution standard deviation tends to the theoretical standard error (SE).

```
## [1] "standard deviation of sample means = 0.7922 || theoretical standard error = 0.7906"
```

3. Show that the distribution is approximately normal.

   The Central Limit Theorem states that the distribution of averages of iid random variables becomes that of a standard normal as the sample size increases. For the sample size of 40, the vector of normalized averages of exponential samples (sn) was plotted superimposed on the standard normal curve so as to show this phenomena.



4. Evaluate the coverage of the confidence interval for 1/lambda: $\bar{X} \pm 1.96 \frac{S}{\sqrt{n}}$

   For all of the 1,000 simulated samples, the mean (vector X) and standard deviation (vector S) were calculated. Here, they were used to determine the requested confidence interval for each sample. After, the coverage was verified by estimating the percentage of these confidence intervals which contains the true population mean 1/lambda (= 5).

```
coverage <- NULL
for(i in 1:nsim){
    ll <- X[i] - 1.96 * S[i]/sqrt(n)
    ul <- X[i] + 1.96 * S[i]/sqrt(n)
    coverage <- c(coverage,ll < 1/lambda & ul > 1/lambda)
}
mean(coverage)
```

```
## [1] 0.931
```

   This result shows that the coverage related to the sample size of 40 is below the expected one (95%). A better coverage could be achieved by using a higher sample size.

P.S.: The markdown file used to generate this report can be found in https://github.com/Fpschwartz1/StatisticalInference_CourseProject/blob/master/CourseProject_Part1.Rmd