COURSERA - Statistical Inference: Assignment

yobos

22 août 2015

SUMMARY

This is the project for the statistical inference class. In it, I will use simulation to explore inference and do some simple inferential data analysis. The project consists of two parts:

1. A simulation exercise.
2. Basic inferential data analysis.

I will create a report to answer the questions. Given the nature of the series, ideally I will use knitr to create the reports and convert to a pdf.

INSTRUCTIONS

In this project I will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. lambda = 0.2 for all of the simulations. I will investigate the distribution of averages of 40 exponentials. Note that I will do a thousand simulations.

I will illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. This document will: 1. Show the sample mean and compare it to the theoretical mean of the distribution. 2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution. 3. Show that the distribution is approximately normal.

Simulation

```
set.seed(19)
sample.size <- 40
num.trials <- 1000
lambda <- 0.2
sim <- t(replicate(num.trials, rexp(sample.size, lambda)))
```

Lets look at the mean distribution's properties of 40 exponentials with parameter lambda=0.2.

1. Sample Mean vs. Theoretical Mean

```
theo.mean<-1/lambda
```

we stated that theoretical mean is 1/lambda = 'r theo.mean' . Lets calculate the sample mean and compare both.

```r
#sample mean = mean of each data rows
sample.means<-rowMeans(sim)
mean.sample.means<-mean(sample.means)
```

One can see that theoretical and sample mean are quite close.

2. Sample Variance vs. Theoretical Variance

```r
theo.variance<-1/lambda^2
```

we stated that theoretical variance is 1/lambda^2 = 'r theo.variance' . Lets calculate the sample variance and compare both.

```r
#As all samples are independent the variance of the whole samples is equal to the variances of the whol

variance<-mean(apply(sim, 1, var))
```

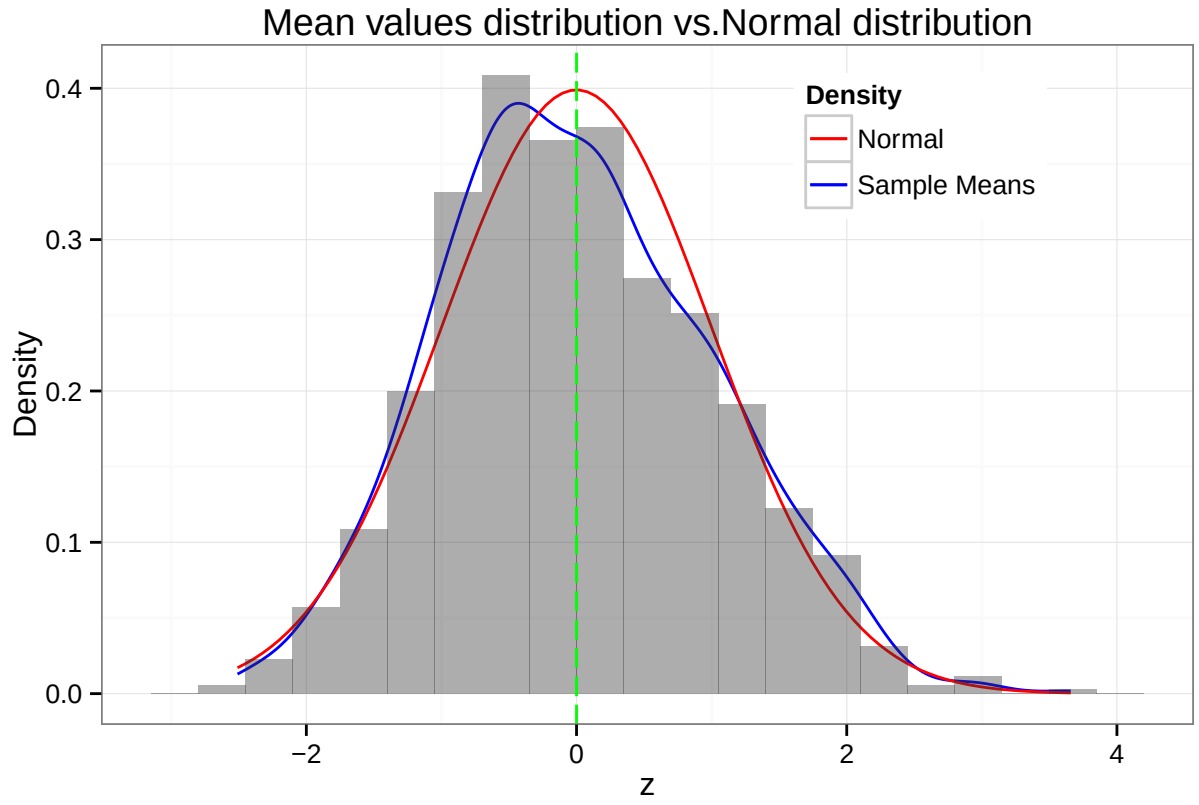
As we can see the theoretical variance and the sample variance are quit close.

3. Sample means distribution vs. normal distribution

In this section, We will figure out if the sample means distribution is approximately a normal distribution.

Lets first normalize the sample means distribution an compare it to a N(0,1) distribution .

```r
library(ggplot2)
xmean <- rowMeans(sim)
zmean <- (xmean - mean(xmean)) / sd(xmean)
qplot(zmean, geom = "blank") +
    geom_line(aes(y = ..density.., colour = 'Sample Means'), stat = 'density') +     stat_function(fun =
    geom_histogram(aes(y = ..density..), alpha = 0.4, binwidth=.35) +
    geom_vline(xintercept=0, colour="green", linetype="longdash") +
    scale_colour_manual(name = 'Density', values = c('red', 'blue')) +
    ylab("Density") + xlab("z") + ggtitle("Mean values distribution vs.Normal distribution") +
    theme_bw() + theme(legend.position = c(0.75, 0.85))
```



Evaluation the confidence interval coverage

To assess whether the samples distribution is approximately normal, we can also show that 95% confidence interval should contain, the mean value for our exponential distribution 95% of the time.

```
set.seed(19)
lambda <- 0.2
# checks for each simulation if the mean is in the confidence interval
inconfint <- function(lambda) {
                x <- rexp(1000, lambda)
                se <- sd(x)/sqrt(1000)
                ll <- mean(x) - 1.96 * se
                ul <- mean(x) + 1.96 * se
                (ll < 1/lambda & ul > 1/lambda)
        }
coverage <- function(lambda) {
    covvals <- replicate(100, inconfint(lambda))
    mean(covvals)
}
simres <- replicate(100, coverage(lambda))
mean(simres)
```

```
## [1] 0.954
```

The confidence interval contains 95.4% which is close to the expected 95%.