Final Assignment for Statistical Inference

Chuck Chekuri

3/4/2017

## Overview

This is the final assignment for Statistical Inference course of Data Science Specialization by John Hopkins University.

This project report describes

1. Simulation exercise to illustrate the Central Limit Theorem
2. Inferential Analysis of Tooth Growth Data

### Part 1. Simulation Exercise

This exercise demonstrates the Central Limit Theorem by simulating the mean and variance of an exponential distribution to compare that with the theoretical distribution. It also shows that the distribution of the mean of an exponential distribution follows a normal distribution.

#### Simulations

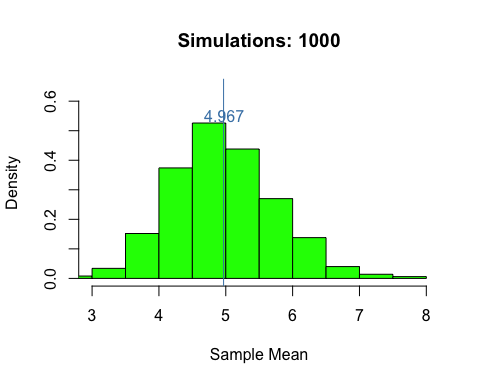
All simulations used a sample of 40 random exponentiall distributed numbers at a rate() of 0.2. Two simulations for calculating the average mean and average variance were performed and the average values from the simulations were compared with the theoretical mean and variance for the given rate() of 0.2

Additional three simulations were done to show the convergence of these averages to the normal distribution illustrating the Central Limit Theorem.

#### Comparing Sample Mean with Theoretical Mean

lambda<-0.2;n<-40  
nsim=1000  
mns = NULL  
for (i in 1 : nsim) mns = c(mns, mean(rexp(n, lambda)))

The above code calculates and stores the simulated sample means in the variable mns. Here is the histogram for the simulation after 1000 simulations. The blue vertical line shows the average sample mean 4.967 is very close the to the theoretical mean 1/ = 5.0.



#### Comparing sample variance with theoretical variance

The code below gets the variances of an exponential distribution for 1000 simulations.

vrs = NULL  
for (i in 1 : nsim) vrs = c(vrs, var(rexp(n, lambda)))

The theoretical variance for an exponential distribution is = 25 and the average variance of the sample means is = 25.2487154.

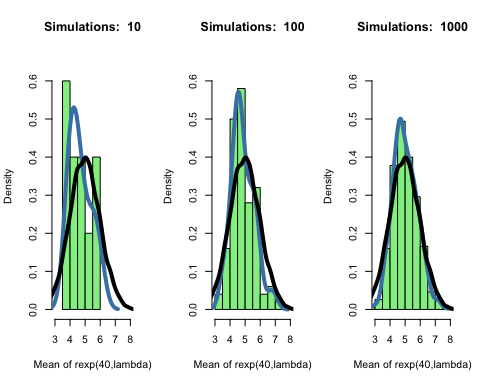
This shows the average variance of sample means 25.2487154 is close the theoretical mean 25.

#### Distribution of sample means is close to normal distribution

Simulating the sample means for 10, 100, 1000 times and drawing the density functions against a normal distribution for a large sample, the graphs show how the distribution (blueline) comes close to the normal distribution (black line).

The following code does the simulation.

set.seed(1999)  
nd <- rnorm(10000, mean=mean(mns)) # shift to align the density plots  
par(mfrow=c(1,3))  
for (nsim in c(10,100,1000)){   
 title<-paste("Simulations: ", nsim)  
 mns = NULL  
 for (i in 1 : nsim) mns = c(mns, mean(rexp(n, lambda)))  
   
 hist(mns, probability = TRUE, col="lightgreen",   
 ylim = c(0, .65), xlim = c(3,8),  
 main=title, xlab="Mean of rexp(40,lambda)")  
 lines(density(mns), lwd=4, col="steelblue")  
 lines(density(nd), col="black", lwd=4)  
}



### Part 2. Inferential Analysis of Tooth Growth Data

As per the rubric, this section covers four specific areas along with the code used.

1. Loading the ToothGrowth data and exploratory data analyses
2. Summary of the data
3. Compare tooth growth by supp and dose
4. Inference along with conclusions and assumptions

#### Loading and exploratory analysis

Tooth Growth Data is a set of observatons of the length of odontoblasts (cells reponsible for tooth growth) in 60 Guinea Pigs that were given Vitamin C using two different supplements.

This data will be explored and inferences drawn on the effect of Vitamin C on these Guinea Pigs.

A quick look at the data reveals that the 60 guinea pigs were equally distributed for each of the two supplements(OJ, VC) and the three dosage levels (0.5,1,2).

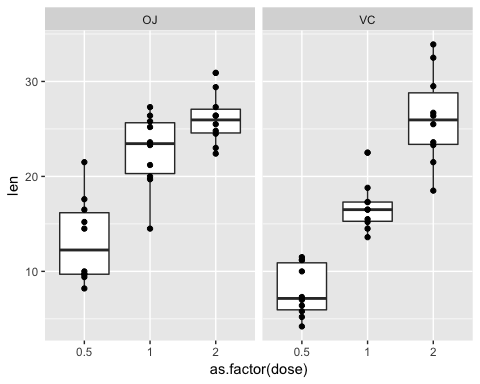
## Supplement  
## Dosage OJ VC  
## 0.5 10 10  
## 1 10 10  
## 2 10 10

## len supp dose   
## Min. : 4.20 OJ:30 Min. :0.500   
## 1st Qu.:13.07 VC:30 1st Qu.:0.500   
## Median :19.25 Median :1.000   
## Mean :18.81 Mean :1.167   
## 3rd Qu.:25.27 3rd Qu.:2.000   
## Max. :33.90 Max. :2.000

How does the length vary by dose and supplement?

library(ggplot2)  
qplot(y=len, x=as.factor(dose), data=ToothGrowth, geom=c("boxplot","smooth","point"), facets=.~supp)

## `geom\_smooth()` using method = 'loess'



From the box plot the mean len of "OJ" group and "VC" group seem to be different. We need to do a t.test to make sure the difference is statistically significant.

Since the observations are on different Guinea Pigs, the observations of len or not paired. a a two sided t.test with paired as false will be used to test the significance.

The code below creates to groups an does a un paied t.test and it's two sided because we are testing if the means are statisitcal same.

g\_vc <- ToothGrowth$len[ToothGrowth$supp=="OJ"];  
g\_oj <- ToothGrowth$len[ToothGrowth$supp=="OJ"];  
t.test(g\_vc, g\_oj, paired = FALSE, alternative = "two.sided")

##   
## Welch Two Sample t-test  
##   
## data: g\_vc and g\_oj  
## t = 0, df = 58, p-value = 1  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -3.414026 3.414026  
## sample estimates:  
## mean of x mean of y   
## 20.66333 20.66333

The t.test shows that the difference in the mean is statistically significant.

We can now do the hypotheis testing with confidenct intervals for the means.