# Statistical Inference Course Project

### Course: statinference-006

### Author: Dennis Barger

### Date: 10/22/2014

## Project Assignment

The project will explore inference by using simulation and simple inferential data analysis. The project consists of two sections:

1. Simulation Exercises
2. Basic inferential data analysis

The project deliverable should be a 2 page report in pdf format that uses knitr to combine text, code and analysis results.

Simulation Exercises: The simulation exercises will utilize R to explore exponential distribution. The R function rexp(n,lambda) should be used to simulate data. Lambda is the rate paramenter. The mean of the exponential distribution is 1/lambda and the standard deviation will also be 1/lambda. Set lambda = 0.2 for all simulations. The exercise will investigate the distsribution of averages of 40 exponential distributions. The exercise should generate at least 1000 simulation for average of 40 exponentials.

The simulation exercise should answer the following questions:

* Show where the distribution is centered and compare to the theoretical center of distribution.
* Show istributions variability and compare to the theoretical variance.
* Show how the distribution is approximately normal.
* Evaluate the coverage of the confidence interval for 1/lambda: X¯±1.96Sn???.

## Simulation Exercises

### Simulation Exercise: Creating Simulation Dataset

The first step in the exercise is to generate a dataset of simulated mean values based on an exponential distribution. The following R code describes the logic for initialization and generation of simulated values.

# Initialize required packages and variables used in the analysis  
library(ggplot2)  
lambda = 0.2  
n = 40  
set.seed(100)  
  
#Generate 1000 mean simulation based on 40 exponentials  
df <- data.frame(replicate(1000,mean(rexp(40,.2))))  
names(df) <- "mean\_value"

A quick summary of the dataset indicates the data is similar to a normal distribution with some slight variation.

summary(df)

## mean\_value   
## Min. :3.04   
## 1st Qu.:4.46   
## Median :4.92   
## Mean :5.00   
## 3rd Qu.:5.51   
## Max. :8.04

g <- ggplot(data=df, aes(x=mean\_value))  
g <- g + geom\_histogram(aes(y=..density..), binwidth=0.20, fill="light blue", color="black")

m <- round(mean(df$mean\_value),2)  
s <- round(sd(df$mean\_vallue),2)  
v <- round(var(df$mean\_value),2)  
print(paste("The mean value is ",m,". The standard deviation is ", s, " and the variance is ",v,sep=""))

## [1] "The mean value is 5. The standard deviation is NA and the variance is 0.64"

### Simulation Exercise: Show distribution center and compare to theoretical center of distribution

The theoretical center of distribution is represented as a normal distribution, therefore the mean is calculated as 1/lambda. Since lambda equals 0.20, then the theoretical center (mean) is equal to 5.

The center of this dataset, which is represented as an exponential distribution, is equal to 4.99. The logic for calculating both centers is as follows:

nm\_mean <- round(1/lambda,2)  
ex\_mean <- round(mean(df$mean\_value),2)  
print(paste("The theoretical mean is ",nm\_mean," and the exponential mean is ", ex\_mean, sep=""))

## [1] "The theoretical mean is 5 and the exponential mean is 5"

s <- round((1/lambda)/sqrt(40),2)  
v <- round((1/lambda)/sqrt(40) ^.2,2)

### Simulation Exercise: Show variability of distribution and compare to theoretical variance

The exponential standard deviation is 0.80 and the normal distribution of 0.79. The exponential variance is 0.64 in comparison to the normal variance of 0.63.

It's easy to see that a large number of exponential simulations such as 1000 can have a distribution similar to a theoretical or normal distribution. The following graph further highlights the similarities of distribution models.

nm\_sd <- round((1/lambda)/sqrt(40),2)  
nm\_vr <- round((1/lambda)/sqrt(40) ^.2,2)  
  
ex\_sd <- round(sd(df$mean\_value),2)  
ex\_vr <- round(var(df$mean\_value),2)  
print(paste("The theoretical standard deviation is ",nm\_sd, " in comparison to the exponential standard deviation of ",ex\_sd,sep=""))

## [1] "The theoretical standard deviation is 0.79 in comparison to the exponential standard deviation of 0.8"

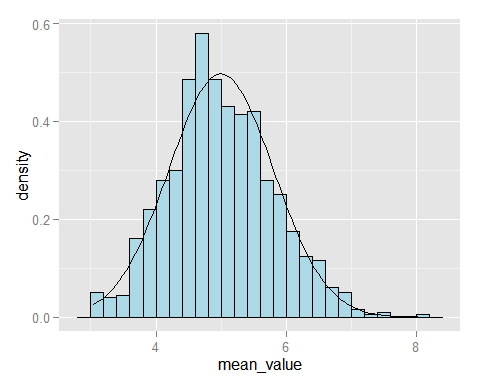
print(paste("The theoretical variance is ",nm\_vr, " in comparison to the exponential variance of ",ex\_vr,sep=""))

## [1] "The theoretical variance is 3.46 in comparison to the exponential variance of 0.64"

### Simulation Exercise: Show how the distribution is approximately normal

The following graph shows a histogram of the exponential distribution mean values with a density overlay of the normal distribution. Once again, we see the mean for both distributions are closely aligned.

g <- ggplot(data=df, aes(x=mean\_value))  
g <- g + geom\_histogram(aes(y=..density..), binwidth=0.20, fill="light blue", color="black")  
g <- g + stat\_function(fun=dnorm, arg=list(mean=1/lambda, sd=sd(df$mean\_value)))  
g



### Simulation Exercise: Evaluate coverage of the confidence interval for 1/lambda

The confidence interval for 1/lambda can be calculated in R using the logic mean(val) + c(-1,1) \* 1.96 \* sd(val)/sqrt(nbr of observations). Based on the code listed below, the 95% confidence intervalue for the mean of the means is 4.95 - 5.05.

mean(df$mean\_value) + c(-1,1) \* 1.96 \* sd(df$mean\_value)/sqrt(nrow(df))

## [1] 4.950 5.049

## Basic Inferential Data Analysis

### Project Assignment

This exercise involves analyzing the ToothGrowth data in the R datasets package. The analysis should include the following deliverables: \* Basic summary data \* Confidence intervals and hypothesis tests to compare tooth growth by supp and dose \* Conclusion of findings include assmptions

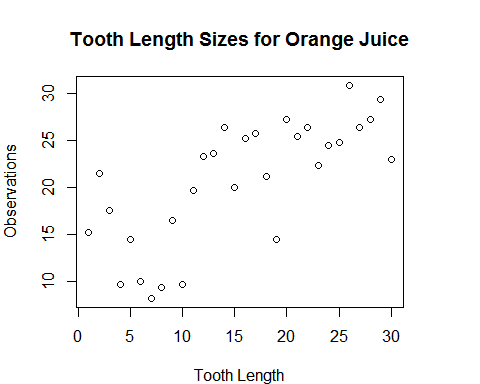
### Basic Summary of Data

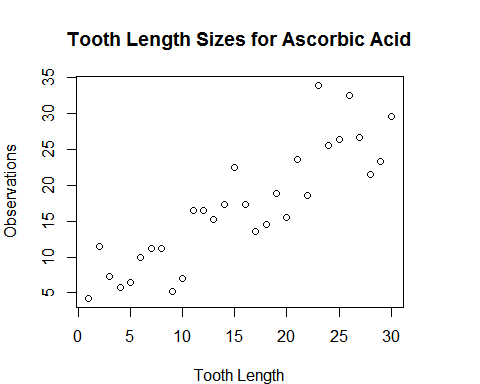
The ToothGrowth dataset represents the effect of Viamin C on Tooth Growth in Guinea Pigs. The dataset contains 60 observations on 3 variables:

* len - Tooth length
* supp - Supplement type (VC = ascorbic acid, OJ = Orange juice)
* dose - dose in milligrams

There is an equal number of by supplement type and dose:

## dose  
## supp 0.5 1 2  
## OJ 10 10 10  
## VC 10 10 10

The following graph plots tooth length values for Orange Juice supplement observations: 

Tooth length values for Ascorbic Acid supplement observations are displayed in this graph:  ### Confidence Intervals Tooth Growth using Orange Juice Supplements: \* Mean = 20.66 \* Standard Deviation = 6.61 \* Variance = 43.63

Tooth Growth using Ascorbic Acid Supplements: \* Mean = 16.96 \* Standard Deviation = 8.27 \* Variance = 68.33

The following code confirms measurements for Orange Juice:

df <- ToothGrowth[ToothGrowth$supp=="OJ",]  
mean(df$len)

## [1] 20.66

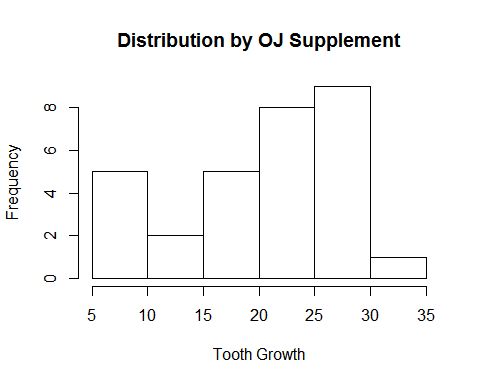
sd(df$len)

## [1] 6.606

var(df$len)

## [1] 43.63

hist(df$len, main="Distribution by OJ Supplement", xlab="Tooth Growth")



The following code confirms measurements for Ascorbic Acid:

df <- ToothGrowth[ToothGrowth$supp=="VC",]  
mean(df$len)

## [1] 16.96

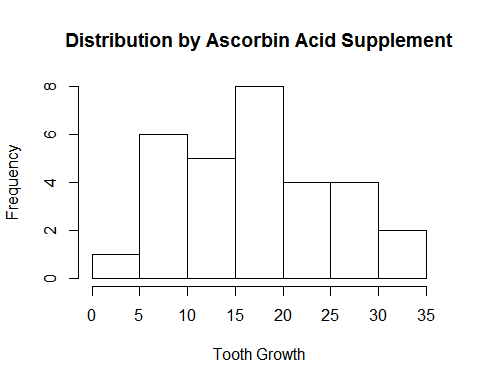
sd(df$len)

## [1] 8.266

var(df$len)

## [1] 68.33

hist(df$len, main="Distribution by Ascorbin Acid Supplement", xlab="Tooth Growth")

 The