Coursera Data Scientist
Statistical Inference: Week 3 Assignment
Author: Michelle Gitgit

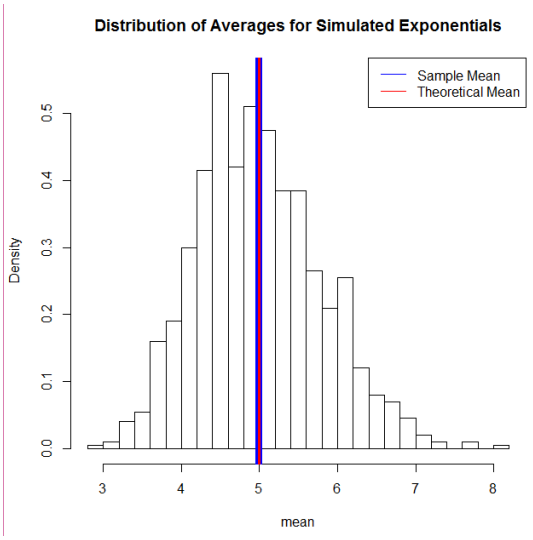
Overview

The information in this document represents the beginning analysis skills of the author. There are two parts in this report. In part 1, a simulation exercise is completed and reported on. Part 2 is a basic inferential data analysis and is submitted separately.

Part 1: Simulations

This simulation consists of three components: comparison of means, comparisons of variance, and comparisons of curves. Several parameters were initially set before data analysis: (a) *set.seed* to ensure consistency with data, (b) *lambda,* (c) *simNum* set to 1000, the number of simulations, and (d) *expNum* set to 40, the number of exponentials. The detailed code for Part 1 is located in Appendix 1

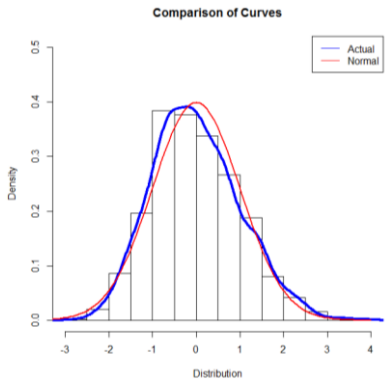
Comparison of Means Brief Summary: A data frame was created to hold data, followed by an appropriate sample based on the set.seed. The mean of each instance was created (means1). The means of all the means was created (meanofmean1). The theoretical mean was determined using 1/lambda. A histogram was made to show the distribution of the simulated exponentials. The thick blue line represents the sample mean (4.995215); the thinner red line represents the theoretical mean (5). A heading for the chart was created, and the legend denotes the meaning of each colored line.



Comparison of Variances Brief Summary: The theoretical variance was determined by the following formula: ((1/lambda)^2)/expNum where expNum is the number of exponentials.

Theoretical Variance: 0.625
Actual Variance: 0.631

The values were similar by comparison.



Demonstrate Distribution is Approximately Normal Summary: A normal curve was drawn over the data in red. The actual curve was drawn in blue. A comparison of the curves demonstrate consistent similarities with peak values around 0.

Appendix 1: Part 1 R Code

```
##This project reflects the work of Michelle Ihrig

##The following section sets the parameters for the project
set.seed(1997) ##ensures consistency with data
lambda<-0.2 ##sets lamda as 0.2
simNum<-1000 ##number of simulations in this project
expNum<-40 ##number of exponentials

##creates frame to store data
frames<-data.frame(mean=numeric(s))

##1 Sample mean compared to theoretical mean
samp1 <-replicate(simNum, rexp(expNum, lambda)) ##creates sample
means1 <-apply(samp1, 2, mean) ##creates mean of instances
meanofmean1 <-mean(means1) ##determines mean of means
meanofmean1 ##prints mean of means
lmbdamean1 <- 1/lambda ##determines theoretical mean
lmbdamean1  ##prints theoretical mean

##Produces visual for sample mean compared to theoretical mean
hist(means1, breaks=30, probability = T, xlab= "mean", main="Distribution of Averages for Simulated
Exponentials")
abline(v=meanofmean1, lwd="8", col="blue") ##lwd makes line noticeable
abline(v=lmbdamean1, lwd="3", col="red") ##lwd makes line noticeable
legend("topright", col=c("blue", "red"), c("Sample Mean", "Theoretical Mean"), lty=c(1,1), lwd=c(1,1))


##2 Compare variances of sample data and theoretical data
thvar2<-round(c(((1/lambda)^2)/expNum), 3) ##determines theoretical variance, rounds three places
thvar2 ##prints theoretical variance for part 2
acvar2 <- round(c(var(means1)), 3) ##determines actual variance, rounds three
acvar2 ##prints actual variance for part 2

##3 Shows distribution is approximately normal
par(mfrow=c(1,1))
hist(scale(means1), probability=T, main="Comparison of Sample and Theoretical Variances",
        xlab="Distribution", ylim=c(0,0.5))
lines(density(scale(means1)), lwd="4", col="blue")
curve(dnorm(x,0,1), -4, 4, lwd="2", col="red", add=T)
legend("topright", col=c("blue", "red"), c("Actual", "Normal"), lty=c(1,1), lwd=c(1,1))
```