Exponential Distribution Simulation; Tooth Growth Analysis

Ash Chakraborty

Wednesday, June 17, 2015

# PART 1: SIMULATION - Sampling the Exponential Distribution

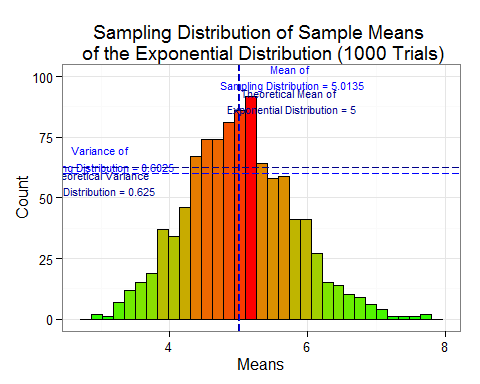
Part 1 of this report investigates the *Exponential Distribution* and compares it to the *Central Limit Theorem*. We conduct simulations on this distribution with a sample size of 40 exponentials by generating a sampling distribution of sample means. The consequences of this sampling distribution will be evaluated for adherence to the central limit theorem.

set.seed(123)   
sample <- rexp(40, rate=0.2)

We conduct a 1000 simulations, and extract the mean of each sample:

samp.dist <- NULL  
for(i in 1:1000) {samp.dist <- c(samp.dist, mean(rexp(40, rate=0.2)))}

## Comparing Variance and Mean to Theoretical Variance and Mean



The distribution has begun to resemble a Gaussian distribution.

## [1] "Mean of Sampling Distribution: 5.0135"

## [1] "Variance of Distribution: 0.6025"

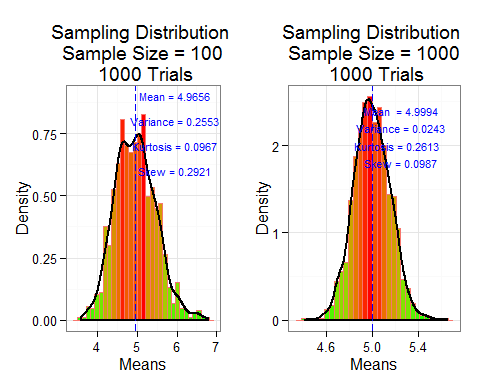
The theoretical mean of the exponential distribution is given by () = 5. **The sampling distribution's mean is a fairly good approximation**. The theoretical variance of the sampling distribution: = 0.625. **The variance obtained from the sampling distribution is a good approximation**.

## Verifying Distribution Normality

We will conduct two simulations with larger sample sizes (n= 100 & 1000), conduct a 1000 trials and compute the *variance* and *skew* :

#sim 1, sample size = 100  
samp.dist1 <- NULL  
for(i in 1:1000){samp.dist1 <- c(samp.dist1, mean(rexp(100, rate=0.2)))}  
#sim 2, sample size=1000  
samp.dist2 <- NULL  
for(i in 1:1000){samp.dist2 <- c(samp.dist2, mean(rexp(1000, rate=0.2)))}

Plotting the resulting sampling distributions side by side:



* The mean is approximated more closely as the sample size increases (in line with CLT)
* Variance of sampling distribution significantly decreases (in line with CLT)
* The skew of the density curve get smaller, suggesting a normal curve

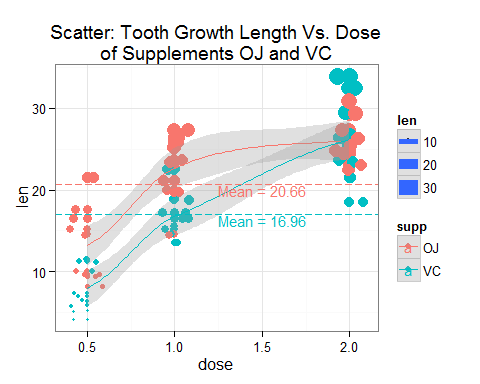
**END OF PART 1**

# PART 2: Analyzing Tooth Growth Data

We run some summary analysis on this dataset, in addition to testing hypotheses to extract any relationships between supplement, dosage, and tooth length.

## Data Summary

The scatter below shows the dataset spread out into groups by supplement and dosage:



* There is a positive correlation between length and dosage amount 0.5 and 1.0 for both supplements.
* There is *no data* for either supplement at dosage amount 1.5.
* Any effect on length at dosage amount of 2.0 for either supplement is unclear.
* Overall, the average sample tooth growth length is greater for "OJ" than it is for "VC".

## Hypothesis Testing

In the first test (which serves as the template for the subsequent tests), we compare the effect on tooth growth between supplements OJ and VC at dose = 0.5. We **assume** that the NULL hypothesis, , is true; independent groups, and unequal variance.

# Function to take in the dosage group and perform a t-test on supplement effect on length  
doseTest <- function(d) {  
 testdf <- NULL  
 testdf <- subset(ToothGrowth, dose==d, select=c('len', 'supp'))  
 # apply t.test to compare the mean lengths by supplement at the dosage group  
 t.test(len ~ supp, paired=FALSE, data=testdf)}  
doseTest(0.5)

##   
## Welch Two Sample t-test  
##   
## data: len by supp  
## t = 3.1697, df = 14.969, p-value = 0.006359  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 1.719057 8.780943  
## sample estimates:  
## mean in group OJ mean in group VC   
## 13.23 7.98

Here, we see that the 95% confidence interval is above zero. The p value of 0.64% is very unlikely, thus causing us to *reject* the NULL hypothesis. **Under a similar assumption, we repeat the *t-test* to get the probability of the alternate hypothesis for the other dosage groups**:

doseTest(1)$p.value

## [1] 0.001038376

doseTest(2)$p.value

## [1] 0.9638516

We note that for dose = 1.0, the p value of 0.1% is unlikely, thus causing us to **reject** the NULL hypothesis. However, for dose = 2.0, the p value of 96.39% is well within the 95th percentile of values that comply with the NULL hypothesis. In this case, then, the NULL hypothesis **stands**.

# Assumptions and Conclusions

The following assumptions must be restated for the tests conducted above:

* The subjects in each dosage groups are independent (not paired)
* Unequal variance between groups has been assumed

The scatter plot and the tests performed on the dataset help us conclude the following:

* At dose=0.5, OJ results in higher tooth growth than VC
* At dose=1.0, OJ results in higher tooth growth than VC
* At dose=2.0, there is no discernible difference in tooth growth between the supplements
* Overall, it seems that OJ is more effective than VC at increasing tooth growth when administered at doses 0.5 or 1.0.

**END OF REPORT**

# APPENDIX 1: GIT CODE

The entire markdown file can be found in this [github repo](https://github.com/ashirwad08/StatisticalInference.git).

# APPENDIX 2: PLOT CODE BLOCKS

## (PART 1) Comparing Variance and Mean to Theoretical Variance and Mean

#sampling distribution of sample means  
g3 <- ggplot()   
g3+geom\_histogram(aes(x=samp.dist, fill=..count..), col="black")+  
 scale\_fill\_gradient("Count",   
 low = "green",   
 high = "red")+  
   
 #mean of sampling dist.  
 geom\_vline(aes(xintercept=mean(samp.dist)),  
 linetype="longdash",   
 col="blue")+  
 geom\_text(aes(x=mean(samp.dist)\*1.15,   
 y=100,  
 label=paste0("Mean of \n Sampling Distribution = ",  
 round(mean(samp.dist), 4))  
 ),   
 col="blue", size=3  
 )+  
 #theoretical mean  
 geom\_vline(aes(xintercept=1/0.2),  
 linetype="longdash",  
 col="darkblue"  
 )+  
 geom\_text(aes(x=(1/0.2)\*1.15,   
 y=90,  
 label=paste0("Theoretical Mean of \n Exponential Distribution = ",   
 round(1/0.2, 4))  
 ),   
 col="darkblue", size=3  
 )+  
   
 #variance of sampling dist.  
 geom\_hline(aes(yintercept=var(samp.dist)\*100),  
 linetype="longdash",  
 col="blue"  
 )+  
 geom\_text(aes(y=var(samp.dist)\*110,   
 x=3,  
 label=paste0("Variance of \n Sampling Distribution = ",   
 round(var(samp.dist), 4))  
 ),   
 col="blue", size=3  
 )+  
 #theoretical variance = population variance/n  
 geom\_hline(aes(yintercept=((1/(0.2^2))/40)\*100),  
 linetype="longdash",  
 col="darkblue")+  
 geom\_text(aes(y=((1/(0.2^2))/40)\*90,   
 x=3,  
 label=paste0("Theoretical Variance \n of Distribution = ",   
 round(((1/(0.2^2))/40), 4))  
 ),   
 col="darkblue", size=3  
 )+   
 labs(title="Sampling Distribution of Sample Means \n of the Exponential Distribution (1000 Trials)",  
 x="Means", y="Count")+  
 theme\_bw()+  
 theme(legend.position="none")

## (PART 1) Verifying Normality

This code chunk is quite large and may be found in its entirety [in github: markdown](https://github.com/ashirwad08/StatisticalInference.git).

## (PART 2) ToothGrowth Summary

g1 <- ggplot(data=ToothGrowth, aes(x=dose, y=len, group=supp, col=supp, size=len) )  
g1 + geom\_point()+  
 geom\_jitter(position=position\_jitter(width=0.1))+  
 geom\_smooth(alpha=0.3, method="loess")+   
 geom\_hline(data=means.df, aes(yintercept=c(means[[1]], means[[2]]), col=supp), linetype="longdash")+  
 geom\_text(data=means.df, aes(x=1.5, y=c(means[[1]], means[[2]]), col=supp,   
 label=c(paste0("Mean = ", round(means[[1]], 2)),   
 paste0("Mean = ", round(means[[2]], 2)))),   
 size=4, vjust=1)+  
 labs(title="Scatter: Tooth Growth Length Vs. Dose\nof Supplements OJ and VC") +  
 theme\_bw()