Exponential Distribution Simulation and Tooth Growth Analysis

Ash Chakraborty

Wednesday, June 17, 2015

Overview

This report consists of two parts. Part 1 conducts a number of simulations on the exponential distribution in order to compare it to the central limit theorem, whereas Part 2 analyzes and tests certain hypotheses about Tooth growth's relationship with certain supplements.

PART 1: SIMULATION - Sampling the Exponential Distribution

Synopsis

Part 1 of this report investigates the *Exponential Distribution* and compares it to the *Central Limit Theorem*. The exponential distribution is given by the probability distribution function, $P(x) = \lambda e^{-\lambda x}$. We also know that it has a mean, $\mu = 1/\lambda$, and variance, $\sigma^2 = 1/\lambda^2$. This report will conduct an appropriate number of 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.

Exponential Distribution

Sample size, n = 40 with a rate parameter, $\lambda = 0.2$. We first generate a sample of 40 random exponentials:

```
library(PerformanceAnalytics)
library(ggplot2)
library(gridExtra)
set.seed(123)
sample <- rexp(40, rate = 0.2)
```

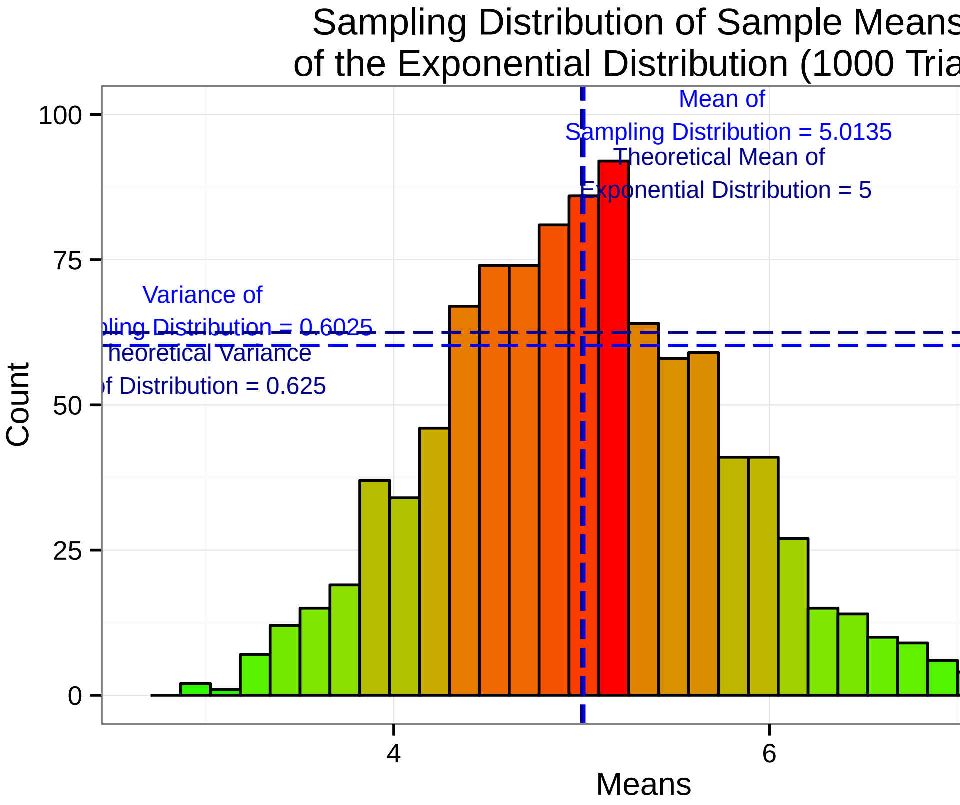
Simulating 1000 Trials

In order to understand the properties of the distribution of the mean of 40 exponentials, we conduct a 1000 simulations, and extract the mean of each sample:

```
samp.dist <- NULL
for(i in 1:1000){
        #extract sample of 40 and take mean of that sample
        samp.dist <- c(samp.dist, mean(rexp(40, rate=0.2)))
}
summary(samp.dist)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.943   4.478   4.982   5.014   5.524   7.717
```

Comparing Variance and Mean to Theoretical Variance and Mean



Now, we plot this sampling distribution:

The distribution of a 1000 means of 40 random exponentials has begun to resemble a Gaussian distribution.

We continue to note from the earlier plot that the sampling distribution's mean and variance may be given by:

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

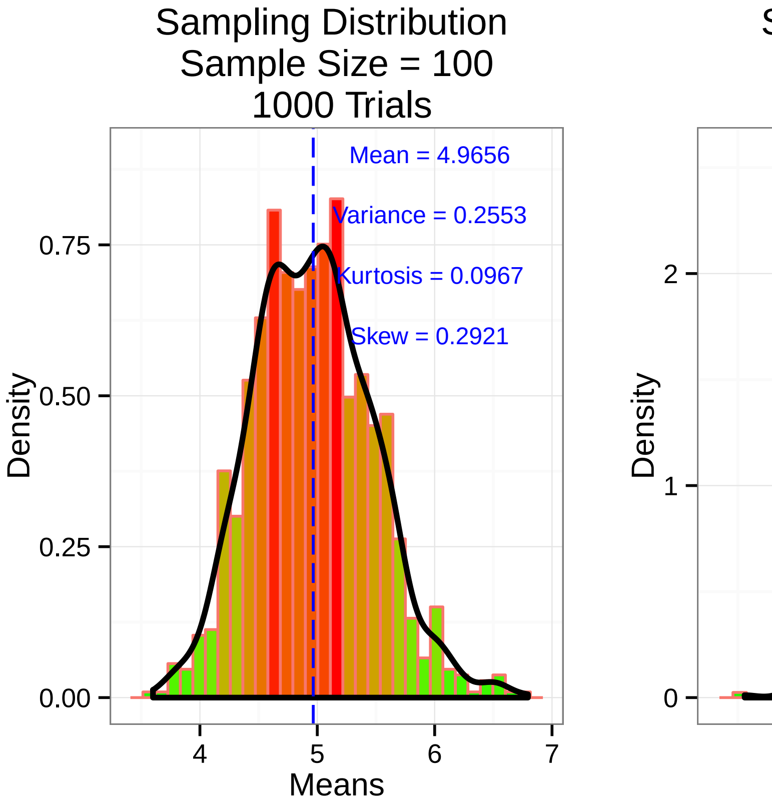
```
## [1] "Variance of Distribution: 0.6025"
```

The theoretical mean of the exponential distribution is given by $1/\lambda$ ($\lambda = 0.2$), which is equal to 5. When we compare the sampling distribution's mean to this theoretical mean, we see that it is a fairly good approximation. Furthermore, the theoretical variance of the sampling distribution is given by the variance of the population mean/sample size, or $(1/\lambda^2)/n$, which = 0.625. Here, it seems that the variance obtained from the sampling distribution of 0.6025, is a fair approximation of the theoretical variance.

Verifying Distribution Normality

In order to confirm that the sampling distribution obtained above is indeed normal, we will test a consequence of the central limit theorem, i.e. *as the sample size approaches infinity, the variation of the sampling distribution gets smaller* and thus results in a closer approximation of the population mean. We will conduct two simulations with greater sample sizes, n= 100 and n = 1000. We will conduct a 1000 trials and compute the variance, kurtosis and skew to give us an idea of the density curve.

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



Plotting the resulting sampling distributions side by side:

Consequently, we observe the following:

- 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

Synopsis

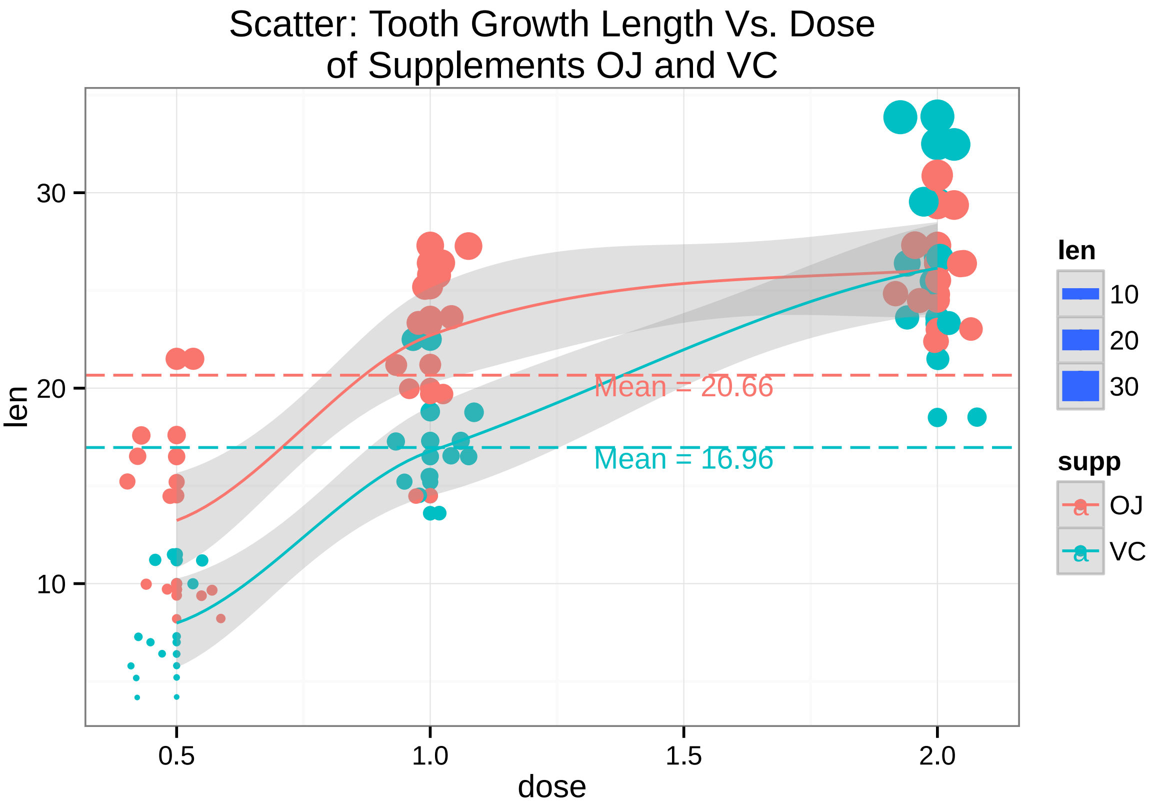
We are going to load the ToothGrowth dataset that's part of the "datasets" package in R. The dataset contains 60 observations of tooth length measured while under the administration of supplements "OJ" and "VC" in various dosages. We run some summary analysis on this dataset, in addition to testing hypotheses to extract any relationships between supplement, dosage, and tooth length.

Data Summary

We see that the dataset is setup as follows:

```
##       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  
```

Each observation in this dataset seems to record the length of the tooth when administered with a certain dosage of supplement OJ or VC. Doses are discrete levels of 0.5, 1.0, 1.5, and 2.0 units; there are two supplements administered: "OJ" and "VC". The mean length of tooth growth when OJ is administered is 20.66 units, and when VC is administered the mean length is 16.96 units. *Visually:*



The plot *seems to* suggest the following:

- 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

Given the summary above, we setup the following tests to help determine the effect of the 2 supplements and their doses on tooth growth. For the tests, the following assumptions are made:

- These are independent groups
- The variances are unequal

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, $H_0 : \mu_{xOJ} = \mu_{xVC}$, is true.

```
# 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, suggesting that the mean of supplement OJ at dose=0.5 is larger than the mean of VC at the same dose. The p value of 0.64% is very unlikely, thus causing us to **reject** the NULL hypothesis.

Similarly, we repeat the *t-test* to compare the 2 supplements in dose groups 1.0 and 2.0, respectively:

```
doseTest(1)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 4.0328, df = 15.358, p-value = 0.001038
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.802148 9.057852
## sample estimates:
## mean in group OJ mean in group VC
##            22.70            16.77
```

```
doseTest(2)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = -0.0461, df = 14.04, p-value = 0.9639
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.79807  3.63807
## sample estimates:
## mean in group OJ mean in group VC
##            26.06            26.14
```

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
- 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.

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 Trial
                x="Means", y="Count")+
        theme_bw()+
        theme(legend.position="none")
```

(PART 1) Verifying Normality

```
# PLOT SIM 1
g4 <- ggplot() + geom_histogram(aes(x=samp.dist1,
                                    y=..density..,
                                    fill=..count..,
                                    col="black"))+
        scale_fill_gradient("Count",
                            low = "green",
                            high = "red")+
        geom_density(aes(samp.dist1), col="black", size=1)+

        #mean of sampling dist.
        geom_vline(aes(xintercept=mean(samp.dist1)),
                   linetype="longdash",
                   col="blue")+
        geom_text(aes(x=mean(samp.dist1)*1.2,
                      y=.9,
                      label=paste0("Mean = ",
                                   round(mean(samp.dist1), 4))
                      ),
                  col="blue", size=3
                  )+

        #variance of sampling dist.
        geom_text(aes(y=.8,
                      x=mean(samp.dist1)*1.2,
                      label=paste0("Variance = ",
                                   round(var(samp.dist1), 4))
                      ),
                  col="blue", size=3
                  )+

        #kurtosis
        geom_text(aes(y=.7,
                      x=mean(samp.dist1)*1.2,
                      label=paste0("Kurtosis = ", round(kurtosis(samp.dist1), 4))
                      ),
                  col="blue", size=3
                  )+
```

```r
        #skew
        geom_text(aes(y=.6,
                      x=mean(samp.dist1)*1.2,
                      label=paste0("Skew = ", round(skewness(samp.dist1), 4))
                      ),
                  col="blue", size=3
                  )+

        labs(title="Sampling Distribution \n Sample Size = 100 \n 1000 Trials",
                x="Means", y="Density")+
        theme_bw()+
        theme(legend.position="none")

# PLOT SIM 2
g5 <- ggplot() + geom_histogram(aes(x=samp.dist2,
                                    y=..density..,
                                    fill=..count..,
                                    col="black"))+
        scale_fill_gradient("Count",
                            low = "green",
                            high = "red")+
        geom_density(aes(samp.dist2), col="black", size=1)+

        #mean of sampling dist.
        geom_vline(aes(xintercept=mean(samp.dist2)),
                   linetype="longdash",
                   col="blue")+
        geom_text(aes(x=mean(samp.dist2)*1.05,
                      y=.9+1.5,
                      label=paste0("Mean  = ",
                                   round(mean(samp.dist2), 4))
                      ),
                  col="blue", size=3
                  )+

        #variance of sampling dist.
        geom_text(aes(y=.8+1.4,
                      x=mean(samp.dist2)*1.05,
                      label=paste0("Variance = ",
                                   round(var(samp.dist2), 4))
                      ),
                  col="blue", size=3
                  )+

        #kurtosis
        geom_text(aes(y=.7+1.3,
                      x=mean(samp.dist1)*1.05,
                      label=paste0("Kurtosis = ", round(kurtosis(samp.dist2), 4))
                      ),
                  col="blue", size=3
                  )+

        #skew
        geom_text(aes(y=.6+1.2,
                      x=mean(samp.dist2)*1.05,
                      label=paste0("Skew = ", round(skewness(samp.dist2), 4))
                      ),
                  col="blue", size=3
                  )+

        labs(title="Sampling Distribution \n Sample Size = 1000 \n 1000 Trials",
                x="Means", y="Density")+
        theme_bw()+
        theme(legend.position="none")

grid.arrange(g4, g5, ncol=2)
```

(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="longdas
        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()
```