Statistical Inference Peer Reviewed Project

Erin Stein

January 5, 2017

Overview

The project consists of two parts:

1. **A simulation exercise.** In this project I investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution is simulated in R with rexp(n, lambda) where lambda = 0.2 is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. I investigate the distribution of averages of 40 exponentials under a thousand simulations, and find that the sample means and sample variances closely approximate the expected values of both under an approximately normal distribution as expected by the CLT.
2. **Basic inferential data analysis.** In this project, I analyze the ToothGrowth data in the R datasets package and use confidence intervals to compare tooth growth by supp and dose. I find that the effects of the supplement on tooth growth are equal under the 95% confidence interval, and that an increase in dose size results in an increase in tooth growth under the 95% confidence interval.

Part I: Simulation Exercise

Simulations

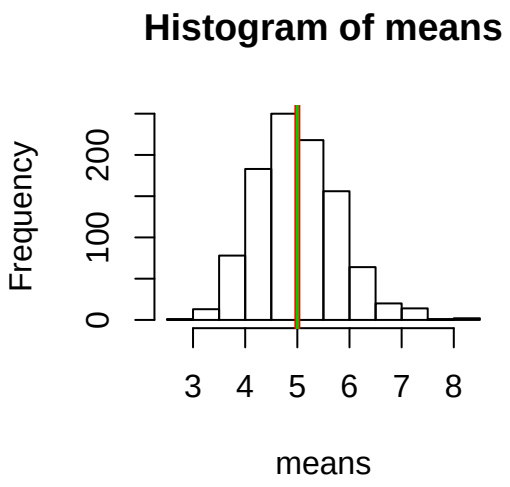
I first simulated 1000 samples of 40 exponentials by using R's exponential distribution and a lambda of 0.2. These samples were then loaded into a 1000 x 40 matrix, where each row corresponds to a sample of 40 exponentials.

```
lam = .2
mn <- 1/.2
sd <- 1/.2
var <- (1/.2)**2
se <- sd/sqrt(40)
sample <- rexp(40000, .2)
mat <- matrix(sample, 1000, 40)
```

Sample Mean versus Theoretical Mean

Next, the mean for each sample (row) was calculated and entered into a vector called 'means' which was then visualized using a simple histogram. The red line indicates the mean of these sample means, and the green line indicates the expected mean of 5.

```
means <- apply(mat, 1, mean)
hist(means)
mns <- mean(means)
abline(v = mean(mns), col = 2, lwd = 3)
abline(v = 5, col = 3, lwd = 1.5)
```



It is easy to see these means are extremely close. In fact, we can do some basic analysis on this sample value to further illustrate.

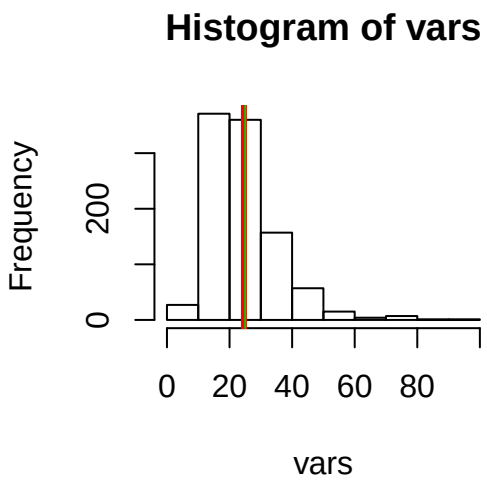
```
mns ##The exact value of the mean of the sample means.
## [1] 5.00026

qu <- (mns-mn)/se  ##The quantile of the mean of sample means from the expected mean of 5.
qu
## [1] 0.0003294424
```

Sample Variance versus Theoretical Variance

Now to analyze the variance of the samples. The variance for each sample (row) was calculated and entered into a vector called 'vars' which was then visualized. The red line indicates the mean of these sample variances, and the green line indicates the expected variance of 25.

```
vars <- apply(mat, 1, var)
hist(vars)
abline(v=mean(vars), col = 2, lwd = 3)
abline(v=var, col = 3, lwd = 1)
```



Once again, it is easy to see these variances are extremely close. The mean of the sample variances below closely matches the expected value of 25.

```
vs <- mean(vars)
vs ##The exact value of the mean of the sample variances.
```

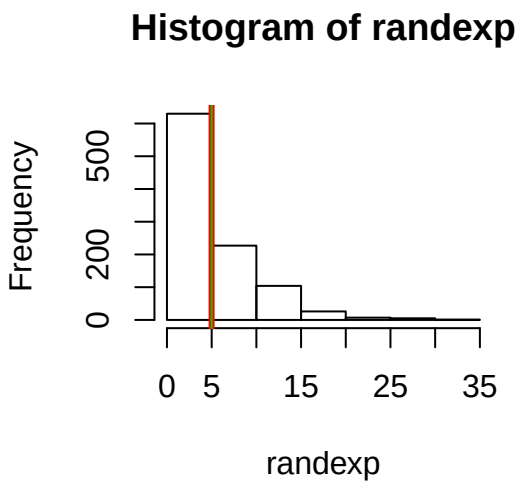
```
## [1] 24.59255
```

Distribution

Note that the histogram of means shown on p.2 is approximately symmetrical about the value 5, which is the expected mean of the distribution.

In contrast, take the following sample, which is simply a histogram of 1000 values drawn randomly from an exponential distribution of lambda 0.2:

```
randexp <- rexp(1000, .2)
hist(randexp)
abline(v = mean(randexp), col = 2, lwd = 3)
abline(v = 5, col = 3, lwd = 1)
```



Note that this distribution is clearly asymmetrical, with a skew to the right. We do not see the even bell-shaped curve of an approximately normal distribution as we do in the first figure.This simple visual exercise shows the strength of the Central Limit Theorem in approximating expected values.

Part II: Basic Inferential Data Analysis Exercise

Load the data and required packages

```
library(ggplot2)
library(dplyr)
data("ToothGrowth")
```

Basic summary of the data

Next, we'll provide a basic summary of the data given in the ToothGrowth dataframe:

```
str(ToothGrowth)
```

```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

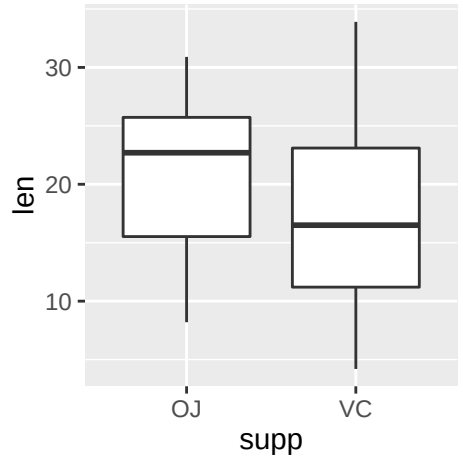
```
ToothGrowth$dose <- as.factor(as.character(ToothGrowth$dose))
summary(ToothGrowth)
```

```
##       len        supp     dose
##  Min.   : 4.20   OJ:30   0.5:20
##  1st Qu.:13.07   VC:30   1  :20
##  Median :19.25           2  :20
##  Mean   :18.81
##  3rd Qu.:25.27
##  Max.   :33.90
```

Exploratory analyses

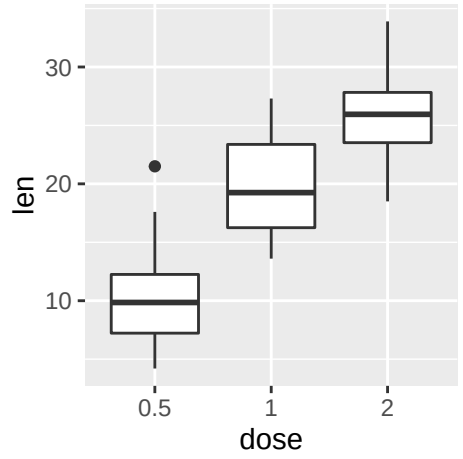
And continue by performing basic exploratory data analyses through plots:

```
ggplot(ToothGrowth, aes(x = supp, y = len)) + geom_boxplot()
```



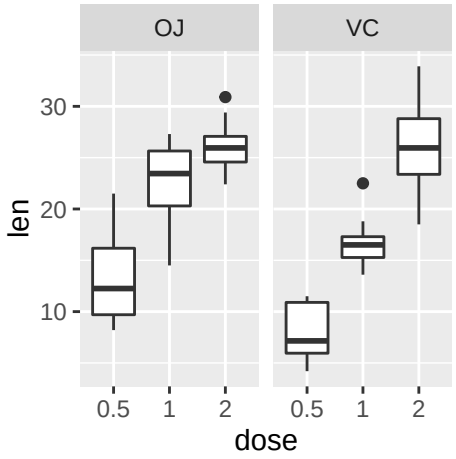

```
##Boxplot of the lengths as a function of the supplement.

ggplot(ToothGrowth, aes(x = dose, y = len)) + geom_boxplot()
```




```
##Boxplot of the lengths as a function of the dosage.

ggplot(ToothGrowth, aes(x = dose, y = len)) + geom_boxplot() + facet_grid(.~supp)
```




```
##Boxplot of the lengths as a function of the dosage and split up by supplement.
```

Compare tooth growth by supp and dose

Let's first determine the mean and standard deviation of the length of tooth growth by supplement.

```
suppgrp <- group_by(ToothGrowth, supp)
supplement <- summarise(suppgrp, Mean.Length = mean(len), SD.Length = sd(len), n = length(supp))
supplement
```

```
## # A tibble: 2 × 4
##     supp Mean.Length SD.Length     n
##   <fctr>       <dbl>     <dbl> <int>
## 1     OJ    20.66333  6.605561    30
## 2     VC    16.96333  8.266029    30
```

Suppose tooth growth length is equivalent despite the supplement taken. That is, Loj = Lvc. Then our alternative hypothesis is that Loj > Lvc.

Let's assume that variance is equivalent across the supplement groups and find the 95% Confidence Interval of the difference in the mean of Loj and Lvc.

```
x2 <- as.numeric(supplement[2, "Mean.Length"])
x1 <- as.numeric(supplement[1, "Mean.Length"])
sd2 <- as.numeric(supplement[2, "SD.Length"])
sd1 <- as.numeric(supplement[1, "SD.Length"])
n2 <- as.numeric(supplement[2, "n"])
n1 <- as.numeric(supplement[1, "n"])
x1-x2+c(-1,1)*qt(.975, n1+n2-2)*sqrt(((n1-1)*sd1**2+(n2-2)*sd2**2)/(n1+n2-2))*sqrt(1/n1+1/n2)
```

```
## [1] -0.1261012  7.5261012
```

Thus, the 95% Confidence Interval runs from -0.13 to 7.5. Since 0 is included in this interval, we fail to reject the null hypothesis that tooth growth length is equivalent across the supplements OJ and VC (Loj = Lvc).

Now, let's consider tooth growth length as a function of dose size. First, determine the mean and standard deviation of the length of tooth growth by dose.

```
dosegrp <- group_by(ToothGrowth, dose)
dosage <- summarise(dosegrp, Mean.Length = mean(len), SD.Length = sd(len), n = length(dose))
dosage
```

```
## # A tibble: 3 × 4
##     dose Mean.Length SD.Length     n
##   <fctr>       <dbl>     <dbl> <int>
## 1    0.5      10.605  4.499763    20
## 2      1      19.735  4.415436    20
## 3      2      26.100  3.774150    20
```

Suppose tooth growth length is equivalent despite the dose taken. That is, Ldose1 = Ldose0.5. Then our alternative hypothesis is that Ldose1 > Ldose0.5.

Let's assume that variance is equivalent across the dose groups and find the 95% Confidence Interval of the difference in the mean of Ldose1 and Ldose0.5.

```
19.735-10.605+c(-1,1)*qt(.975, 38)*sqrt((19*4.415436**2+19*4.499763**2)/38)*sqrt(1/20+1/20)
```

```
## [1]  6.276252 11.983748
```

Both endpoints of the 95% confidence interval are positive; thus, we reject the null hypothesis and conclude that the length taken under dose (1) is greater than the length taken under dose (0.5).

Now, repeat the calculation under the null hypothesis that Ldose2 = Ldose1, with an alternative hypothesis that Ldose2 > Ldose1. Again, assume that variance is equivalent across the dose groups.

```
26.1-19.735+c(-1,1)*qt(.975, 38)*sqrt((19*4.415436**2+19*3.77415**2)/38)*sqrt(1/20+1/20)
```

```
## [1] 3.735613 8.994387
```

Again, both endpoints of the 95% confidence interval are positive; thus, we reject the null hypothesis and conclude that the length taken under dose (2) is greater than the length taken under dose (1).

By the transitive property, a dosage of 2 also results in more growth than a dosage of 0.5.

Conclusion

With the data provided, we maintain the null hypothesis that the supplements OJ and VC result in equivalent tooth growth, and accept the alternative hypothesis that the greater the dose, the longer the tooth growth.

More research should be executed to break down dose by supplement.

‘‘‘