Investigate the exponential distribution in R and compare it with the Central Limit Theorem

Umesh Moolchandani

May 04 2022

Overview

The purpose of this data analysis is to investigate the exponential distribution and compare it to the Central Limit Theorem. For this analysis, the lambda will be set to 0.2 for all of the simulations. This investigation will compare the distribution of averages of 40 exponentials over 1000 simulations.

Simulations

First, we need to generate random data. The below code samples random variables from samples n = 40 and a lambda rate of 0.2 one thousand times. We then store the means and variances of this data in the appropriate variable.

```
ECHO=TRUE
set.seed(1863)
lambda <- 0.2; n <- 40; numSims <- 1000
simData <- data.frame(matrix(rexp(numSims*n,lambda),numSims,n))
simVariances <- data.frame(variances = apply(simData,1, var))
meanVarianceMeans <- mean(simVariances$variances)
```

Sample Mean versus Theoretical Mean

A property of the Exponential Distribution is that the true mean is equal to the inverse of lambda. So with our rate of 0.2, our true mean should be equivalent to 5. Calculating the mean from the simulations with give the sample mean.

```
simMeans <- data.frame(means = apply(simData,1, mean))
meanSampleMeans <- mean(simMeans$means)
theoryMean <- 1/lambda
test <- t.test(simMeans$means, mu = theoryMean)
print(test)
```

```
##
##  One Sample t-test
##
## data:  simMeans$means
## t = 0.24461, df = 999, p-value = 0.8068
## alternative hypothesis: true mean is not equal to 5
```

```
## 95 percent confidence interval:
##  4.955918 5.056637
## sample estimates:
## mean of x
##  5.006277
```

As we can see, 5.01 is roughly equivalent to our theoretical mean of 5. We also show this with a confidence interval of 95%. Given our data, if the true value is equal to 5 we were 80.68% likely to see this result in the data.

Sample Variance versus Theoretical Variance

We repeat the above for seeing how our sample variance compares to the true variance. In an exponential distribution, the variance is equal to the inverse square of lambda. Calculating the variance from the simulation means with give the sample variance.

```
simVariances <- data.frame(variances = apply(simData,1, var))
meanVarianceMeans <- mean(simVariances$variances)
theoryVariance <- (1/lambda)^2
```

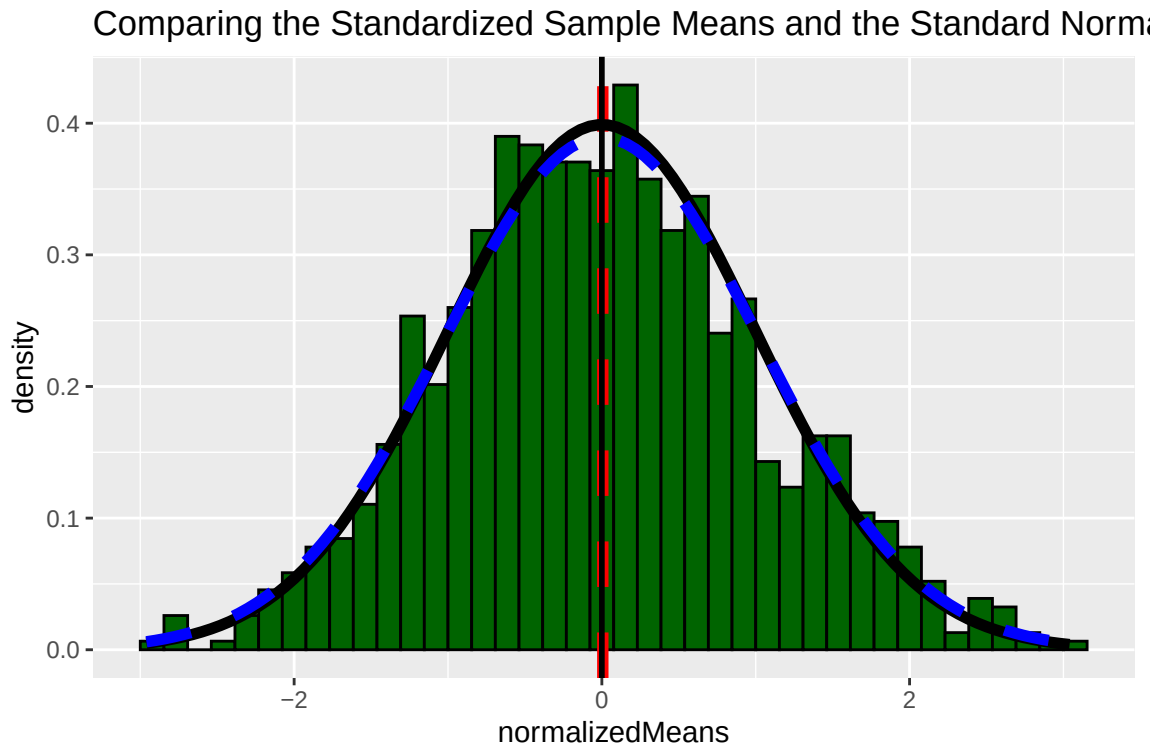
Distribution

To show how the distribution of sample means is normally distributed, we use the below figure. The figure plots the standardized histogram of the sample means, overlaying the density function and a standard normal distribution bell curve.

The sample values for the mean and distribution are given by the dashed lines, while those belonging to the standard normal are given in solid black. As we can see, the means almost exactly overlap and the distribution of sample means almost 1-to-1 follows the standard normal distribution.

```
standardError <- meanSampleMeans/sqrt(n)
standFunc <- function(b){
        round((b - meanSampleMeans)/standardError,2)
}
library(ggplot2)
normalizedMeans <- apply(simMeans,1,FUN = standFunc)
normalizedHist <- ggplot() +
        geom_histogram(aes(x = normalizedMeans, y = ..density..),
                       bins = 40, color = "black", fill = 'darkgreen'
        ) +
        labs(   title = 'Comparing the Standardized Sample Means and the Standard Normal',
                xlab = 'Mean',
                ylab = 'Frequency'
        ) +
        geom_vline( xintercept = meanSampleMeans - (1/lambda),
                    color = "red", linetype = 'dashed', size = 2
        ) +
        geom_vline( xintercept = 0,
                color = "black", size = 1
        ) +
        stat_function( fun = dnorm,
                args = list(mean = 0,sd = 1), colour ='black', size =2
```

```
        ) +
        stat_function( args = list(
                        mean = mean(normalizedMeans),
                        sd = sd(normalizedMeans)),
                        fun = dnorm, color = 'blue', linetype = 'dashed', size = 2
        )
print(normalizedHist)
```



Basic Inferential Data Analysis

Summary

The purpose of the this data analysis is to analyze the ToothGrowth data set by comparing the guinea tooth growth by supplement and dose. First, I will do exploratory data analysis on the data set. Then I will do the comparison with confidence intervals in order to make conclusions about the tooth growth.

1. Load the ToothGrowth data and perform exploratory data analyses

```
library(datasets)
data(ToothGrowth)
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 ...
```

```
head(ToothGrowth)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

```
summary(ToothGrowth)
```

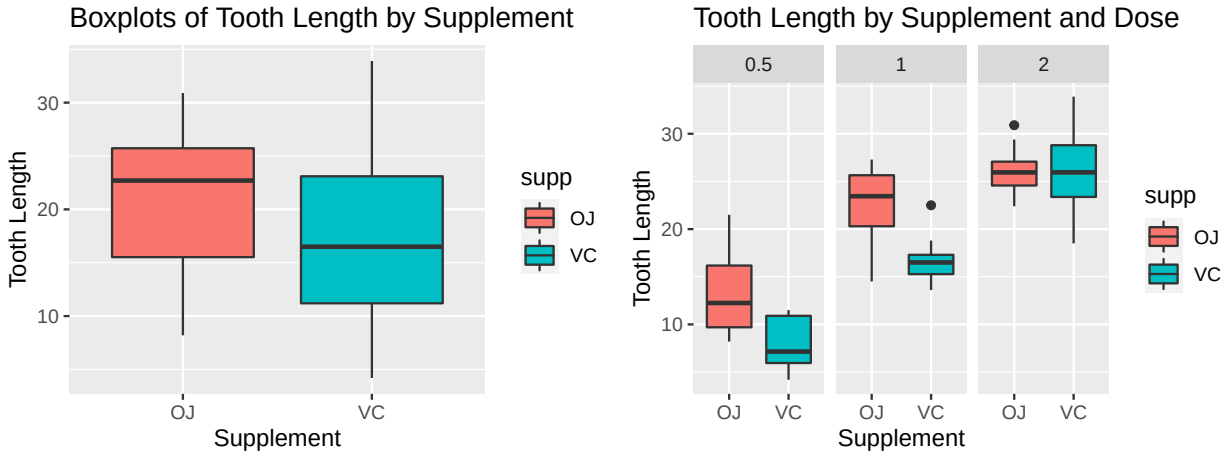
```
##       len        supp         dose
##  Min.   : 4.20   OJ:30   Min.   :0.500
##  1st Qu.:13.07   VC:30   1st Qu.:0.500
##  Median :19.25           Median :1.000
##  Mean   :18.81           Mean   :1.167
##  3rd Qu.:25.27           3rd Qu.:2.000
##  Max.   :33.90           Max.   :2.000
```

2. Exploratory Data Analysis

First, we look at general boxplots of our variables of concern.

```
library(gridExtra)
suppBox <- ggplot(data = ToothGrowth, aes(x = supp, y = len)) + geom_boxplot(aes(fill = supp)) +
        labs(title = "Boxplots of Tooth Length by Supplement", x = "Supplement", y = "Tooth Length")

suppDoseBox <- ggplot(data = ToothGrowth, aes(x = supp, y = len)) + geom_boxplot(aes(fill = supp)) +
        facet_wrap(~ dose) +
        labs(title = "Tooth Length by Supplement and Dose",
             x = "Supplement", y = "Tooth Length")
grid.arrange(arrangeGrob(suppBox,suppDoseBox,ncol=2))
```



3. Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose.

Before we can do some 2 sample t-testing on the dataset we need to split the data into groups with a level of 2 by supplement OJ and VC:

```
dose0.5 <- filter(ToothGrowth, ToothGrowth$dose == 0.5)
dose1.0 <- filter(ToothGrowth, ToothGrowth$dose == 1.0)
dose2.0 <- filter(ToothGrowth, ToothGrowth$dose == 2.0)
t.test(len ~ supp, paired = FALSE, var.equal = FALSE, data = ToothGrowth)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means between group OJ and group VC is not equal to 0
## 95 percent confidence interval:
##  -0.1710156  7.5710156
## sample estimates:
## mean in group OJ mean in group VC
##         20.66333         16.96333
```

All types of doses (0.5 - 2.0 mg/mL) have a p-value lower than 0.05 which means there is a difference in means. The null hypothesis can be rejected (when p is low H0 must go. . .) and there is a significant difference in supplement type.

Then we can test whether OJ or VC per similar dosis of x mg/mL have statistical significant differences in mean length (tooth growth):

```
t.test(ToothGrowth$len ~ ToothGrowth$supp, var.equal = F, data = dose0.5)
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len by ToothGrowth$supp
## t = 1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means between group OJ and group VC is not equal to 0
## 95 percent confidence interval:
##  -0.1710156  7.5710156
## sample estimates:
## mean in group OJ mean in group VC
##         20.66333         16.96333
```

Dose of 0.5 mg/mL have a p-value lower than 0.05 which means there is a difference in means. The null hypothesis can be rejected (when p is low H0 must go. . .) and there is a significant difference in supplement type with the chosen dose.

```
t.test(ToothGrowth$len ~ ToothGrowth$supp, var.equal = F, data = dose1.0)
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  ToothGrowth$len by ToothGrowth$supp
## t = 1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means between group OJ and group VC is not equal to 0
## 95 percent confidence interval:
##  -0.1710156  7.5710156
## sample estimates:
## mean in group OJ mean in group VC
##         20.66333         16.96333
```

Dose of 1.0 mg/mL have a p-value lower than 0.05 which means there is a difference in means. The null hypothesis can be rejected (when p is low H0 must go...) and there is a significant difference in supplement type with the chosen dose.

```
t.test(ToothGrowth$len ~ ToothGrowth$supp, var.equal = F, data = dose2.0)
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len by ToothGrowth$supp
## t = 1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means between group OJ and group VC is not equal to 0
## 95 percent confidence interval:
##  -0.1710156  7.5710156
## sample estimates:
## mean in group OJ mean in group VC
##         20.66333         16.96333
```

Dose of 2.0 mg/mL have a p-value greater than 0.05 which means there is NOT a difference in means. The null hypothesis can **NOT** be rejected and there is NOT a significant difference in supplement type with the chosen dose.

4. State your conclusions and the assumptions needed for your conclusions.

The t-test assumes random and independent sampling (paired = FALSE), normality of data distribution, adequacy of sample size, and equality of variance (var.equal = TRUE). From the tests it seems that supplement type have a significant difference in mean tooth length (growth) except when dose is high (2.0 mg/mL).

The conclusion is when the dose is 0.5 or 1.0 there is a difference between the teeth growth after taking OJ and VC, while when the dose is 2.0, there is no difference between the teeth growth after taking OJ and VC. The assumption needed is we first assumed the whole population is normally distributed, then we assumed the population is normally distributed under each dose.