Simulating Exponentials

Jade Forlani-Brennan

10/02/2020

Overview

This project report is in 2 sections:

1. Exponential Distrubtion Sampling Simulations
2. Analysing Tooth Growth Data

In the first section of this report I simulate an exponential distrubtion with parameter lambda of 0.2. And repeat this simulation 40 times, taking the mean and variance of each of the 40 simulations. The theoretical and sample mean is compared. And the same for the theoretical and sample variance. Lastly the normal distrubtion shape of the samples is shown.

In the second section Tooth Growth Data is explored and summarised. Finally, tooth growth by supp and dose is compared with confidence intervals.

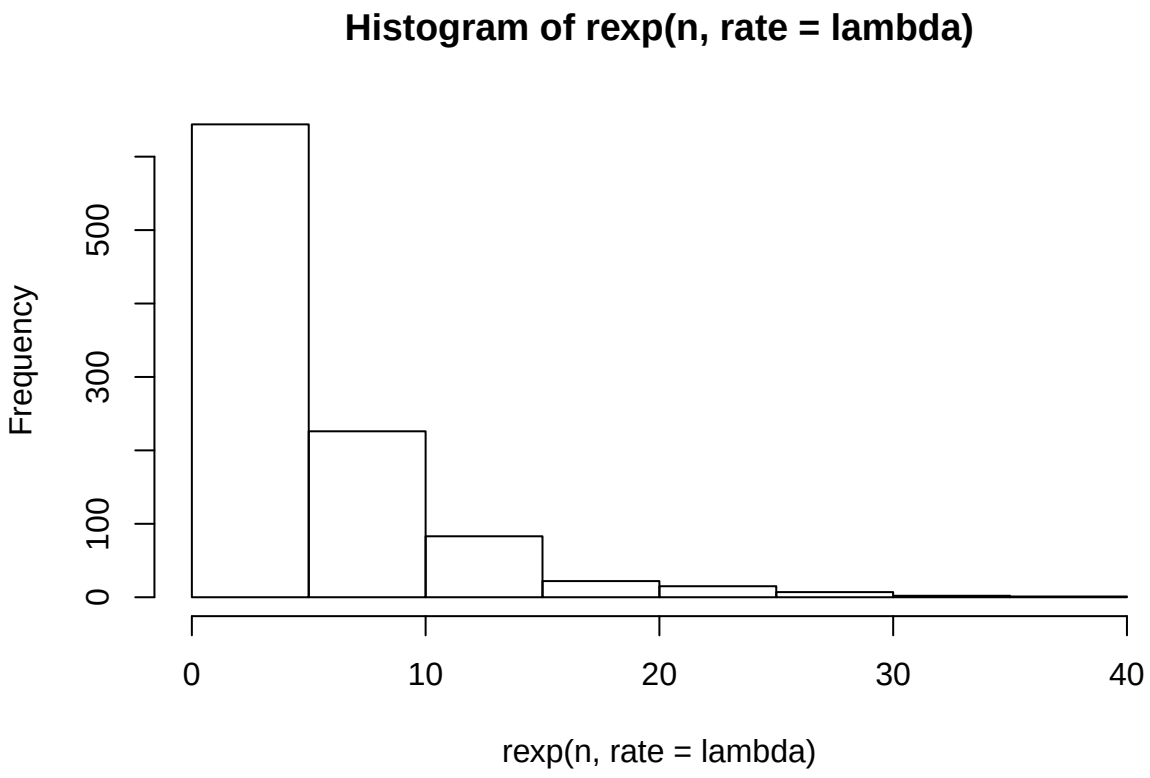
Exponential Distrubtion Sampling Simulations

```
lambda = 0.2 # number
n = 1000 #  size of each distribution, number of data points
B = 40 # number of simulations
```

Simulation

First, creating the histogram of a single simulation consisting of **1000** data points.

```
hist(rexp(n,rate = lambda))
```



We see the tell tale shape of an expoentially decaying distrubtion. Most of the values generated are closer to 0, with less and less larger numbers.

Next, repeating this simulation **40** times and looking at the mean for each simulation and storing ths result in the means variable.

```
means = NULL
for (i in 1:n) means = c(means, mean(rexp(B,rate=lambda)))
```

This is then repeated again **40** times this time looking at the variance of each simulation, storing this result in the vars variable.

```
vars = NULL
for (i in 1:n) vars = c(vars, var(rexp(B,rate=lambda)))
```

Sample Mean vs. Theoretical Mean

```
meanTheory = 1/lambda # theoretical mean
meanSample = means[1] # sample mean
meanAvg = mean(means) # average of sample means
```

The theoretical mean is **5** vs the sample mean of **5.5793661**. Which is very close with a single sample of **1000** data points but not exact. However, if we average all the samples out the average simulated mean is **4.9758972** which is very close to the theoretical number. However, still slightly difference because of simulation or *"Monte Carlo"* error.

Sample Variance vs. Theoretical Variance

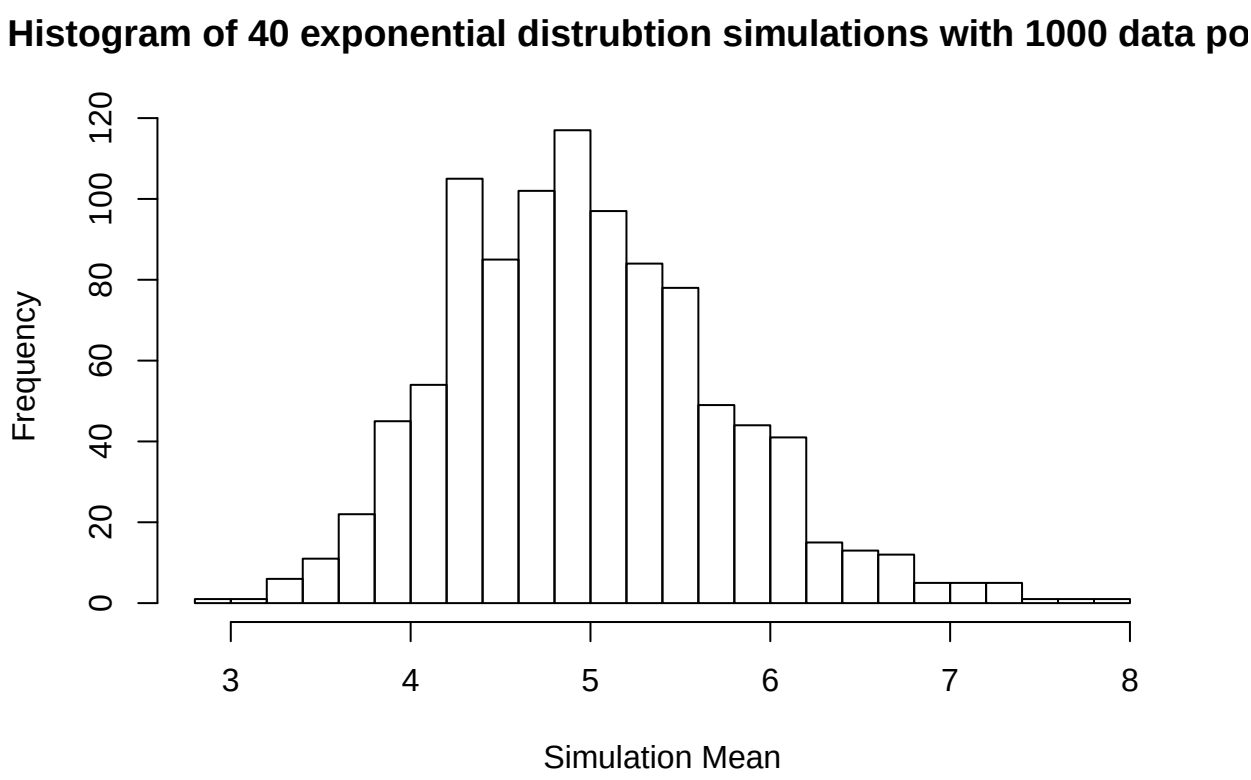
```
sdTheory = 1/lambda # theoretical stmeanTheoryandard deviation
varTheory = sdTheory^2 # theoretical variance
varSample = vars[1] # sample variance
varAvg = mean(vars) # average of sample variances
```

The theoretical variance is the sample standard deviation squared equal to **25**. Compare this with the sample variance of **34.0916556**. Which is close but not the sample despite a single sample of **1000** data points. However, if we average all the samples out the average simulated variane is **24.7935172** which is very close to the theoretical number. However, still slightly difference because of simulation or *"Monte Carlo"* error.

Sample Means Normally Distrubted

Looking at a histogram of the mean of each of the **40** simulations we see what looks much closer to a normal distrubtion shape.

```
hist(means,
     breaks = 20,
     main = paste("Histogram of", B, "exponential distrubtion simulations with",n,"data points"),
     xlab = "Simulation Mean")
```



We can tell thi from the tell tale bell curve shape centered at approximatly the theoretical mean of **5** which thins out at the tails.

Analysing Tooth Growth Data

Exploratory Data Analysis

We begin by loading the ToothGrowth data from the R datasets package.

```
library(datasets) # load data sets
toothData <- ToothGrowth # load Tooth Growth Dataset
```

Summary

Let's have a look at the dimensions of the dataset first.

```
dims <- dim(toothData)
```

The dataset consts of 60 observations and 3 variables. Let's have a look at the variables next.

```
str(toothData)
```

```
## '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 ...
```

We see there is a numeric len, 2 factors for supp as either "OJ" or "VC". And a dose.

And the first 10 rows.

```
head(toothData)
```

```
##    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
```

Having a quick look at the doses which are listed as a numeric.

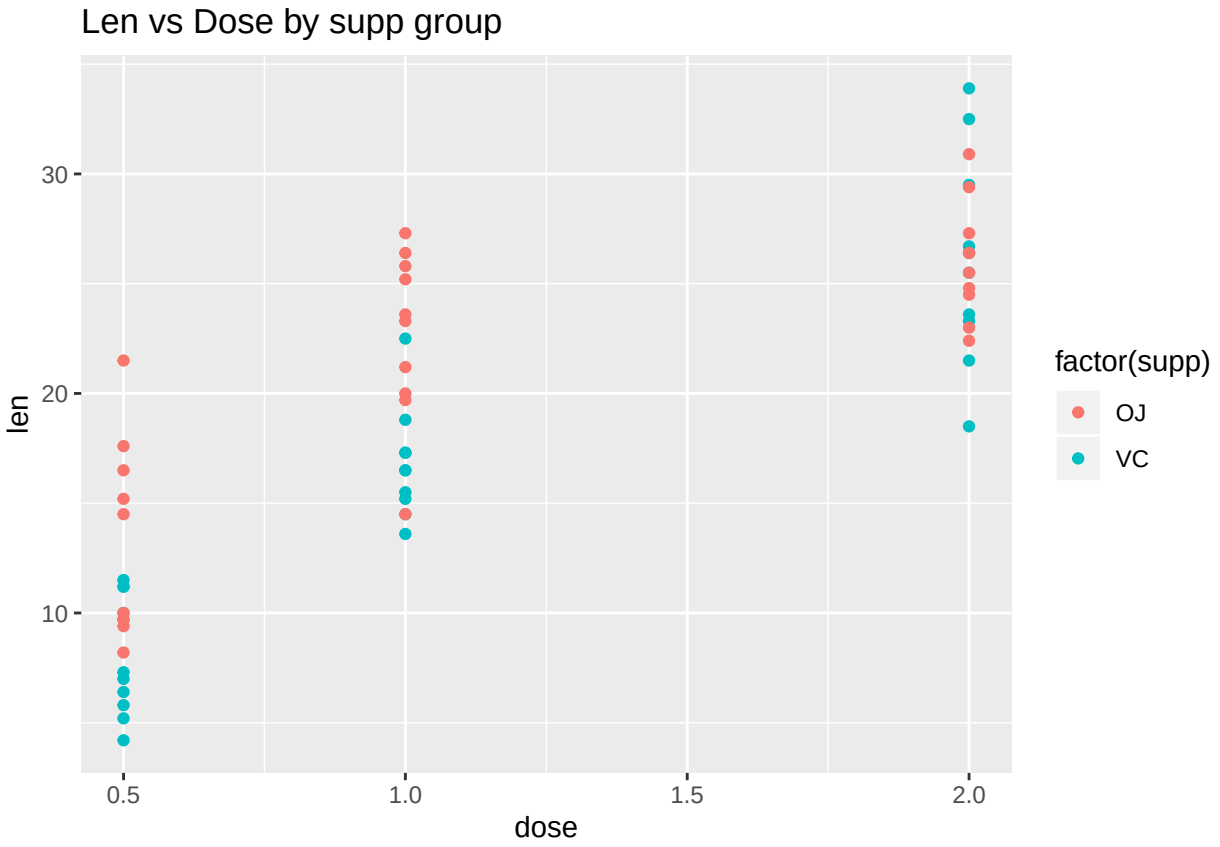
```
table(toothData$dose)
```

```
##
## 0.5   1   2
##  20  20  20
```

But in fact we only see one of 3 options, 0.5, 1, or 2. And each appears 20 times.

Next let's look at each group seperately on the same plot.

```
library(ggplot2)
g <- ggplot(toothData, aes(dose,len))
g + geom_point(aes(color=factor(supp))) + labs(title = "Len vs Dose by supp group")
```




```
splitTooth <- split(toothData, toothData$supp)
OJmean <- mean(splitTooth$OJ$len)
VCmean <- mean(splitTooth$VC$len)
meanDiff <- OJmean-VCmean
```

Looking at the means of both (averaged across all doses) we see

- OJ had a mean of **20.66**
- VC hada mean of **16.96**

And the difference between the two is that OJ was **3.7** larger than VC

Hypothesis Testing

For a hypthesis test I will use the "Permutation Test" method covered in week 4 module 13: Resampling.

The null hypothesis then is that VC and OJ are the same. I.e. any variation in length is random. The alternative hypothesis is that there is a difference.

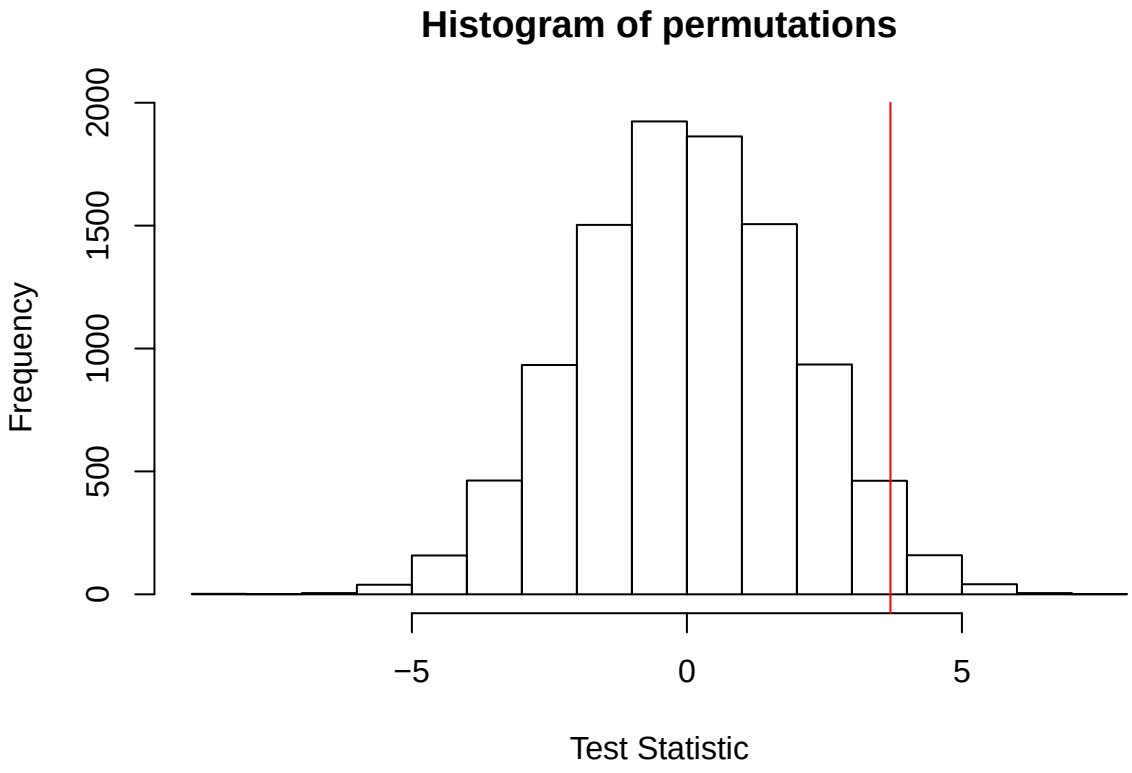
Our test statistic is the **3.7** we found before, the difference in means.

We will next permuate the group labels and recalculate this test statistic each time. We reject the null hypothesis if we fail to find a permutation that offers us a more extreme test statistic than **3.7**.

```
B <- 10000 # number of permutations to simulate
y <- toothData$len
label <- as.character(toothData$supp)
testStat <- function(df,l) mean(df[l=="OJ"]) - mean(df[l=="VC"])
observedStat <- testStat(y,label)
permutations <- sapply(1 : B, function(i) testStat(y,sample(label)))
```

Let's jump in and see how our permutations went, adding the observed statistic as a red vertical line at **3.7**.

```
hist(permutations, xlab = "Test Statistic")
abline(v=observedStat,col="red")
```



We can see that yes it is included within the bounds of permutations. So it is possible that there is not significant difference between OJ and VC in terms of tooth growth.

More formally, we estimate the p-value as the proportion of permutations which were more extreme (greater than) the observerd statistic amoung all the simulations.

```
pvalue <- mean(permutations > observedStat)
```

This gives a p-value of **0.0307** or **'r round(pvalue*100,0)`** percent. Meaning at a 90% confidence interval we would reject and at a 95% confidence interval we would accept.

Or in plain english we could say we are 90% confidence that OJ is better for tooth growth than VC. But we could **not say* we are 95% confident.

Assumptions & Conclusions

My conclusion is that I am 90% confdient OJ is better for tooth growth than VC. However, this is predicated on the following assumptions

- Len is in the same units for all observations
- Dose is in the same units for all observations
- Len is correlated only to supp and there is not a hidden confounding variable
- Each observation was independant of one another
- Len can be averaged across all dosages to get a total average

In particular this last assumption is one worth investigating further. Further investigation could look at each dosage rate seperately rather than aggregating and averaging them together.