# Statistical Inference Assignment

*Brian Stroh*

*September 26, 2018*

```
library(dplyr)
library(reshape2)
library(ggplot2)
library(gridExtra)
set.seed(51)
```

## Part 1: Exponential Simulation

### Overview

In the Exponential Simulation section of this report, we will explore how the Central Limit Theorem can be applied to a non-normal variable. We will create 1000 samples of 40 random instances from an exponential distribution. Then we will plot a histogram of the mean and variance of each of the 1000 samples and compare these to the equivalent normal curves.

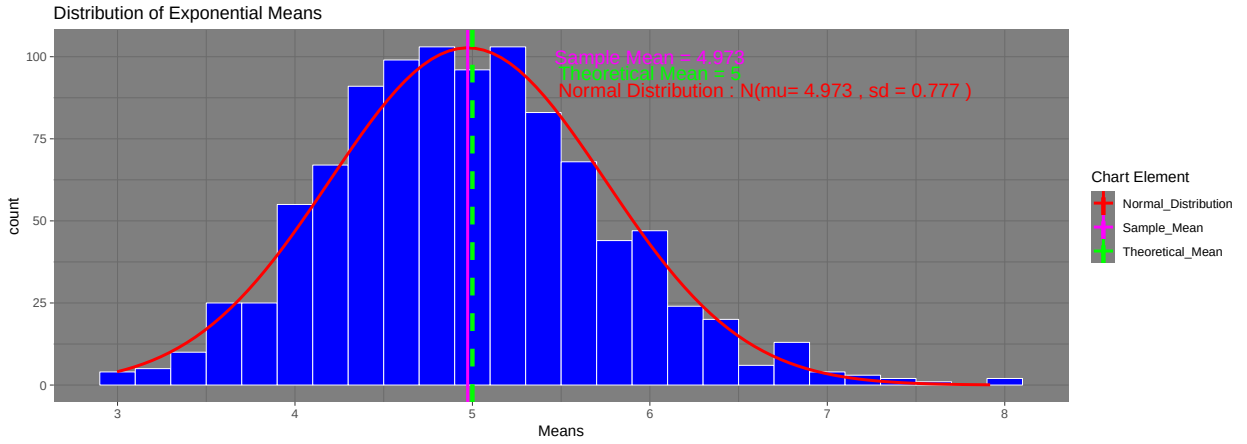
### Simulations

```
mySample<-matrix(rexp(n=40*1000, rate = .2),40,1000)
Means<-colSums(mySample)/40
#Variance calculated using the formula : Var[x] = E[X^2] - (E[x])^2
#source: https://en.wikipedia.org/wiki/Variance
Variances<-colSums(mySample^2)/40-(colSums(mySample)/40)^2
mySampleStats<-tbl_df(cbind(Means,Variances))
```

### Sample Mean versus Theoretical Mean

```
ggplot(mySampleStats, aes(x = Means, mean = mean(mySampleStats$Means),
                          sd = sqrt(var(mySampleStats$Means)),
                          binwidth = .2, n = nrow(mySampleStats))) +
      theme_dark() +
      geom_histogram(binwidth = .2, colour = "white", fill = "blue", size = 0.1) +
      stat_function(fun = function(x) dnorm(x, mean = mean(mySampleStats$Means),
                                            sd = sqrt(var(mySampleStats$Means))) *
                          nrow(mySampleStats) * .2,
                    aes(color = "Normal_Distribution"), size =1) +
      geom_vline(aes(xintercept = (1/.2), color = "Theoretical_Mean"), linetype="dashed",
                 size=1.5) +
      geom_vline(aes(xintercept = mean(mySampleStats$Means), color = "Sample_Mean"),
                 size=1) +
      scale_color_manual(name = "Chart Element", values = c(Theoretical_Mean = "green",
                                                            Sample_Mean = "magenta",
                                                            Normal_Distribution = "red"))+
      ggtitle("Distribution of Exponential Means") +
      annotate("text",label=paste("Sample Mean =",round(mean(mySampleStats$Means),3)),
               x=6,y=100,size=5,color="magenta")+
      annotate("text",label=paste("Theoretical Mean =",1/.2),
               x=6,y=95,size=5,color="green")+
```

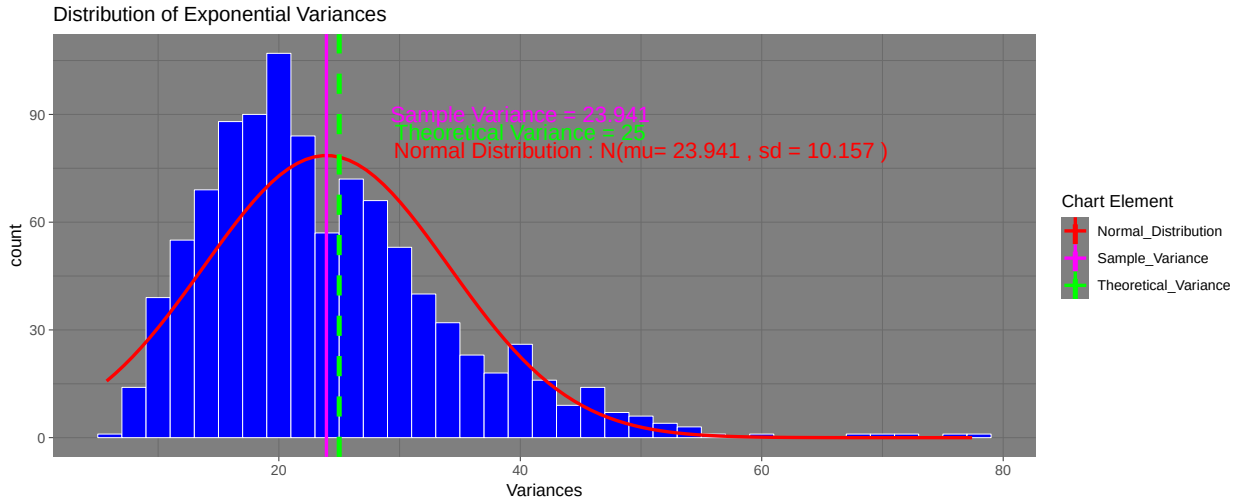
```
      annotate("text",label=paste("Normal Distribution : N(mu=",
                                  round(mean(mySampleStats$Means),3),
                                  ", sd =",round(sqrt(var(mySampleStats$Means)),3),")"),
               x=6.65,y=90,size=5,color="red")
```




It can be seen from the figure above that the Central Limit Theorem applies well to our exponential sample means. The Theoretical Mean is quite close to the Sample Mean and the shape of the histogram of the means is very Gaussian. This distribution is approximately normal.

### Sample Variance versus Theoretical Variance

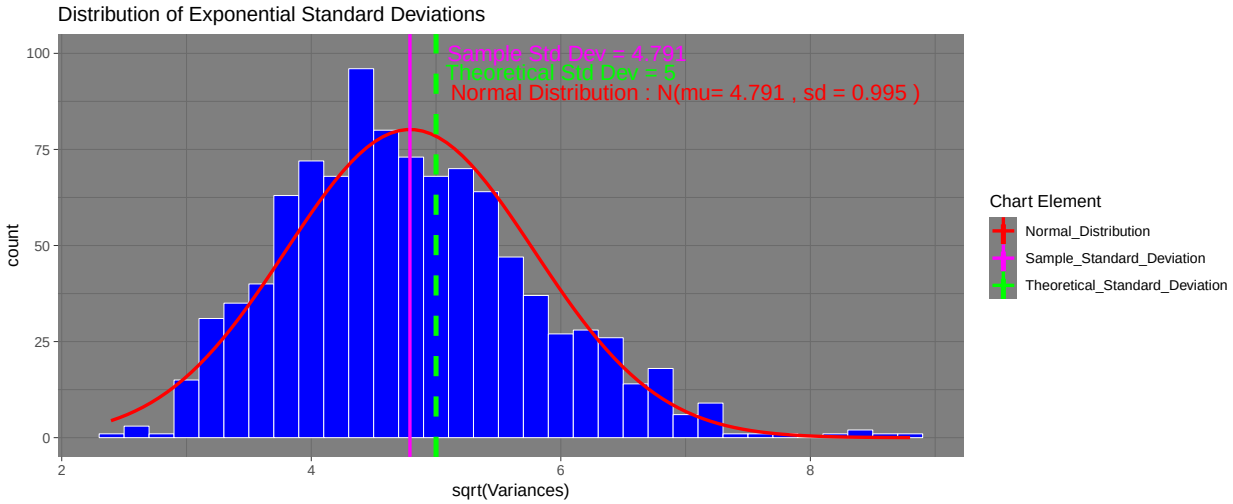
```
ggplot(mySampleStats, aes(x = Variances, mean = mean(mySampleStats$Variances),
                          sd = sqrt(var(mySampleStats$Variances)),
                          binwidth = 2, n = nrow(mySampleStats))) +
      theme_dark() +
      geom_histogram(binwidth = 2, colour = "white", fill = "blue", size = 0.1) +
      stat_function(fun = function(x) dnorm(x, mean = mean(mySampleStats$Variances),
                                            sd = sqrt(var(mySampleStats$Variances))) *
                          nrow(mySampleStats) * 2,
                    aes(color = "Normal_Distribution"), size = 1) +
      geom_vline(aes(xintercept = (1/.2)^2, color = "Theoretical_Variance"),
                 linetype="dashed", size=1.5) +
      geom_vline(aes(xintercept = mean(mySampleStats$Variances),
                     color = "Sample_Variance"), size=1) +
      scale_color_manual(name = "Chart Element", values = c(Theoretical_Variance= "green",
                                                            Sample_Variance= "magenta",
                                                            Normal_Distribution= "red")) +
      ggtitle("Distribution of Exponential Variances") +
      annotate("text",label=paste("Sample Variance =",
                                  round(mean(mySampleStats$Variances),3)),
               x=40,y=90,size=5,color="magenta")+
      annotate("text",label=paste("Theoretical Variance =",(1/.2)^2),
               x=40,y=85,size=5,color="green") +
      annotate("text",label=paste("Normal Distribution : N(mu=",
                                  round(mean(mySampleStats$Variances),3),
                                  ", sd =",
                                  round(sqrt(var(mySampleStats$Variances)),3),")"),
               x=50,y=80,size=5,color="red")
```




Comparing the figure above to the equivalent figure for the means shows that the Central Limit Theorem does not apply as well to the variance of our exponential samples. The Theoretical Variance is still close to the Sample Variance, but the shape of the histogram of the variances is not nearly as Gaussian shaped as the histogram of the means was. Perhaps this is due to the unit scale of the variance relative to the mean. We will attempt to reduce some of the skew shown in the above figure by instead plotting the standard deviations of each of the samples.

### Sample StandardDeviation versus Theoretical StandardDeviation

```
ggplot(mySampleStats, aes(x = sqrt(Variances), mean = mean(sqrt(mySampleStats$Variances)),
                          sd = sqrt(var(sqrt(mySampleStats$Variances))),
                          binwidth = .2, n = nrow(mySampleStats))) +
      theme_dark() +
      geom_histogram(binwidth = .2, colour = "white", fill = "blue", size = 0.1) +
      stat_function(fun = function(x) dnorm(x, mean = mean(sqrt(mySampleStats$Variances)),
                                            sd = sqrt(var(sqrt(mySampleStats$Variances))))
                    * nrow(mySampleStats) * .2,
                    aes(color = "Normal_Distribution"), size = 1) +
      geom_vline(aes(xintercept = (1/.2), color = "Theoretical_Standard_Deviation"),
                 linetype="dashed", size=1.5) +
      geom_vline(aes(xintercept = mean(sqrt(mySampleStats$Variances)),
                     color = "Sample_Standard_Deviation"), size=1) +
      scale_color_manual(name = "Chart Element",
                         values = c(Theoretical_Standard_Deviation = "green",
                                    Sample_Standard_Deviation = "magenta",
                                    Normal_Distribution = "red")) +
      ggtitle("Distribution of Exponential Standard Deviations")+
      annotate("text",label=paste("Sample Std Dev =",
                                  round(mean(sqrt(mySampleStats$Variances)),3)),
               x=6.05,y=100,size=5,color="magenta") +
      annotate("text",label=paste("Theoretical Std Dev =",1/.2),
               x=6,y=95,size=5,color="green")+
      annotate("text",label=paste("Normal Distribution : N(mu=",
                                  round(mean(sqrt(mySampleStats$Variances)),3),
                                  ", sd =",
                                  round(sqrt(var(sqrt(mySampleStats$Variances))),3),")"),
               x=7,y=90,size=5,color="red")
```




The Standard Deviation distribution appears to be quite Gaussian and is approximately normal. This means that the Central Limit Theorem can be applied to summary statistics other than just the mean.

## Part 2: ToothGrowth Inferential Data Analysis

### ToothGrowth Data Summaries

```
#~~~~~~~~~~~~~~~~~~~~~~~~RESAMPLING~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

data("ToothGrowth")

#Creating subsets for hypothesis testing and resampling
d.5<-filter(ToothGrowth,dose==.5)
d1<-filter(ToothGrowth,dose==1)
d2<-filter(ToothGrowth,dose==2)
sOJ<-filter(ToothGrowth,supp=="OJ")
sVC<-filter(ToothGrowth,supp=="VC")

#Summary for full dataset
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
```

```
#Summary for supplement dose of .5 mg
summary(d.5)
```

```
##       len         supp         dose
##  Min.   : 4.200   OJ:10   Min.   :0.5
##  1st Qu.: 7.225   VC:10   1st Qu.:0.5
##  Median : 9.850           Median :0.5
##  Mean   :10.605           Mean   :0.5
##  3rd Qu.:12.250           3rd Qu.:0.5
```

```
##  Max.   :21.500           Max.   :0.5
```

```
#Summary for supplement dose of 1 mg
summary(d1)
```

```
##       len        supp         dose
##  Min.   :13.60   OJ:10   Min.   :1
##  1st Qu.:16.25   VC:10   1st Qu.:1
##  Median :19.25           Median :1
##  Mean   :19.73           Mean   :1
##  3rd Qu.:23.38           3rd Qu.:1
##  Max.   :27.30           Max.   :1
```

```
#Summary for supplement dose of 2 mg
summary(d2)
```

```
##       len        supp         dose
##  Min.   :18.50   OJ:10   Min.   :2
##  1st Qu.:23.52   VC:10   1st Qu.:2
##  Median :25.95           Median :2
##  Mean   :26.10           Mean   :2
##  3rd Qu.:27.82           3rd Qu.:2
##  Max.   :33.90           Max.   :2
```

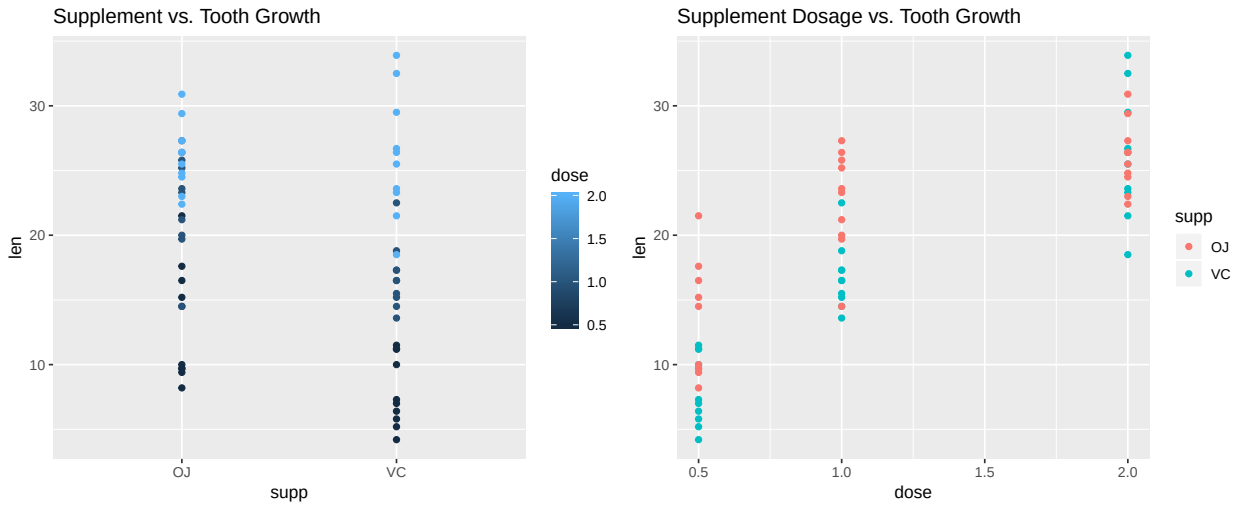
```
#Summary for supplement OJ
summary(sOJ)
```

```
##       len        supp         dose
##  Min.   : 8.20   OJ:30   Min.   :0.500
##  1st Qu.:15.53   VC: 0   1st Qu.:0.500
##  Median :22.70           Median :1.000
##  Mean   :20.66           Mean   :1.167
##  3rd Qu.:25.73           3rd Qu.:2.000
##  Max.   :30.90           Max.   :2.000
```

```
#Summary for supplement VC
summary(sVC)
```

```
##       len        supp         dose
##  Min.   : 4.20   OJ: 0   Min.   :0.500
##  1st Qu.:11.20   VC:30   1st Qu.:0.500
##  Median :16.50           Median :1.000
##  Mean   :16.96           Mean   :1.167
##  3rd Qu.:23.10           3rd Qu.:2.000
##  Max.   :33.90           Max.   :2.000
```

```
p1<-ggplot(ToothGrowth, aes(x= supp, y=len, color=dose), geom = "dotplot") + geom_point()+
      ggtitle("Supplement vs. Tooth Growth")
p2<-ggplot(ToothGrowth, aes(x= dose, y=len, color=supp), geom = "dotplot") + geom_point()+
      ggtitle("Supplement Dosage vs. Tooth Growth")
grid.arrange(p1, p2, ncol = 2)
```




In the Supplement vs. Tooth Growth plot above, we can see that the range for the OJ supplement lies entirely within the range of the VC supplement. This will likely cause problems when trying to differentiate the impacts of each of the supplements on tooth growth.
In the Supplement Dosage vs. Tooth Growth plot above, so can see a noticeable shift in tooth gowth as the dosage increases, although there is still a noticeable amount of overlap in the ranges.
Conducting a quick resampling exercise should help to clear up whether the distributions are in fact distinct or not.

### ToothGrowth Resampling Visualizations

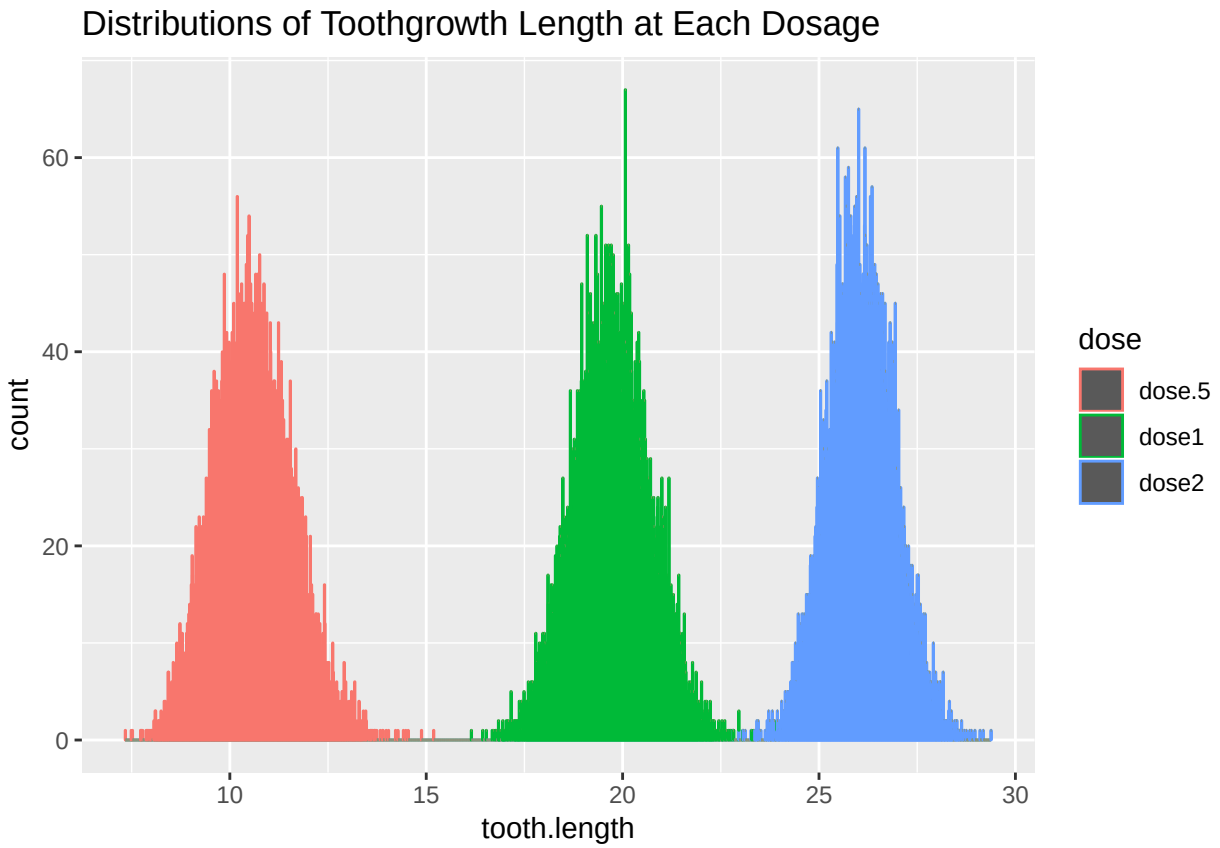
```
reSample_Dose<-function(n=1000,yVar=3){
      rows.5<-sample(1:nrow(d.5),size=n,replace=TRUE)
      rows1<-sample(1:nrow(d1),size=n,replace=TRUE)
      rows2<-sample(1:nrow(d2),size=n,replace=TRUE)

      dSamp.5<-d.5[rows.5,c(1,yVar)]
      dSamp1<-d1[rows1,c(1,yVar)]
      dSamp2<-d2[rows2,c(1,yVar)]

      d.5_avg<-mean(dSamp.5$len)
      d1_avg<-mean(dSamp1$len)
      d2_avg<-mean(dSamp2$len)

      return(c(d.5_avg,d1_avg,d2_avg))
}

mySample<-data.frame(t(replicate(10000,reSample_Dose(n=20))))
names(mySample) <- c("dose.5","dose1","dose2")
mySampleMelt<-melt(mySample)
names(mySampleMelt)<-c("dose","tooth.length")
ggplot(mySampleMelt,aes(x=tooth.length, color = dose)) +
      geom_histogram(binwidth = .01) +
      ggtitle("Distributions of Toothgrowth Length at Each Dosage")
```




As we can see from the 10,000 resamples, it is likely that there is a different average tooth growth at each dose level. There is very little overlap in these resampled averages, so will probably reject a hypothesis that the average tooth growth is the same across all dose levels. A resample size of 20 was chosen for this exercise because there are 20 data points for each dose level in the original dataset.

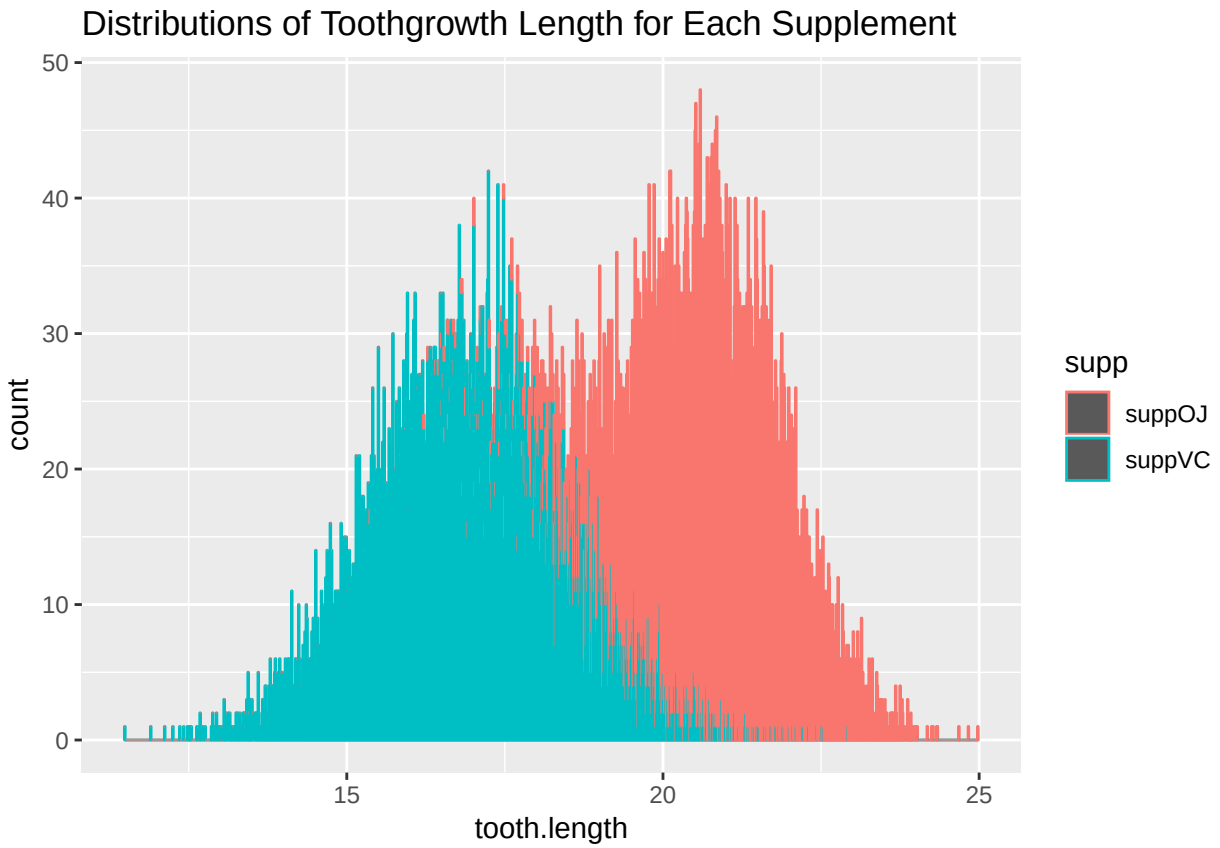
```
reSample_Supp<-function(n=1000,yVar=2){
      rowsOJ<-sample(1:nrow(sOJ),size=n,replace=TRUE)
      rowsVC<-sample(1:nrow(sVC),size=n,replace=TRUE)

      sSampOJ<-sOJ[rowsOJ,c(1,yVar)]
      sSampVC<-sVC[rowsVC,c(1,yVar)]

      OJ_avg<-mean(sSampOJ$len)
      VC_avg<-mean(sSampVC$len)

      return(c(OJ_avg,VC_avg))
}

mySample2<-data.frame(t(replicate(10000,reSample_Supp(n=30))))
names(mySample2) <- c("suppOJ","suppVC")
mySampleMelt2<-melt(mySample2)
names(mySampleMelt2)<-c("supp","tooth.length")
ggplot(mySampleMelt2,aes(x=tooth.length, color = supp)) +
      geom_histogram(binwidth = .01) +
      ggtitle("Distributions of Toothgrowth Length for Each Supplement")
```




As we can see from the 10,000 resamples, it is unlikely that we will reject a hypothesis that the average tooth growth is the same across both supplement levels. There is significant overlap in the resampled tooth growth distributions for these two supplements. A resample size of 30 was chosen for this exercise because there are 30 data points for each supplement level in the original dataset.

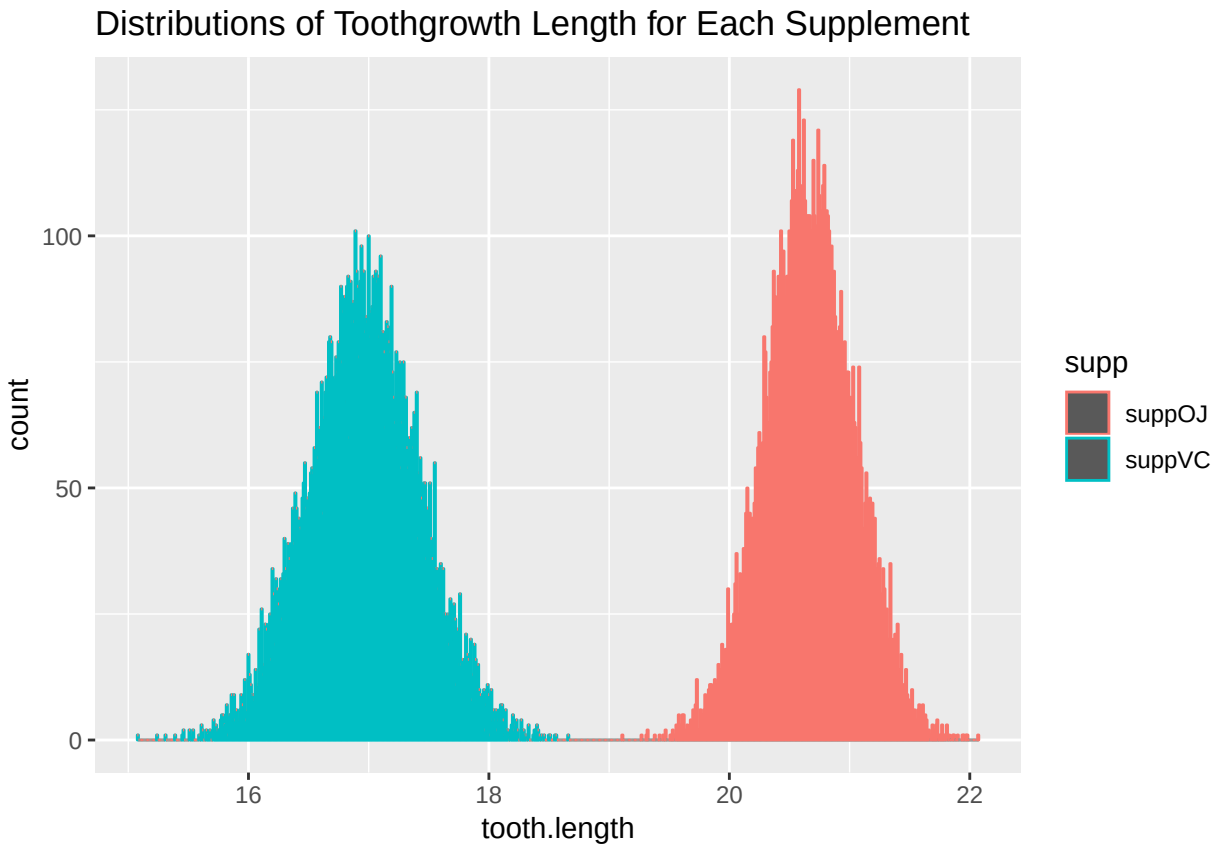
```
reSample_Supp<-function(n=1000,yVar=2){
      rowsOJ<-sample(1:nrow(sOJ),size=n,replace=TRUE)
      rowsVC<-sample(1:nrow(sVC),size=n,replace=TRUE)

      sSampOJ<-sOJ[rowsOJ,c(1,yVar)]
      sSampVC<-sVC[rowsVC,c(1,yVar)]

      OJ_avg<-mean(sSampOJ$len)
      VC_avg<-mean(sSampVC$len)

      return(c(OJ_avg,VC_avg))
}

mySample2<-data.frame(t(replicate(10000,reSample_Supp(n=300))))
names(mySample2) <- c("suppOJ","suppVC")
mySampleMelt2<-melt(mySample2)
names(mySampleMelt2)<-c("supp","tooth.length")
ggplot(mySampleMelt2,aes(x=tooth.length, color = supp)) +
      geom_histogram(binwidth = .01) +
      ggtitle("Distributions of Toothgrowth Length for Each Supplement")
```




As evidenced from this figure, it is likely that we would reject the null hypothesis of the two supplements having the same average tooth growth length **if our sample size was 10 times larger** than in the original dataset, assuming that the same distribution of tooth growth would result.

### ToothGrowth T-Testing

**Assumptions:**

1. The ToothGrowth dataset contains all available information.
2. The experiment conducted was done in a scientific manner.
    - This includes the assumption that there was no overlap in test subjects being given varying doses or varying supplements.

```
#~~~~~~~~~~~~~~~~~~~~~~~~HYPOTHESIS TESTING~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
#Null Hypothesis is: The average tooth growth is identical across all doses and all supplements

#T-testing the alternative hypothesis: The average tooth growth changes as supplement dosage changes.
t.test(len~dose,paired=FALSE,data=rbind(d1,d.5))
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -6.4766, df = 37.986, p-value = 1.268e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.983781  -6.276219
## sample estimates:
```

```
## mean in group 0.5   mean in group 1
##            10.605            19.735
```

```
t.test(len~dose,paired=FALSE,data=rbind(d2,d1))
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -4.9005, df = 37.101, p-value = 1.906e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.996481 -3.733519
## sample estimates:
## mean in group 1 mean in group 2
##          19.735          26.100
```

```
t.test(len~dose,paired=FALSE,data=rbind(d2,d.5))
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -11.799, df = 36.883, p-value = 4.398e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.15617 -12.83383
## sample estimates:
## mean in group 0.5   mean in group 2
##            10.605            26.100
```

### Dosage Conclusions

We tested each of the dosage subsets against each other, and with each set of doses we can be 95% confident that the average tooth growth is different between the dose amounts. Our P value is much smaller than 5% in each case. Thus, we reject the null hypothesis in favor of the alternative hypothesis that average tooth growth varies by supplement dose amount.

```
#T-testing the alternative hypothesis: The average tooth growth changes as supplement changes.
t.test(len~supp,paired=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 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
```

### Supplement Conclusions

We tested the two supplement subsets against each other, and we cannot be 95% confident that the average tooth growth is different between the supplements. This is because:

1. Our P-Value for this test is greater than our alpha 5% (1 - 95%).
2. Our 95% confidence interval contains the possibility of a 0 difference in average tooth growth as a function of supplement type.
Thus, we fail to reject the null hypothesis in favor of the alternative hypothesis that average tooth growth varies by supplement.