JHU Statistical Inference - course project

Harm Lammers

22 september 2016

Contents

Overview

This report describes the outcome of my search in R and the course material to submit a solution to the course project as described above.

Part 1: Simulation Exercise Instructions

In this project you will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. Set lambda = 0.2 for all of the simulations.

You will investigate the distribution of averages of 40 exponentials. Note that you will need to do a thousand simulations.

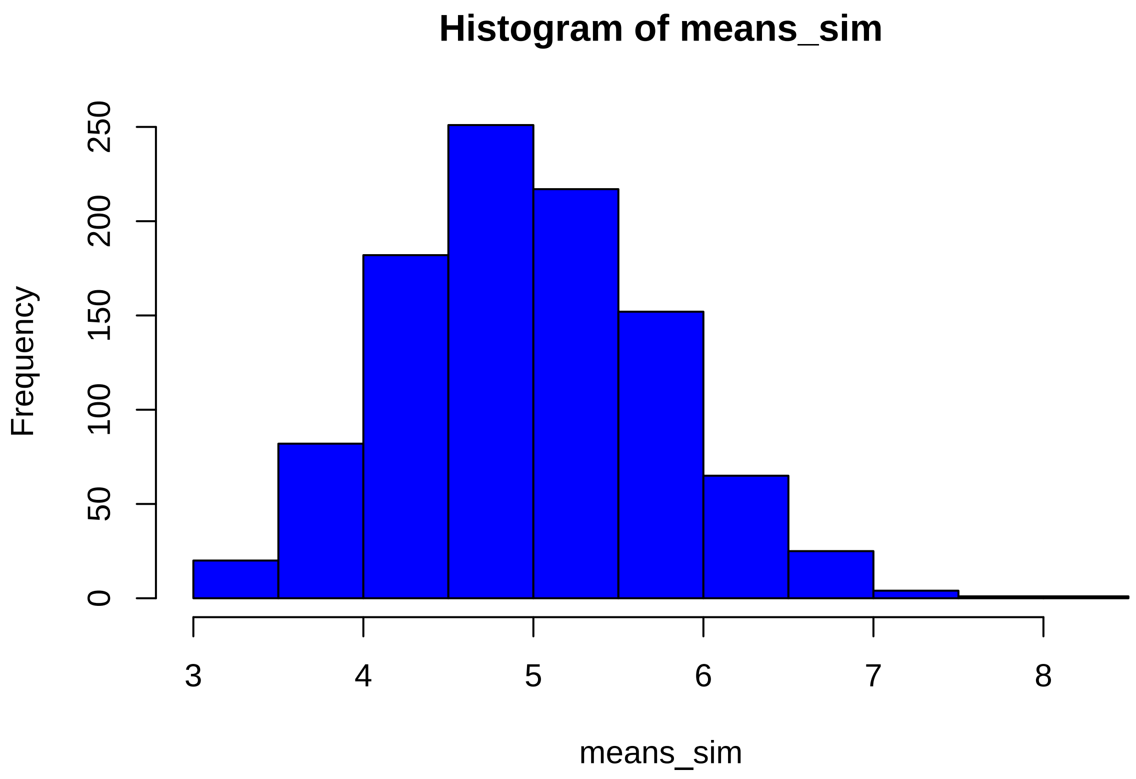
Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. You should - Show the sample mean and compare it to the theoretical mean of the distribution. - Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution. - Show that the distribution is approximately normal.

In point 3, focus on the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials.

Simulations and the R-code

We need to do 1000 simulations in which we derive a sample mean out of 40 simulated values for an exponential distribution with rate 0.2 (lambda). Then we plot the means in a Histogram "means_sim".

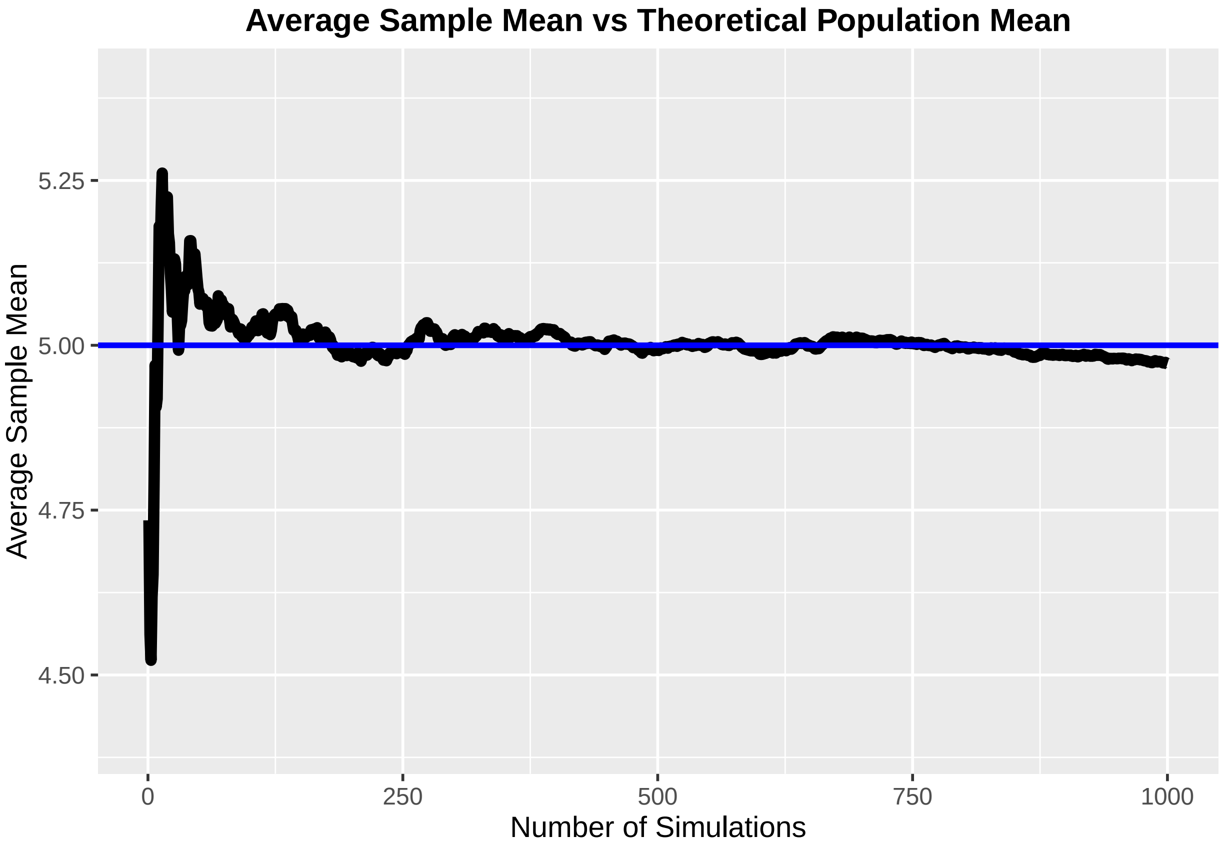
```
#Definition of variables
n_sim     <- 1000
n         <- 40
lambda    <- 0.2
set.seed(12345)

#Derived distribution parameters
mu        <- 1/lambda
sd        <- 1/lambda

#Create a matrix with n_sim rows and n columns corresponding to random simulation n times
matrix_sim  <- matrix(rexp(n_sim * n, rate=lambda), n_sim, n)

#Create a vector with rowmeans
means_sim   <- rowMeans(matrix_sim)

#Plot the means in a histogram
library(ggplot2)
hist(means_sim, col="blue")
```



Sample Mean versus Theoretical Mean

In order to show what happens we can plot the average sample mean for i iterations, where i = 1 … 1000. The average sample mean should converge to the distribution mean according to the Central Limit Theorem.

```r
#Define variables
means_sum   <- vector("numeric")
means_avg   <- vector("numeric")

#For any iteration i compute the average mean of the i sample means
means_sum[1] <- means_sim[1]
for (i in 2:n_sim) { means_sum[i] <- means_sum[i-1] + means_sim[i] }
for (i in 1:n_sim) { means_avg[i] <- means_sum[i]/i }

#Plot the computed average_iteration_mean against the theoretical mean
library(ggplot2)

g <- ggplot(data.frame(x = 1:n_sim, y = means_avg), aes(x = x, y = y))
g <- g + geom_hline(yintercept = 0) + geom_line(size = 2)
g <- g + geom_abline(intercept = 1 / lambda, slope = 0, color = "blue", size = 1)
g <- g + scale_y_continuous(breaks=c(4.50, 4.75, 5.00, 5.25), limits=c(4.4, 5.4))
g <- g + theme(plot.title = element_text(size=12, face="bold", vjust=2, hjust=0.5))
g <- g + labs(title="Average Sample Mean vs Theoretical Population Mean")
g <- g + labs(x = "Number of Simulations", y = "Average Sample Mean")
print(g)
```



The sample means converges with increasing iterations 4.971972.
Which gets close to the theoretical mean of the exponential distribution (1/lambda with lambda = 0.2): 5.

Sample Variance versus Theoretical Variance

According to the Central Limit Theorem the variance of the mean is sigma / squareroot(n).
So the variance of the distribution can be estimated by n * variance(means).

The theoretical variance of the exponential distribution is (1/lambda)^2: 25.
The estimated variance of the distribution is n * variance(means) : 24.6317049.

The reported values are quite close; suggesting support for the CLT :-).

Distribution

The Law of Large Numbers states that averages of iid samples converge to population means that they are estimating.
The Central Limit Theorem states that averages are aproximately normal, with distributions

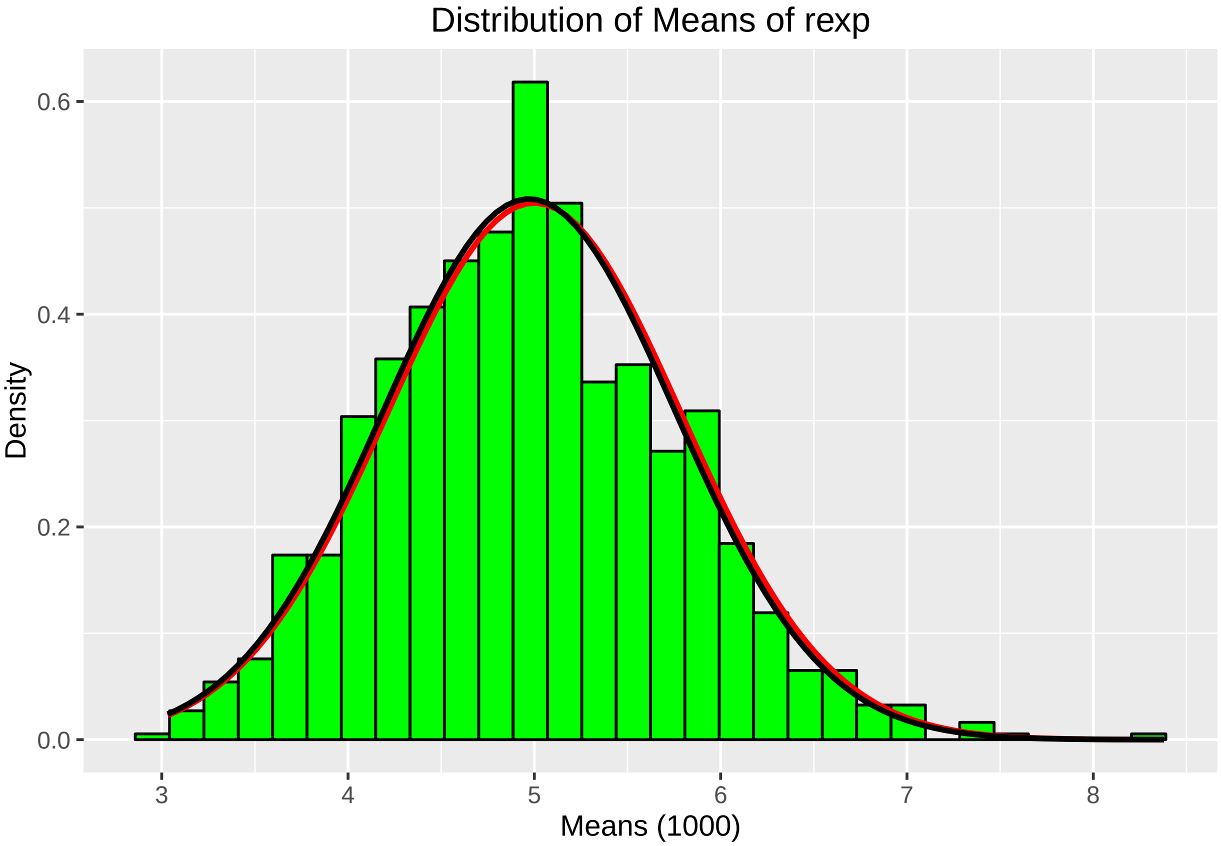
- centered at the population mean
- with standard deviation equal to the standard error of the mean
- CLT gives no guarantee that n is large enough.

We can illustrate this with plots of the standard normal distribution (red) overlayed by the distribution of the means with increasing number of simulations. This report only contains the plots for the original value (n_sim0 = 1000) and (n_sim3 = 30.000).

```
#Definition of variables
n_sim0   <- 1000
n_sim3   <- 30000

n        <- 40
lambda   <- 0.2

set.seed(12345)

#Derived distribution parameters
mu       <- 1/lambda
sd       <- 1/lambda

#Create a matrix with n_sim rows and nx columns corresponding to random simulation nx times
matrix_sim0 <- matrix(rexp(n_sim0 * n, rate=lambda), n_sim0, n)
matrix_sim3 <- matrix(rexp(n_sim3 * n, rate=lambda), n_sim3, n)

#Create a vector with rowmeans
means_sim0  <- rowMeans(matrix_sim0)
means_sim3  <- rowMeans(matrix_sim3)

library(ggplot2)
X = means_sim0

plotdata <- data.frame(X)
plot1 <- ggplot(plotdata,aes(x = X))
plot1 <- plot1 +geom_histogram(aes(y=..density..), colour="black",fill="green")
plot1 <- plot1+labs(title="Distribution of Means of rexp", x="Means (1000)", y="Density")
```

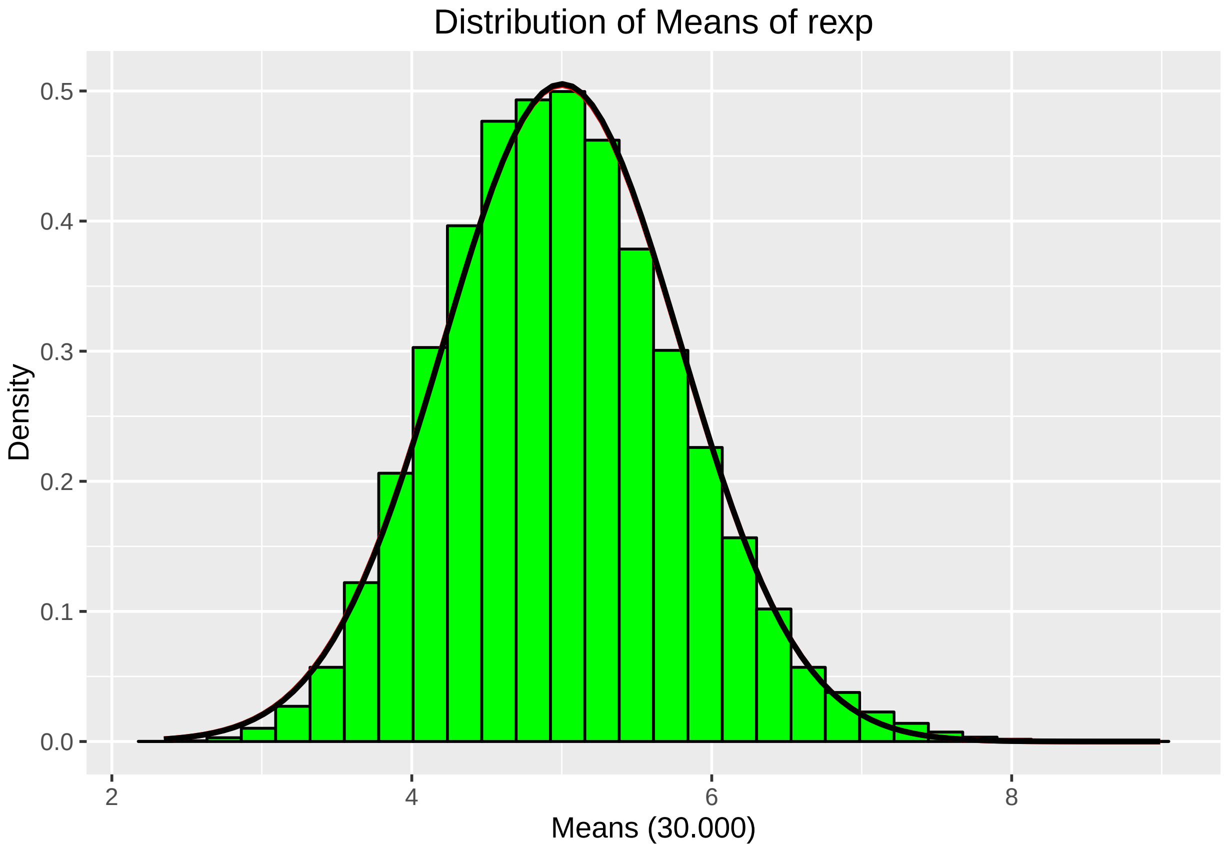
```
plot1 <- plot1 +stat_function(fun=dnorm,args=list( mean=1/lambda, sd=sqrt((1/lambda)^2/n)), color="red"
plot1 <- plot1 +stat_function(fun=dnorm,args=list( mean=mean(X), sd=sqrt(var(X))),color="black", size=1
print(plot1)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
#Similar set of statements for X=means_sim3.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



As you can see, when the number of simulations increases, the fit seems to be better and the estimated distribution follows the normal distribution more closely.

Part 2: Basic Inferential Data Analysis Instructions

Now in the second portion of the project, we're going to analyze the ToothGrowth data in the R datasets package. Load the ToothGrowth data and perform some basic exploratory data analyses

```
#Load data
library(datasets)
data(ToothGrowth)

#Explore a bit
dim(ToothGrowth)     #Size of the dataset
str(ToothGrowth)     #Structure of the data
summary(ToothGrowth) #Basic summary of the dataset
head(ToothGrowth)    #First rows
tail(ToothGrowth)    #Last rows

#Statistics
for (i in 1:dim(ToothGrowth)[2]) { TG_mean[i] <- mean(ToothGrowth[,i])}
TG_mean
TG_var <- c(var(ToothGrowth[,1]), 1, var(ToothGrowth[,3]))
TG_var
TG_sd = sqrt(TG_var)
```

```
TG_sd
```

Provide a basic summary of the data.

A basic summary of the data consists of the structure and its contents.

```r
str(ToothGrowth)                                #Structure of the data
```

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

```r
summary(ToothGrowth)                            #Basic summary of the dataset
```

```
##       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  
```

Since ToothGrowth contains data on measured ToothLength on different cases, it's nice to know how the cases have been organised.

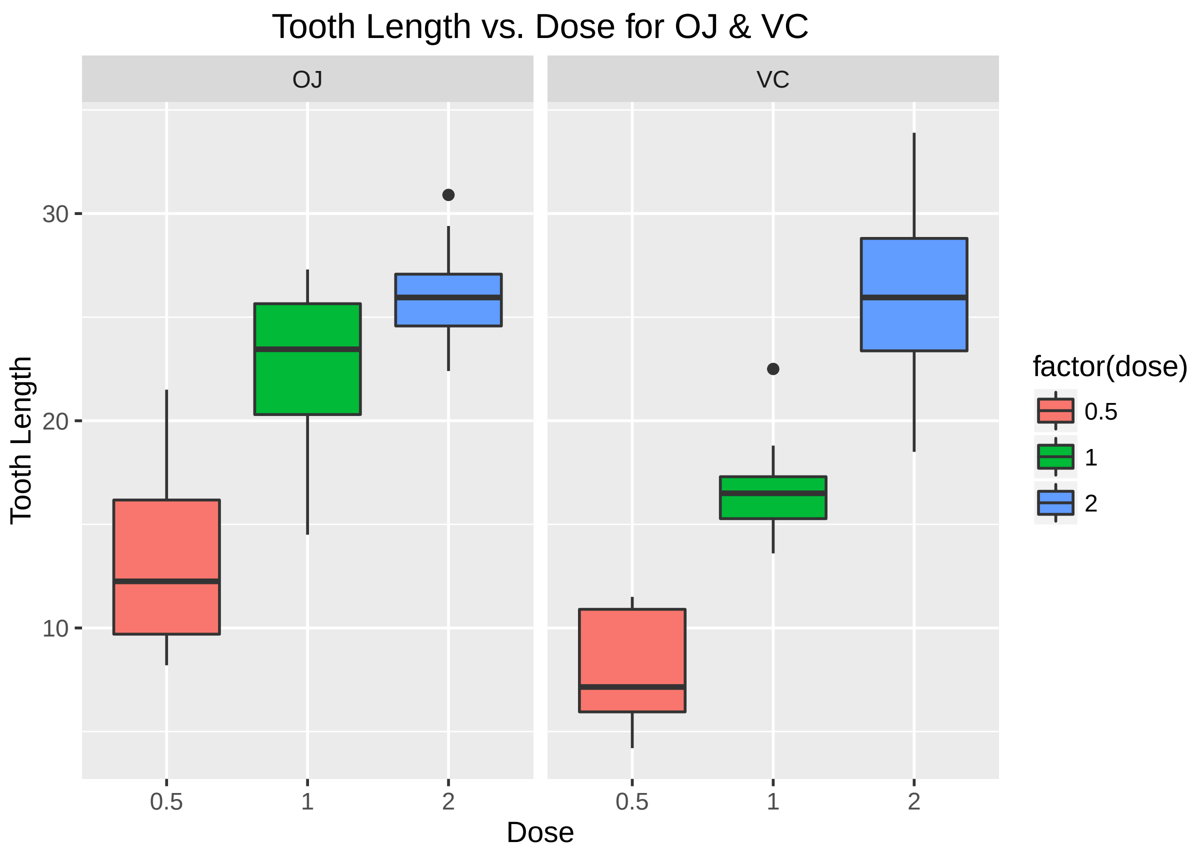
```r
table(ToothGrowth$supp,ToothGrowth$dose)  #Len is the dependent variable, so let's check the structure
```

```
## 
##      0.5  1  2
##   OJ  10 10 10
##   VC  10 10 10
```

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

Let's plot the data first, explaining ToothLength by the two variables Supp and Dose.

```r
#Boxplot graph of the tooth length vs the dose
ggplot(ToothGrowth, aes(x = factor(dose), y = len, fill = factor(dose)))+
      geom_boxplot()+
      facet_grid(.~supp)+
      labs(title = "Tooth Length vs. Dose for OJ & VC",
      x = "Dose", y = "Tooth Length")
```



The Box plot suggests that
- for supp OJ there's a non-linear relationship between ToothLength and the doubling of the dose (decreasing merits);
- for supp VC there seems to be a linear relationship betweeen ToothLength and the doubling of te dose;
- OJ seems to support ToothGrowth better than VC with dose smaller than 2.

Let's focus on the last one.

Effect of supp VC and OJ on ToothLength given dose x

We need to test for each dose wether there's a difference in effect between the two supplements. Given the Boxplot we should assume different variance between the two experiments.

So we need a two-sided t-test on wether the difference of the means equals zero or not.

Dose 0.5

```
#Test dose 0.5
dose05 <- ToothGrowth[ToothGrowth$dose == 0.5, ]
t.test(len ~ supp, paired=FALSE, var.equal=FALSE, data=dose05)
```

```
##
##  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
```

We have a 95% confidence interval that does not contain 0, telling us to believe that supp. OJ performs statistically significantly better than supp. VC with dose 0.50.

Dose 1.0

```
#Test dose 1.0
dose10 <- ToothGrowth[ToothGrowth$dose == 1.0, ]
t.test(len ~ supp, paired=FALSE, var.equal=FALSE, data=dose10)
```

```
##
##  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
```

We have a 95% confidence interval that does not contain 0, telling us to believe that supp. OJ performs statistically significantly better than supp. VC with dose 1.0.

Dose 2.0

```
#Test dose 2.0
dose20 <- ToothGrowth[ToothGrowth$dose == 2.0, ]
t.test(len ~ supp, paired=FALSE, var.equal=FALSE, data=dose20)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = -0.046136, 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
```

Now we have a 95% confidence interval that does contain 0, telling us to believe that supp. OJ and supp. VC perform statistically equally with dose 2.0.

Effect of doubling the dose on ToothLength given supp y

We need to test for each supplement wether there's a difference in effect in doubling the dose. Given the Boxplot we should assume different variance between the two experiments.

So we need a two-sided t-test on wether the difference of the means equals zero or not.

Supp OJ

```
#Create subsets of the data
OJ05 <- ToothGrowth[ToothGrowth$supp == 'OJ' & ToothGrowth$dose == 0.5, ]
OJ10 <- ToothGrowth[ToothGrowth$supp == 'OJ' & ToothGrowth$dose == 1.0, ]
OJ20 <- ToothGrowth[ToothGrowth$supp == 'OJ' & ToothGrowth$dose == 2.0, ]

t.test(OJ05$len, OJ10$len, paired=FALSE, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  OJ05$len and OJ10$len
## t = -5.0486, df = 17.698, p-value = 8.785e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -13.415634  -5.524366
## sample estimates:
## mean of x mean of y
##     13.23     22.70
```

```
t.test(OJ10$len, OJ20$len, paired=FALSE, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  OJ10$len and OJ20$len
## t = -2.2478, df = 15.842, p-value = 0.0392
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.5314425 -0.1885575
## sample estimates:
## mean of x mean of y
##     22.70     26.06
```

We have 95% confidence intervals that do not contain 0, telling us to believe that doubling the dose with supp. OJ performs statistically significantly better starting with dose 0.50 up to dose 2.0.

Supp VC

```
#Create subsets of the data
VC05 <- ToothGrowth[ToothGrowth$supp == 'VC' & ToothGrowth$dose == 0.5, ]
VC10 <- ToothGrowth[ToothGrowth$supp == 'VC' & ToothGrowth$dose == 1.0, ]
VC20 <- ToothGrowth[ToothGrowth$supp == 'VC' & ToothGrowth$dose == 2.0, ]

t.test(VC05$len, VC10$len, paired=FALSE, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  VC05$len and VC10$len
## t = -7.4634, df = 17.862, p-value = 6.811e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.265712  -6.314288
## sample estimates:
## mean of x mean of y
##      7.98     16.77
```

```
t.test(VC10$len, VC20$len, paired=FALSE, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  VC10$len and VC20$len
## t = -5.4698, df = 13.6, p-value = 9.156e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -13.054267  -5.685733
## sample estimates:
## mean of x mean of y
##     16.77     26.14
```

We have 95% confidence intervals that do not contain 0, telling us to believe that doubling the dose with supp. VC performs statistically significantly better starting with dose 0.50 up to dose 2.0.

Conclusions and Assumptions

Conclusions

- Pigs given the OJ supplement at 0.5 and 1.0 dosages have significantly faster tooth growth than guinea pigs given VC at the same doses;

- Pigs given OJ or VC at a dose of 2.0 do not have significantly different tooth growth;

- Doubling supplement dosage significantly increases tooth growth (proven untill dose 2.0).

Assumptions

- The variances between the sample popluations are not equal;

- The sample data is not paired;

- The sample population distribution is mound shaped and not skewed.