Simulation Exercise and Basic Data Analysis

Jean-Paul Courneya

8/6/2017

```r
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
```

This class project consists of two parts:

1. A simulation exercise.

2. Basic inferential data analysis.

First you will set a working directory and load dependencies

```r
setwd()
library(knitr)
library(ggplot2)
```

Part 1: Simulation Exercise

In this part of the project we are investigating the exponential distribution in R and comparing 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 = s for all of the simulations.

Investigating the distribution of averages of 40 exponentials.

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials.

Use set.seed so that the simulation can be repeated. Setting the seed provides the same starting value to the random number generating function.

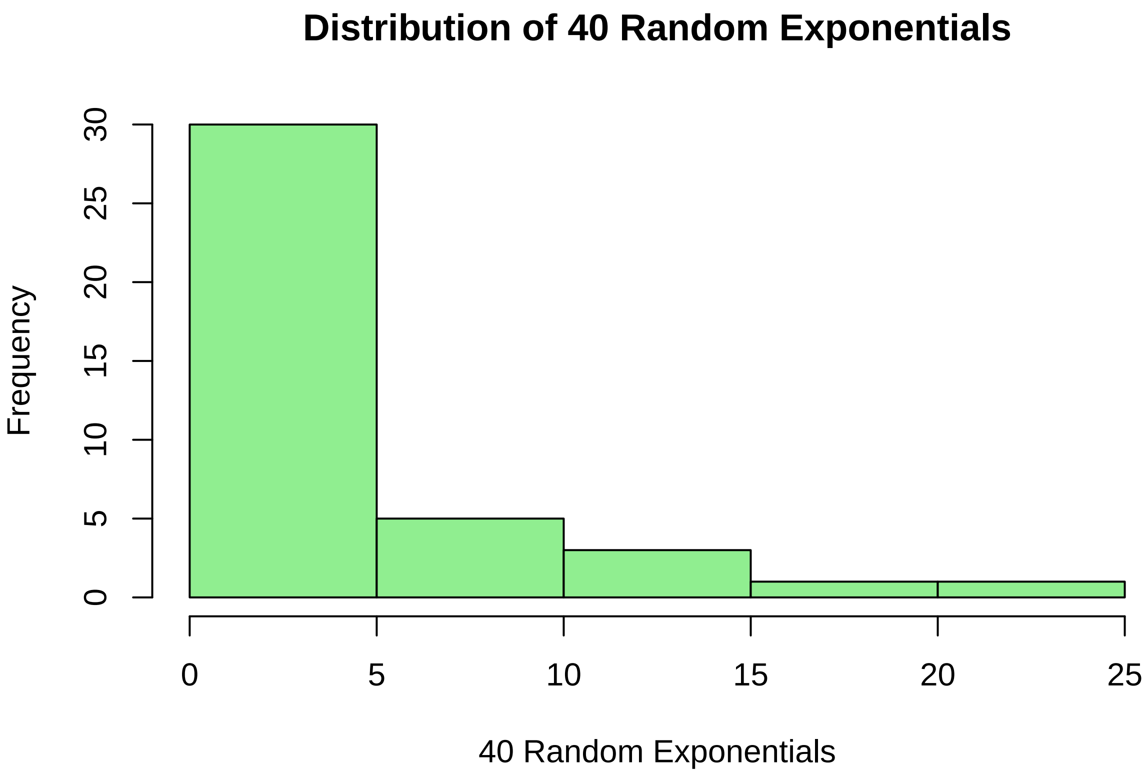
```r
set.seed(15)
```

set some constants

```r
lambda <- 0.2
sampleSize <- 40
simulations <- 1000
```

Start by taking a look at the distribution of a sample of 40 random exponentials

```r
SimDist <- rexp(sampleSize, 0.2)
hist(SimDist, xlab = "40 Random Exponentials", ylab = "Frequency",
    main = "Distribution of 40 Random Exponentials", col = "lightgreen")
```



Clearly we see that 40 random exponentials chosen look exponentially distributed.

Calculate the Sample vs Theoretical Mean

Next we will evaluate the mean of the sample data set from above.

```r
Sample.Mean <- mean(SimDist)
cat("Sample Mean: ", Sample.Mean)
```

```
## Sample Mean:  4.506131
```

The theoretical mean for an eponential distribution as stated in the instructions is 1/lambda

```r
Theoretical.Mean <- 1/lambda
cat("Theoretical Mean: ", Theoretical.Mean)
```

```
## Theoretical Mean:  5
```

The sample mean approximates the theoretical mean.

Calculate the Sample vs Theoretical Variance

```
Sample.Variance <- var(SimDist)
cat("Sample Variance: ", Sample.Variance)
```

```
## Sample Variance:  21.66123
```

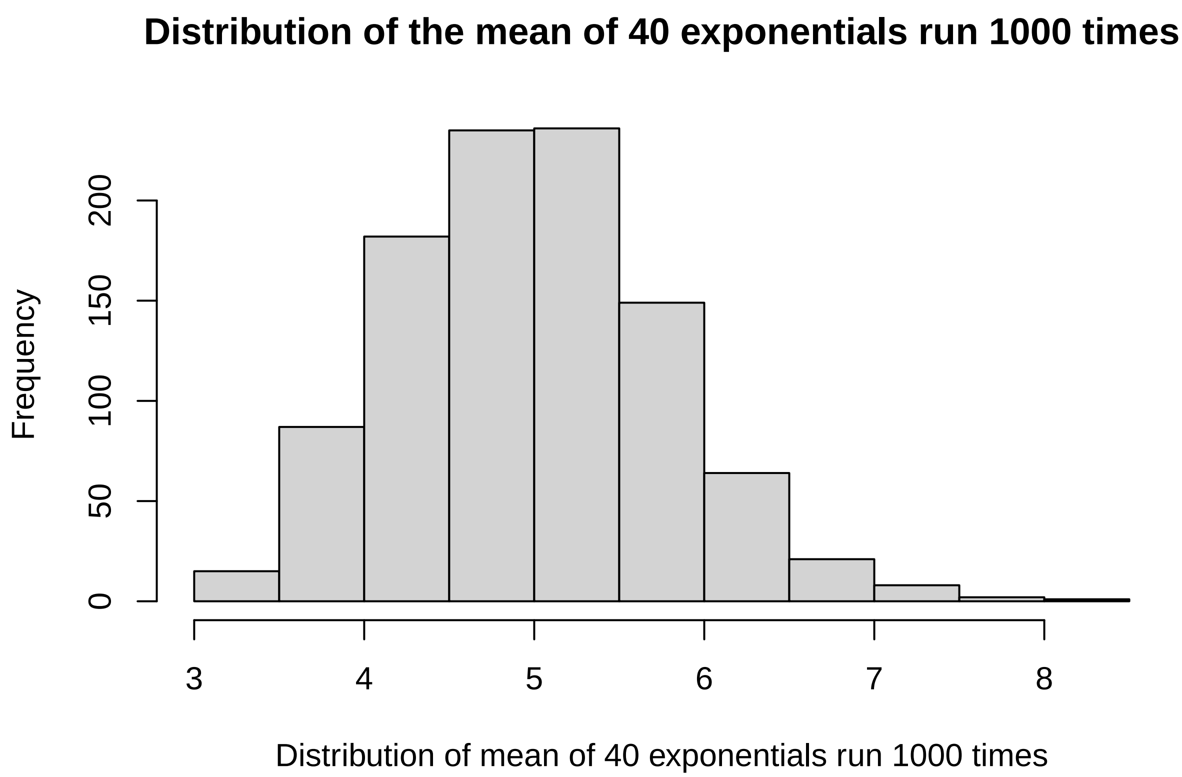
```
Theoretical.Variance <- (1/lambda)^2/sampleSize
cat("Theoretical Variance: ", Theoretical.Variance)
```

```
## Theoretical Variance:  0.625
```

There is a big difference between the sample variance and theoretical variance.

Now that we have seen the mean and variance of 40 random exponentials and compared it to the theoretical mean and variance. Next we will look at the distribution of the mean of 40 exponentials run 1000 times.

```
Simulation <- replicate(1000, mean(rexp(40, 0.2)))
hist(Simulation, xlab = "Distribution of mean of 40 exponentials run 1000 times",
    ylab = "Frequency", main = "Distribution of the mean of 40 exponentials run 1000 times",
    col = "lightgrey")
```



The data show a distribution around a central point of 5. Next we will evaluate the mean of the replicated sample data set.

```
mean(replicate(1000, rexp(40, 0.2)))
```

```
## [1] 4.976328
```

This distribution looks far more Gaussian than the original exponential distribution!

Part 2: Basic inferential data analysis.

```
library(ggplot2)
```

Load the ToothGrowth data and perform some basic exploratory data analyses

```
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 ...
```

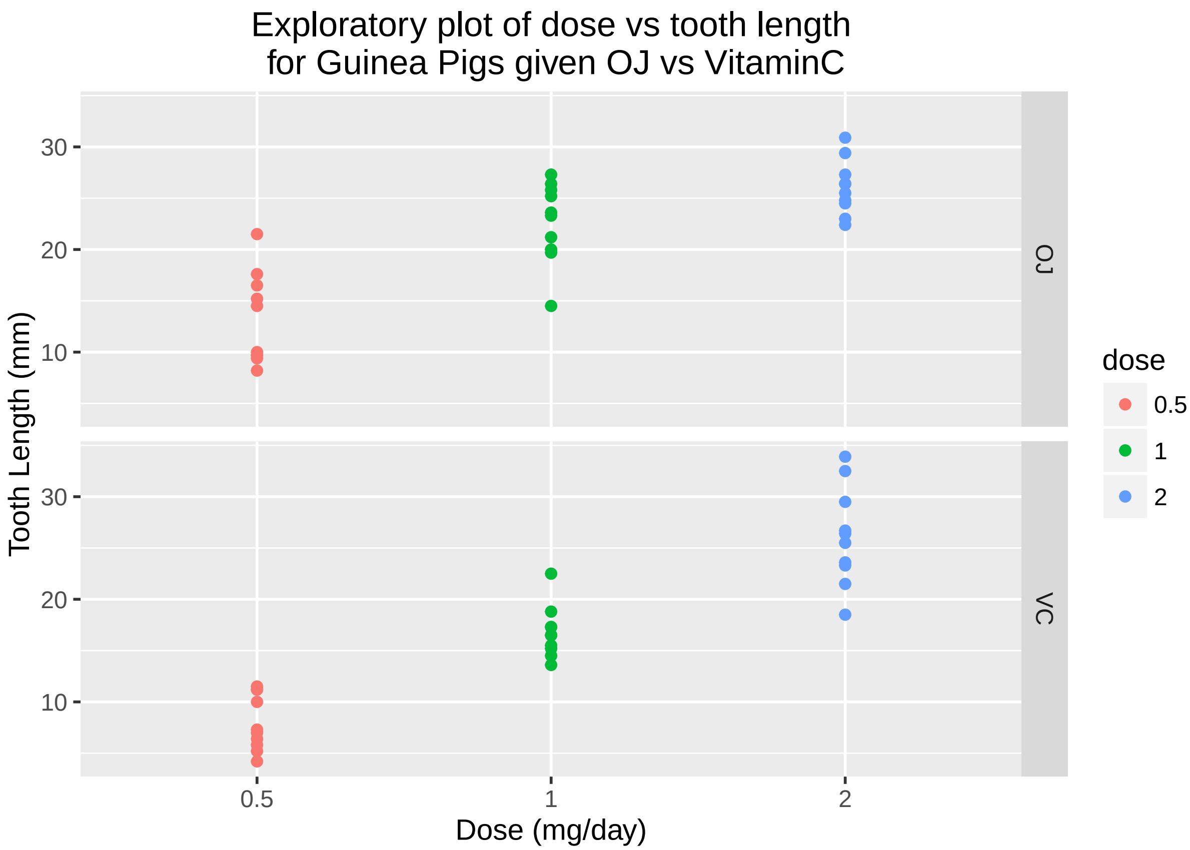
The data for the dose is currently numeric. Since the doses are factors they will be changed to factors.

```
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
str(ToothGrowth)
```

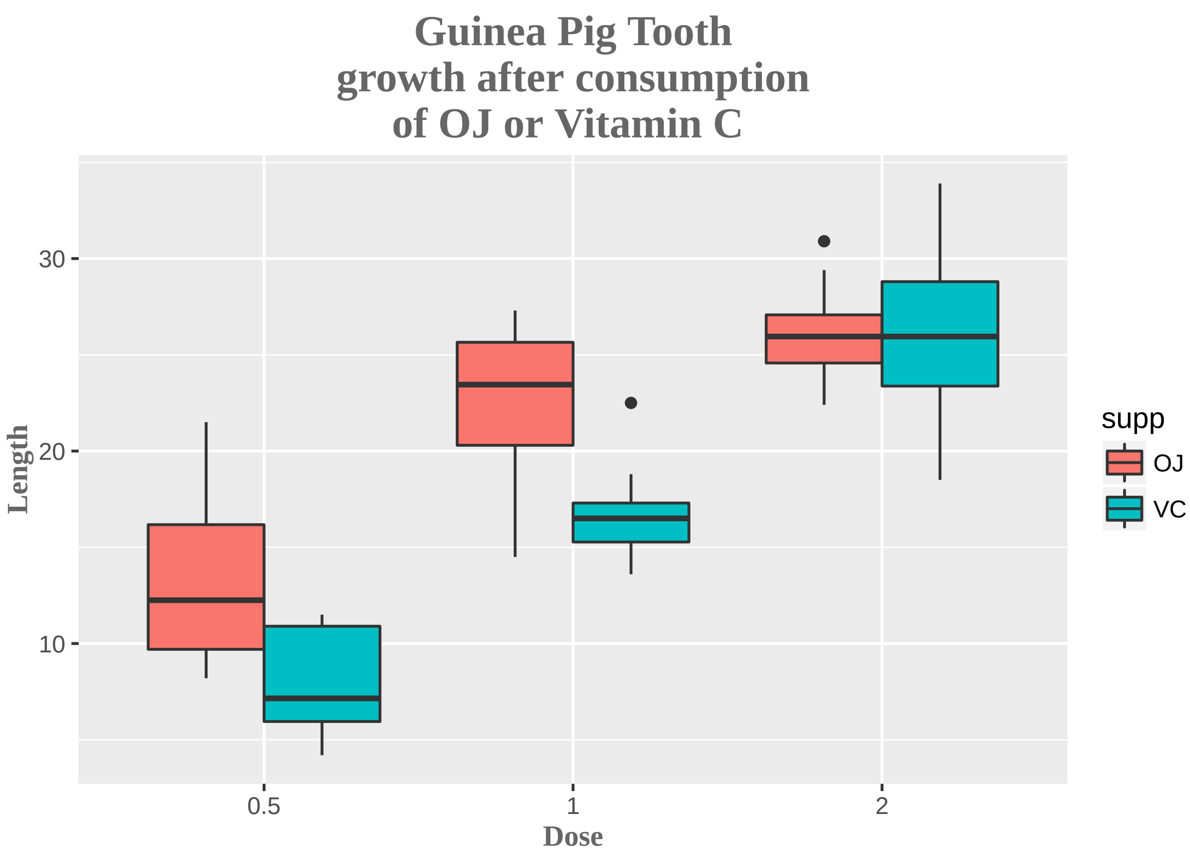
```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: Factor w/ 3 levels "0.5","1","2": 1 1 1 1 1 1 1 1 1 1 ...
```

What does the data look like?

```
qplot(dose, len, data = ToothGrowth, facets = supp ~ ., color = dose,
    xlab = "Dose (mg/day)", ylab = "Tooth Length (mm)", main = "Exploratory plot of dose vs tooth lengt
```




```r
f <- ggplot(ToothGrowth, aes(x = dose, y = len, fill = supp))
f + geom_boxplot() + ggtitle("Guinea Pig Tooth\ngrowth after consumption\nof OJ or Vitamin C ") +
    labs(x = "Dose", y = "Length") + theme(plot.title = element_text(family = "serif",
    color = "#666666", face = "bold", size = 16, hjust = 0.5)) +
    theme(axis.title = element_text(family = "serif", color = "#666666",
        face = "bold", size = 11))
```



A basic summary of the data

```
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
```

compare tooth growth by supp and dose using confidence intervals and/or hypothesis tests

The overall goal of understanding this data is achieved by determining if vitamin C affects tooth growth and if it matters how the vitamin C is delivered.

A t.test can be used to compare the tooth len growth between supplements (VitaminC or OJ) and each dose.

Comparison at ($p < 0.05$) tooth len growth at a dose of 0.5 (VitaminC or OJ)

```
t.test(ToothGrowth$len[ToothGrowth$supp == "OJ" & ToothGrowth$dose ==
    0.5], ToothGrowth$len[ToothGrowth$supp == "VC" & ToothGrowth$dose ==
    0.5])
```

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

```
##
## data:  ToothGrowth$len[ToothGrowth$supp == "OJ" & ToothGrowth$dose ==  and ToothGrowth$len[ToothGrow
## 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 of x mean of y
##     13.23      7.98
```

OJ has a higher effect on tooth growth at this dose

Comparison at ($p < 0.05$) tooth len growth at a dose of 1.0 (VitaminC or OJ)

```r
t.test(ToothGrowth$len[ToothGrowth$supp == "OJ" & ToothGrowth$dose ==
    1], ToothGrowth$len[ToothGrowth$supp == "VC" & ToothGrowth$dose ==
    1])
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[ToothGrowth$supp == "OJ" & ToothGrowth$dose ==  and ToothGrowth$len[ToothGrow
## 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 of x mean of y
##     22.70     16.77
```

OJ has a higher effect on tooth growth at this dose

Comparison at ($p < 0.05$) tooth len growth at a dose of 2.0 (VitaminC or OJ)

```r
t.test(ToothGrowth$len[ToothGrowth$supp == "OJ" & ToothGrowth$dose ==
    2], ToothGrowth$len[ToothGrowth$supp == "VC" & ToothGrowth$dose ==
    2])
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[ToothGrowth$supp == "OJ" & ToothGrowth$dose ==  and ToothGrowth$len[ToothGrow
## 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 of x mean of y
##     26.06     26.14
```

Ive just demonstrated that there is a difference in tooth growth between supplements. at the low doses (0.5, 1.0) OJ has more of an effect on Tooth growth but not at 2.0

Next I will compare tooth growth based on dose of supplement (OJ or VC).

First I will compare 0.5 vs 1 of OJ Comparison at ($p < 0.05$).

```
t.test(ToothGrowth$len[ToothGrowth$supp == "OJ" & ToothGrowth$dose ==
    0.5], ToothGrowth$len[ToothGrowth$supp == "OJ" & ToothGrowth$dose ==
    1])
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[ToothGrowth$supp == "OJ" & ToothGrowth$dose ==  and ToothGrowth$len[ToothGrow
## 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
```

There is a significant difference.

Here is the comparison 0.5 vs 2 of OJ Comparison at ($p < 0.05$).

```
t.test(ToothGrowth$len[ToothGrowth$supp == "OJ" & ToothGrowth$dose ==
    0.5], ToothGrowth$len[ToothGrowth$supp == "OJ" & ToothGrowth$dose ==
    2])
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[ToothGrowth$supp == "OJ" & ToothGrowth$dose ==  and ToothGrowth$len[ToothGrow
## t = -7.817, df = 14.668, p-value = 1.324e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -16.335241  -9.324759
## sample estimates:
## mean of x mean of y
##     13.23     26.06
```

There is a significant differnce

Here is the comparison 1.0 vs 2 of OJ Comparison at ($p < 0.05$).

```
t.test(ToothGrowth$len[ToothGrowth$supp == "OJ" & ToothGrowth$dose ==
    1], ToothGrowth$len[ToothGrowth$supp == "OJ" & ToothGrowth$dose ==
    2])
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[ToothGrowth$supp == "OJ" & ToothGrowth$dose ==  and ToothGrowth$len[ToothGrow
## 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
```

The difference is significant but not the same maginitude of effect as between 0.5 and (1 or 2)

Next I will compare the effect of Vitamin C

First I will compare 0.5 vs 1 of VC Comparison at ($p < 0.05$).

```
t.test(ToothGrowth$len[ToothGrowth$supp == "VC" & ToothGrowth$dose ==
    0.5], ToothGrowth$len[ToothGrowth$supp == "VC" & ToothGrowth$dose ==
    1])
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[ToothGrowth$supp == "VC" & ToothGrowth$dose ==  and ToothGrowth$len[ToothGrow
## 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
```

There is a significant difference.

Here is the comparison 0.5 vs 2 of VC Comparison at ($p < 0.05$).

```
t.test(ToothGrowth$len[ToothGrowth$supp == "VC" & ToothGrowth$dose ==
    0.5], ToothGrowth$len[ToothGrowth$supp == "VC" & ToothGrowth$dose ==
    2])
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[ToothGrowth$supp == "VC" & ToothGrowth$dose ==  and ToothGrowth$len[ToothGrow
## t = -10.388, df = 14.327, p-value = 4.682e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -21.90151 -14.41849
## sample estimates:
## mean of x mean of y
##      7.98     26.14
```

There is a significant difference

Here is the comparison 1.0 vs 2 of VC Comparison at ($p < 0.05$).

```
t.test(ToothGrowth$len[ToothGrowth$supp == "VC" & ToothGrowth$dose ==
    1], ToothGrowth$len[ToothGrowth$supp == "VC" & ToothGrowth$dose ==
    2])
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[ToothGrowth$supp == "VC" & ToothGrowth$dose ==  and ToothGrowth$len[ToothGrow
## 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
```

I have just demonstrated that not only is there a difference between the supplements effect on tooth growth by dose but that also tooth growth is significantly affected by OJ and Vitamin C. There is a maximum benefit which starts to level off at a dose of 2.0 since we see the percent difference between a dose of 1 and 2 and effect of teeth growth is decreasing regardless of supplement given.

Overall Guinea Pig tooth growth is most affected by OJ supplement.There is a noticible effect between doses of Vitamin C however within doses.