SI Course Project

Yigit Ozan Berk

6/27/2019

Overview

In a few (2-3) sentences explain what is going to be reported on.

Simulations

Initiation of required packages:

```
library(ggplot2)
```

Simulating 1000 sets of 40 exponential random variables:

```
set.seed(121212)
my_lambda <- 0.2
my_n <- 40
nsim <- 1000

samples <- matrix(nrow = 40, ncol = 1000)
for(i in 1:1000) samples[,i] = rexp(my_n, my_lambda)
```

Means, Standard Deviations, and Variances of simulated 40 exponential random variables:

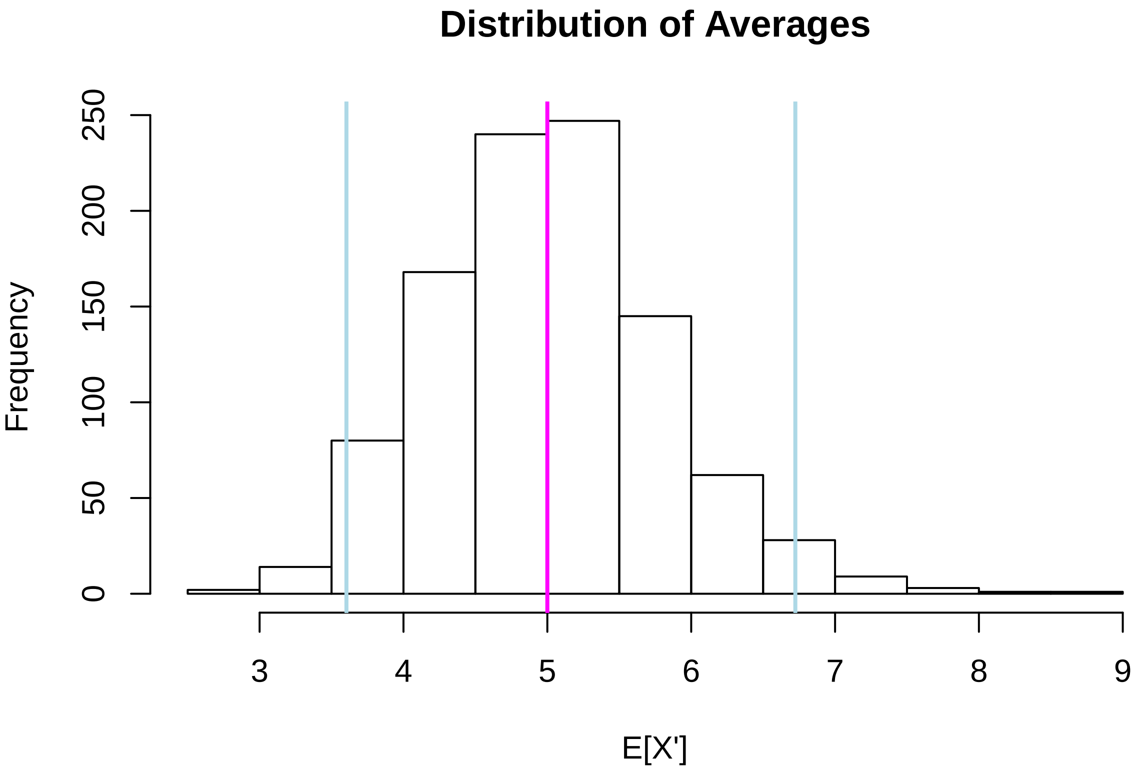
```
mns <- apply(samples, 2, mean)

sdevs <- apply(samples, 2, sd)

vars <- apply(samples, 2, var)
```

Sample Mean versus Theoretical Mean

The theoretical mean of the distribution is lambda = 1/0.2 = `1/.2`

The overall average of 1000 simulated 40 exponential random variables are `mean(mns)`

```
hist(mns, main = "Distribution of Averages",
     xlab = "E[X']")
abline(v = 5, col = "magenta", lwd = 2)
abline(v = quantile(mns, c(.025, .975)), col = "lightblue", lwd = 2)
```



The distribution of averages is centered around the theoretical mean as expected; with 95% confidence interval:

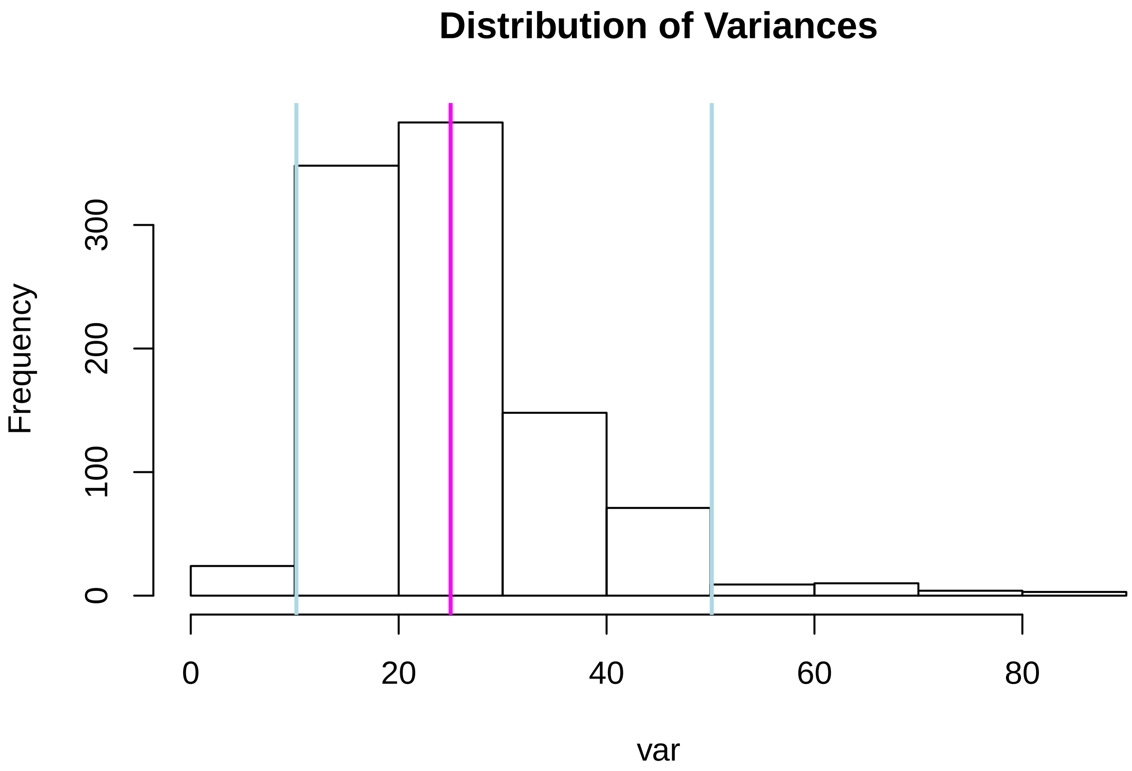
```
quantile(mns, c(.025, .975))
```

```
##     2.5%    97.5%
## 3.603528 6.723835
```

Sample Variance versus Theoretical Variance

With a theoretical variance of 25 *(1/lambda^2)*, the distribution of observed variances are as follows

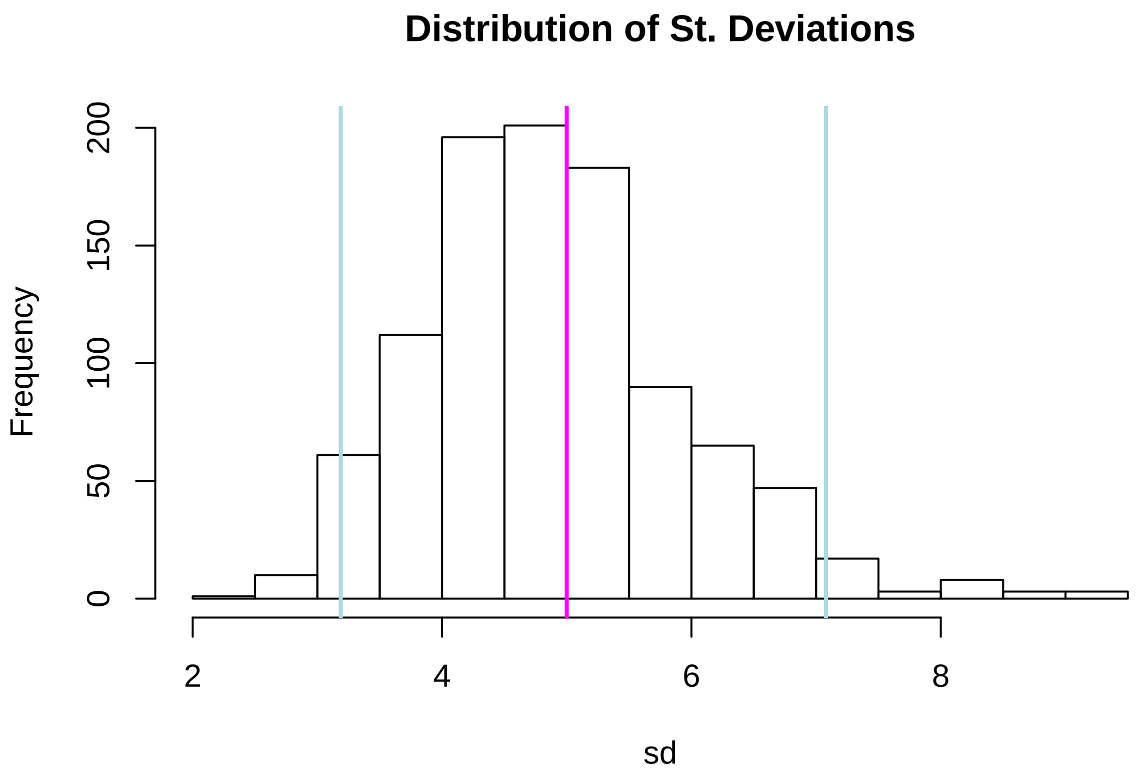
```
hist(vars, main = "Distribution of Variances",
     xlab = "var")
abline(v = 25, col = "magenta", lwd = 2)
abline(v = quantile(vars, c(.025,.975)), col = "lightblue", lwd = 2)
```



The average variation in the simulations is observed as 25.03814 which is close to theoretical variance.

If we look at the distribution of standard deviations, we can see that it's centered at the theoretical standard deviation with an approximately normal distribution:

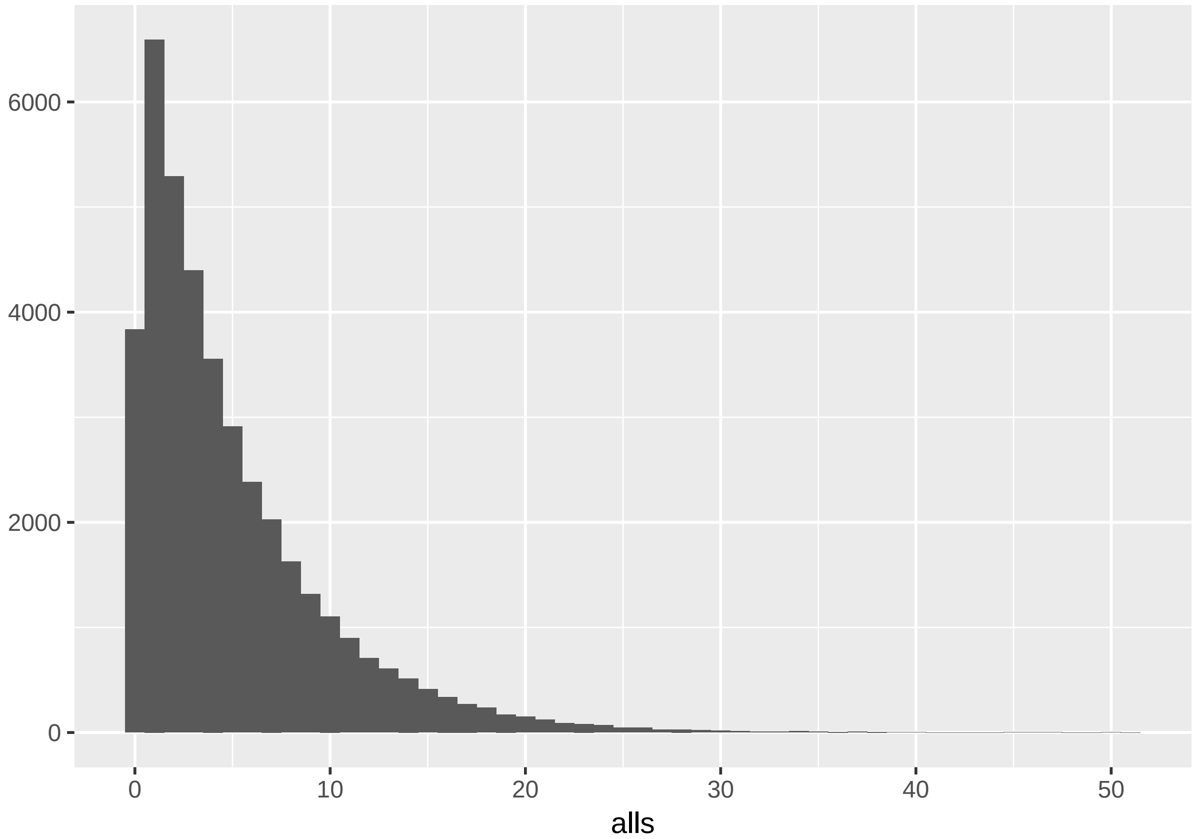
```
hist(sdevs, main = "Distribution of St. Deviations",
     xlab = "sd")
abline(v = 5, col = "magenta", lwd = 2)
abline(v = quantile(sdevs, c(.025,.975)), col = "lightblue", lwd = 2)
```



Distribution

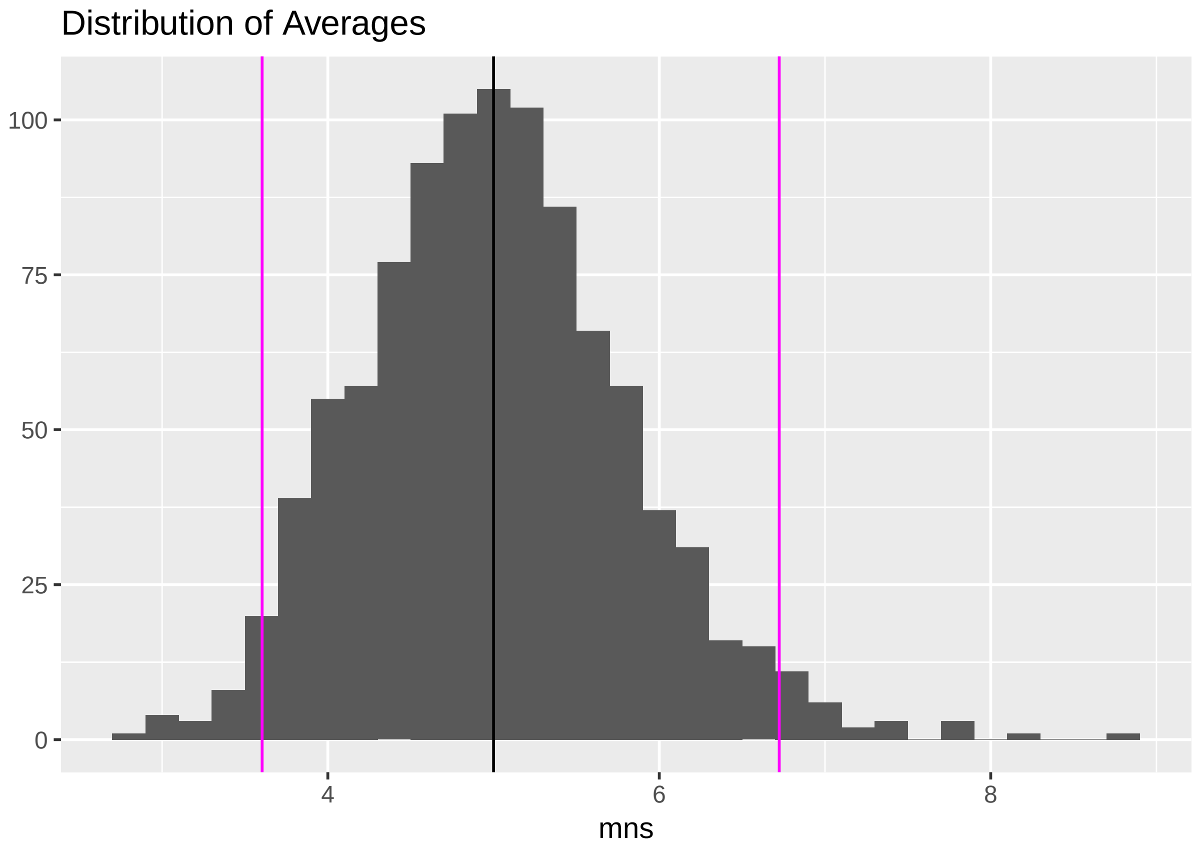
If we look at all the random draws of the exponential random variables, we can see the graph of the frequencies constitutes an exponential distribution. By CLT as the size of the variables increases, the frequency density will transform into a finer exponential distribution density graph.

```
alls <- NULL
for(i in 1:40) {
        alls <- c(alls, samples[i, ])
}
qplot(alls, binwidth = 1)
```



However, if we look at the distribution of averages of 40 exponential random variables over the 1000 simulations, we can clearly see the frequency distribution is not equal to an exponential distribution:

```
g <- qplot(mns, binwidth = .2)
g <- g + geom_vline(xintercept = 5)
g <- g + geom_vline(xintercept = quantile(mns, c(.025, .975)), col = "magenta")
g <- g + ggtitle("Distribution of Averages")
g
```



Instead, the distribution of averages is roughly a normal distribution with mean centered around the theoretical mean(5) of the population. One can tell the distribution is roughly normal by the bell shape curve of the frequency density.

By CLT as the number of simulations increases, the distribution of averages of these 40 random draws will look more and more like a normal distribution with mean equal to theoretical mean of the population.

Part 2 - ToothGrowth Data Analysis

At first look, the data has 60 observations with 3 variables, tooth growth length, suplement type, used dosage:

```
library(datasets)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
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
```

When we look at the composition of the observed data, we can see that there are length variables for 10 observations of half dosage, 10 obs of 1 dose, and 10 obs of 2 doses for each OJ supplement and VC supplement.

```
table(ToothGrowth[ToothGrowth$supp == "OJ", 3])
```

```
##
## 0.5   1   2
##  10  10  10
```

```
table(ToothGrowth[ToothGrowth$supp == "VC", 3])
```

```
##
## 0.5   1   2
##  10  10  10
```

We can compare the difference in tooth growth between different supplements and different doses of supplements.

Firstly, let's see the range of the tooth growth in each supplement type

```
myVC <- ToothGrowth %>% filter(supp == "VC")
myOJ <- ToothGrowth %>% filter(supp == "OJ")

range(myVC$len)
```

```
## [1]  4.2 33.9
```

```
range(myOJ$len)
```

```
## [1]  8.2 30.9
```

VC results seem to have a vider range than OJ.

The overall difference of means is as follows:

```
mean(myOJ$len) - mean(myVC$len)
```

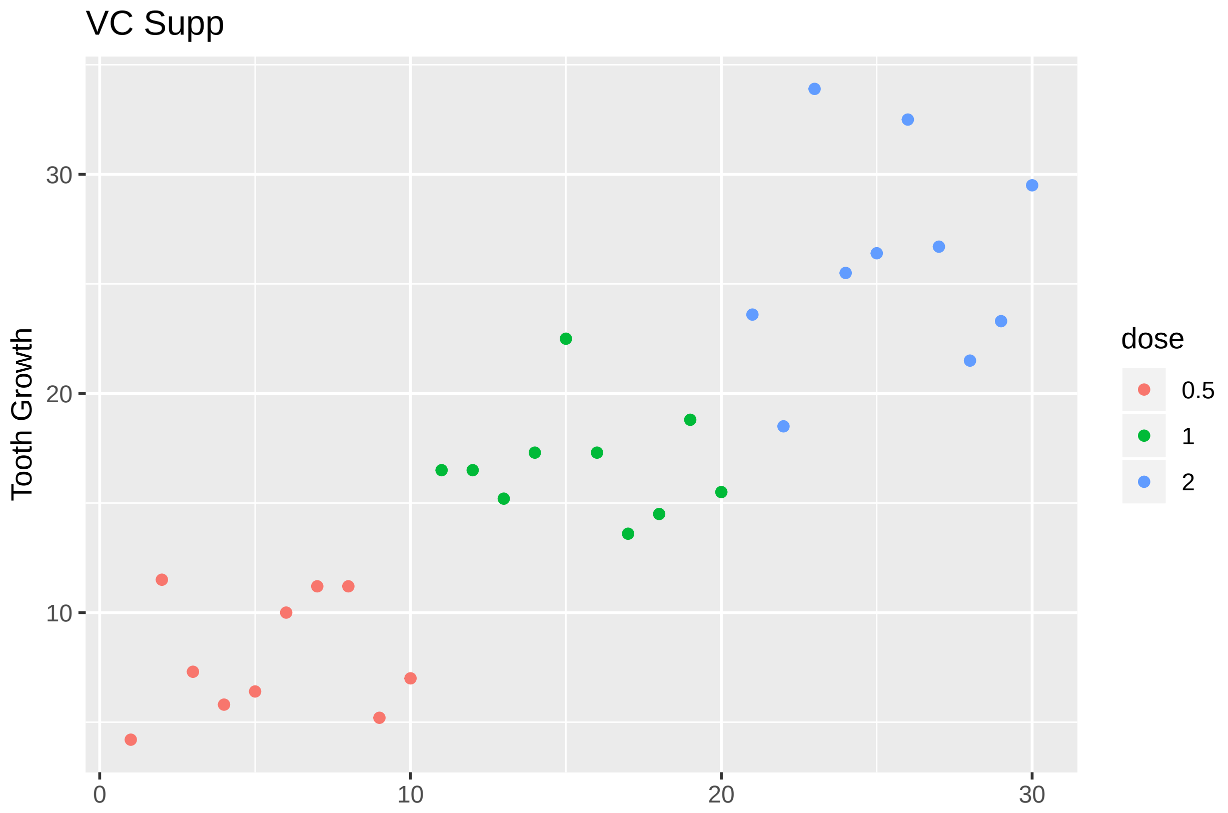
```
## [1] 3.7
```

On average, it looks like the tooth with OJ supplement grew 3.7 units more than tooth with VC supplement.

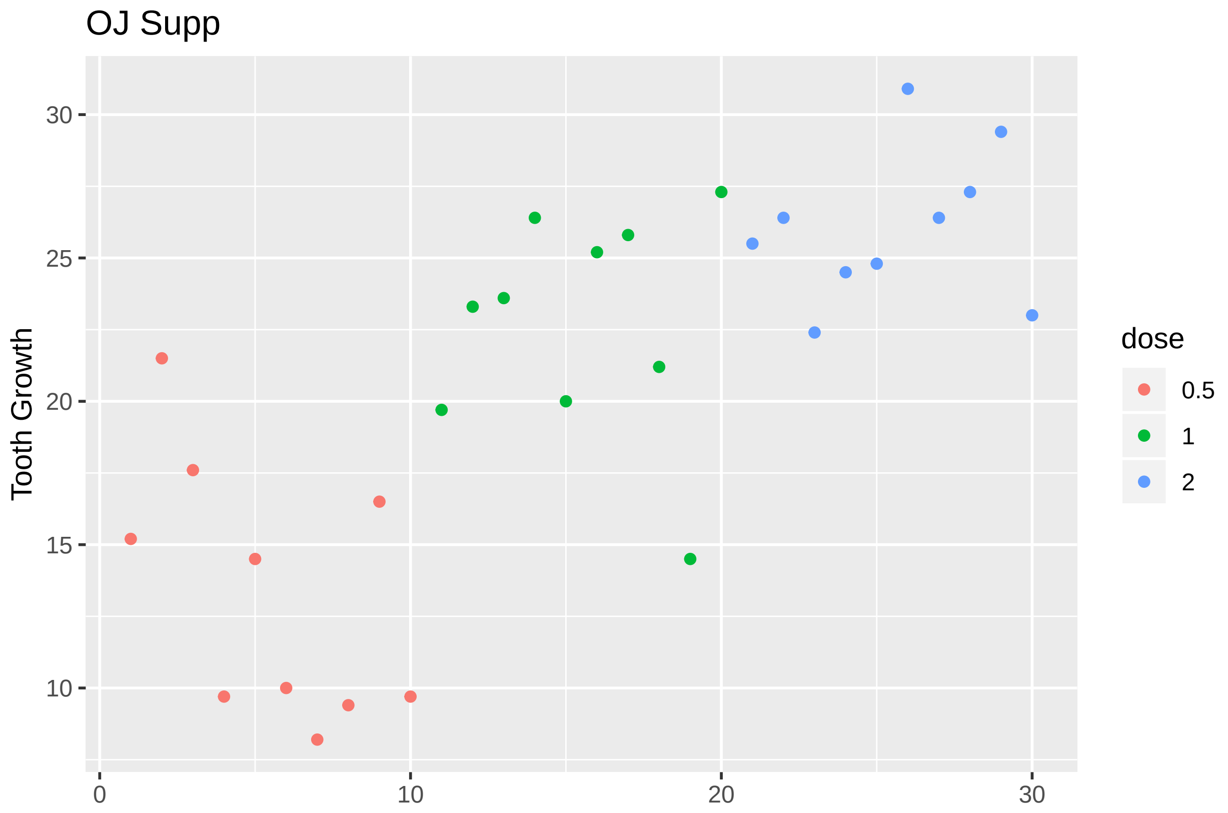
However, this generalisation is definitely not unbiased, because we know that the data contains length results of different doses. The average difference of tooth growth may be variable depending on the dosage.

As the sample size is not very large, we can check all the data we have on a scatterplot:

```
myVC$dose <- as.factor(myVC$dose)
qplot(y = len, data = myVC, col = dose, ylab = "Tooth Growth", main = "VC Supp")
```



```
myOJ$dose <- as.factor(myOJ$dose)
qplot(y = len, data = myOJ, col = dose, ylab = "Tooth Growth", main = "OJ Supp")
```



At first look, VC supp results had a vider range than OJ supp results, but upon closer look, we see that there is greater variation in OJ supplements.

We can test this with a hypothesis.

H_0: sd_oj > sd_vc ,for every dose subset

But first, let's see the more basic hypothese until this point:

mu_VC0.5 < mu_VC1 mu_VC1 < mu_VC2

mu_OJ0.5 < mu_OJ1 mu_OJ1 < mu_OJ2

and see the confidence intervals in which these are true.

Analyzing VC Supp Effects on Tooth Growth

assumptions

To calculate the confidence interval of this hypothesis, equal variance among groups is assumed.

The sample distribution is assumed to be normal. As the sample size is small, T distribution methods are applied.

H_0

The null hypothesis is that the mean of dose 0.5 is equal to the mean of dose 1.

The second null hpytohesis is that the mean of dose 1 is equal to the mean of dose 2.

calculation

First we need to organize the groups into subsets

```
VC0.5 <- myVC %>% filter(dose == "0.5")
VC1 <- myVC %>% filter(dose == "1")
VC2 <- myVC %>% filter(dose == "2")
#example outlook
VC1
```

```
##     len supp dose
## 1  16.5   VC    1
## 2  16.5   VC    1
## 3  15.2   VC    1
## 4  17.3   VC    1
## 5  22.5   VC    1
## 6  17.3   VC    1
## 7  13.6   VC    1
## 8  14.5   VC    1
## 9  18.8   VC    1
## 10 15.5   VC    1
```

Then, calculating the confidence intervals:

Testing dose 0.5 and dose 1:

H_0 : The mean of dose 0.5 is equal to the mean of dose 1.

```
t.test(VC1$len - VC0.5$len, paired = F, var.equal = T)
```

```
##
##  One Sample t-test
##
## data:  VC1$len - VC0.5$len
## t = 6.1364, df = 9, p-value = 0.0001715
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   5.549601 12.030399
## sample estimates:
## mean of x
##      8.79
```

According to the T test, p value is smaller than .05 so we reject the null hypothesis that the mean differences is equal to zero. The mean of VC0.5 is 6.13 estimated standard errors away from the mean of VC1. Of which the probability of happening if the means are equal is 0.0001715.

The difference of means is of 8.79 units.

Our related power is

```
power.t.test(n = 20, delta = mean(VC1$len) - mean(VC0.5$len), sd = sd(myVC$len), type = "one.sample", al
```

```
## [1] 0.9983221
```

Testing dose 2 and dose 1:

H_0 : The mean of dose 1 is equal to the mean of dose 2.

```
t.test(VC2$len - VC1$len, paired = F, var.equal = T)
```

```
##
##  One Sample t-test
##
## data:  VC2$len - VC1$len
## t = 5.346, df = 9, p-value = 0.0004648
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   5.405082 13.334918
## sample estimates:
## mean of x
##      9.37
```

According to the T test, p value is smaller than 0.05 so we reject the null hypothesis that the mean differences is equal to zero. The mean of VC2 is 5.346 estimated standard errors away from the mean of VC1. Of which the probability of happening if the means are equal is 0.0004648

The difference of means is of 9.37 units.

Our related power is

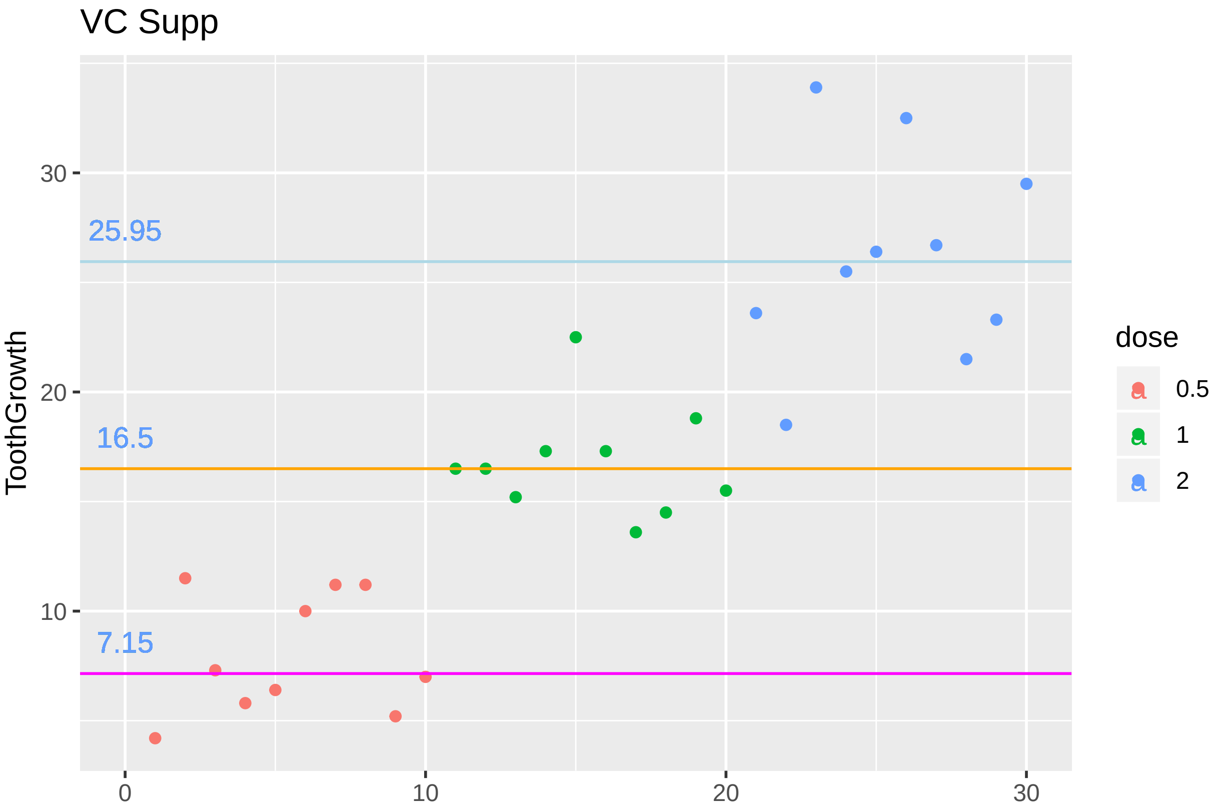
```
power.t.test(n = 20, delta = mean(VC2$len) - mean(VC1$len), sd = sd(myVC$len), type = "one.sample", alt
```

```
## [1] 0.9993905
```

To visualize the medians in a graph together:

```
g <- qplot(y = myVC$len, data = myVC, col = dose, ylab = "ToothGrowth", main = "VC Supp")
g <- g + geom_hline(yintercept = median(VC0.5$len), col = "magenta")
g <- g + geom_hline(yintercept = median(VC1$len), col = "orange")
```

```
g <- g + geom_hline(yintercept = median(VC2$len), col = "lightblue")
g <- g + geom_text(aes(0,median(VC0.5$len),label = median(VC0.5$len), vjust = -1))
g <- g + geom_text(aes(0,median(VC1$len),label = median(VC1$len), vjust = -1))
g <- g + geom_text(aes(0,median(VC2$len),label = median(VC2$len), vjust = -1))
g
```



The median values depending on the dosage is given in the graph above. With %95 confidence we can say the the tooth growth increases with the increase of dose of VC supp.

The average unit growths depending on the dosage with %95 confidence interval are given below:

```
rbind(
        t.test(VC0.5$len)$conf,
        t.test(VC1$len)$conf,
        t.test(VC2$len)$conf
)
```

```
##             [,1]      [,2]
## [1,]  6.015176  9.944824
## [2,] 14.970657 18.569343
## [3,] 22.707910 29.572090
```

In conclusion, according to our analysis, with %95 confidence, we can say that:

The tooth growth with VC supp dose 0.5 is between 6.01 and 9.94 units;

The tooth growth with VC supp dose 1 is between 14.97 and 18.56 units;

the tooth growth with VC supp dose 2 is between 22.70 and 29.57 units.

Analyzing OJ Supp Effects on Tooth Growth

assumptions

To calculate the confidence interval of this hypothesis, unequal variance among groups is assumed.

As the variances are unequal, the data do not follow a T distribution. But it can be estimated by a T distribution.

H_0

The null hypothesis is that the mean of dose 0.5 is equal to the mean of dose 1.

The second null hpytohesis is that the mean of dose 1 is equal to the mean of dose 2.

calculation

First we need to organize the groups into subsets:

```
OJ0.5 <- myOJ %>% filter(dose == "0.5")
OJ1 <- myOJ %>% filter(dose == "1")
OJ2 <- myOJ %>% filter(dose == "2")
```

Then, calculating the confidence intervals:

H_0 : The mean of dose 0.5 is equal to the mean of dose 1.

```
t.test(OJ1$len - OJ0.5$len, paired = F, var.equal = F)
```

```
##
##  One Sample t-test
##
## data:  OJ1$len - OJ0.5$len
## t = 4.1635, df = 9, p-value = 0.002435
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   4.324616 14.615384
## sample estimates:
## mean of x
##      9.47
```

The p value is smaller than 0.05. So we reject the null hypotesis. The observed mean difference is 4.1635 estimated standard deviations away from the OJ1 mean. Of which the probability of happening if the null hypothesis is true is 0.002435.

The difference of means is 9.47 units.

H_0 : The mean of dose 1 is equal to the mean of dose 2.

```
t.test(OJ2$len - OJ1$len, paired = F, var.equal = F)
```

```
##
##  One Sample t-test
##
## data:  OJ2$len - OJ1$len
## t = 1.9435, df = 9, p-value = 0.08384
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
```

```
## -0.5509376  7.2709376
## sample estimates:
## mean of x
##      3.36
```

The p value is larger than 0.05.

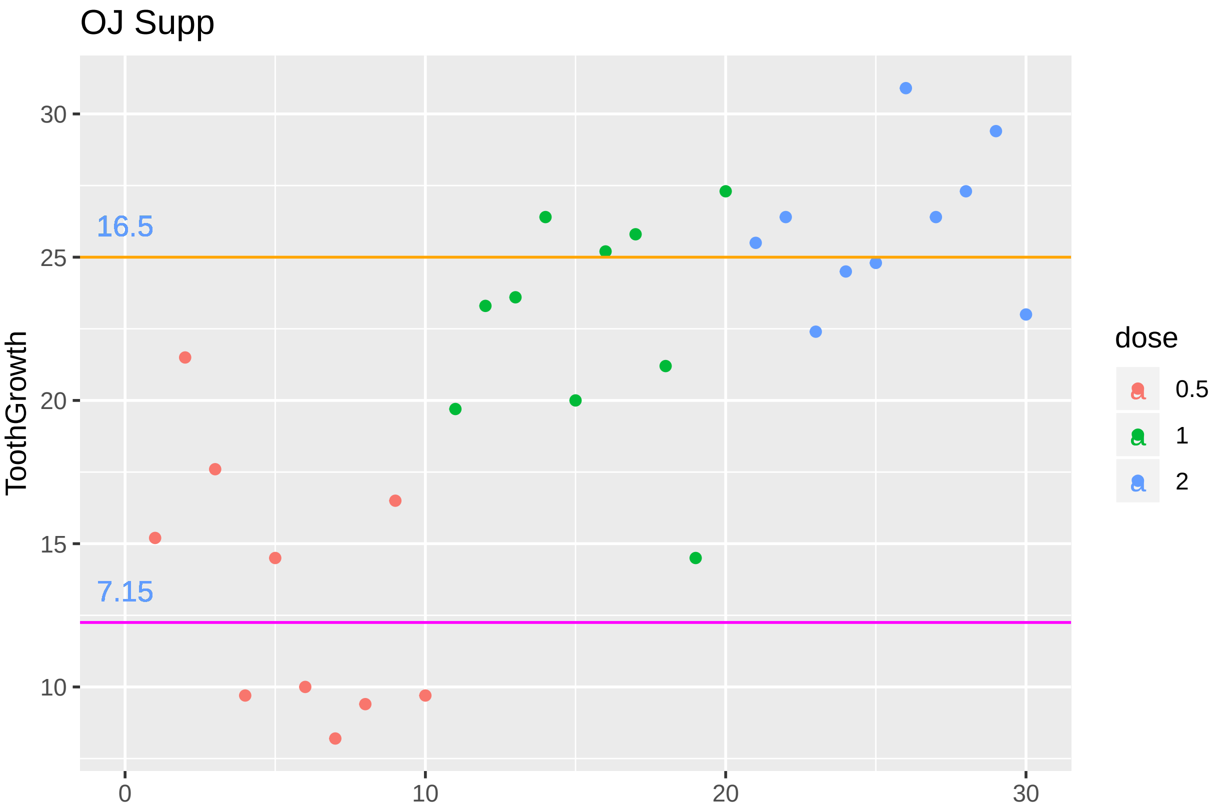
So we fail to reject the null hypothesis.

The observed mean difference is 1.9435 estimated standard deviations away from the OJ2 mean. Of which the probability of happening if the null hypothesis is true is 0.08384.

The difference of means is 3.36 units.

Let's visualize the medians on a graph

```
myOJ12 <- rbind(OJ1, OJ2)
g2 <- qplot(y = myOJ$len, data = myOJ, col = dose, ylab = "ToothGrowth", main = "OJ Supp")
g2 <- g2 + geom_hline(yintercept = median(OJ0.5$len), col = "magenta")
g2 <- g2 + geom_hline(yintercept = median(myOJ12$len), col = "orange")
g2 <- g2 + geom_text(aes(0,median(OJ0.5$len),label = median(VC0.5$len), vjust = -1))
g2 <- g2 + geom_text(aes(0,median(myOJ12$len),label = median(VC1$len), vjust = -1))
g2
```



The median values depending on the dosage is given in the graph above. With %95 confidence we can say the the tooth growth increases if dosage of OJ is >0.5.

The average unit growths depending on the dosage with %95 confidence interval are given below:

```
rbind(
        t.test(OJ0.5$len)$conf,
        t.test(myOJ12$len)$conf
)
```

```
##           [,1]     [,2]
## [1,] 10.03972 16.42028
## [2,] 22.65688 26.10312
```

In conclusion, according to our analysis, with %95 confidence, we can say that:

The tooth growth with OJ supp dose 0.5 is between 10.03 and 16.42 units;

The tooth growth with OJ supp dose 1 or dose 2 is between 22.65 and 26.10 units.