Part two

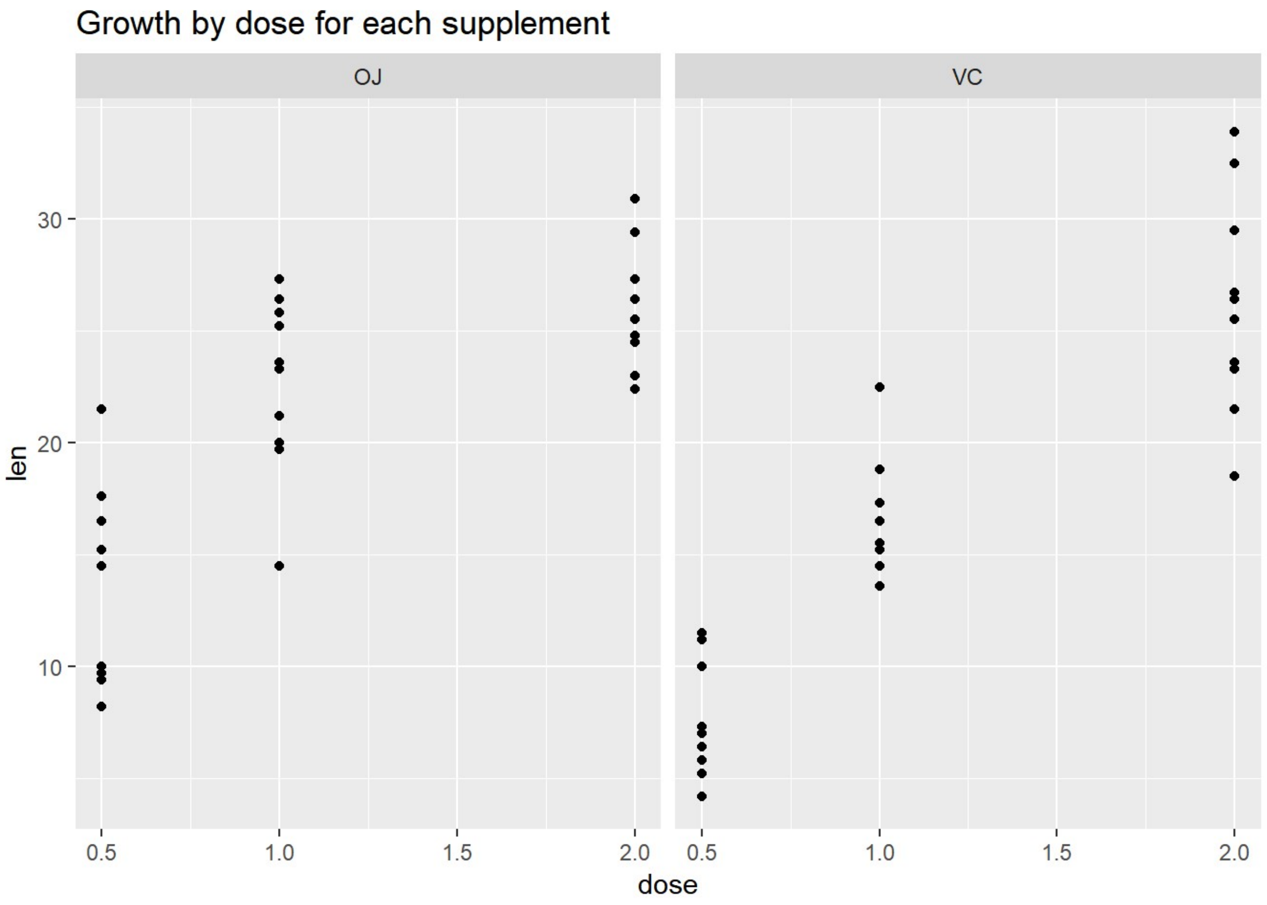
Rohil

February 19, 2019

Basic Inferential Data Analysis

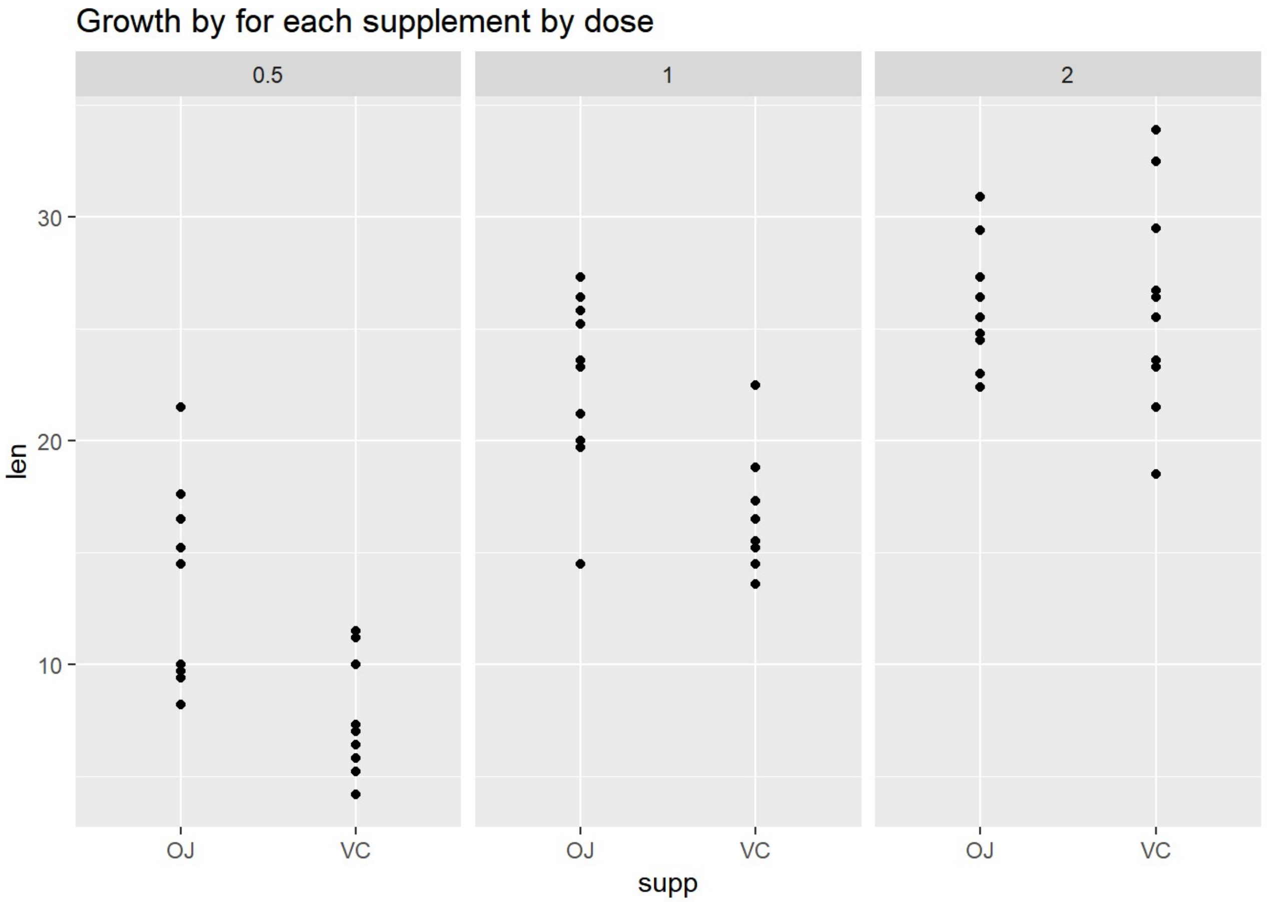
1. Let us first look at a few exploratory charts

```
grouped_data<-group_by(ToothGrowth,supp,dose)
a<-ggplot(grouped_data,aes(dose,len))+geom_point()+labs(title="Growth by dose for eac
h supplement", ylab = "Length")+facet_grid(.~supp)
a
```



Looking at the above chart it seems the median length for each dose with the OJ supplement, is higher than VC.

```
a<-ggplot(grouped_data,aes(supp,len))+geom_point()+labs(title="Growth by for each supp
lement by dose", ylab = "Length")+facet_grid(.~dose)
a
```



Again looking at the length, seperated by dose, we see the median of the length for OJ is higher than VS for each dose. We also see as the dose increases as does the median length for both supplements.

2. Now we will do a quick summary of the data

```
summarise(grouped_data,mean(len),sd(len),median(len))
```

```
## # A tibble: 6 x 5
## # Groups:   supp [?]
##   supp   dose `mean(len)` `sd(len)` `median(len)`
##   <fct> <dbl>       <dbl>     <dbl>         <dbl>
## 1 OJ      0.5       13.2       4.46         12.2
## 2 OJ      1         22.7       3.91         23.5
## 3 OJ      2         26.1       2.66         26.0
## 4 VC      0.5        7.98      2.75          7.15
## 5 VC      1         16.8       2.52         16.5
## 6 VC      2         26.1       4.80         26.0
```

We have calculated the mean, standard deviation and median for each combination of supplement and dose, and we see that indeed the OJ supplement has a higher length across the doses compared to VC. The standard deviation varies does not seem to show a pattern.

3. There are several hypothesis tests we can do, first we can see if there is any difference in length between doses 0.5,1 and 2, for each supplement, then we can also see if there is any difference is length by supplements for by dose. Our null hypothesis for all these tests is that there is no difference in length. The alternative can either be the mean in difference is not equal to 0, or that the difference is in a specific direction.

Thus now we split the data by dose and supplement.

```
splitdata<-split(ToothGrowth,list(ToothGrowth$supp,ToothGrowth$dose))
print(splitdata)
```

```
## $OJ.0.5
##     len supp dose
## 31 15.2   OJ  0.5
## 32 21.5   OJ  0.5
## 33 17.6   OJ  0.5
## 34  9.7   OJ  0.5
## 35 14.5   OJ  0.5
## 36 10.0   OJ  0.5
## 37  8.2   OJ  0.5
## 38  9.4   OJ  0.5
## 39 16.5   OJ  0.5
## 40  9.7   OJ  0.5
## 
## $VC.0.5
##     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
## 7  11.2   VC  0.5
## 8  11.2   VC  0.5
## 9   5.2   VC  0.5
## 10  7.0   VC  0.5
## 
## $OJ.1
##     len supp dose
## 41 19.7   OJ    1
## 42 23.3   OJ    1
## 43 23.6   OJ    1
## 44 26.4   OJ    1
## 45 20.0   OJ    1
## 46 25.2   OJ    1
## 47 25.8   OJ    1
## 48 21.2   OJ    1
## 49 14.5   OJ    1
## 50 27.3   OJ    1
## 
## $VC.1
##     len supp dose
## 11 16.5   VC    1
## 12 16.5   VC    1
## 13 15.2   VC    1
## 14 17.3   VC    1
## 15 22.5   VC    1
## 16 17.3   VC    1
## 17 13.6   VC    1
## 18 14.5   VC    1
## 19 18.8   VC    1
## 20 15.5   VC    1
## 
## $OJ.2
##     len supp dose
## 51 25.5   OJ    2
## 52 26.4   OJ    2
## 53 22.4   OJ    2
## 54 24.5   OJ    2
## 55 24.8   OJ    2
## 56 30.9   OJ    2
## 57 26.4   OJ    2
## 58 27.3   OJ    2
## 59 29.4   OJ    2
## 60 23.0   OJ    2
## 
## $VC.2
##     len supp dose
## 21 23.6   VC    2
## 22 18.5   VC    2
## 23 33.9   VC    2
## 24 25.5   VC    2
## 25 26.4   VC    2
## 26 32.5   VC    2
## 27 26.7   VC    2
## 28 21.5   VC    2
## 29 23.3   VC    2
## 30 29.5   VC    2
```

In the first t-test we llok at the difference in length for each supplement for a dose of 0.5, from the exploratory charts and the summary it seems that the length for OJ for 0.5 is greater than VC, hence we will used a one tail test to see if the the mean length for OJ is indeed greater than the mean length for VC. We will use a signifgance of 0.05 to reject the null hypothesis

```
t.test(splitdata[["OJ.0.5"]]$len,splitdata[["VC.0.5"]]$len, alternative=c("greater"))
```

```
##
##  Welch Two Sample t-test
##
## data:  splitdata[["OJ.0.5"]]$len and splitdata[["VC.0.5"]]$len
## t = 3.1697, df = 14.969, p-value = 0.003179
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  2.34604     Inf
## sample estimates:
## mean of x mean of y
##     13.23      7.98
```

Thus we see the p-values is 0.003, which suggests if the null hypothesis were true, our observation would be very unlikely, hence we reject it. Looking at the confidence interval we see it does not contain 0, hence again we reject the null hypothesis. Thus our t-test indicates that the mean legnth for the OJ suppliment is higher than the VC suppliment.

Now we perform the same t-test, excpet for a dose of 1

```
t.test(splitdata[["OJ.1"]]$len,splitdata[["VC.1"]]$len, alternative=c("greater"))
```

```
##
##  Welch Two Sample t-test
##
## data:  splitdata[["OJ.1"]]$len and splitdata[["VC.1"]]$len
## t = 4.0328, df = 15.358, p-value = 0.0005192
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  3.356158      Inf
## sample estimates:
## mean of x mean of y
##     22.70     16.77
```

Again our p-value is far less than 0.05, so again we reject the null hypothesis and again this suggests the mean growth in OJ is better than VC.

We now perfrom the t-test for a dose of 2, however from our exploratory charts and summary, the mean length is actually identical, so instead of a one-sided hyptohesis test, we will do a two-sided hypothesis test.

```
t.test(splitdata[["OJ.2"]]$len,splitdata[["VC.2"]]$len)
```

```
##
##  Welch Two Sample t-test
##
## data:  splitdata[["OJ.2"]]$len and splitdata[["VC.2"]]$len
## 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
```

Here we see that the p-value is 0.9369, which suggests that the difference we see is highly likely given the null hypothesis. Hence we do not reject the null hypothesis, so for a dose of 2, there is no difference between the supliments.

Now we do hypothesis tests for just dosage, here we test if there is any difference in length when the dosage is 0.5 vs 1, 1 vs 2 and 0.5 vs 2. First we split the data by dosage

```
splitdata<-split(ToothGrowth,list(ToothGrowth$dose))
print(splitdata)
```

```
## $`0.5`
##     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
## 7  11.2   VC  0.5
## 8  11.2   VC  0.5
## 9   5.2   VC  0.5
## 10  7.0   VC  0.5
## 31 15.2   OJ  0.5
## 32 21.5   OJ  0.5
## 33 17.6   OJ  0.5
## 34  9.7   OJ  0.5
## 35 14.5   OJ  0.5
## 36 10.0   OJ  0.5
## 37  8.2   OJ  0.5
## 38  9.4   OJ  0.5
## 39 16.5   OJ  0.5
## 40  9.7   OJ  0.5
##
## $`1`
##     len supp dose
## 11 16.5   VC    1
## 12 16.5   VC    1
## 13 15.2   VC    1
## 14 17.3   VC    1
## 15 22.5   VC    1
## 16 17.3   VC    1
## 17 13.6   VC    1
## 18 14.5   VC    1
## 19 18.8   VC    1
## 20 15.5   VC    1
## 41 19.7   OJ    1
## 42 23.3   OJ    1
## 43 23.6   OJ    1
## 44 26.4   OJ    1
## 45 20.0   OJ    1
## 46 25.2   OJ    1
## 47 25.8   OJ    1
## 48 21.2   OJ    1
## 49 14.5   OJ    1
## 50 27.3   OJ    1
##
## $`2`
##     len supp dose
## 21 23.6   VC    2
## 22 18.5   VC    2
## 23 33.9   VC    2
## 24 25.5   VC    2
## 25 26.4   VC    2
## 26 32.5   VC    2
## 27 26.7   VC    2
## 28 21.5   VC    2
## 29 23.3   VC    2
## 30 29.5   VC    2
## 51 25.5   OJ    2
## 52 26.4   OJ    2
## 53 22.4   OJ    2
## 54 24.5   OJ    2
## 55 24.8   OJ    2
## 56 30.9   OJ    2
## 57 26.4   OJ    2
## 58 27.3   OJ    2
## 59 29.4   OJ    2
## 60 23.0   OJ    2
```

Now we do a t.test for a dosage of 0.5 vs 1, from the exploratory data, we see that the mean length is higher for a dosage of 1, thus we will do a one-sided t-test where the alternate hypothesis is that the mean length of dose of 0.5 is less than 1

```
t.test(splitdata[["0.5"]]$len,splitdata[["1"]]$len, alternative=c("less"))
```

```
## 
##  Welch Two Sample t-test
## 
## data:  splitdata[["0.5"]]$len and splitdata[["1"]]$len
## t = -6.4766, df = 37.986, p-value = 6.342e-08
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -6.753323
## sample estimates:
## mean of x mean of y 
##    10.605    19.735
```

The p-value for this is test is basically 0, thus this suggests we shoudl reject the null hypothesis, and the mean length from dosage 1 is higher than 0.5.

Now we do the test for 1 vs 2.

```
t.test(splitdata[["1"]]$len,splitdata[["2"]]$len, alternative=c("less"))
```

```
##
##  Welch Two Sample t-test
##
## data:  splitdata[["1"]]$len and splitdata[["2"]]$len
## t = -4.9005, df = 37.101, p-value = 9.532e-06
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf -4.17387
## sample estimates:
## mean of x mean of y
##    19.735    26.100
```

Again the p-value for this is test is basically 0, thus this suggests we shoudl reject the null hypothesis, and the mean length from dosage 2 is higher than 1.

From the t-tests we have done we see that as we increase the dosage the mean length increase. We also see that for lower dosages OJ provides a higher mean length than VC, but for a dosage of 2, there seems to be no difference. Our assumptions to reach this conclusion is that the sample average will be distributed according to the central limit theorem and that the signifigance of 0.05 is a good threshold for these tests.