Course Project Inference - 2

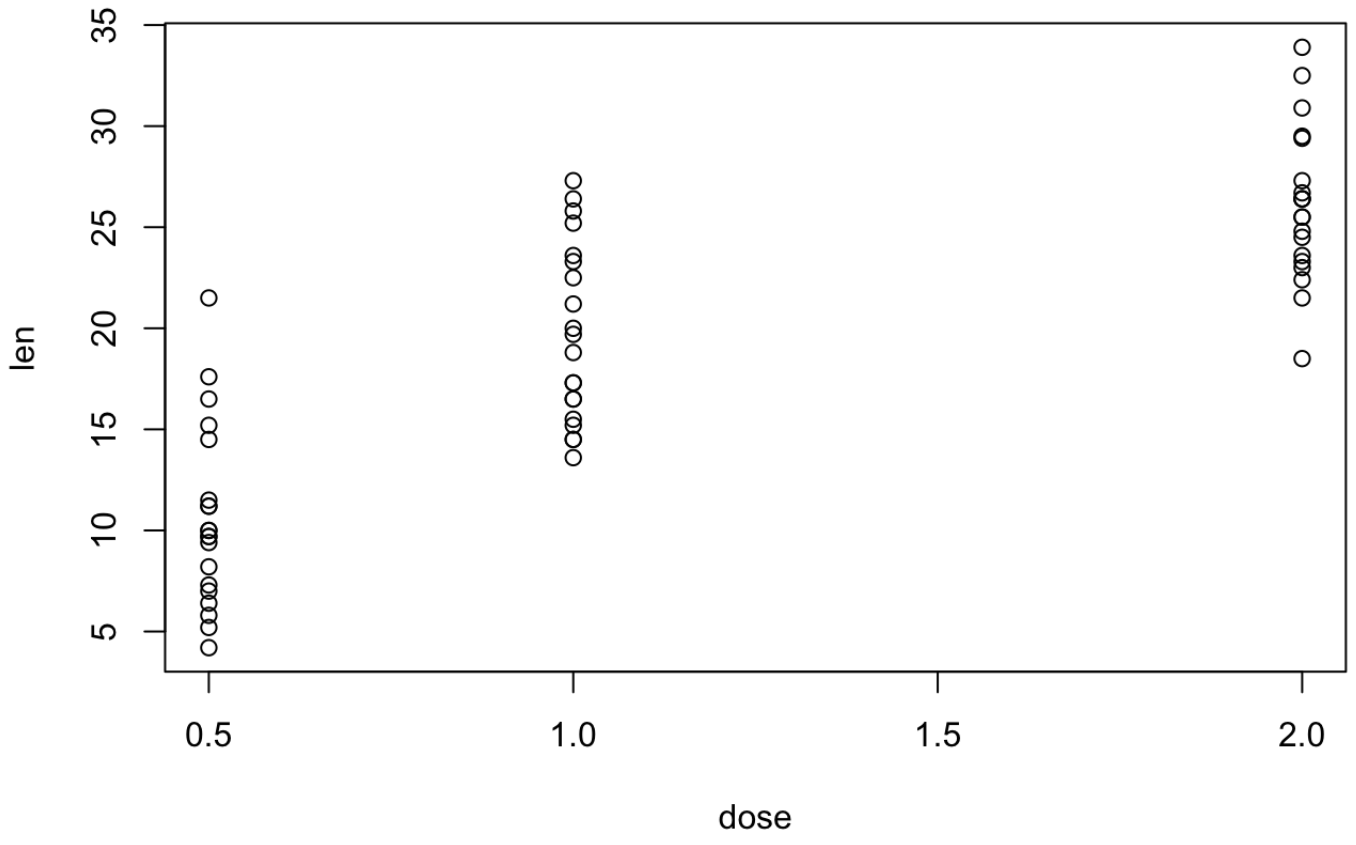
Question 1 - Load the Tooth Growth data and perform analysis

To begin let's load in the data set and compute some basic stats:

```
data(ToothGrowth); 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
```

Given that there are only two numeric variables we can do a quick scatter plot



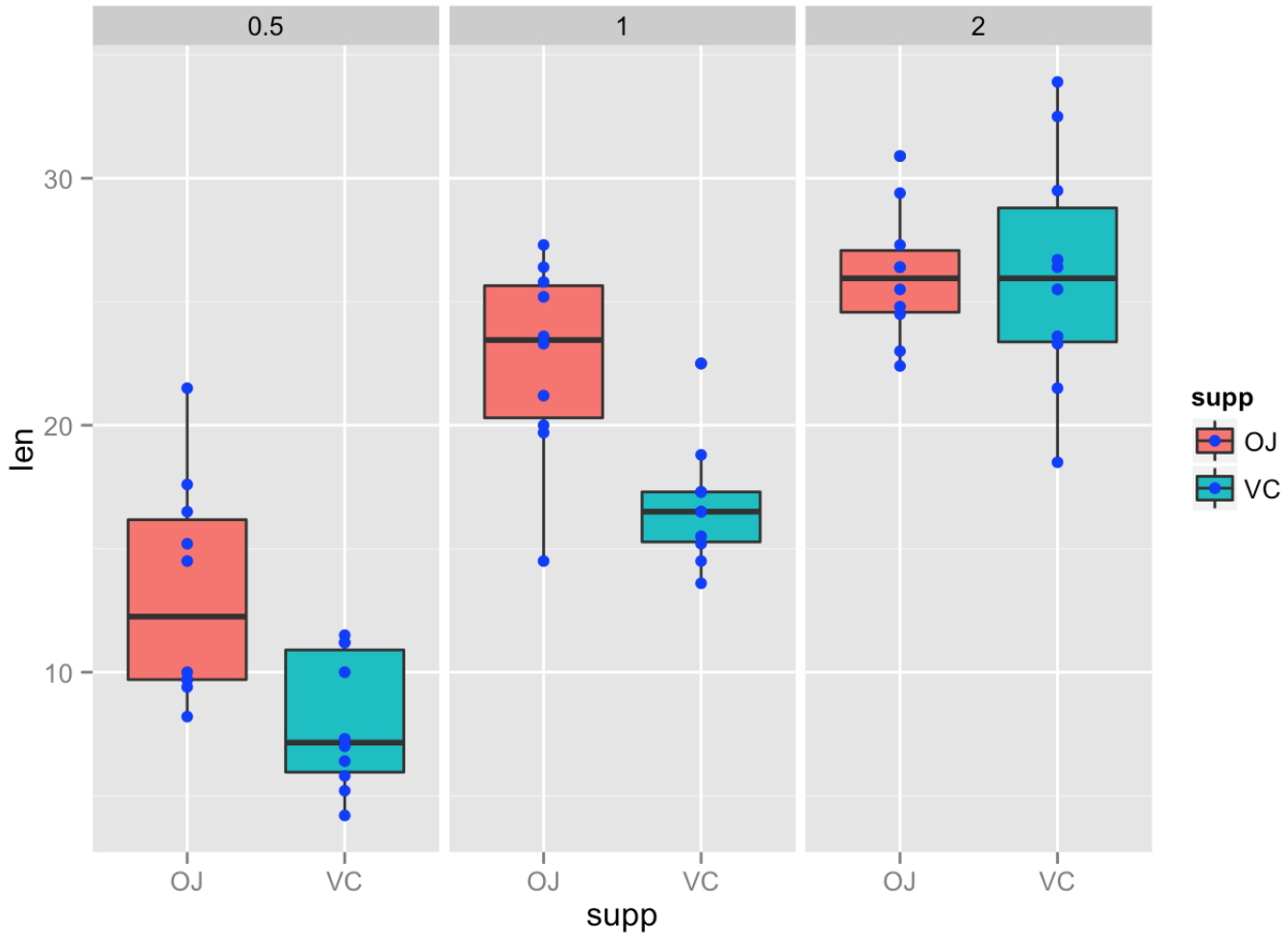
It's clear that the dose variable is also a factor:

```
ToothGrowth$dose <- factor(ToothGrowth$dose)
```

Given that 2 of the variables are categorical and only 1 is numeric let's plot the data with boxplot seperated by facets using the ggplot2 library.

Here we can clearly see the distribution of the len variable based on each of the two factors

```
library(ggplot2)
p <- ggplot(ToothGrowth, aes(supp, len, fill = supp)) + geom_boxplot() + geom_point(color = "
blue")
p + facet_grid(. ~ dose)
```



Question 2- provide a basic summary of the data

In order to do a basic summary we'll just look at some key stats relative to the categories of interest: dose and Supplemnt Type

```
with(ToothGrowth, tapply(len, list(dose=dose, supp=supp), mean))
```

```
##      supp
## dose     OJ    VC
##   0.5 13.23  7.98
##   1   22.70 16.77
##   2   26.06 26.14
```

```
with(ToothGrowth, tapply(len, list(dose=dose, supp=supp), sd))
```

```
##      supp
## dose        OJ       VC
##   0.5 4.459709 2.746634
##   1   3.910953 2.515309
##   2   2.655058 4.797731
```

```
with(ToothGrowth, tapply(len, list(dose=dose, supp=supp), median))
```

```
##      supp
## dose     OJ    VC
##   0.5 12.25  7.15
##   1   23.45 16.50
##   2   25.95 25.95
```

Question 3 - Use confidence intervals and hypothesis tests to compare tooth growth by supp and dose.

To begin we'll perform an unpaired T-Test comparing the OJ treatment and the VC treatment for each dose level. We'll also assume that the variances are different in the different groups(this seems to be fair assumption given the sd based on the samples for dose vs supp). The t.test function handles the calculations of the confidence interval as well as the hypothesis test.

```
for (i in levels(ToothGrowth$dose)){
  t_results <- t.test(len ~ supp,
                      data = subset(ToothGrowth, ToothGrowth$dose == i),
                      var.equal = FALSE,
                      paired = FALSE
                      )
  print(paste("Dose level:", i))
  print(t_results)
  }
```

```
## [1] "Dose level: 0.5"
##
##  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
##
## [1] "Dose level: 1"
##
##  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
##
## [1] "Dose level: 2"
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = -0.0461, 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
```

Question 4 - State your conclusions and the assumptions needed for your conclusions.

The results of the t-test indicate that for doses of 0.5 or 1 milligrams there is a meaningful difference in the average outcome and that it is likely that those who use the OJ supplement will see greater lengths than those who use Vitamin C. For the two cases where the doses are 0.5 and 1 mg we reject the null hypothesis, that the means are equal, because of the low p-value associated with that hypothesis. Furthermore, we concluded that OJ is likely to have a positive impact on tooth length, because the confidence intervals for the difference of the means are greater than zero in both cases.

For the 2 mg case there is insuffucuent cause to reject the null hypothesis that the means are equal. Additionally, the confidence interval passes through zero and the boxplot indicates that the two applications have nearly identical means and symmetric distributions.