IST 772 Homework 4

Maya Mileva

due date: Oct 31th, 2019

I did this homework by myself, with help from the book and the professor.

```
## Run these functions to get a clean test of homework code
## dev.off() # Clear the graph window
cat('\014')  # Clear the console
```

```
rm(list=ls()) # Clear user objects from the environment
```

Exercises

7. The built-in PlantGrowth data set contains three different groups, each representing a different plant food diet (you may need to type data(PlantGrowth) to activate it). The group labeled “ctrl” is the control group, while the other two groups are each a dif- ferent type of experimental treatment. Run the summary() command on PlantGrowth and explain the output. Create a histogram of the ctrl group. As a hint about R syntax, here is one way that you can access the ctrl group data:

PlantGrowth*weight*[*PlantGrowth*group==”ctrl”]

Also create histograms of the trt1 and trt2 groups. What can you say about the differ- ences in the groups by looking at the histograms?

```
## Activate the data set
data("PlantGrowth")

dim(PlantGrowth) # A data frame of 30 cases on 2 variables
```

```
## [1] 30  2
```

```
head(PlantGrowth, n=3)
```

```
##   weight group
## 1   4.17  ctrl
## 2   5.58  ctrl
## 3   5.18  ctrl
```

```
summary(PlantGrowth)
```

```
##      weight       group   
##  Min.   :3.590   ctrl:10  
##  1st Qu.:4.550   trt1:10  
##  Median :5.155   trt2:10  
##  Mean   :5.073            
##  3rd Qu.:5.530            
##  Max.   :6.310            
```

In this data set we can find results from an experiment to compare yields(measured by dried weight of plants) obtained under a control and two different treatment conditions.

summary() reveals some overlapping information to the str command about the “weight” variable(numeric) in the data set. The “weight” in and lb, and tha manimum weight is 3.590 and the maximum is 6.310(range is fro 3.590 to 6.310 - very small).

1st Qu. refers to the dividing line at the top of the first quartile. If we look at all the weights and line them up in order, we can divide up the whole into 4 groups, where each group had the same number of observations(smallest on the left and largest on the right). So the 1st Qu is the value of the weight (**4.550**) that divides the first quarter of the cases from the other three quarters.

Median refers to the value of the weight that splits the whole weight group in half(half having highter values and half - lower). We can also state that the median is the line that separates the second and third Qu. The value for our variable is **5.155**.

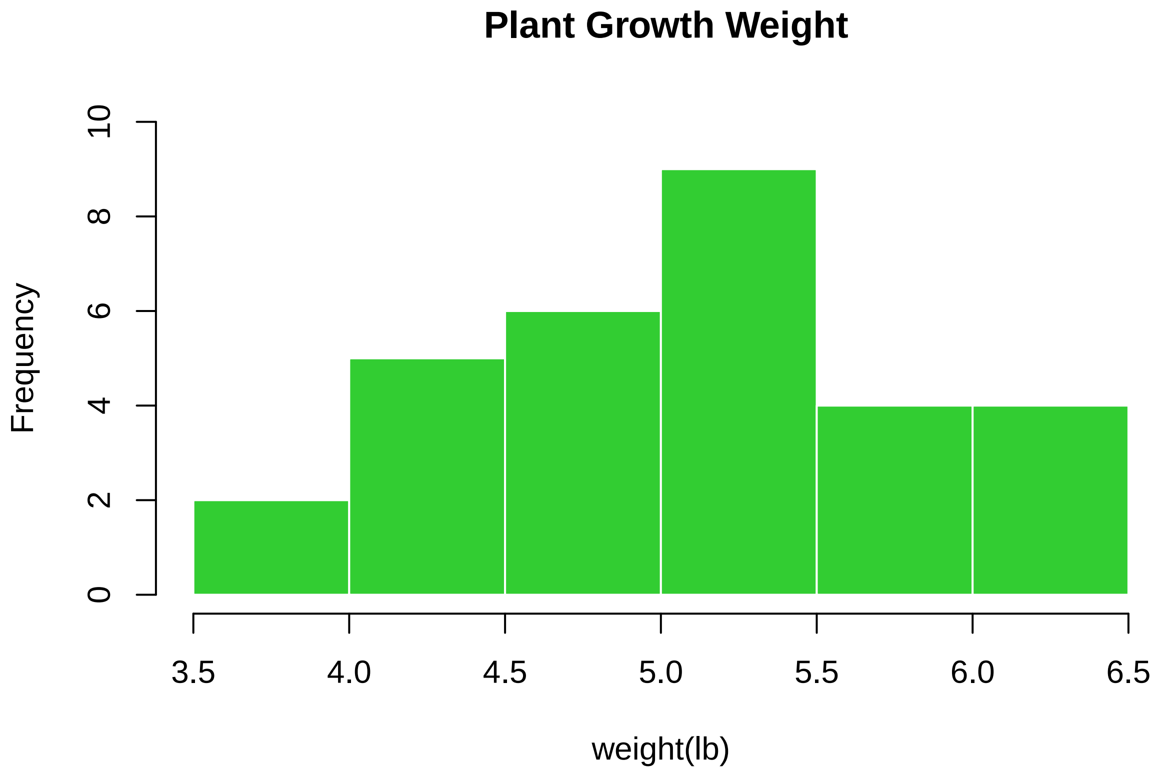
Mean shows the average weight - **5.073**.

3rd Qu is the third quartile(**5.530**). It represents the third final dividing line that splits all the cases into 4 equal parts. Quartiles give a sense of the shape of the distrubution, can be used for comparisons too(if we want to know if a sample is drawn by specific data set).

“group” variable is categor. and consist of 3 equal groups of 10(representing a different plant food diet): “ctrl: control group,”trt1" and “trt2”: different type of experimantal treatment.

```r
hist(PlantGrowth$weight

     , ylim = c(0,10)
     #, xlim = c(0,10)
     , xlab = "weight(lb)"
     , main = " Plant Growth Weight"
     , border = "white"
     , col = "limegreen"
     )
```



```
## Find unique values of the group variable
unique(PlantGrowth$group)
```

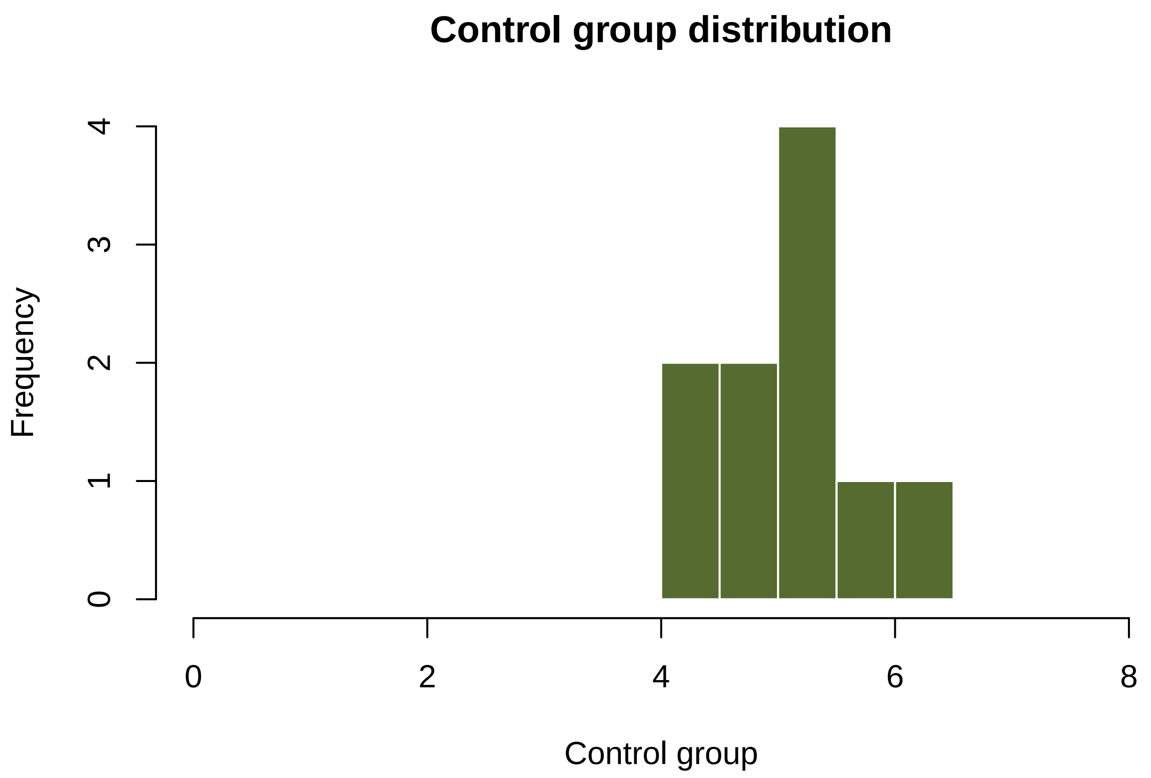
```
## [1] ctrl trt1 trt2
## Levels: ctrl trt1 trt2
```

```
## Subseting only the "ctrl"
c_gr <-PlantGrowth$weight[PlantGrowth$group=="ctrl"]

summary(c_gr) # very small range
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.170   4.550   5.155   5.032   5.293   6.110
```

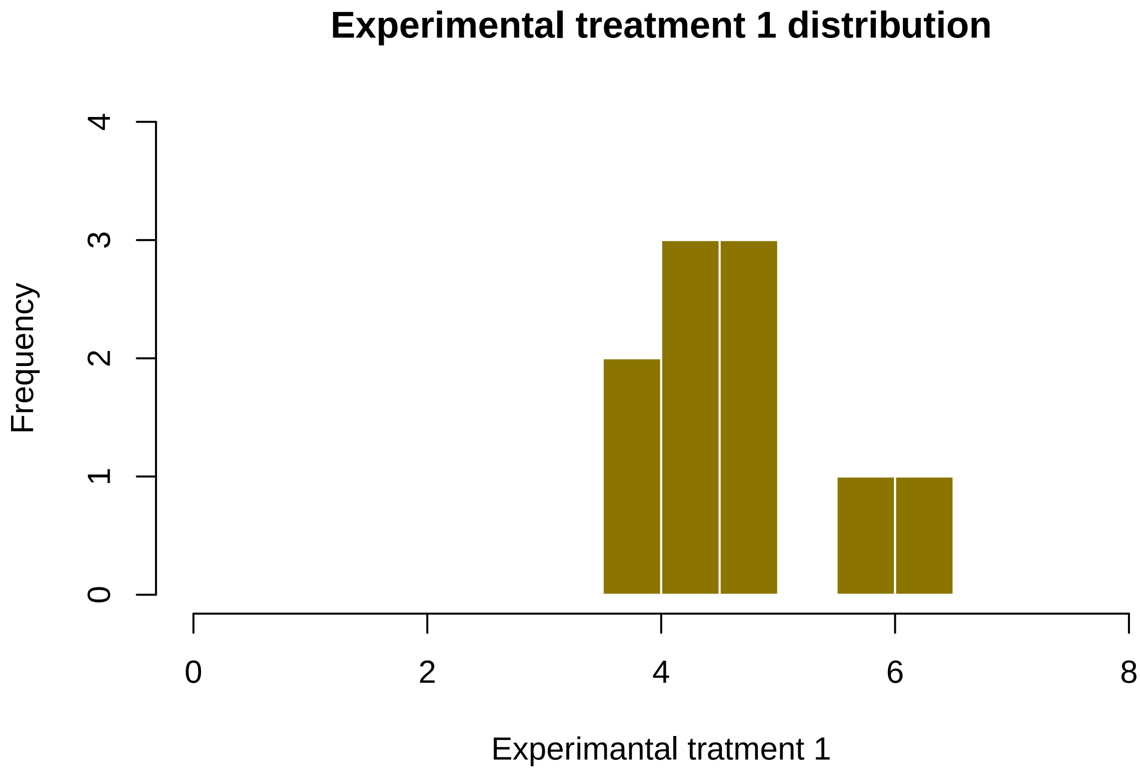
```
hist(c_gr
     , main = "Control group distribution"
     , xlab = "Control group"
     , xlim = c(0,8)
     , col = "darkolivegreen"
     , border = "white"
     )
```



```
## Subseting only the "trt1"
tr1_gr <-PlantGrowth$weight[PlantGrowth$group=="trt1"]
summary(tr1_gr) # wider range than ctrl
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.590   4.207   4.550   4.661   4.870   6.030
```

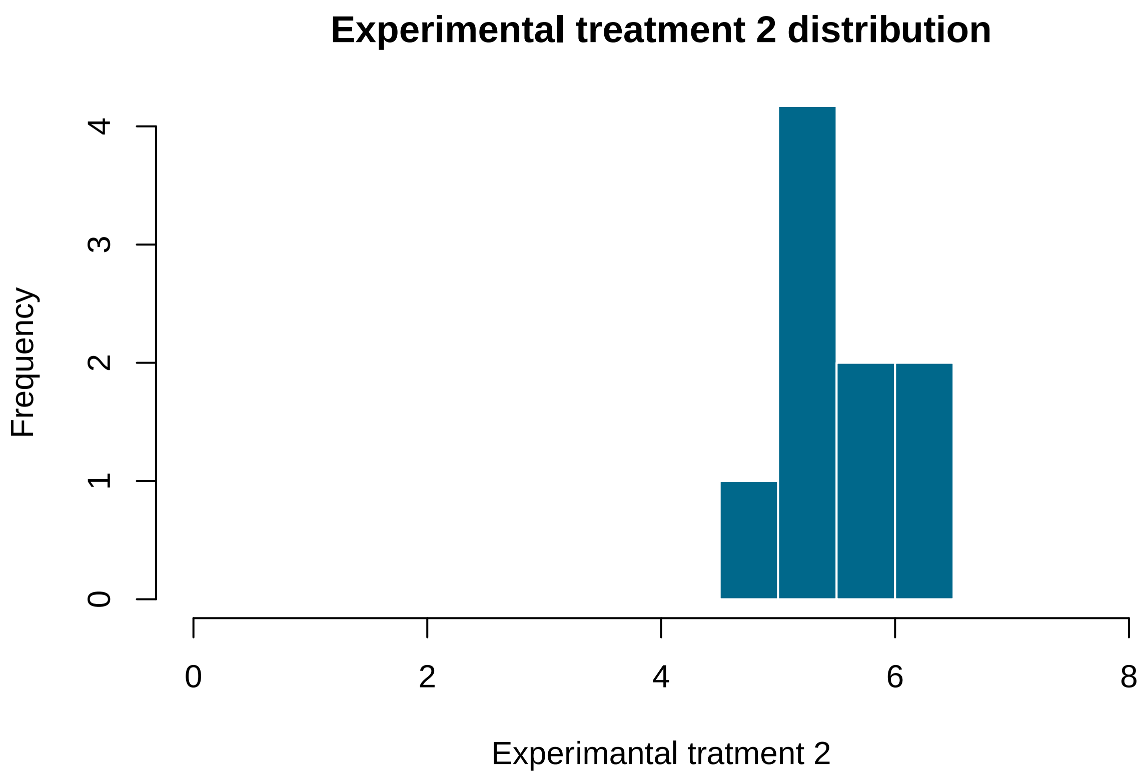
```
hist(tr1_gr
     , main = "Experimental treatment 1 distribution"
     , xlab = "Experimantal tratment 1"
     , xlim = c(0,8)
     , ylim = c(0,4)
     , col = "gold4"
     , border = "white"
     )
```



```
## Subseting only the "trt2"
tr2_gr <-PlantGrowth$weight[PlantGrowth$group=="trt2"]
summary(tr2_gr) # wider range than ctrl
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.920   5.268   5.435   5.526   5.735   6.310
```

```
hist(tr2_gr
     , main = "Experimental treatment 2 distribution"
     , xlab = "Experimantal tratment 2"
     , xlim = c(0,8)
     , ylim = c(0,4)
     , col = "deepskyblue4"
     , border = "white"
     , breaks=3
     )
```

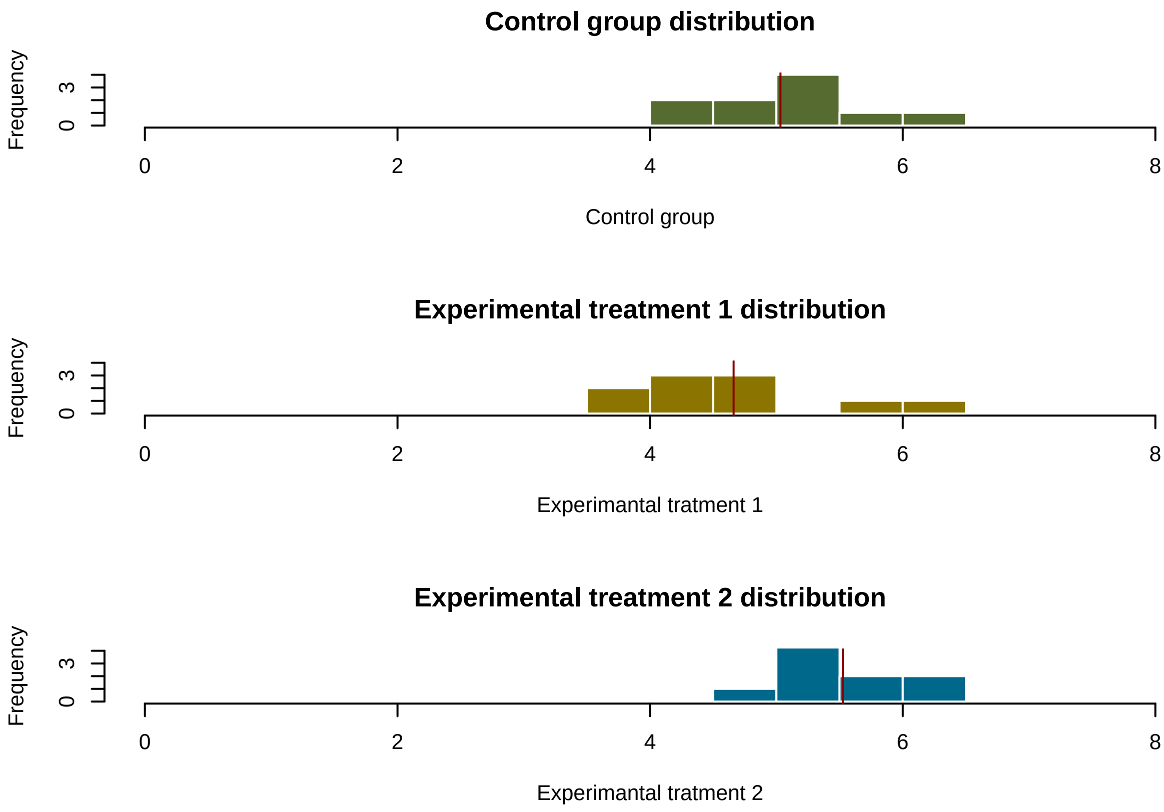


```
par(mfrow = c(3,1))
hist(c_gr
     , main = "Control group distribution"
     , xlab = "Control group"
     , xlim = c(0,8)
     , col = "darkolivegreen"
     , border = "white"
     )
m1 <- mean(c_gr)
abline(v = m1, col = 'darkred')

hist(tr1_gr
     , main = "Experimental treatment 1 distribution"
     , xlab = "Experimantal tratment 1"
     , xlim = c(0,8)
     , ylim = c(0,4)
     , col = "gold4"
     , border = "white"
     )
m2 <- mean(tr1_gr)
abline(v = m2, col = 'darkred')

hist(tr2_gr
     , main = "Experimental treatment 2 distribution"
     , xlab = "Experimantal tratment 2"
     , xlim = c(0,8)
```

```
      , ylim = c(0,4)
      , col = "deepskyblue4"
      , border = "white"
      , breaks=3
      )
m3 <- mean(tr2_gr)
abline(v = m3, col = 'darkred')
```

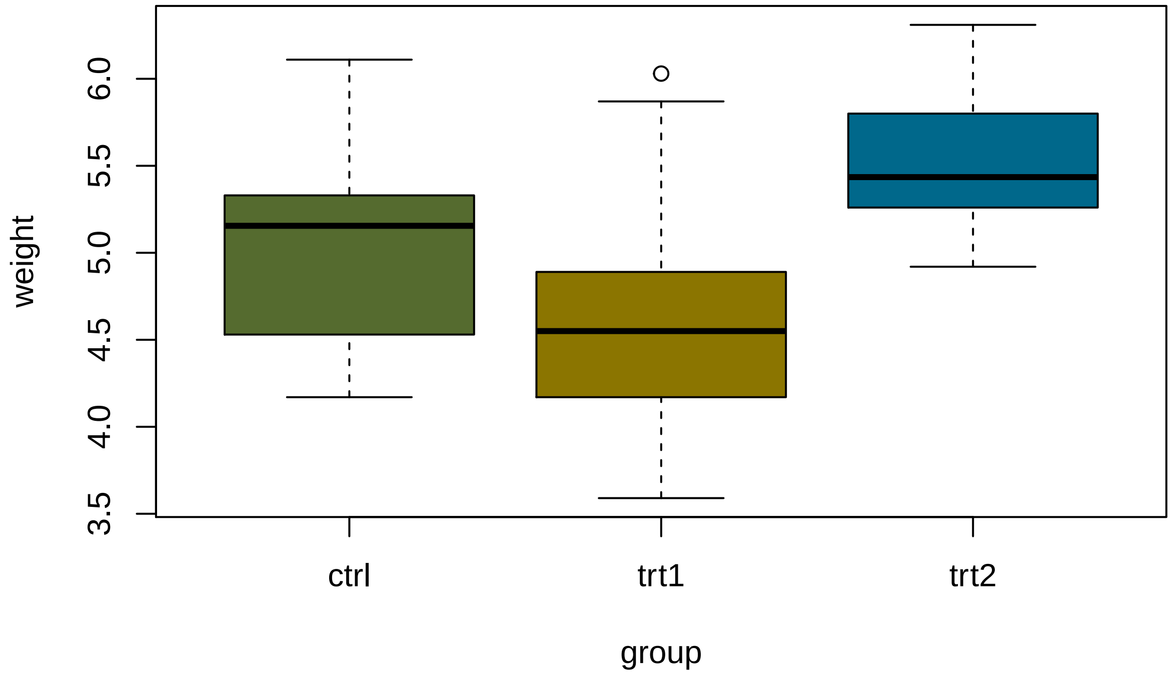


We can see that treatment 2 has much less variance. The control group shows the most variance from the mean. The are all really tight and not normali ditributed. With smaller sample sizes tails are higher than the normal curve , representing greater uncertainty.

8. Create a boxplot of the plant growth data, using the model “weight ~ group.” What can you say about the differences in the groups by looking at the boxplots for the dif- ferent groups?

```
## Creating boxplot wight grouped by group
boxplot(weight~group, data = PlantGrowth
        , col = c("darkolivegreen", "gold4", "deepskyblue4")

)
```



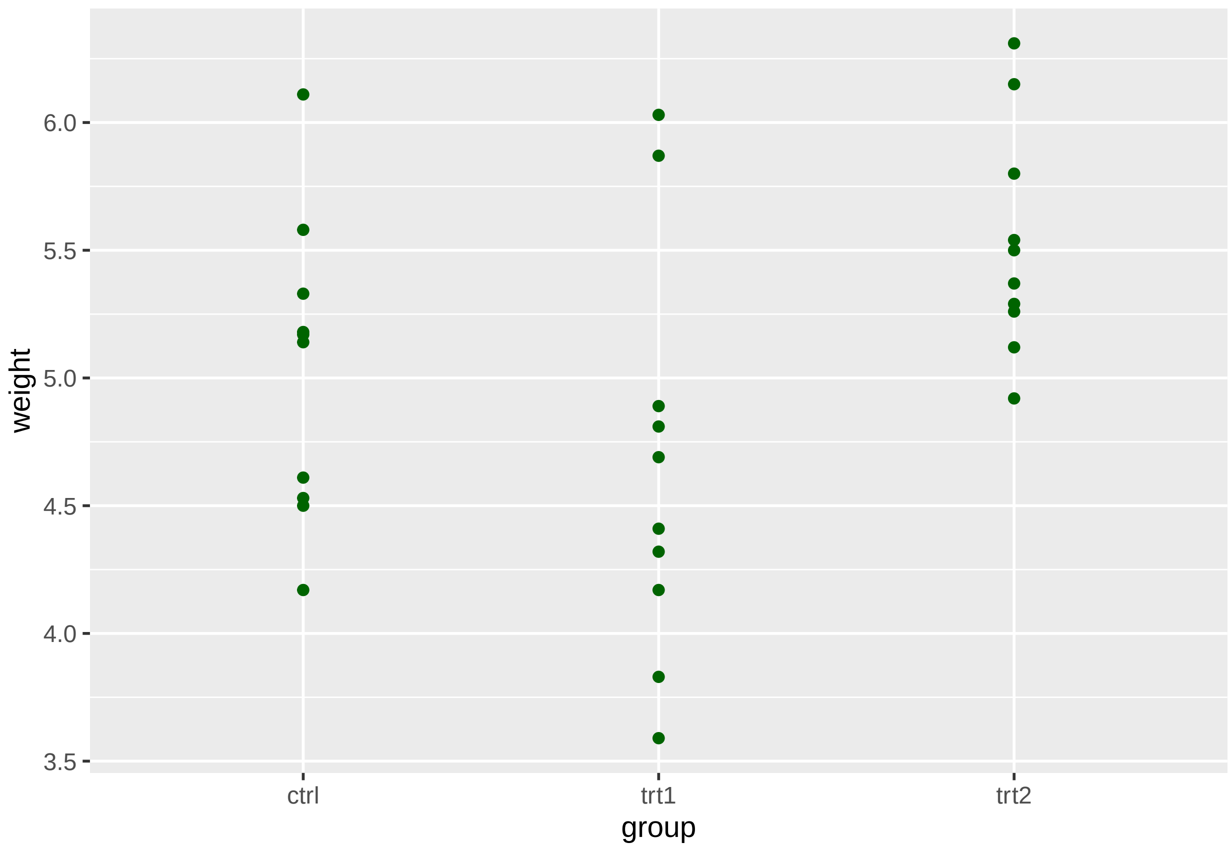
Box plots pack a lot of information into a small peice. In each case upper and lower boundaries of the box represent the first and third quartiles respectively(25% of the cases are above the box and 25% are bolow the box). The dark band in the middle represents the median for each variable. We can see that in the "ctrl" , the median is quite clese to the third quartile, indicating that 25% of cases are clustering in that small region of 5.15lb and 5.29lb. in these boxplots wiskers represents the position of the max and min values, respectively. In the "trt1" we can notices highest "extreme" value represented by outlier. The boxplots fror the three groups overlap which indicate similarity between these groups.In comparing the hight of the boxes, we can notice "trt2" is considered less variable compared to the other two. The wiskers show that very lowest value fro "ctrl" falls at the first quartile for the "trt1" and the lowest value of "trt2" falls at the third quartile of "trt1" reafirming their difference. We are working with samples so we can't be certain that the difference between these three groups can be trusted.

"trt1" has an outlier, which can skew the data distribution. "trt2" has higher yeilds in general and like we notices from the histograms less variance. We can see again in "ctrl" gr that the mean is toward the upper end so that means bigger variance from it. It looks like trt1 performed worse.

"trt1" has higher values and higher values on average so that can be the best treatment to recomend, but we can never be completely sure.

If we want to see better picture on their performance we can create different plot.

```
library(ggplot2)
best <- ggplot(PlantGrowth, aes(x=group, y=weight))+geom_point(color="darkgreen")
best
```



___ 9. Run a t-test to compare the means of ctrl and trt1 in the PlantGrowth data. Report and interpret the confidence interval. Make sure to include a carefully worded statement about what the confidence interval implies with respect to the population mean differ- ence between the ctrl and trt1 groups.___

```r
## Inferential test
t.test(PlantGrowth$weight[PlantGrowth$group=="ctrl"],
       PlantGrowth$weight[PlantGrowth$group=="trt1"])
```

```
##
##  Welch Two Sample t-test
##
## data:  PlantGrowth$weight[PlantGrowth$group == "ctrl"] and PlantGrowth$weight[PlantGrowth$group == "
## t = 1.1913, df = 16.524, p-value = 0.2504
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2875162  1.0295162
## sample estimates:
## mean of x mean of y
##     5.032     4.661
```

t.test() invokes the “Student‘s t-Test”. We analyzed the difference in the results between control group n=10 and treatment 1 n=10. Results showed mean difference of -1.629, indicating that on average “trt1” reach the wieght before the “ctrl” gr.
The t-test has used the two samples(ctrl and trt1), to calculate confidence interval ranging from a mean difference of -0.29 (lower end) to 1.03(upper end). We constructed 95% confidence interval around this mean difference.95% chance that the confidence interval would contain pop mean difference in the long run

(**This confidence interval may or may not contain the true population value.**) The width of the confidence band, about plus minus 0.37, gives some indication of the amount of uncertainty around the point of estimate(0.66). To reduce this uncertainty , we would have to increase sample size, reduce variability in within groups or both.

10. Run a t-test to compare the means of ctrl and trt2 in the PlantGrowth data. Report and interpret the confidence interval.

```
## Inferential test
t.test(PlantGrowth$weight[PlantGrowth$group=="trt1"],
       PlantGrowth$weight[PlantGrowth$group=="trt2"])
```

```
##
##  Welch Two Sample t-test
##
## data:  PlantGrowth$weight[PlantGrowth$group == "trt1"] and PlantGrowth$weight[PlantGrowth$group == "t
## t = -3.0101, df = 14.104, p-value = 0.009298
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.4809144 -0.2490856
## sample estimates:
## mean of x mean of y
##     4.661     5.526
```

The t-test calculated confidence interval ranging froma mean difference of -1.48 to -0.25lb. 95% confidence interval means that out of a 100 full replications of a study 95% of the confidence intervals will contain the mean difference(95 of those 100 times we would get a canfidence interval, where the population value would fall somewhere inside of the boundaries range) or 95% chance that the confidence interval would contain pop mean difference in the long run. This particular confidence interval DOES NOT necessarily contain the true population value of the mean difference. 5% falls out of the range we we can never be sure what we will get as result.

From t-test we got interval estimate of the population value (of -1.48 to -0.25lb). That range represent uncertaitly, we don't know and we can not know what the population value is. We have kind narrow interval which indicates lower uncertainty, but still only give us an idea where the true population mean difference lie, just adds to the weight of evidence. We can conclude that the span of -1.48 up to -0.25 strenghtens the weight of evidence that the population difference between "trt1" and "trt2" is a negative number somewhere in the region of -0.86 plus minus 0.62. The confidence interval doesn not **prove** that there is a difference in the treatments and the weight result, but it suggerst that possibility, and give us a sense of uncertainty of that conclusion(+ - 0.62).

You can not prove anything from samples or by using statistical inference.