Statistical Inference Part I Homework

Tim Hulak

I produced the material below with no assistance

Information on the PlantGrowth dataset was obtained from https://www.rdocumentation.org/packages/datasets/versions/3.6.2/topics/PlantGrowth

T-statistic interpretation was obtained from https://blog.minitab.com/en/statistics-and-quality-data-analysis/what-are-t-values-and-p-values-in-statistics

Degrees of freedom interpretation was obtained from https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/degrees-of-freedom/

Exercises 7-10 on page 66 of *Reasoning with Data: An Introduction to Traditional and Bayesian Statistics Using R*

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 different 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 differences in the groups by looking at the histograms?

```
data("PlantGrowth")
head(PlantGrowth)
```

```
##   weight group
## 1   4.17  ctrl
## 2   5.58  ctrl
## 3   5.18  ctrl
## 4   6.11  ctrl
## 5   4.50  ctrl
## 6   4.61  ctrl
```

```
dim(PlantGrowth)
```

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

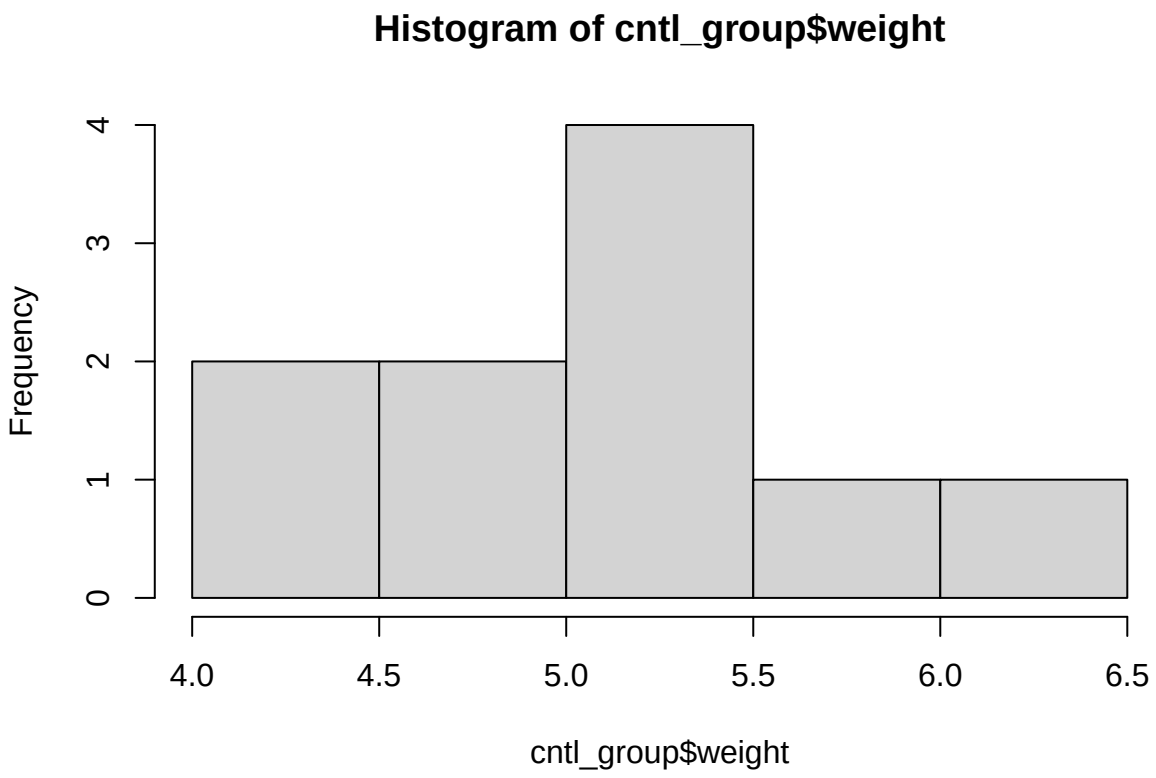
```
summary(PlantGrowth)
```

```
##      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
```

Answer: The PlantGrowth dataset are the results from an experiment to compare yields (as measured by dried weight of plants) obtained under a control and two different treatment conditions. Therefore, the output of summary(PlantGrowth) shows the minimum dried weight of all of the groups, the 1st quartile (or the value in which 25% of the data points fall under), the median dried weight, the mean (or average) dried weight, the 3rd quartile (or the value in which 75% of the data points fall under), and the maximum value dried weight of all of the groups. In addition, the summary() function show the 3 unique values (or groups) in the "groups" column and how many records there are for each.

```
cntl_group <- PlantGrowth[PlantGrowth$group == "ctrl",]
hist(cntl_group$weight)
```




```
trt1_group <- PlantGrowth[PlantGrowth$group == "trt1",]
hist(trt1_group$weight)
```

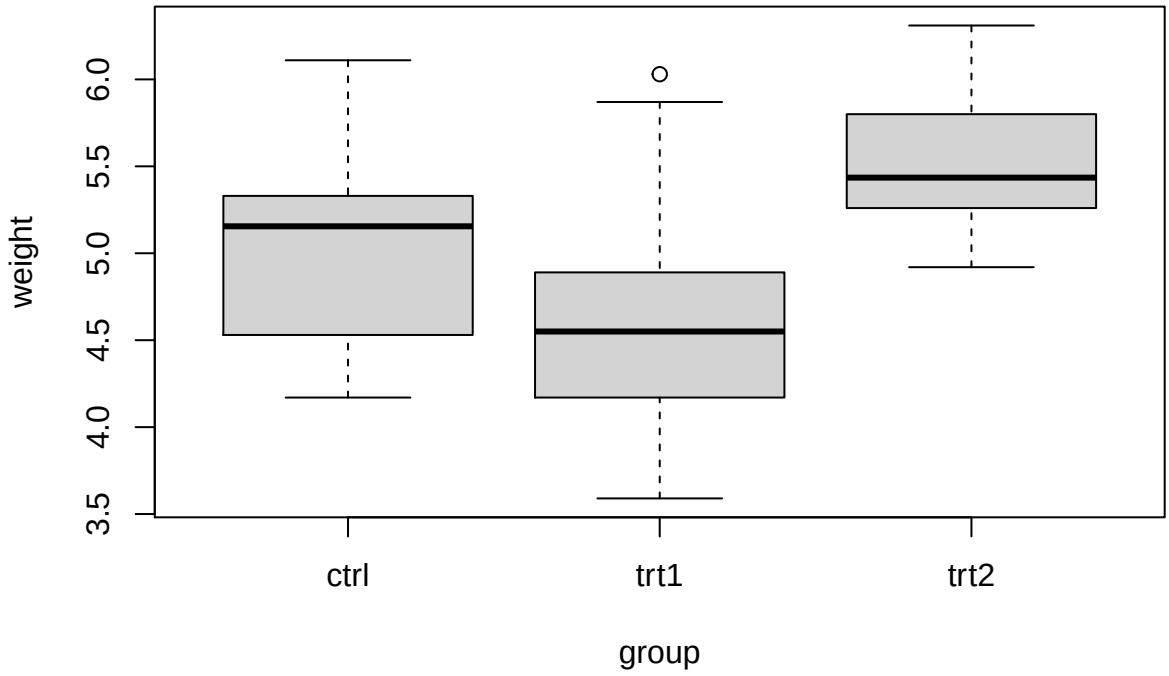

```
trt2_group <- PlantGrowth[PlantGrowth$group == "trt2",]
hist(trt2_group$weight)
```

Answer 2: Looking at the histograms for the ctrl group, trt1 group, and the trt2 group it can be seen that each of the histograms is right skewed in some way. In addition, the ctrl group and the trt2 group seem to

have the bulk of their weights between the values of 5.0 and 5.5 whereas the largest bar in the trt1 group is between 4.0 and 5.0. This means that the plants in the trt2 group are likely more similar to the ctrl group, whereas the plants in the trt1 group are likely smaller. Finally, there are gaps in the trt1 group and the trt2 group (perhaps they are bi-modal)

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 different groups?

```
boxplot(weight ~ group, data=PlantGrowth)
```



Ansswer When observing the boxplots for the different groups, it can be seen that that the most of the data points in the ctrl group are below the median weight of that group. The plot for the trt1 group shows the median weight is almost exactly in the middle of the distribution and there appears to be statistical outliers on the higher end of the group. The plot for the trt1 group shows that that group has a higher median weight than the other two groups and appears to have much less variability between the largest and smallest values in that group.

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 difference between the ctrl and trt1 groups.

```
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
```

Answer: The T-statistic for the *ctrl* and *trt1* groups is **1.1913**. Since the T-statistic is somewhat close to 0, there is evidence that there may be no (or little) difference between the means.

The degrees of freedom value is **16.524**. This means that there were **17.524** pieces of information (number of items in the sample) that went into calculating the estimate (formula: *df = n - 1*)

The p-value of the T-Test between the *ctrl* and *trt1* groups is rather high at **0.2504** (well above the conventional 0.05 threshold). This means that we can reject the alternative hypothesis (the means are equal) in favor of the null hypothesis (the true difference in means is not equal to 0). In other words, the means of the *ctrl* group are larger than the means in the *trt1* group.

The lower bound of the confidence interval is approximately **-0.288** and the upper bound of the confidence interval is approximately **1.029**. This means that if the study were run 100 times, then 95 times out of 100 the the confidence interval would contain the population mean within it's range

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

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

```
##
##  Welch Two Sample t-test
##
## data:  PlantGrowth$weight[PlantGrowth$group == "ctrl"] and PlantGrowth$weight[PlantGrowth$group == "
## t = -2.134, df = 16.786, p-value = 0.0479
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.98287213 -0.00512787
## sample estimates:
## mean of x mean of y
##     5.032     5.526
```

Answer: The T-statistic for the *ctrl* and *trt2* groups is **-2.134**. Since the T-statistic is somewhat further away from 0, there is evidence that there is a difference between the means.

The degrees of freedom value is **16.786**. This means that there were **17.786** pieces of information (number of items in the sample) that went into calculating the estimate (formula: *df = n - 1*)

The p-value of the T-Test for the *ctrl* and *trt2* groups is very low (just below 0.05 at **0.0479**). This means that we can reject the null hypothesis (the means are equal) in favor of the alternative hypothesis (the true difference in means is not equal to 0). In other words, the means of the *ctrl* group are smaller than the means in the *trt1* group.

The lower bound of the confidence interval is approximately **-0.982** and the upper bound of the confidence interval is approximately **-0.005**. This means that if the study were run 100 times, then 95 times out of 100 the the confidence interval would contain the population mean within it's range.