Course Project for 'Statistical Inference' on coursera.org
Part 2 - Study of the ToothGrowth example data

R. Rodríguez

The statement for this part of the project says:

> Now in the second portion of the class, we're going to analyze the ToothGrowth data in the R datasets package.
>
> 1. Load the ToothGrowth da2. ta and perform some basic exploratory data analyses
> 2. Provide a basic summary of the data.
> 3. Use confidence intervals and hypothesis tests to compare tooth growth by supp and dose. (Use the techniques from class even if there's other approaches worth considering)
> 4. State your conclusions and the assumptions needed for your conclusions.

NOTE: Sourcecode for this document available at https://github.com/ricrogz/StatisticInference

For this project we are going to work with data that is built into R. This data is available through the variable `ToothGrowth`. Also, a help page is available, describing the data (?ToothGrowth).

As stated in the documentation, this data comes from an essay testing the effect of Vitamin C on tooth growth in Guinea Pigs. The data consists of the measured tooth length in 6 groups of 10 animals, divided into two groups depending on the delivery method of the Vitamin C (in orange juice or as ascorbic acid): Each of the two groups is also divided into 3 subgroups according to the dosage, 0.5, 1 or 2 mg per dose.

Peeking at the data we can confirm that it matches the description. Also, as the data is quite small, we can have a look at the full table, confirming that we have 60 data samples, 10 for each combination of delivery method and dose.

Since the dosages are standard (0.5, 1, 2), we transform this variable into factor, and then we look at some summary data and make a quick exploratory plot. The code for both the transformation and the peek at the data can be seen in Appendix 2. The generated bar plot ("Figure 1") can be found in Appendix 1.

From this quick peek at the data, we can formulate some observations:

1. Vitamin C dosage seems to have an effect on teeth length, since in both delivery methods a higher Vitamin C dose is associated with longer teeth.
2. Delivery method also has an influence on tooth length, since orange juice delivery seems to be associated with longer teeth, but with the exception of the highest dosage, where tooth lengths are very close.
3. The exception mentioned in the previous point might indicate a threshold level between 1 and 2 mg, which marks a "maximum effect point", over which higher doses of Vitamin C produce no additional effect on the tooth growth.

Based on these observations, we propose the following hypotheses, and check them by means of two-group t-tests and p-values. We make no assumptions about the population variances (whether constant or different), and therefore we will use safest approach considering different variances in the populations. In any case, guinea pigs in each essay are different, and therefore we will not consider the data as "paired". The code for the tests and the R outcomes are listed in the Appendix 2.

In relation to the first observation about the dosage, our hypothesis will be that the mean tooth length corresponding to two different dosages is the same. This means checking the hypotheses pairwise. We will

make the t-test using the populations from both delivery methods together. First, we compare the 0.5 mg and 1 mg dosages.

In this case, we get an all-negative confidence interval ([-11.984, -6.276]), as well as a very small ($< 10^6$) p-value. This indicates that our hypothesis has to be rejected (our interval does not include 0, so means cannot be equal; p-value is below our α of 5%), and confirms our observation about the 1 mg-dosed population having longer teeth than the 0.5 mg one.

If we repeat the test for the 1 and 2 mg doses, we get a similar result, with a slightly higher p-value ($< 10^4$), which still leads us to reject the hypothesis of equal means, and thus confirming the hypothesis of the 2 mg dosage producing even longer teeth than the 1 mg doses.

We could do a third test with the 0.5 and 2 mg doses, but the outcome would be the same as the two previous ones, as it is a logical consequence of them, so we will skip it this time.

As for our second observation about the delivery method, in spite of our third observarion, we will make three separate tests, one for each dosage level, since we have observed different behaviors of the lower dosage levels and the higher one.

In each case, we will estabilish the null hypothesis that mean tooth growths for both delivery methods are the same, and checking it with the following tests, one for each dosage level:

The resulting p-values (0.006, 0.001 and 0.964) indicate that we have to reject our hypothesis for the lower dosages, but we fail to reject it for the 2 mg dosage. This indicates that our first observation is right in these two cases, and the delivery method indeed has an influence on the teeth length.

For the 2 mg dose, we get quite the opposite result, a very high p-value, which does not allow us to reject our null hypothesis, meaning that the tooth length means are the same, and confirming that, according to our third observation, something is happening indeed, since the clear trend in the two lower dosages is lost in the third one.

So, we can conclude that:

1. Dosage has influence on tooth length. The higher the dosage (up to 2 mg, no data is available beyond this point), the longer the teeth.
2. The delivery method is important when the dosage is between 0.5 and 1 mg, but is no longer of influence when dosing 2 mg (more data would be needed to get a more accurate threshold), with orange juice being more effective than ascorbic acid.

Appendix 1 : Figures

Generate exploratory plot:

```
ggplot(data, aes(x = dose, y = len)) + geom_boxplot(aes(fill = dose)) + facet_wrap(~ supp)
```

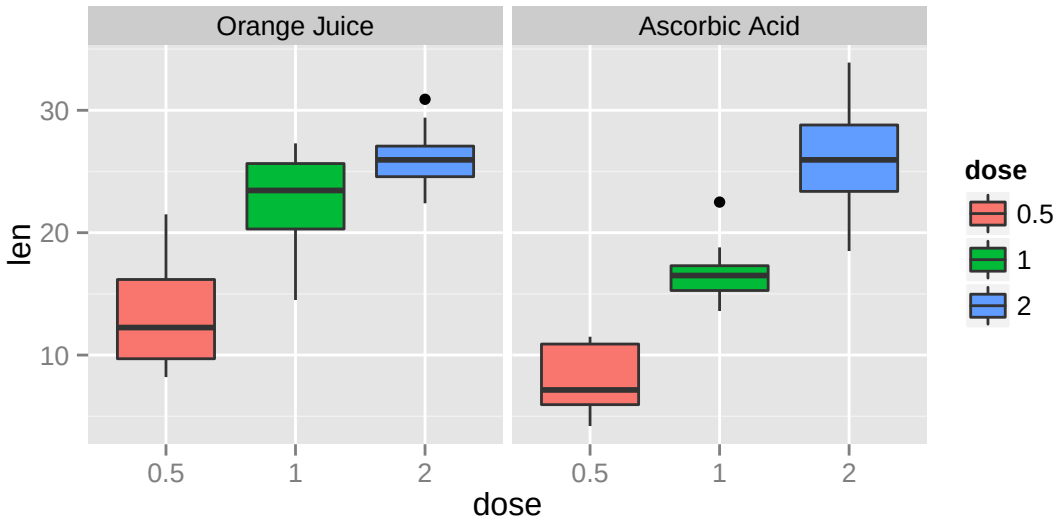

Figure 1: Exploratory plot

Appendix 2 : Code and output

```
#  Dose transformation to factor and data peek:

data <- ToothGrowth
data$dose <- factor(data$dose)
levels(data$supp) <- c("Orange Juice", "Ascorbic Acid")
str(data)
```

```
## '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 "Orange Juice",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: Factor w/ 3 levels "0.5","1","2": 1 1 1 1 1 1 1 1 1 1 ...
```

```
summary(data)
```

```
##       len                  supp     dose
##  Min.   : 4.2   Orange Juice :30   0.5:20
##  1st Qu.:13.1   Ascorbic Acid:30   1  :20
##  Median :19.2                      2  :20
##  Mean   :18.8
##  3rd Qu.:25.3
##  Max.   :33.9
```

```r
# Dosage testing: 0.5 mg vs 1 mg.

data_05_1 <- subset(data, dose %in% c(0.5,1))
t.test(len ~ dose, paired=F, var.equal=F, data=data_05_1)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -6.477, df = 37.99, p-value = 1.268e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.984  -6.276
## sample estimates:
## mean in group 0.5   mean in group 1
##             10.61             19.73
```

```r
# Dosage testing: 1 mg vs 2 mg.

data_1_2 <- subset(data, dose %in% c(1,2))
t.test(len ~ dose, paired=F, var.equal=F, data=data_1_2)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -4.901, df = 37.1, p-value = 1.906e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.996 -3.734
## sample estimates:
## mean in group 1 mean in group 2
##           19.73           26.10
```

```r
# Delivery method testing

t.test(len ~ supp, paired=F, var.equal=F, data=data[data$dose==0.5,])
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 3.17, df = 14.97, p-value = 0.006359
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.719 8.781
## sample estimates:
##  mean in group Orange Juice mean in group Ascorbic Acid
##                       13.23                       7.98
```

```r
t.test(len ~ supp, paired=F, var.equal=F, data=data[data$dose==1,])
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 4.033, df = 15.36, p-value = 0.001038
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.802 9.058
## sample estimates:
##  mean in group Orange Juice mean in group Ascorbic Acid
##                       22.70                       16.77
```

```r
t.test(len ~ supp, paired=F, var.equal=F, data=data[data$dose==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:
##  -3.798  3.638
## sample estimates:
##  mean in group Orange Juice mean in group Ascorbic Acid
##                       26.06                       26.14
```