Statistical Inference Project Part 2: Basic Inferential Data Analysis

Jared Kramer

1/22/2021

Introduction

In this report I perform basic exploratory data analysis on a dataset reflecting the effect of Vitamin C on Tooth Growth in Guinea Pigs. I also perform several T-tests analyzing whether there is any statistically significant difference in tooth growth as between (1) the delivery methods of vitamin C (orange juice or ascorbic acid) or (2) between the dose levels of vitamin C.

Exploratory Data Analysis

First, I load the `ToothGrowth` dataset, and observe using `str` that it has 60 data records across three columns (which I rename for clarity).
(1) length, a numeric data set with a variety of values, (2) supplement (indicating the type of vitamin C supplementation, per R documentation) which takes on two values ("OJ" for orange juice and "VC" for ascorbic acid) and (3) dose, indicating the vitamin C dosage, which is numeric among the set of values 0.5, 1.0 and 2.0 (mg/day).

```
library(datasets); library(dplyr)
names(ToothGrowth) <- c("length", "supplement", "dose")
str(ToothGrowth)
```

```
## 'data.frame':    60 obs. of  3 variables:
##  $ length    : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supplement: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose      : num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

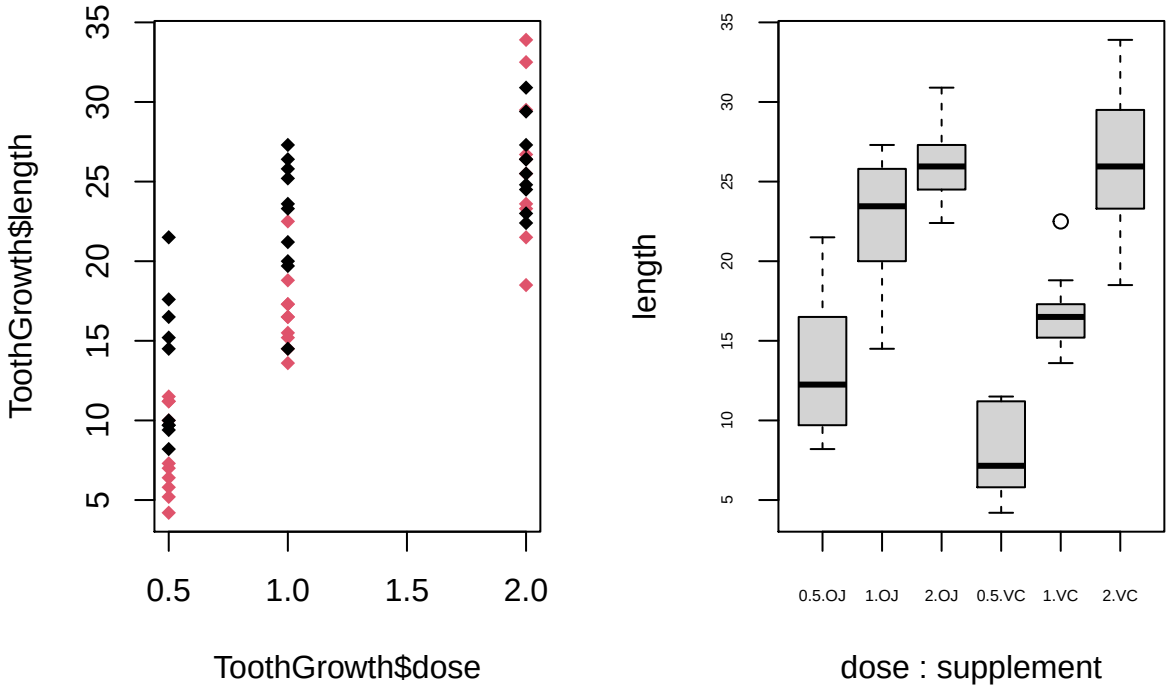
With `supplement` and `dose` indicating the treatment types that subjects received, and `length` apparently intended to be the observed variable, I also observe (using `table`) that the different doses / supplements all had an equal number (10 measurements, for each dose / supplement pairing).

```
table(ToothGrowth$supplement, ToothGrowth$dose)
```

```
##
##      0.5  1  2
##   OJ  10 10 10
##   VC  10 10 10
```

As such it seemed reasonable to explore the data by presenting a box and whisker plot of the range of length values for each supplement / dose treatment pair.

```r
par(mfrow = c(1, 2))
plot(x = ToothGrowth$dose, y = ToothGrowth$length, col = ToothGrowth$supplement, pch = 18)
boxplot(length ~ dose * supplement, data = ToothGrowth, cex.axis = 0.5)
```



Hypothesis Testing 1: OJ vs. VC Supplement

First, we use a two-sample T-test to test whether there is a statistically significant difference in means between the tooth growth where a subject is given "OJ" vs. "VC" supplement. The null hypothesis in this t-test will be that the means are equal; the alternative hypothesis is that they are not equal (we will use a two sided test because the means may differ in either direction).

```r
vc <- ToothGrowth[ToothGrowth$supplement == "VC", 1]
oj <- ToothGrowth[ToothGrowth$supplement == "OJ", 1]
```

Because we have no indication that the individual observations within the OJ and VC subsets of the data have a one-to-one correspondence relationship (e.g., same subject), we set paired to FALSE. While there is no indication whether it makes sense to assume variance is equal between the two distributions, from the observed data set we note sample standard deviations of 8.2660287 and 6.605561 for the different supplements. These are different enough to suggest the variances may not be the same, and so we leave var.equal as its default value of FALSE.

```r
result_supp <- t.test(vc, oj, alternative = "two.sided", paired = FALSE)
```

We can extract from the t-test that the 95% confidence interval for the difference between the means of the two supplement groups is -7.5710156, 0.1710156, which includes zero, indicating that at an alpha = 5% significance threshold, we should NOT reject the null hypothesis that the means of the distributions of OJ vs. VC supplement are equal. The P value of the T-Test of: 0.0606345, which at a P-value in excess of 0.05 similarly indicates that we do NOT reject the the null hypothesis of equal means.

Hypothesis Testing 2: Dosage Level

Our next hypothesis whether the means of tooth growth at different dosage levels are different from one another. Since there are three groups of doses (0.5, 1.0, and 2.0), we will test the three pairwise t-tests (0.5 vs. 1.0, 1.0 vs. 2.0 and 0.5 vs. 2.0).

```
d_0_5 <- ToothGrowth[ToothGrowth$dose == 0.5, 1]
d_1_0 <- ToothGrowth[ToothGrowth$dose == 1.0, 1]
d_2_0 <- ToothGrowth[ToothGrowth$dose == 2.0, 1]
```

Once again, without any indication of any one-to-one correspondence of subjects, we default to paired = FALSE, and despite the standard deviations of these three data subsets being closer together at 4.4997632, 4.4154364 and 3.7741503, we will not assume the variances of the underlying distributions are the same (i.e., same as before).

```
result_dose1 <- t.test(d_0_5, d_1_0, alternative = "two.sided", paired = FALSE)
result_dose2 <- t.test(d_1_0, d_2_0, alternative = "two.sided", paired = FALSE)
result_dose3 <- t.test(d_0_5, d_2_0, alternative = "two.sided", paired = FALSE)
```

The results are as follows:

- 0.5 vs. 1.0: the confidence interval of the difference in means is (-11.9837813, -6.2762187), which does not contain zero, indicating that can REJECT the null hypothesis that the means of growth at the 0.5 and 1.0 dosage levels are the same. Similarly we see that the P-value of this t-test, 1.2683007×10^{-7} is below 0.05, indicating the same result.
- 1.0 vs. 2.0: the confidence interval of the difference in means is (-8.9964805, -3.7335195), which does not contain zero, indicating that can REJECT the null hypothesis that the means of growth at the 1.0 and 2.0 dosage levels are the same. Similarly we see that the P-value of this t-test, 1.9064295×10^{-5} is below 0.05, indicating the same result.
- 0.5 vs. 2.0: finally, we have the confidence interval in the difference of means is (-18.1561665, -12.8338335), which does not contain zero, indicating that we can REJECT the null hypothesis that the means of growth at the 0.5 and 2.0 dosage levels are the same. Similarly we see this in the P-value of this t-test, 4.397525×10^{-14}, which is below 0.05.

Results

Based on the T-tests performed above, the conclusion is: - At a 5% significance level, we CANNOT REJECT the null hypothesis that the mean tooth growth levels are the same for the two different types of supplement (ascorbic acid and orangge juice). - At a 5% significance level, we CAN REJECT the null hypothesis that the mean tooth growth levels are the same for the dosage levels studied. This is true comparing between all three levels of dosage: 0.5, 1.0 and 2.0 mg/day (in all cases, the higher dose results in a mean tooth growth level for the higher dose that is higher than that of the lower dose).