Statistical Inference Project Part 2: Basic Inferential Data Analysis

Eric Erkela

11/16/2020

Overview

In this section of the two-part statistical inference project, we will explore the "ToothGrowth" data set, which is packaged in the base R installation. Our objective will be to establish which (if any) of the provided vitamin C delivery vectors/dosages have a measurable impact on tooth growth in guinea pigs, using formal hypothesis testing.

Dependencies

Below are all the packages used in this analysis, produced here for convenience:

```
library(datasets)
library(dplyr, warn.conflicts = FALSE)
library(ggplot2)
```

Part 2: Inferential Data Analysis

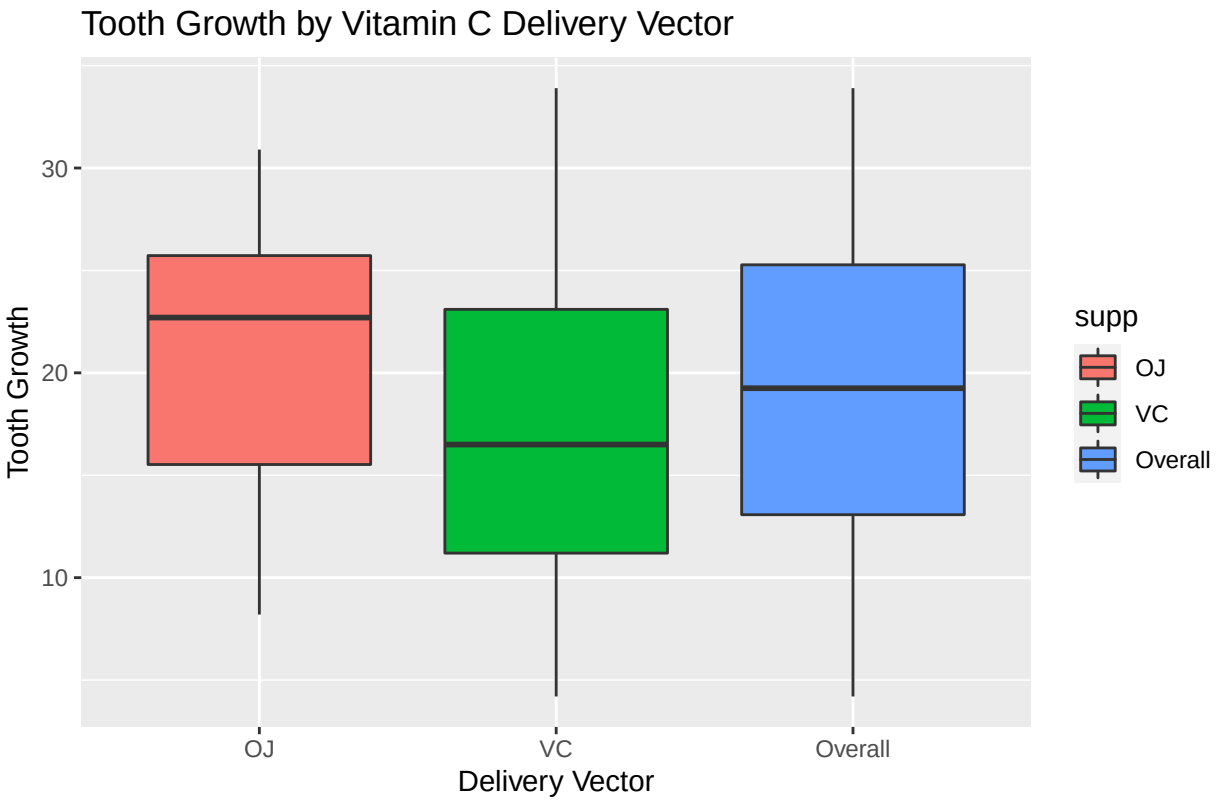
First, we have to load our data:

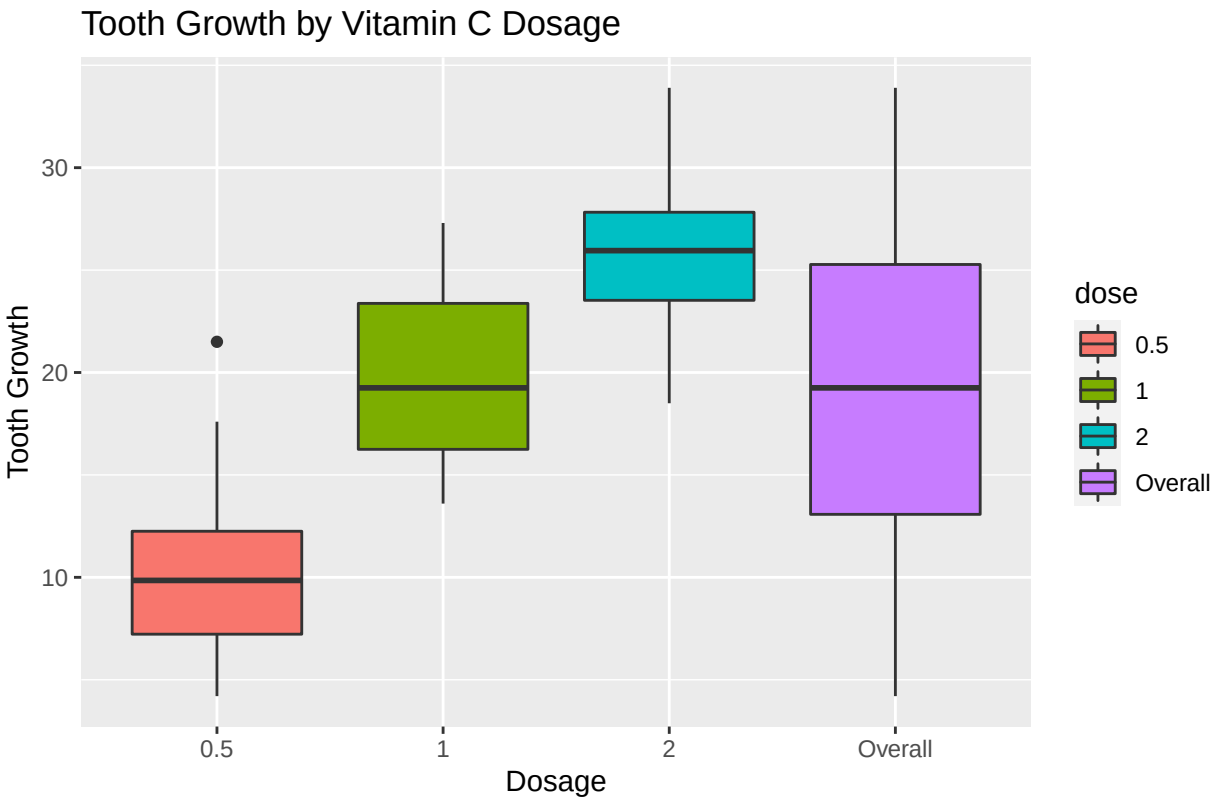
```
data("ToothGrowth")
str(ToothGrowth)
```

```
## '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 "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 ...
```

As we can see, the data set contains 60 observations of three variables: len (numeric), supp (Factor), and dose (numeric). A look at this particular data set's help page offers a helpful summary of its contents. From it, we can see that its purpose is to investigate the effect of vitamin C on tooth growth in guinea pigs. The variable len corresponds to the length of the odontoblast cells (which are responsible for tooth growth) in each subject, while supp and dose correspond to the delivery vector (orange juice - "OJ" or ascorbic acid - "VC") and vitamin C dosage (0.5, 1, or 2) respectively. It is also important to note that the number of subjects counted in the study (60) is the same as our total number of observations, which indicates that none of the collected data points are paired in any way.

We can perform a quick, exploratory analysis to investigate the relationship between our delivery vector/dosage combinations and tooth growth by making a couple simple box-and-whisker plots, as shown.





Where we've included an "Overall" delivery vector category so that we might compare the results for a specific vector against the sample as a whole. This will be important for developing our hypotheses.

As we can see, there does seem to be some measurable deviation in tooth growth due to vitamin C delivery vector/dose combination. In order to answer this question quantitatively, however, we will need to apply hypothesis testing. In order to do so, we need to come up with a set of hypotheses to describe our data, one null and one alternative for each of our supp/dose levels - 5 sets in total. Luckily, each of these will be identical, namely (in mathematical notation):

$$h_0 : \mu_{supp/dose} = \mu_{total}$$
$$h_a : \mu_{supp/dose} \neq \mu_{total}$$

Since we are only testing inequality between the means, we will use a two sided t test to determine a p-value for each set of hypotheses. If any of these values are below our threshold value $\alpha = 0.05$, then we can safely reject the null hypothesis and conclude that that delivery vector does in fact have a measurable effect on tooth growth.

```
options(scipen = 999)
all <- ToothGrowth$len
OJ <- subset(ToothGrowth, supp == "OJ")$len
VC <- subset(ToothGrowth, supp == "VC")$len
d0.5 <- subset(ToothGrowth, dose == 0.5)$len
d1.0 <- subset(ToothGrowth, dose == 1.0)$len
d2.0 <- subset(ToothGrowth, dose == 2.0)$len
p_vals <- c(pOJ = t.test(OJ, all)$p.value,
            pVC = t.test(VC, all)$p.value,
            p0.5 = t.test(d0.5, all)$p.value,
            p1.0 = t.test(d1.0, all)$p.value,
            p2.0 = t.test(d2.0, all)$p.value)
round(p_vals, digits = 8)
```

```
##        pOJ        pVC       p0.5       p1.0       p2.0
## 0.23951888 0.30955709 0.00000029 0.51187729 0.00000043
```

As we can see, only two p-values pass our threshold criteria $\alpha = 0.05$. Those that do correspond to dose $= 0.5$ and dose $= 2.0$, for which we have strong evidence to reject the null hypothesis in favor of the alternative, that these dosage levels do seem to have an effect on tooth growth in guinea pigs. For both delivery vectors and the 1.0 dosage level, however, we lack the evidence we need to safely reject the null hypothesis, meaning we have no evidence to suggest that these factors have an effect on tooth growth in guinea pigs.

Before we end this section, it's worth noting that we've performed more than one p-test in order to answer our research questions. Do we need to worry about controlling false positives? Well, since we've arranged our p-values into a single vector, it's easy for us to do so by applying a conservative, family-wise error correction.

```
p.adjust(p_vals, method = "bonferroni") < 0.05
```

```
##   pOJ   pVC  p0.5  p1.0  p2.0
## FALSE FALSE  TRUE FALSE  TRUE
```

As we can see, introducing this family-wise cut does not change our results in any way.