Part 2: Basic Inferential Data Analysis

Akhil Kota

2022-07-22

Overview

In this project, we will carry out some statistical inference on the Tooth Growth data set, which tracks the growth of teeth in a set of guinea pigs given certain doses of supplements. First, we will carry out some exploratory analysis. Then, we will carry out statistical inference to compare tooth growth by the dose and supp variables in the data set (whether they make a difference or not).

Exploratory Analysis

First, let's look at some summary data for the Tooth Growth data.

```
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 ...
```

```
summary(ToothGrowth)
```

```
##       len        supp         dose
##  Min.   : 4.20   OJ:30   Min.   :0.500
##  1st Qu.:13.07   VC:30   1st Qu.:0.500
##  Median :19.25           Median :1.000
##  Mean   :18.81           Mean   :1.167
##  3rd Qu.:25.27           3rd Qu.:2.000
##  Max.   :33.90           Max.   :2.000
```

It looks like the dose variable is given as a continuous variable, even though it's probably a discrete variable in the context of this experiment. Let's look at the unique values of dose to see if this is true, and make it a factor variable if it is. This will help with plotting.

```
unique(ToothGrowth$dose)
```

```
## [1] 0.5 1.0 2.0
```

```
ToothGrowth$dose<-as.factor(ToothGrowth$dose)
```

Reprinting the summary:

```
summary(ToothGrowth)
```

```
##       len        supp     dose
##  Min.   : 4.20   OJ:30   0.5:20
##  1st Qu.:13.07   VC:30   1  :20
##  Median :19.25           2  :20
##  Mean   :18.81
```

```
##  3rd Qu.:25.27
##  Max.   :33.90
```

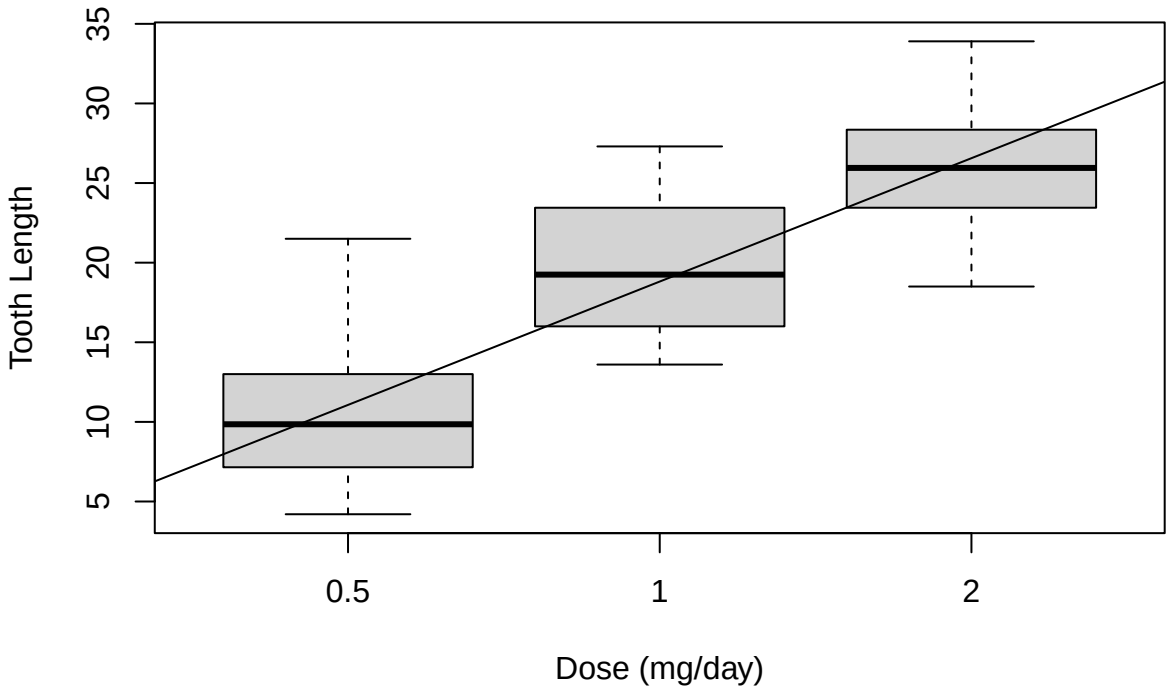
And just for good measure, let's make sure that the doses are evenly distributed across supps.

```
summary(subset(ToothGrowth, supp=="VC"))
```

```
##       len        supp     dose
##  Min.   : 4.20   OJ: 0   0.5:10
##  1st Qu.:11.20   VC:30   1  :10
##  Median :16.50           2  :10
##  Mean   :16.96
##  3rd Qu.:23.10
##  Max.   :33.90
```

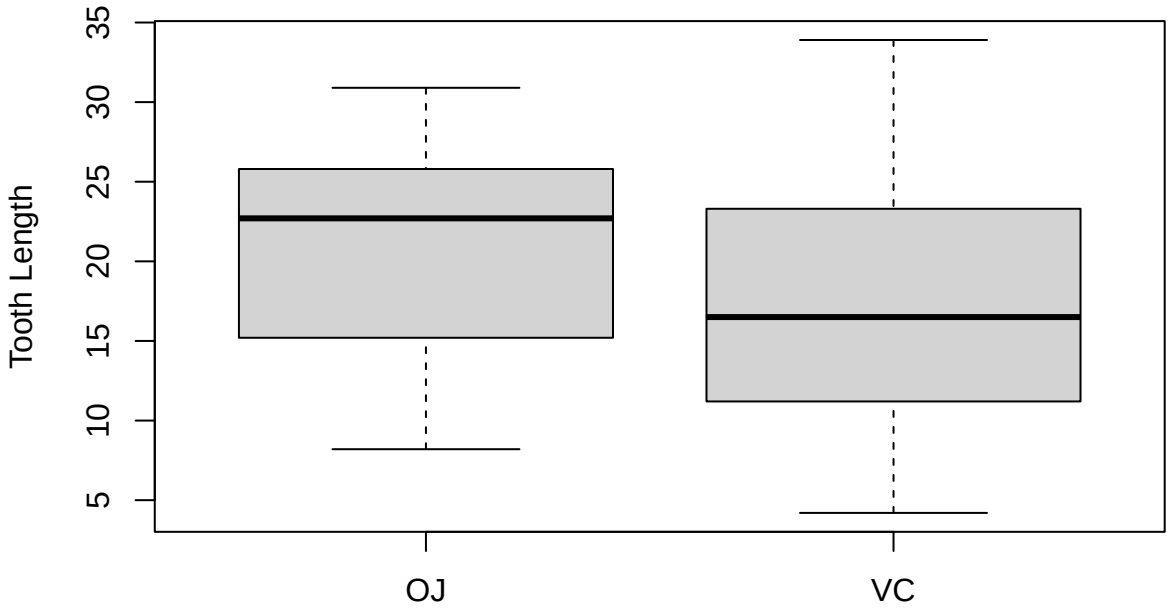
Great! Now let's visualize the data. Since these are discrete variables, we will opt to visualize with boxplots. First, let's look at the data by dose.

```
boxplot(len ~ dose, data=ToothGrowth, xlab = "Dose (mg/day)", ylab = "Tooth Length")
abline(lm(ToothGrowth$len ~ as.numeric(ToothGrowth$dose)))
```



There seem to be a general positive correlation for tooth length vs dosage of supplement (the best fit line has been provided as well). We will test this later with inferential analysis. For now, let's also look at the relationship between supp and len.

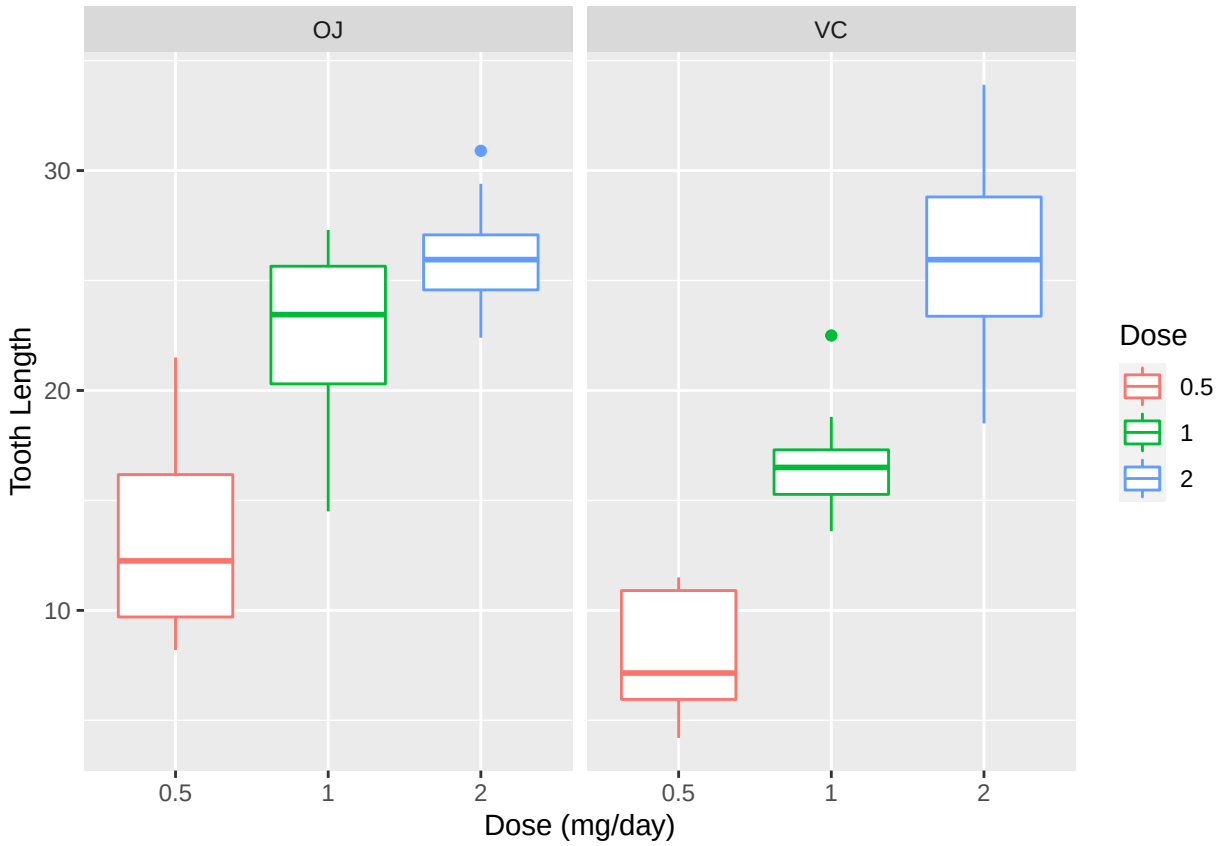
```
boxplot(len ~ supp, data=ToothGrowth, xlab = "Supplement Type", ylab = "Tooth Length")
```



These seem to be much more similar, not showing much difference on the boxplot (again, will be tested with inference).

Lastly, let's check for relationships between dose and tooth length for each supplement individually.

```
g <- ggplot(ToothGrowth, aes(dose, len, color=dose))
g <- g + geom_boxplot() + facet_grid(cols = vars(supp))
g + labs(x = "Dose (mg/day)", y = "Tooth Length", color="Dose")
```



There seems to be a clear difference between distributions of data between dose groups. Also, the VC data appears more extreme than the OJ data, showing a steeper uptrend than the OJ data.

More Summary Data: Distributions Across Dose and Supp

Let's get the means, variances, and standard deviations in the data, grouping by dose first, and then supp.

```
ToothGrowth %>% group_by(dose) %>%
    summarize(mean=mean(len), variance=var(len), stdev=sd(len)) %>% print
```

```
## # A tibble: 3 x 4
##   dose   mean variance stdev
##   <fct> <dbl>    <dbl> <dbl>
## 1 0.5    10.6     20.2  4.50
## 2 1      19.7     19.5  4.42
## 3 2      26.1     14.2  3.77
```

```
ToothGrowth %>% group_by(supp) %>%
    summarize(mean=mean(len), variance=var(len), stdev=sd(len)) %>% print
```

```
## # A tibble: 2 x 4
##   supp   mean variance stdev
##   <fct> <dbl>    <dbl> <dbl>
## 1 OJ     20.7     43.6  6.61
## 2 VC     17.0     68.3  8.27
```

Inferential Analysis

Now, we will perform some statistical inference to consolidate our hypotheses. Let's first test across supp. This will be an unpaired two-sided t-test testing with $\alpha = 0.05$ for whether or not there is a difference between the two. We will assume unequal population variances across the supp, since our summary data above shows a difference in variances.

```
vclens <- subset(ToothGrowth, supp=="VC")$len
ojlens <- subset(ToothGrowth, supp=="OJ")$len
t<-t.test(vclens, ojlens, alternative = "two.sided", mu = 0)
t
```

```
##
##  Welch Two Sample t-test
##
## data:  vclens and ojlens
## t = -1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.5710156  0.1710156
## sample estimates:
## mean of x mean of y
##  16.96333  20.66333
```

The 95% confidence interval does barely contain the null hypothesis μ of 0, so we fail to reject that the null hypothesis (that there is no difference between the means of the two supp groups). The p-value of 0.0606345>0.05 also indicates this same conclusion.

Next, we look at the dose. Since there are 3 doses, we perform 3 different t-tests to find if there are significant differences between the doses. This time, since we already saw a trend in our data, we will perform one-sided tests, where we will try to infer whether the pop. means for the higher doses are greater than the pop. means for the lower doses ($H_a = \mu > 0$, when subtracting higher dose mean from lower dose mean). The variances are relatively similar across these, so we will assume equal variances this time.

```
halflens <- subset(ToothGrowth, dose==0.5)$len
onelens <- subset(ToothGrowth, dose==1)$len
twolens <- subset(ToothGrowth, dose==2)$len
onehalft <- t.test(onelens, halflens, alternative = "greater", mu = 0, var.equal = TRUE)
twoonet <- t.test(twolens, onelens, alternative = "greater", mu = 0, var.equal = TRUE)
twohalft <- t.test(twolens, halflens, alternative = "greater", mu = 0, var.equal = TRUE)
```

Comparing dose=1 mg/day to dose=0.5 mg/day,

```
onehalft
```

```
##
##  Two Sample t-test
##
## data:  onelens and halflens
## t = 6.4766, df = 38, p-value = 6.331e-08
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  6.753344      Inf
## sample estimates:
## mean of x mean of y
##    19.735    10.605
```

Comparing dose=2 mg/day to dose=1 mg/day,

```
twoonet
```

```
##
##  Two Sample t-test
##
## data:  twolens and onelens
## t = 4.9005, df = 38, p-value = 9.054e-06
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  4.175196      Inf
## sample estimates:
## mean of x mean of y
##    26.100    19.735
```

Finally, comparing dose=2 mg/day to dose=0.5 mg/day,

```
twohalft
```

```
##
##  Two Sample t-test
##
## data:  twolens and halflens
## t = 11.799, df = 38, p-value = 1.419e-14
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  13.28093      Inf
## sample estimates:
## mean of x mean of y
##    26.100    10.605
```

In all cases, we see extremely small p-values and confidence intervals that do not contain zero. Thus, for almost any reasonable α (we used 0.05), it seems that we can reject the null hypothesis (that the means are not different across doses) in favor of the alternative (that the means increase as doses increase, between these three tested doses).

Conclusions

In conclusion, we see that there is not enough evidence to reject that there is no difference in tooth length across different supplements. However, there is strong evidence to reject that there is no difference in tooth length across various dosage levels of supplement. This supports the general trends we saw in the exploratory plots.

Several assumptions were made in regards to the study and our analysis. Firstly, we assumed that the study was conducted in the appropriate manner, using a random collection of guinea pigs representative of the population and minimizing biases across the study. Also, we assume that the distributions of sample means of guinea pig tooth lengths along the different supplements and dosage levels are normally distributed, or approximately so, for a sample size of 20 (for doses) or 30 (for supplements). This is generally a reasonable assumption for large enough sample sizes. To further this analysis, we could include a resampling technique (bootstrap or permutation) to provide us with some more clues about the sample mean distribution, and strengthen our findings.