Exploring the ToothGrowth dataset

Spiros Baxevanakis

May 08, 2021

Contents

Load and Explore Data **1**

Hypothesis & Permutation Tests **4**

Supplement Tests 4

Dose Tests 5

Assumptions **5**

Load and Explore Data

One of the standard learning data sets included in R is the "ToothGrowth" data set. The tooth growth data set is the length of the odontoblasts (teeth) in each of 10 guinea pigs at three Vitamin C dosage levels (0.5, 1, and 2 mg) with two delivery methods (orange juice or ascorbic acid).

The file contains 60 observations of 3 variables:

- len : Tooth length
- supp : Supplement type (VC or OJ)
- dose : Dose in milligrams

```
library(datasets)
# convert df to tibble for tidyverse compatibility
tooth = tibble(ToothGrowth)
# convert dose column to factor
tooth$dose = parse_factor(as.character(tooth$dose))
```

```
glimpse(tooth) # print the first few values
```

```
## Rows: 60
## Columns: 3
## $ len  <dbl> 4.2, 11.5, 7.3, 5.8, 6.4, 10.0, 11.2, 11.2, 5.2, 7.0, 16.5, 16.5,~
## $ supp <fct> VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, V~
## $ dose <fct> 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 1, 1, 1, 1, 1, ~
```

```
summary(tooth) # calcuate some statistics for each variable
```

```
##       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
```

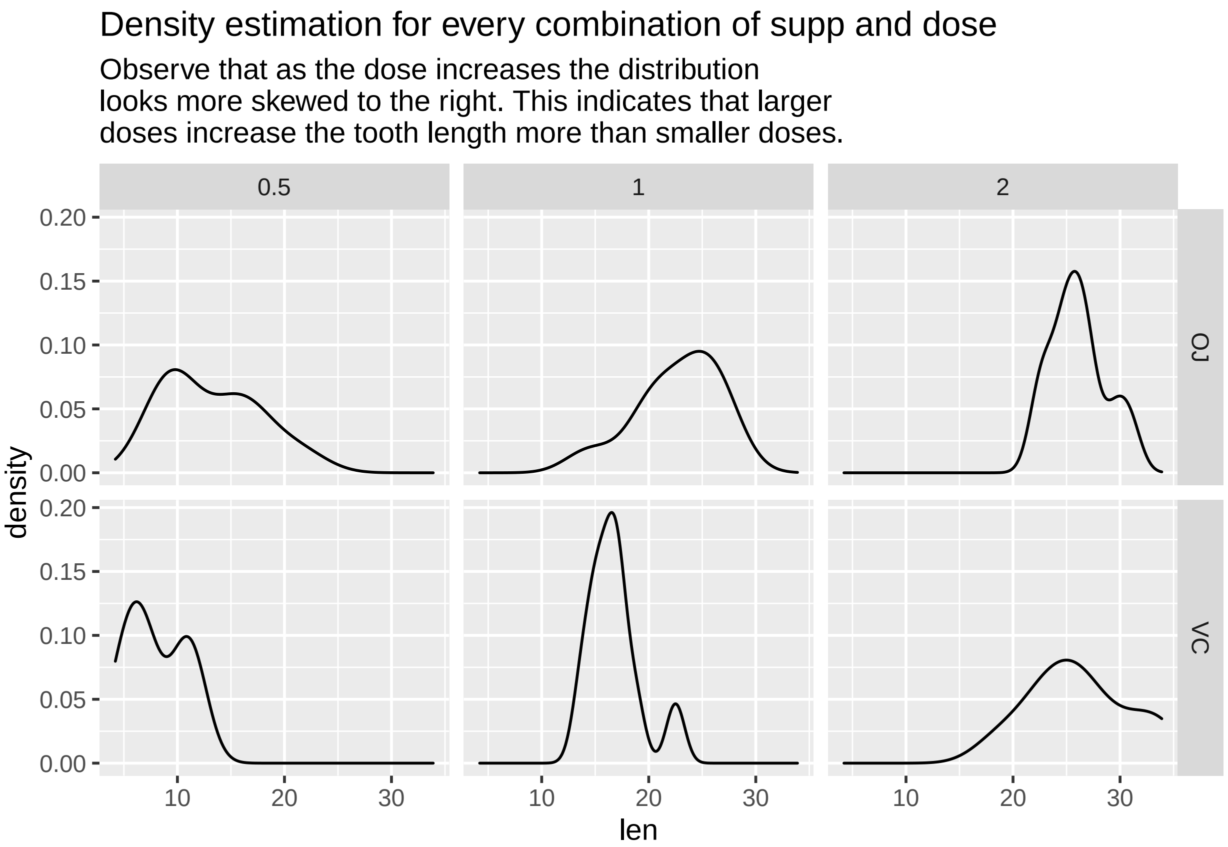
```
count(tooth, supp, dose) # group by supp and dose, then count their observations
```

```
## # A tibble: 6 x 3
##   supp  dose      n
##   <fct> <fct> <int>
## 1 OJ    0.5      10
## 2 OJ    1        10
## 3 OJ    2        10
## 4 VC    0.5      10
## 5 VC    1        10
## 6 VC    2        10
```

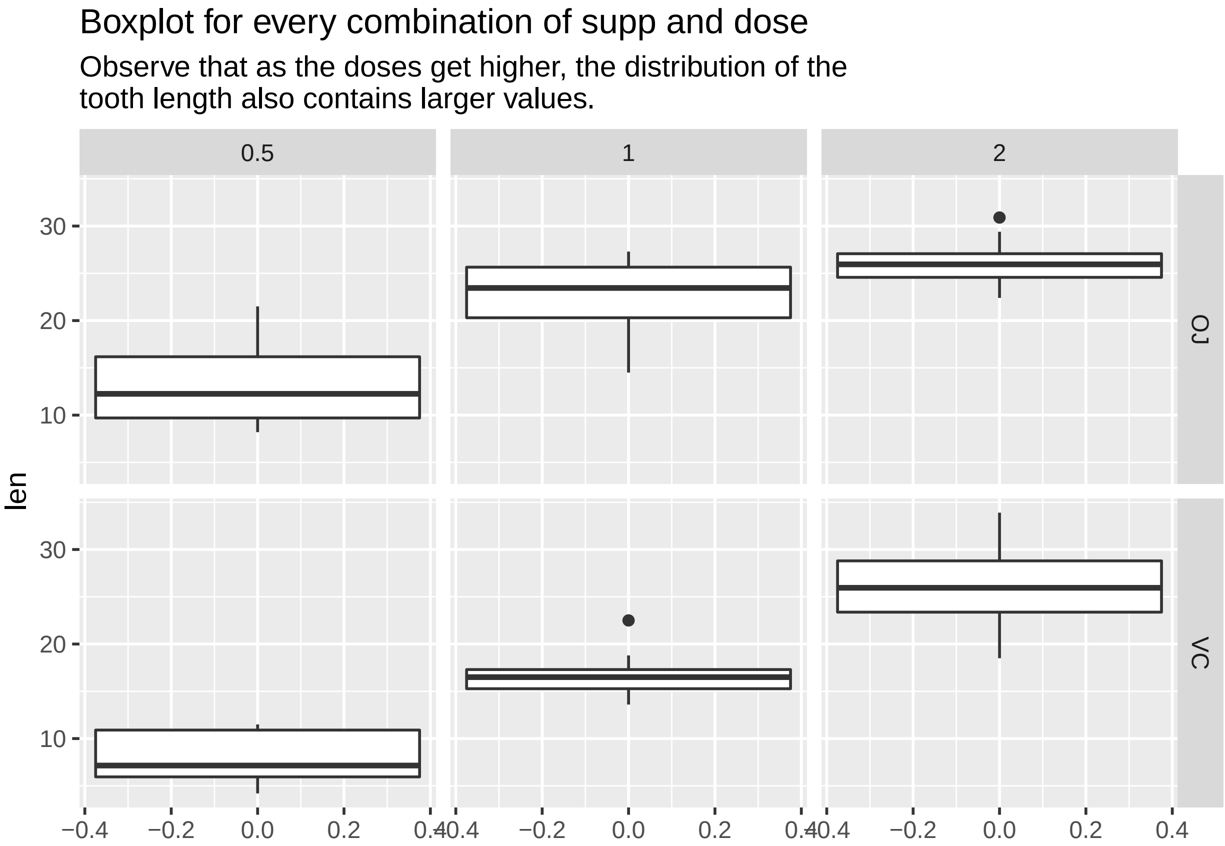
For every combination of supp and dose there are 10 observations (rows).

Now let's do some plots!

```
# histogram by supp and dose
ggplot(data=tooth, aes(x=len)) +
    geom_density() +
    facet_grid(supp ~ dose) +
    labs(title = 'Density estimation for every combination of supp and dose',
         subtitle = 'Observe that as the dose increases the distribution
looks more skewed to the right. This indicates that larger
doses increase the tooth length more than smaller doses.')
```



```
# boxplot by supp and dose
ggplot(data=tooth, aes(y=len)) +
    geom_boxplot() +
    facet_grid(supp ~ dose) +
    labs(title = 'Boxplot for every combination of supp and dose',
         subtitle = 'Observe that as the doses get higher, the distribution of the
tooth length also contains larger values.')
```



Boxplot for every combination of supp and dose

Observe that as the doses get higher, the distribution of the tooth length also contains larger values.

```
# mean and var by supp
tooth %>%
  group_by(supp) %>%
  summarize(mean=mean(len),var=var(len))
```

```
## # A tibble: 2 x 3
##   supp   mean   var
##   <fct> <dbl> <dbl>
## 1 OJ     20.7  43.6
## 2 VC     17.0  68.3
```

```
# mean and var by dose
tooth %>%
  group_by(dose) %>%
  summarize(mean=mean(len),var=var(len))
```

```
## # A tibble: 3 x 3
##   dose   mean   var
##   <fct> <dbl> <dbl>
## 1 0.5    10.6  20.2
## 2 1      19.7  19.5
## 3 2      26.1  14.2
```

```
# mean and var for every combination of supp and dose
tooth %>%
    group_by(supp,dose) %>%
    summarize(mean=mean(len),var=var(len))
```

```
## `summarise()` has grouped output by 'supp'. You can override using the `.groups` argument.
```

```
## # A tibble: 6 x 4
## # Groups:   supp [2]
##   supp  dose   mean   var
##   <fct> <fct> <dbl> <dbl>
## 1 OJ    0.5   13.2  19.9
## 2 OJ    1     22.7  15.3
## 3 OJ    2     26.1   7.05
## 4 VC    0.5    7.98  7.54
## 5 VC    1     16.8   6.33
## 6 VC    2     26.1  23.0
```

The mean tooth length was larger when taking the OJ supplement with .5 and 1 doses. With the 2mg dose however there is no difference in the mean tooth length between supplement types. Although, at 2mg dose, the OJ supplement is 3 times less variable on tooth length than the VC supplement.

Hypothesis & Permutation Tests

Supplement Tests

Test the null hypothesis that the supplement type does not affect tooth length. H_0 is that the difference in means is 0 and H_a that the difference in means is not 0.

```
# select len column for every supplement type
toothoj = tooth %>% filter(supp == 'OJ') %>% select(len)
toothvc = tooth %>% filter(supp == 'VC') %>% select(len)
```

```
t.test(toothoj, toothvc, alternative = 'two.sided')
```

```
##
##  Welch Two Sample t-test
##
## data:  toothoj and toothvc
## 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:
##  -0.1710156  7.5710156
## sample estimates:
## mean of x mean of y
##  20.66333  16.96333
```

The p value is very close to the rejection region and although normally we would fail to reject the H_0, we know from the exploratory data analysis we performed, that the H_a is true. Since the p value is calculated in the direction of the H_a we hypothesize that if we perform the same test using $H_a > 0$ instead of the original $H_a \neq 0$ the p value would be smaller allowing us to reject the H_0. This is because we observe from the data that the difference in means is greater than 0 (mean of OJ - mean of VC).

```
t.test(toothoj, toothvc, alternative = 'greater')
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  toothoj and toothvc
## t = 1.9153, df = 55.309, p-value = 0.03032
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.4682687       Inf
## sample estimates:
## mean of x mean of y
##  20.66333  16.96333
```

Indeed, our hypothesis is correct. The p value is smaller when testing on Ha > 0. Thus we can reject the Ho.

Dose Tests

```
# get all the observations that have doses 1 or 2
doses12 = tooth %>% filter(dose %in% c(1,2)) %>% select(len, dose)
y = doses12$len # get only len variable
dose = doses12$dose # get only dose variable
# function to calcuate differnece in means
testStat = function(y,d) mean(y[d==2]) - mean(y[d==1])
observedStat = testStat(y,dose)
# sample dose labels and calculate testStat 10000 times.
permutations = sapply(1:10000, function(i) testStat(y, sample(dose)))
observedStat
```

```
## [1] 6.365
```

```
# number of permutations that are more extreme than the observed stat
mean(permutations > observedStat)
```

```
## [1] 0
```

No permutation is as or more extreme than our observed statistic (except the observed statistic itself)! We can safely conclude that a dose of 2 is associated with a larger tooth length. This is also observed in the box and density plots of the previous sections.

We will repeat the same procedure between .5 and 1 doses.

```
doses51 = tooth %>% filter(dose %in% c(.5,1)) %>% select(len, dose)
y = doses51$len
dose = doses51$dose
testStat = function(y,d) mean(y[d==1]) - mean(y[d==.5])
observedStat = testStat(y,dose)
permutations = sapply(1:10000, function(i) testStat(y, sample(dose)))
observedStat
```

```
## [1] 9.13
```

```
mean(permutations > observedStat)
```

```
## [1] 0
```

We reach the same conclusion as with the comparison between 2 and 1 doses.

Assumptions

In order to draw the conclusions in the hypothesis test section, we assume that the data is not skewed and follows a normal distribution. This is a requirement of the Student's t-test and t-confidence interval.