ToothGrowth Dataset: Exploratory Analysis and Basic Statistical Inferences

KSHITIJ MISHRA

Overview

In this report I perform an exploratory analysis of the ToothGrowth dataset and I infer some basic conclusions based on statistical hypothesis testing.

Background

The ToothGrowth dataset summarizes the effect on odontoblasts (cells responsible for tooth growth) of different dose levels of Vitamin C. This study was performed in 60 guinea pigs, divided into six equal groups. Each animal received one of three dose levels of vitamin C (0.5, 1, or 2 mg/day) by one of two suplement types, orange juice (OJ) or ascorbic acid (a form of vitamin C and coded as VC).

The data is availabl as a data frame with 60 observations on 3 variables.

Column	Name	Type	Description
[,1]	len	numeric	Tooth length
[,2]	supp	factor	Supplement type (VC or OJ)
[,3]	dose	numeric	Dose in milligrams/day

(Source: C. I. Bliss (1952). The Statistics of Bioassay. Academic Press.)

ToothGrowth Dataset: Exploratory Data analysis

The environment is initialized an the dataset is loaded with the following commands:

```
#Initialization and load packages
library(ggplot2)
library(dplyr)
library(datasets)

data("ToothGrowth")
```

For convenience, the data frame is coerced into a *tibble*; its structure is displayed below.

```
ToothGrowth <- as_tibble(ToothGrowth)
str(ToothGrowth)
```

```
## tibble [60 x 3] (S3: tbl_df/tbl/data.frame)
## $ len : num [1:60] 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 [1:60] 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

As the dose levels are discrete, they can be coerced into factors. Below, is a the result of the function *summary()* applied to this dataset:

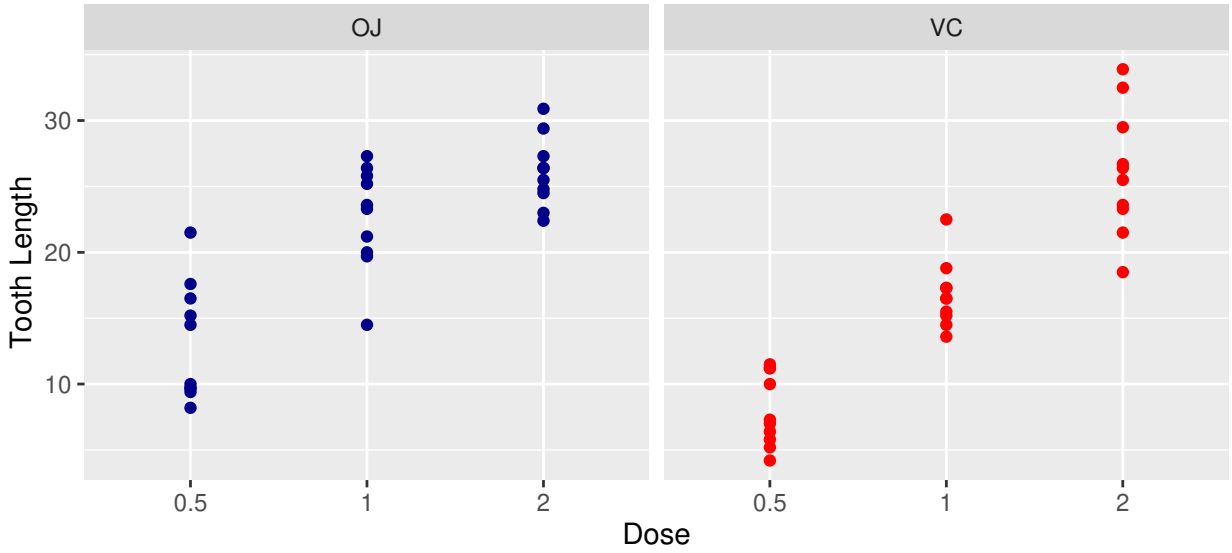
```
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
summary(ToothGrowth)
```

```
##        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
```

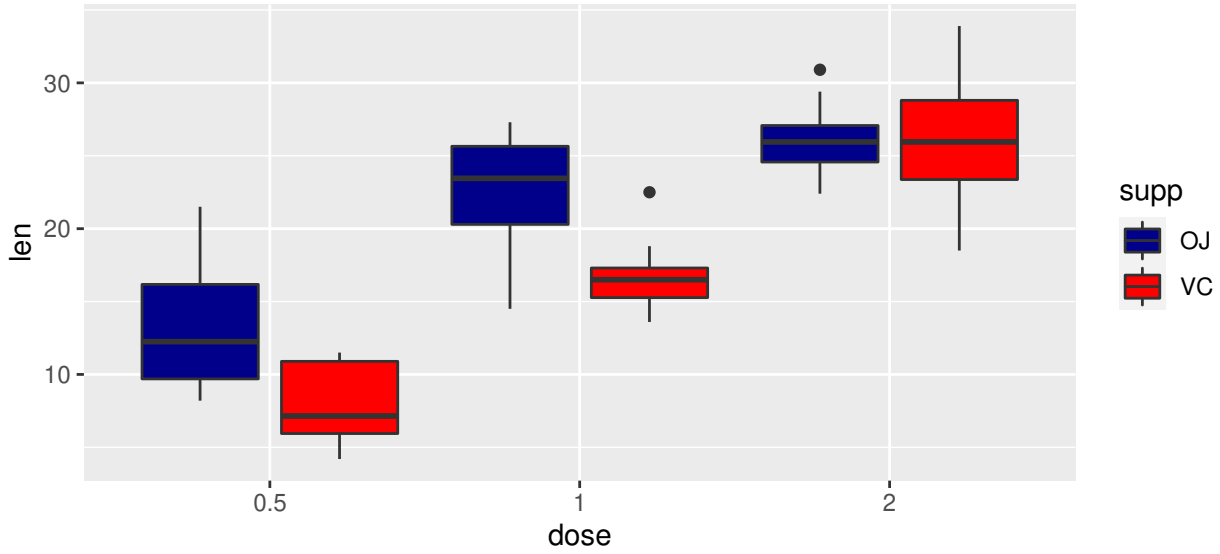
In fhe figure below, the tooth-length data is presented as a function of dose for both supplement types.

```
ggplot(ToothGrowth, aes(x = dose, y = len)) +
        geom_point(aes(color = supp)) + facet_grid(.~supp) +
        scale_color_manual(values = c("darkblue", "red"))+
        xlab("Dose") + ylab("Tooth Length")+
        theme(legend.position="None")
```



Generally speaking, increasing vitamin-C doses appear to have an increasing beneficial effect on tooth length. Below, the data is presented in boxplot form:

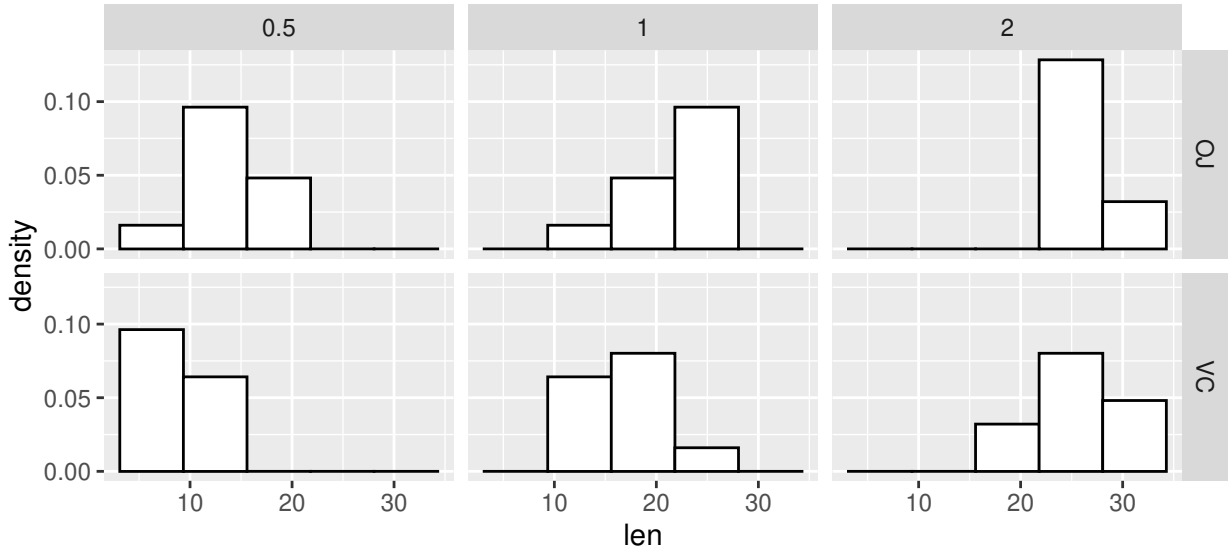
```
ggplot(ToothGrowth, aes(x = dose, y = len)) +
        geom_boxplot(aes(fill = supp), position = position_dodge(0.9)) +
        scale_fill_manual(values = c("darkblue", "red"))
```



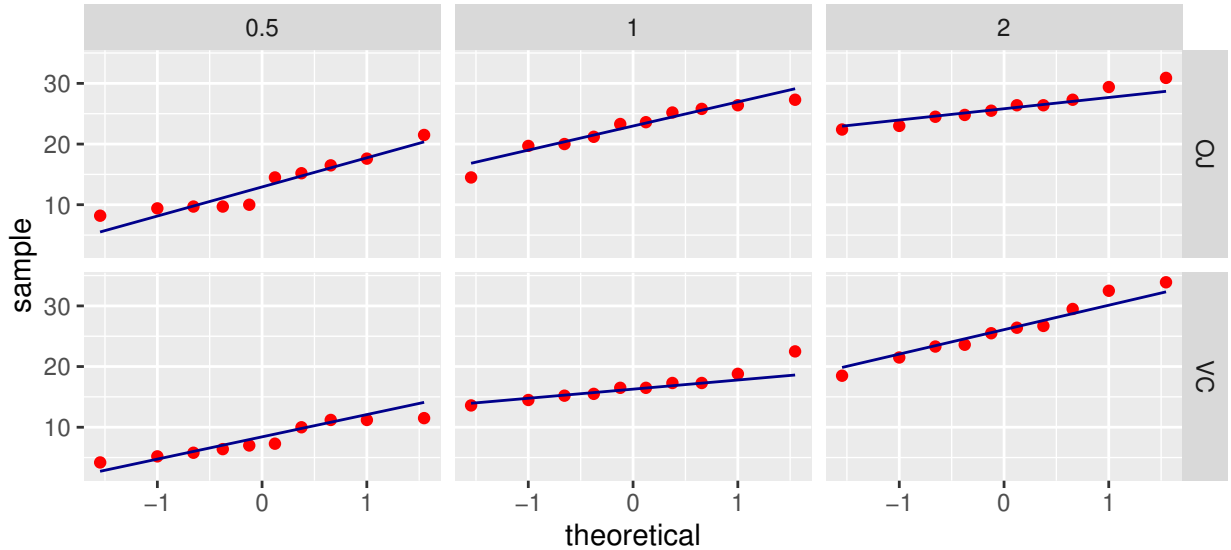
For the lower doses, OJ seems to have a higher beneficial effect on tooth growth, compared to VC. We will explore this question formally in the following section.

We now turn to the shape of the data distributions. The figures below show histograms for each condition, followed by Q-Q normality plots.

```
# Facet Histogram plot
bw <- 2 * IQR(ToothGrowth$len) / length(ToothGrowth$len)^(1/3) # Freedman–Diaconis rule for binwidth
ggplot(ToothGrowth, aes(x=len))+
        geom_histogram(aes(y=..density..),color="black", fill="white", binwidth=bw)+
        facet_grid(supp ~ dose)
```




```
# Facet Q-Q plot
ggplot(ToothGrowth, aes(sample=len))+
        stat_qq(distribution=qnorm, color="red") +
        stat_qq_line(distribution=qnorm, color="darkblue") +
        facet_grid(supp ~ dose)
```



From the histograms above, it is hard to assess whether the data can be approximated by a normal or a T-distribution; there are too few data points for each condition. However, the Q-Q plots suggest that a normal or a T-distribution modeling assumption is reasonable. Furthermore, it will be assumed that the means of the data 10 points available for each test condition can be approximated by a T-distribution.

Tooth length as a function of Supplement Type and Dose.

Pairwise T-test comparisons for the means for the lengths observed in each of the six test conditions will be performed. The following code constructs the testing framework. The *condition* array describes the test parameters for each of the six testing conditions. The *pairs* array lists the fifteen possible pairwise comparisons between these six conditions.

```
var1 <- c("OJ","VC")
var2 <- c("0.5","1","2")
condition <- matrix(as.character(NA),nrow=length(var1)*length(var2),ncol=2)
k<-0
for(i in 1:length(var1)) {
        for(j in 1:length(var2)) {
                k<-k+1
                condition[k,1]<-var1[i]
                condition[k,2]<-var2[j]
        }
}
pair <- matrix(as.numeric(NA),nrow=choose(k,2),ncol=2)
l<-0
for(i in 1:(k-1)) {
        for(j in (i+1):k) {
                l<-l+1
                pair[l,1]<-i
                pair[l,2]<-j
        }
}
condition
```

```
##      [,1] [,2]
## [1,] "OJ" "0.5"
## [2,] "OJ" "1"
## [3,] "OJ" "2"
## [4,] "VC" "0.5"
## [5,] "VC" "1"
## [6,] "VC" "2"
```

```
pair
```

```
##       [,1] [,2]
## [1,]    1    2
## [2,]    1    3
## [3,]    1    4
## [4,]    1    5
## [5,]    1    6
## [6,]    2    3
## [7,]    2    4
## [8,]    2    5
## [9,]    2    6
## [10,]   3    4
## [11,]   3    5
## [12,]   3    6
## [13,]   4    5
## [14,]   4    6
## [15,]   5    6
```

The following code performs these pairwise T-test comparisons, storing the results in the *comparisons*

data frameThe tests are performed using the Welch Two Sample t-test with unequal variances assumed. Significance is determined by *p<0.05*.

```
pSignif <-0.05
comparisons <- data.frame(condADel = as.character(NULL), condADose = as.character(NULL),
          meanA = as.numeric(NULL),condBDel=as.character(NULL), condBDose=as.character(NULL),
          meanB = as.numeric(NULL),meanDiffBA = as.numeric(NULL),pValue=as.numeric(NULL),
          significant=as.logical(NULL))
for(i in 1:l) {
set1 <- ToothGrowth %>% filter(supp==condition[pair[i,1],1],dose==condition[pair[i,1],2]) %>% select(len)
set2 <- ToothGrowth %>% filter(supp==condition[pair[i,2],1],dose==condition[pair[i,2],2]) %>% select(len)
          tt <- t.test(set1,set2,alternative="two.sided")
          comparisons <- add_row(comparisons,condADel=condition[pair[i,1],1],
    condADose=condition[pair[i,1],2],meanA=tt$estimate[1],condBDel=condition[pair[i,2],1],
      condBDose=condition[pair[i,2],2],meanB=tt$estimate[2],meanDiffBA=meanB-meanA,
        pValue=tt$p.value,significant=pValue<pSignif)
}
comparisons %>% knitr::kable("latex", booktabs = TRUE,align=rep('c', ncol(comparisons)),
                        digits=4) %>% kableExtra::kable_styling(latex_options = "striped")
```

condADel	condADose	meanA	condBDel	condBDose	meanB	meanDiffBA	pValue	significant
OJ	0.5	13.23	OJ	1	22.70	9.47	0.0001	TRUE
OJ	0.5	13.23	OJ	2	26.06	12.83	0.0000	TRUE
OJ	0.5	13.23	VC	0.5	7.98	-5.25	0.0064	TRUE
OJ	0.5	13.23	VC	1	16.77	3.54	0.0460	TRUE
OJ	0.5	13.23	VC	2	26.14	12.91	0.0000	TRUE
OJ	1	22.70	OJ	2	26.06	3.36	0.0392	TRUE
OJ	1	22.70	VC	0.5	7.98	-14.72	0.0000	TRUE
OJ	1	22.70	VC	1	16.77	-5.93	0.0010	TRUE
OJ	1	22.70	VC	2	26.14	3.44	0.0965	FALSE
OJ	2	26.06	VC	0.5	7.98	-18.08	0.0000	TRUE
OJ	2	26.06	VC	1	16.77	-9.29	0.0000	TRUE
OJ	2	26.06	VC	2	26.14	0.08	0.9639	FALSE
VC	0.5	7.98	VC	1	16.77	8.79	0.0000	TRUE
VC	0.5	7.98	VC	2	26.14	18.16	0.0000	TRUE
VC	1	16.77	VC	2	26.14	9.37	0.0001	TRUE

Analysis and Conclusions

OJ Delivery, Effect of Dose

The OJ dose group comparisons are shown on the table below.

```
comparisons %>% filter(condADel=="OJ",condBDel=="OJ") %>%
          knitr::kable("latex", booktabs = TRUE,align=rep('c', ncol(comparisons)),
                  digits=4) %>% kableExtra::kable_styling(latex_options = "striped")
```

condADel	condADose	meanA	condBDel	condBDose	meanB	meanDiffBA	pValue	significant
OJ	0.5	13.23	OJ	1	22.70	9.47	0.0001	TRUE
OJ	0.5	13.23	OJ	2	26.06	12.83	0.0000	TRUE
OJ	1	22.70	OJ	2	26.06	3.36	0.0392	TRUE

The differences between groups are statistically significant (p<0.05, the null hypothesis is rejected for every pairwise comparison), and the data shows a clear increase in benefit with an increase in dose.

VC Delivery, Effect of Dose

```
comparisons %>% filter(condADel=="VC",condBDel=="VC") %>%
        knitr::kable("latex", booktabs = TRUE,align=rep('c', ncol(comparisons)),
                digits=4) %>% kableExtra::kable_styling(latex_options = "striped")
```

condADel	condADose	meanA	condBDel	condBDose	meanB	meanDiffBA	pValue	significant
VC	0.5	7.98	VC	1	16.77	8.79	0e+00	TRUE
VC	0.5	7.98	VC	2	26.14	18.16	0e+00	TRUE
VC	1	16.77	VC	2	26.14	9.37	1e-04	TRUE

The differences between groups are statistically significant (p<0.05, the null hypothesis is rejected for every pairwise comparison), and the data shows a clear increase in benefit with an increase in dose.

Comparison of the Two Supplement Types, Equal Dose

```
comparisons %>% filter(condADel=="OJ",condBDel=="VC", condADose == condBDose) %>%
        knitr::kable("latex", booktabs = TRUE,align=rep('c', ncol(comparisons)),
                digits=4) %>% kableExtra::kable_styling(latex_options = "striped")
```

condADel	condADose	meanA	condBDel	condBDose	meanB	meanDiffBA	pValue	significant
OJ	0.5	13.23	VC	0.5	7.98	-5.25	0.0064	TRUE
OJ	1	22.70	VC	1	16.77	-5.93	0.0010	TRUE
OJ	2	26.06	VC	2	26.14	0.08	0.9639	FALSE

At the lower doses, 0.5 mg/day and 1 mg/day, there is a statistically significant difference (p<0.05, the null hypothesis is rejected for these two pairwise comparison) between the OJ and VC supplement types, with the OJ type resulting in a higher tooth growth benefit compared to the VC type. However, at the largest dose, 2 mg/day, the two supplement types appear to convey the same benefit (p>0.05, the null hypothesis is NOT rejected).

Comparison of the Two Supplement Types, Different Doses

```
comparisons %>% filter(condADel=="OJ",condBDel=="VC", condADose != condBDose) %>%
        knitr::kable("latex", booktabs = TRUE,align=rep('c', ncol(comparisons)),
                digits=4) %>% kableExtra::kable_styling(latex_options = "striped")
```

condADel	condADose	meanA	condBDel	condBDose	meanB	meanDiffBA	pValue	significant
OJ	0.5	13.23	VC	1	16.77	3.54	0.0460	TRUE
OJ	0.5	13.23	VC	2	26.14	12.91	0.0000	TRUE
OJ	1	22.70	VC	0.5	7.98	-14.72	0.0000	TRUE
OJ	1	22.70	VC	2	26.14	3.44	0.0965	FALSE
OJ	2	26.06	VC	0.5	7.98	-18.08	0.0000	TRUE
OJ	2	26.06	VC	1	16.77	-9.29	0.0000	TRUE

Consistent with the observations in the previous section, there is no statistical significant difference in benefit ($p>0.05$, the null hypothesis is NOT rejected) between OJ at the 1 mg/day dose, and VC at the 2 mg/day dose, suggesting that OJ is a more effective vitamin-C supplement, providing the same benefit at a lower dose.

Assumptions in the above analysis

- Results from these small sample sizes can be extrapolated to the general population.
- Group randomization was not biased in any way.
- Sample sizes are large enough, and the means of the data with a T-distribution.