Statistical Inference on Tooth Growth Dataset

Trent Parkinson

December 26, 2017

Overview

The project provides a basic analysis of the `ToothGrowth` data in the R datasets package. The data gives the length of the odontoblast (teeth) in each of 10 guinea pigs at three different dosage levels (0.5, 1, and 2 mg) with two supplements (Vitamin C and Orange Juice). The following will occur;

- Loads the ToothGrowth data and perform some basic exploratory data analyses.
- Provide a basic summary of the data.
- Use confidence intervals and t-tests to compare tooth growth by `supp` and `dose`.
- States conclusions and any assumptions made.

Setting up environment

Necessary libraries for loading, manipulating, and plotting. Reading the `ToothGrowth` dataset into a `data.table`.

```
library(data.table)
library(ggplot2)
library(gridExtra)
library(dplyr)
library(printr)

data("ToothGrowth")
tooth_growth <- data.table(ToothGrowth)
```

Data Structure

```
str(tooth_growth)
```

```
## Classes 'data.table' and '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 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
head(tooth_growth)
```

len	supp	dose
4.2	VC	0.5
11.5	VC	0.5
7.3	VC	0.5
5.8	VC	0.5
6.4	VC	0.5
10.0	VC	0.5

Brief Summary

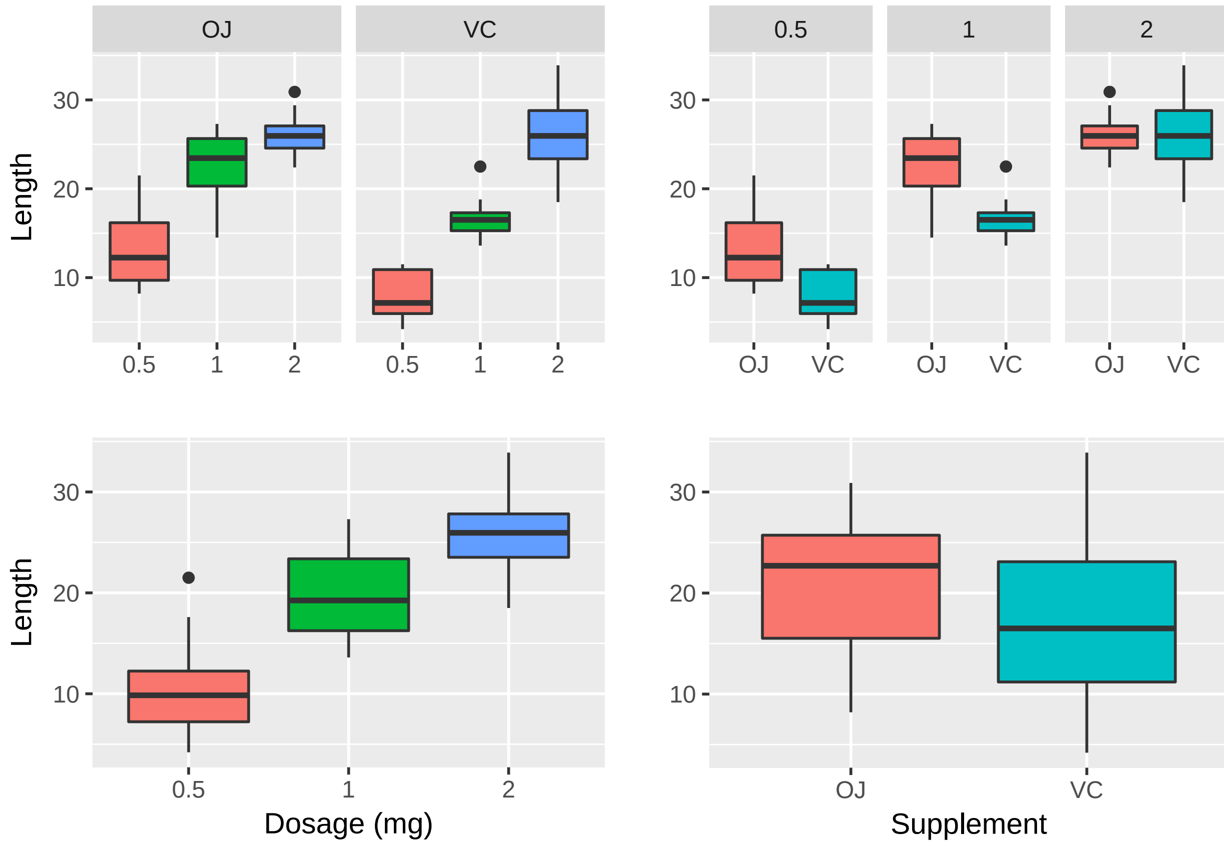
```r
summary(tooth_growth)
```

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

Visualizations

Setting the `dose` variable as factors to make the plotting and t-tests easier. Plotting a grid of box plots for all the different combinations of `len` as it depends on `supp` and `dose`.

```r
tooth_growth$dose <- as.factor(tooth_growth$dose)

plot1 <- ggplot(data = tooth_growth, aes(x = dose, y = len)) +
    geom_boxplot(aes(fill = dose)) +
    facet_grid(.~supp) +
    theme(legend.position = "None") +
    xlab(" ") + ylab("Length")

plot2 <- ggplot(data = tooth_growth, aes(x = supp, y = len)) +
    geom_boxplot(aes(fill = supp)) +
    facet_grid(.~dose) +
    theme(legend.position = "None") +
    xlab(" ") + ylab(" ")

plot3 <- ggplot(data = tooth_growth, aes(x = supp, y = len)) +
    geom_boxplot(aes(fill = supp)) +
    theme(legend.position = "None") +
    xlab("Supplement") + ylab(" ")

plot4 <- ggplot(data = tooth_growth, aes(x = dose, y = len)) +
    geom_boxplot(aes(fill = dose)) +
    theme(legend.position = "None") +
    xlab("Dosage (mg)") + ylab("Length")

grid.arrange(plot1, plot2, plot4, plot3, nrow = 2, ncol = 2)
```



Analysis

The goal of the project is to analyze the effect of the different supplements (Vitamin C and Orange Juice) at all of the dose levels (0.5, 1, 2 mg). The following is the code used to perform all the different combinations of t-tests. All of the test were performed using a 95% confidence level, with unequal variances assumed.

```
OJvsVC <- t.test(len ~ supp, data = tooth_growth)

my.t.test <- function(fac1,fac2) {
    t.test(tooth_growth$len[tooth_growth$dose == fac1],
           tooth_growth$len[tooth_growth$dose == fac2])
}

supp.t.test <- function(supp1,dose1,supp2,dose2) {
    sub_tooth <- filter(tooth_growth, (dose == as.numeric(dose1) & supp == supp1 ) |
                            dose == as.numeric(dose2) & supp == supp2)
    t.test(len ~ supp, data=sub_tooth)
}

same_supp.t.test <- function(supp1,dose1,dose2) {
    sub_tooth <- filter(tooth_growth, (dose == as.numeric(dose1) & supp == supp1 ) |
                            dose == as.numeric(dose2) & supp == supp1)
    t.test(len ~ dose, data=sub_tooth)
}
```

```r
my_tests <- mapply(FUN = my.t.test,c(.5,.5,1.0),c(1.0,2.0,2.0),SIMPLIFY = FALSE)

d1 <- c(0.5,1.0,2.0,0.5,0.5,1.0,2.0,1.0,2.0)
d2 <- c(0.5,0.5,0.5,1.0,2.0,1.0,1.0,2.0,2.0)
supp_tests <- mapply(FUN = supp.t.test,rep("VC",9),d1,rep("OJ",9),d2,SIMPLIFY = FALSE)

d3 <- rep(c("OJ","VC"),each = 3)
d4 <- rep(c(1.0,2.0,2.0), times = 2)
d5 <- rep(c(0.5,0.5,1.0), times = 2)
same_supp_tests <- mapply(FUN = same_supp.t.test,d3,d4,d5,SIMPLIFY = FALSE)
```

Results for t-test

The following code takes all of the t-test results from the previous section and places them in a `data.frame` for easier readability, and quicker viewing. Note that all values have been rounded to four decimal places.

```r
t_table <- list(OJvsVC)
t_table <- append(t_table,my_tests)
t_table <- append(t_table,same_supp_tests)
t_table <- append(t_table,supp_tests)

t.stat = c(); df = c(); lower.CL = c(); upper.CL = c()
p.value = c(); mean.A = c(); mean.B = c()

rnames <- c("VC vs OJ",
            "1.0 vs 0.5","2.0 vs 0.5","2.0 vs 1.0",
            "OJ-1.0 vs 0.5","OJ-2.0 vs 0.5","OJ-2.0 vs 1.0",
            "VC-1.0 vs 0.5","VC-2.0 vs 0.5","VC-2.0 vs 1.0",
            "VC-0.5 vs OJ-0.5","VC-1.0 vs OJ-0.5","VC-2.0 vs OJ-0.5",
            "VC-0.5 vs OJ-1.0","VC-0.5 vs OJ-2.0","VC-1.0 vs OJ-1.0",
            "VC-2.0 vs OJ-1.0","VC-1.0 vs OJ-2.0","VC-2.0 vs OJ-2.0")

for (i in 1:19) {
    t.stat <- append(t.stat,c(t_table[[i]]$statistic))
    df <- append(df,c(t_table[[i]]$parameter))
    lower.CL <- append(lower.CL,c(t_table[[i]]$conf.int[1]))
    upper.CL <- append(upper.CL,c(t_table[[i]]$conf.int[2]))
    p.value <- append(p.value,c(t_table[[i]]$p.value))
    mean.B <- append(mean.B,c(t_table[[i]]$estimate[1]))
    mean.A <- append(mean.A,c(t_table[[i]]$estimate[2]))
}

mean.diff <- mean.A - mean.B

t_results <- data.frame("t.stat" = t.stat, "df" = df,
                        "lower.CL" = lower.CL, "upper.CL" = upper.CL,
                        "p.value" = p.value, "mean.A" = mean.A,
                        "mean.B" = mean.B, "mean.diff" = mean.diff,
                        row.names = rnames)

knitr::kable(round(t_results,4), caption = "Welch Two Sample t-test Results")
```

Welch Two Sample t-test Results

	t.stat	df	lower.CL	upper.CL	p.value	mean.A	mean.B	mean.diff
VC vs OJ	1.9153	55.3094	-0.1710	7.5710	0.0606	16.9633	20.6633	-3.700
1.0 vs 0.5	-6.4766	37.9864	-11.9838	-6.2762	0.0000	19.7350	10.6050	9.130
2.0 vs 0.5	-11.7990	36.8826	-18.1562	-12.8338	0.0000	26.1000	10.6050	15.495
2.0 vs 1.0	-4.9005	37.1011	-8.9965	-3.7335	0.0000	26.1000	19.7350	6.365
OJ-1.0 vs 0.5	-5.0486	17.6983	-13.4156	-5.5244	0.0001	22.7000	13.2300	9.470
OJ-2.0 vs 0.5	-7.8170	14.6678	-16.3352	-9.3248	0.0000	26.0600	13.2300	12.830
OJ-2.0 vs 1.0	-2.2478	15.8424	-6.5314	-0.1886	0.0392	26.0600	22.7000	3.360
VC-1.0 vs 0.5	-7.4634	17.8624	-11.2657	-6.3143	0.0000	16.7700	7.9800	8.790
VC-2.0 vs 0.5	-10.3878	14.3271	-21.9015	-14.4185	0.0000	26.1400	7.9800	18.160
VC-2.0 vs 1.0	-5.4698	13.6000	-13.0543	-5.6857	0.0001	26.1400	16.7700	9.370
VC-0.5 vs OJ-0.5	3.1697	14.9688	1.7191	8.7809	0.0064	7.9800	13.2300	-5.250
VC-1.0 vs OJ-0.5	-2.1864	14.1997	-7.0081	-0.0719	0.0460	16.7700	13.2300	3.540
VC-2.0 vs OJ-0.5	-6.2325	17.9048	-17.2635	-8.5565	0.0000	26.1400	13.2300	12.910
VC-0.5 vs OJ-1.0	9.7401	16.1408	11.5185	17.9215	0.0000	7.9800	22.7000	-14.720
VC-0.5 vs OJ-2.0	14.9665	17.9793	15.5418	20.6182	0.0000	7.9800	26.0600	-18.080
VC-1.0 vs OJ-1.0	4.0328	15.3577	2.8021	9.0579	0.0010	16.7700	22.7000	-5.930
VC-2.0 vs OJ-1.0	-1.7574	17.2972	-7.5643	0.6843	0.0965	26.1400	22.7000	3.440
VC-1.0 vs OJ-2.0	8.0325	17.9476	6.8597	11.7203	0.0000	16.7700	26.0600	-9.290
VC-2.0 vs OJ-2.0	-0.0461	14.0398	-3.7981	3.6381	0.9639	26.1400	26.0600	0.080

Conclusions for t-test

Therefore the tests failed to reject the null-hypothesis (difference in means is equal to zero) in three of the cases. Since all test were held on a 95% confidence level, any values with a p-value greater than 0.05 or the confidence intervals contains the value zero will fail to reject. The results are,

- `VC vs OJ`, p-value of 0.0606
- `VC-2.0 vs OJ-1.0`, p-value of 0.0965
- `VC-2.0 vs OJ-2.0`, p-value of 0.9639

Final Conclusions

Based on the analysis performed we can conclude that as `dose` levels increase the tooth growth also increases. We can also conclude at the lower dose levels there is a statistical difference in the means of the 0.5 and 1.0 mg dose levels for `OJ` and `VC` with `OJ` being more effective on tooth growth, but at 2.0 mg `OJ` and `VC` has no significant difference.

Further study could be conducted to see if higher dosages (above 2.0 mg) would be more effective for `VC` or `OJ`. The mean of `OJ` seems to be decreasing at a higher rate than `VC` and can be seen in the difference of means for the dose levels. `VC` may be more effective than `OJ` at higher levels.

These assumptions are based on the following:

- The sample is representative of the population
- Independent variables were randomly assigned
- The distribution of the means is normal