Statistical Inference Course Project Part2

adrianvs

16 Oct 2015

Introduction

The ToothGrowth dataset examines the effect of vitamin C on the length of odontoblasts (cells responsible for tooth growth) in guinea pigs. 60 animals recieved one of three dose levels of vitamin C (0.5, 1, and 2 mg/day), either in the form of orange juice (OJ) or as ascorbic acid (VC)[1].
This report provides a summary of the data and basic statistical analysis to infer general conclusions.

Exploratory Data Analysis

A. Numeric Summary

The following tables provide a basic summary of the data available. Notice the difference in means between the three dosage groups in table 2 and the smaller variance in the 2 mg group.

n	Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum	Standard Deviation
60	4.20	13.07	19.25	18.81	25.27	33.90	7.65

Table 1: Data Summary

Dose	n	Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum	Standard Deviation
0.5	20	4.20	7.22	9.85	10.61	12.25	21.50	4.50
1	20	13.60	16.25	19.25	19.73	23.38	27.30	4.42
2	20	18.50	23.53	25.95	26.10	27.83	33.90	3.77

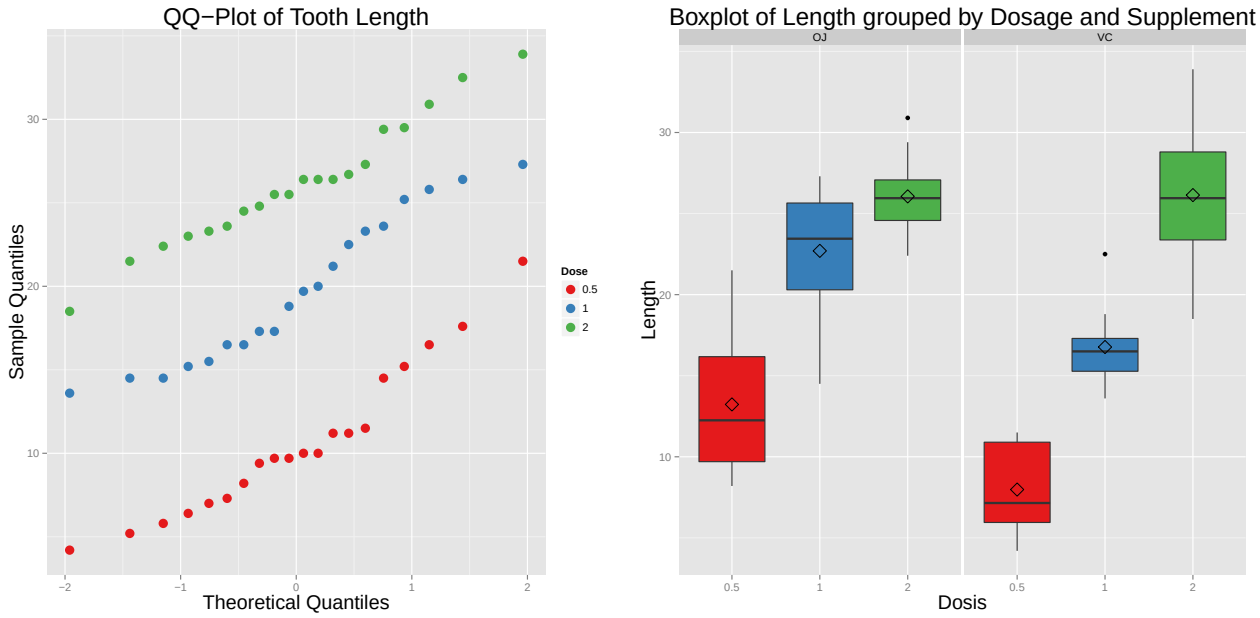
Table 2: Summary grouped by dose [mg/day]

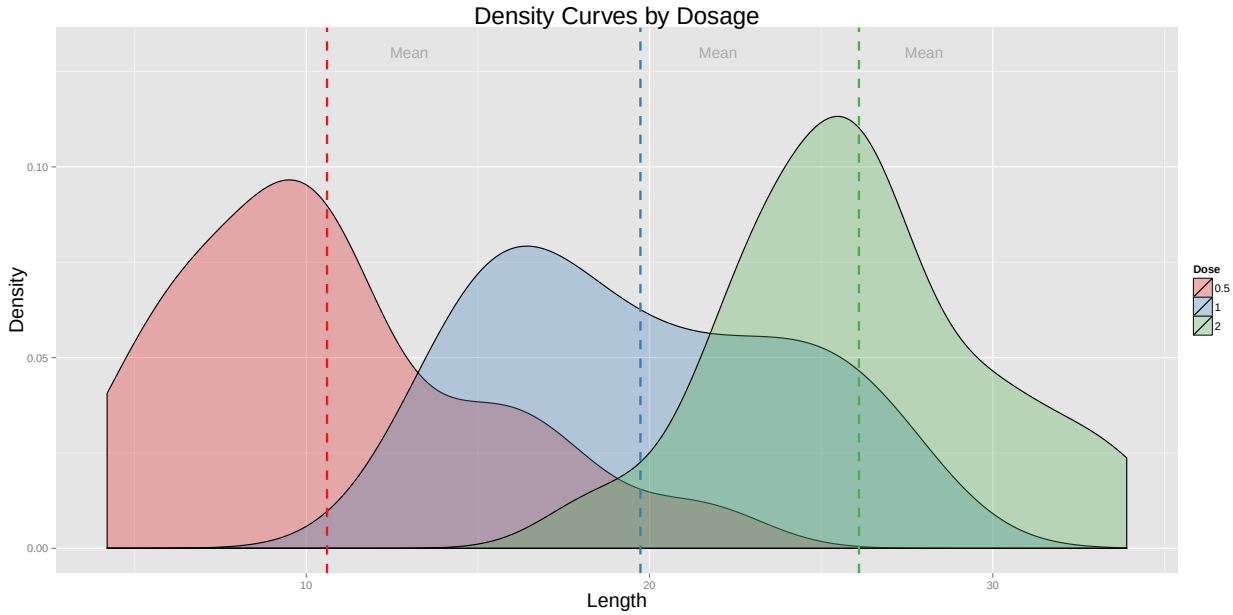
Supplement	n	Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum	Standard Deviation
OJ	30	8.20	15.52	22.70	20.66	25.73	30.90	6.61
VC	30	4.20	11.20	16.50	16.96	23.10	33.90	8.27

Table 3: Summary grouped by Supplement (Orange Juice (OJ), Ascorbic Acid (VC))

B. Distribution

The following graphs show that the sample of odontoblast length is approximately normaly distributed; the quantile-quantile plot lines appear roughly linear. The density curves visualize the difference in mean length and distribution between the three dosage groups. Visual impression suggests three independent groups with distinct but overlapping distributions.





Statistical Analysis

To test whether the observed differences in group means and distribution allow us to infer a correlation in vitamin C dosis and odontoblast length in the general guinea pig populaiton, a two sided, one sample t-test of the difference of means between the different groups was performed. As can be seen from table 4, a higher dose of vitamin C correlates significantly with increased odontoblast length. As could be expected, the confidence intervals show the strongest effect in the comparison of the 0.5mg and 2mg groups. Yet the 0.5mg increase between the 0.5mg and 1mg groups shows a stronger response than the 1mg increase between the 1mg and 2mg groups, suggesting a non-linear relation. These results have p-values smaller than 0.001.
The form of supplement averaged over all dosages correlates to a smaller degree with the observed odontoblast length (CI 1.4 - 6).

	95% CI lower	95% CI upper	p-value
2 mg vs 0.5 mg	12.62	18.37	< 0.001
2 mg vs 1 mg	3.47	9.26	< 0.001
1 mg vs 0.5 mg	6.39	11.87	< 0.001

Table 4: T-test and 95% confidence intervals, grouped by dosage

	95% CI lower	95% CI upper	p-value
Orange Juice vs Ascorbic Acid	1.409	5.991	0.003

Table 5: T-test and 95% confidence intervals, grouped by supplement

Assumptions

The analysis presented here is based on severeal assumptions:

1. The observations are independend and from the same population
2. The distribution of this population approximates the normal distribution
3. The distribustion of the sample approximates the normal or t distribution

As can be seen from the density curves, especially point 3 for the 1mg dosis group can be contested.

Conclusion

1. Each incremental step in vitamin C dosage correlates significantly with an increase in odontoblast length
2. Orange juice as supplement averaged over all dosages correlates positively with the observed odontoblast length.

Appendix - R Code

```
## Loading relevant packages
library(datasets); library(ggplot2); library(dplyr)
library(xtable); library(RColorBrewer); library(gridExtra)

## Naming and arragning the data set
options(xtable.comment = FALSE)
data(ToothGrowth)
df <- as.tbl(ToothGrowth)
names(df) <- c("len","Supplement","Dose")
df$Dose <- as.factor(df$Dose)

## ------------------------------------------------------------------------------##
## Creating the information for numerical summary tables

## for all observations
```

```
table1 <- df %>% summarise("n" = length(len),
                           "Minimum" = min(len),
                           "1st Quantile" = quantile(len,0.25),
                           "Median" = median(len),
                           "Mean" = mean(len),
                           "3rd Quantile" = quantile(len,0.75),
                           "Maximum" = max(len),
                           "Standard Deviation" = sd(len))

## for observations grouped by dosage
table2 <- df %>% group_by(Dose) %>% summarise("n" = length(len),
                                    "Minimum" = min(len),
                                    "1st Quantile" = quantile(len,0.25),
                                    "Median" = median(len),
                                    "Mean" = round(mean(len),2),
                                    "3rd Quantile" = quantile(len,0.75),
                                    "Maximum" = max(len),
                                    "Standard Deviation" = sd(len))

## for observations grouped by supplement type
table3 <- df %>% group_by(Supplement) %>% summarise("n" = length(len),
                                    "Minimum" = min(len),
                                    "1st Quantile" = quantile(len,0.25),
                                    "Median" = median(len),
                                    "Mean" = round(mean(len),2),
                                    "3rd Quantile" = quantile(len,0.75),
                                    "Maximum" = max(len),
                                    "Standard Deviation" = sd(len))

## Creating, formating and printing te tables
print(xtable(table1,align="ccccccccc"), include.rownames=FALSE)
print(xtable(table2,align="cccccccccc"), include.rownames=FALSE)
print(xtable(table3,align="cccccccccc"), include.rownames=FALSE)

## ------------------------------------------------------------------------------------##
## Plots
## Calculating the grouped means
df.mean <- df %>% group_by(Dose) %>% summarise(mean = mean(len))

## Overlaying density curves
DensCurv <-
  ggplot(df, aes(len, fill = Dose)) +
    geom_density(alpha=.3) +
    scale_fill_brewer(palette="Set1") +
    labs(title="Density Curves by Dosage") +
    labs(x="Length", y="Density") +
    theme(plot.title = element_text(size = 22), axis.title = element_text(size = 18)) +
    geom_vline(data=df.mean, aes(xintercept=mean,  colour=Dose),
        linetype="dashed", size=1) +
    scale_colour_brewer(palette="Set1") +
    annotate("text", label = "Mean", x = 13, y = 0.13, size = 5, color="darkgrey") +
    annotate("text", label = "Mean", x = 22, y = 0.13, size = 5, color="darkgrey") +
    annotate("text", label = "Mean", x = 28, y = 0.13, size = 5, color="darkgrey")
```

```
## Quantile-quantile plot, grouped by dosage
qq <-
  ggplot(df, aes(sample = len, color = Dose)) +
    stat_qq(geom="point",size=4) +
    scale_colour_brewer(palette="Set1") +
    labs(title="QQ-Plot of Tooth Length") +
    labs(x="Theoretical Quantiles", y="Sample Quantiles") +
    theme(plot.title = element_text(size = 22), axis.title = element_text(size = 18))

## boxplot grouped by dosage and supplement type
box <-
  ggplot(df, aes(Dose, len)) +
    geom_boxplot(aes(fill = Dose)) +
    facet_grid(. ~ Supplement) +
    scale_fill_brewer(palette="Set1") +
    labs(title="Boxplot of Length grouped by Dosage and Supplement") +
    labs(x="Dosis", y="Length") +
    theme(plot.title = element_text(size = 22), axis.title = element_text(size = 18)) +
    guides(fill=FALSE) +
    stat_summary(fun.y=mean, geom="point", shape=5, size=4)

## printing 3 plots on one canvas
grid.arrange(qq,box,DensCurv, ncol = 2,
             layout_matrix = cbind(c(1,3), c(2,3)))

## ------------------------------------------------------------------------------------##
## Statistical calculations

## Creating 3 groups by dosage
group0.5 <- df %>% filter(Dose == 0.5)
group1   <- df %>% filter(Dose == 1)
group2    <- df %>% filter(Dose == 2)

## Calculating and combining the CIs and p-values for 3 groups in one named matrix
p <- rbind(
        t.test(group2$len-group0.5$len)$p.value,
        t.test(group2$len-group1$len)$p.value,
        t.test(group1$len-group0.5$len)$p.value)

ci <- round(rbind(
        t.test(group2$len-group0.5$len)$conf.int,
        t.test(group2$len-group1$len)$conf.int,
        t.test(group1$len-group0.5$len)$conf.int),2)

ma <- cbind(ci,round(p,9))

## substituting p-value with <0.001 if smaller than that
for(i in 1:3) {
        ifelse (as.numeric(ma[i,3]) < 0.001,ma[i,3] <- "< 0.001",ma[i,3])
}
rownames(ma) <- c("2mg vs 0.5mg","2mg vs 1mg","1mg vs 0.5mg")
colnames(ma) <- c("CI lower","CI upper","p-value")
```

```r
## printing and formating the resulting table
print(xtable(ma,align="lccc"), include.rownames=TRUE)

## Creating groups by supplement
oj <- df %>% filter(Supplement == "OJ")
vc <- df %>% filter(Supplement == "VC")

## Calculating and combining the CIs and p-values for both groups in one named matrix
p2  <- round(t.test(oj$len-vc$len)$p.value,3)
ci2 <- round(t.test(oj$len-vc$len)$conf.int,3)
ma2 <- matrix(c(ci2,p2), ,nrow =1)
colnames(ma2) <- c("CI lower","CI upper","p-value")
rownames(ma2) <- "Orange Juice vs Ascorbic Acid"

## printing and formating the resulting table
print(xtable(ma2,align ="lccc",digits = 3), include.rownames=TRUE)
```

Footnotes 1: ToothGrowth - R Dataset