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

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: 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

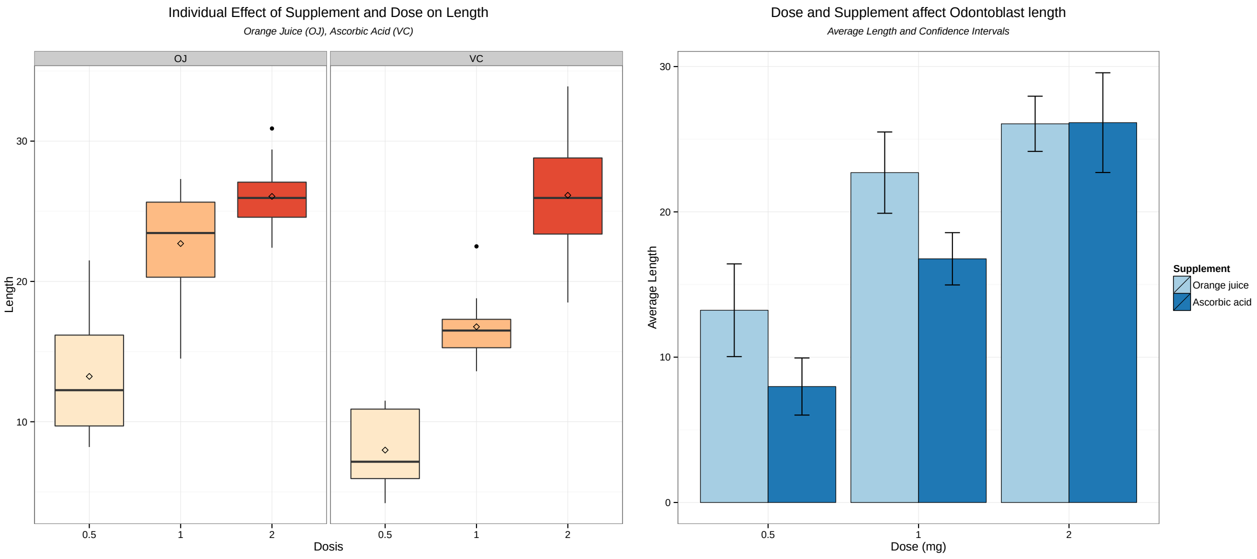
Table 2: Summary by Dose

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 by Supplement

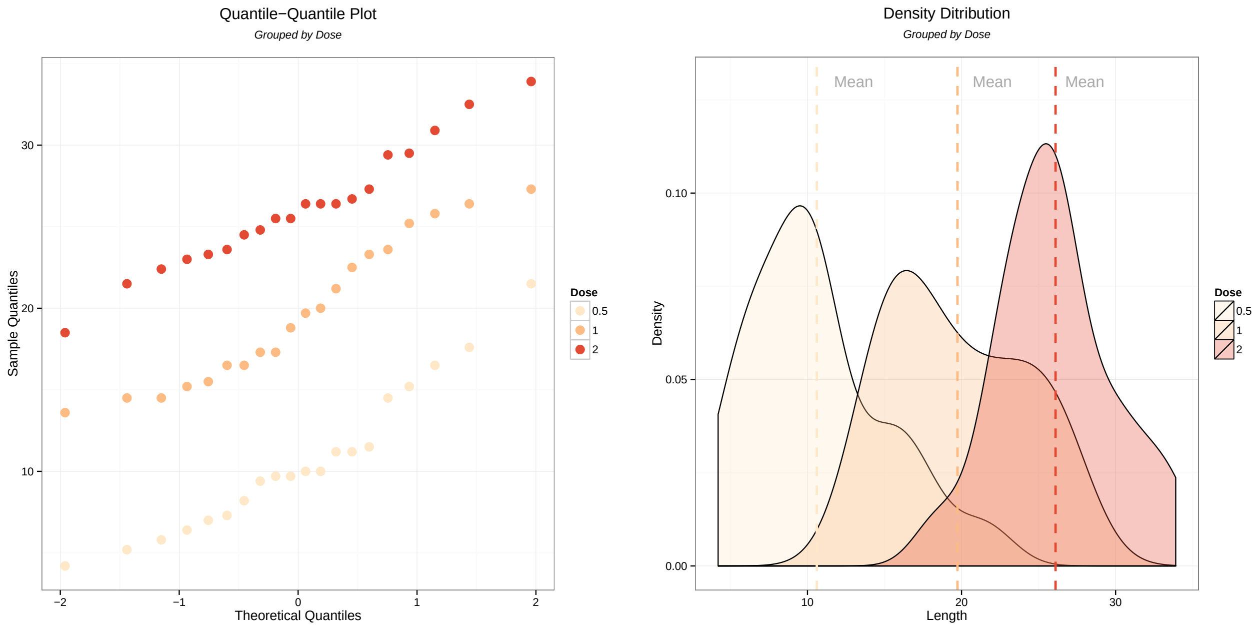
Figure 1 gives a visual summary of the data and shows the effect on the average odontoblast length in relation to the dose of vitamin C and the type of supplement given. Notice the overlapping confidence intervals in the second plot, suggesting that the type of supplement chosen might not influence the average length.

Figure 1: Data Summary



The following graphs (Figure 2) 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.

Figure 2: Distribution



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, two 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, bounded relationship. These results have corresponding p-values smaller than 0.001.
The type of supplement used (Table 5) has no statistical significant effect on average odondoblast length. Tables 6 and 7 show the average length increases by incremental dose for either supllement alone.

	95% CI lower	95% CI upper	p-value
2 mg vs 0.5 mg	12.83	18.16	< 0.001
2 mg vs 1 mg	3.73	9.00	< 0.001
1 mg vs 0.5 mg	6.28	11.98	< 0.001

Table 4: T-test and Confidence Intervals by Dose Groups

	95% CI lower	95% CI upper	p-value
Orange Juice vs Ascorbic Acid	-0.171	7.571	0.061

Table 5: T-test and Confidence Intervals by Supplement Type

	95% CI lower	95% CI upper	p-value
2 mg vs 0.5 mg	9.320	16.340	< 0.001
2 mg vs 1 mg	0.190	6.530	0.039
1 mg vs 0.5 mg	5.520	13.420	< 0.001

Table 6: T-test and Confidence Intervals of Orange Juice by Dose Groups

	95% CI lower	95% CI upper	p-value
2 mg vs 0.5 mg	14.420	21.900	< 0.001
2 mg vs 1 mg	5.690	13.050	< 0.001
1 mg vs 0.5 mg	6.310	11.270	< 0.001

Table 7: T-test and Confidence Intervals of Ascorbic Acid by Dose Groups

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 could be contested.

Conclusion

1. Each incremental step in vitamin C dosage correlates significantly with an increase in odontoblast length
2. The supplement type has no influence on average odontoblast length

Appendix - R Code

```r
## Loading and formating data and packages
library(datasets); library(ggplot2); library(dplyr)
library(xtable); library(RColorBrewer); library(gridExtra)
options(xtable.comment = FALSE)

data(ToothGrowth)
df <- as.tbl(ToothGrowth)
names(df) <- c("len","Supplement","Dose")
df$Dose <- as.factor(df$Dose)

# Creates a statistical summary data frame for plotting of CI error bars
df.sum <- df %>% group_by(Supplement, Dose) %>%
        summarise(n       = length(len),
                  meanlen = mean(len),
                  sd      = sd(len),
                  ci      = qt(0.975, length(len)-1)*sd(len)/sqrt(length(len)))

# Summarises the data frame for table creation
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))


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))

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))

# Creates and prints the summary tables
print(xtable(table1,align="ccccccccc"), include.rownames=FALSE, floating = FALSE)
print(xtable(table2,align="cccccccccc"), include.rownames=FALSE, floating = FALSE)
print(xtable(table3,align="cccccccccc"), include.rownames=FALSE, floating = FALSE)
```

```r
# Creates the boxplot
box <-
  ggplot(df, aes(Dose, len)) +
    geom_boxplot(aes(fill = Dose)) +
    facet_grid(. ~ Supplement) +
    scale_fill_brewer(palette="OrRd") +
    ggtitle(expression(atop("Individual Effect of Supplement and Dose on Length",
                   atop(italic("Orange Juice (OJ), Ascorbic Acid (VC)"), "")))) +
    labs(x="Dosis", y="Length") +
    theme(plot.title = element_text(size = 28), axis.title = element_text(size = 20)) +
    guides(fill=FALSE) +
    stat_summary(fun.y=mean, geom="point", shape=5, size=2) +
    theme_bw()

# Creates the bar plot
bar <-
  ggplot(df.sum, aes(x=Dose, y=meanlen, fill=Supplement)) +
      geom_bar(position=position_dodge(), stat="identity", size=.3, color = "black") +
      scale_fill_brewer(palette="Paired", name="Supplement",
                        breaks=c("OJ", "VC"),
                        labels=c("Orange juice", "Ascorbic acid")) +
      geom_errorbar(aes(ymin=meanlen-df.sum$ci, ymax=meanlen+df.sum$ci),
                   width=.2,
                   position=position_dodge(.9)) +
      xlab("Dose (mg)") +
      ylab("Average Length") +
      ggtitle(expression(atop("Dose and Supplement affect Odontoblast length",
                   atop(italic("Average Length and Confidence Intervals"), "")))) +
      theme(plot.title = element_text(size = 22), axis.title = element_text(size = 18)) +
      theme_bw()

# Prints box and bar plot on one canvas
grid.arrange(box,bar, ncol = 2)

# Creates grouped summary data frame for plotting means
df.mean <- df %>% group_by(Dose) %>% summarise(mean = mean(len))

# Creates overlapping densitiy curves
DensCurv <-
  ggplot(df, aes(len, fill = Dose)) +
    geom_density(alpha=.3) +
    scale_fill_brewer(palette="OrRd") +
    ggtitle(expression(atop("Density Ditribution",
                   atop(italic("Grouped by Dose"), "")))) +
    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="OrRd") +
    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") +
    theme_bw()
```

```r
# Creates the qq-plot
qq <-
  ggplot(df, aes(sample = len, color = Dose)) +
    stat_qq(geom="point", size=4) +
    scale_colour_brewer(palette="OrRd") +
    ggtitle(expression(atop("Quantile-Quantile Plot",
                   atop(italic("Grouped by Dose"), "")))) +
    labs(x="Theoretical Quantiles", y="Sample Quantiles") +
    theme(plot.title = element_text(size = 22), axis.title = element_text(size = 18)) +
    theme_bw()

# Prints density and qq plot on one canvas
grid.arrange(qq,DensCurv, ncol = 2)

# Defines function, input: data frame, output named dataframe with p-values and CIs
testdf <- function(x) {
        p <- rbind(
                with(x, t.test(len[Dose == 2], len[Dose == 0.5])$p.value),
                with(x, t.test(len[Dose == 2], len[Dose == 1])$p.value),
                with(x, t.test(len[Dose == 1], len[Dose == 0.5])$p.value))

        ci <- round(rbind(
                with(x, t.test(len[Dose == 2], len[Dose == 0.5])$conf.int),
                with(x, t.test(len[Dose == 2], len[Dose == 1])$conf.int),
                with(x, t.test(len[Dose == 1], len[Dose == 0.5])$conf.int)),2)

        ma <- as.data.frame(cbind(ci,round(p,9)))
        for(i in 1:3) {
                ifelse (as.numeric(ma[i,3]) < 0.001,ma[i,3] <- "< 0.001", ma[i,3] <-substr(ma[i,3],1,5))
        }
        rownames(ma) <- c("2 mg vs 0.5 mg","2 mg vs 1 mg","1 mg vs 0.5 mg")
        colnames(ma) <- c("95% CI lower","95% CI upper","p-value")

        return(ma)
}

# Prints and creates table 4
print(xtable(testdf(df),align="lccc"),
                        include.rownames=TRUE, floating = FALSE)

# Create table 5
p2  <- round(with(df, t.test(len ~ Supplement)$p.value),3)
ci2 <- round(with(df, t.test(len ~ Supplement)$conf.int),3)
ma2 <- matrix(c(ci2,p2), ,nrow =1)
colnames(ma2) <- c("95% CI lower","95% CI upper","p-value")
rownames(ma2) <- "Orange Juice vs Ascorbic Acid"

# Prints table 5
print(xtable(ma2,align ="lccc",digits = 3, include.rownames=TRUE), floating = FALSE)

# Creates Orange Juice data frame, calls defined function and prints table 6
oj <- df %>% filter(Supplement == "OJ")
print(xtable(testdf(oj),align ="lccc",digits = 3, include.rownames=TRUE), floating = FALSE)
```

```r
# Creates Ascorbic Acid data frame, calls defined function and prints table 7
vc <- df %>% filter(Supplement == "VC")
print(xtable(testdf(vc), align ="lccc",digits = 3, include.rownames=TRUE), floating = FALSE)
```

Footnotes

1: ToothGrowth - R Dataset