Analysis of Guinea Pig Tooth Growth

Steve Myles

July 22, 2015

Contents

Data Summary / Exploratory Analysis

This report provides a comparison of guinea pig odontoblast (tooth) growth (`len`, measured in microns) by supplement (`supp`) and daily dosage (`dose`) using the `ToothGrowth` data from the R `datasets` package. First, the data were loaded into a data frame (replacing `dose`'s values with factors as no calculations were performed on dosage) and relevant packages were loaded as shown in Appendix 1. The data were then evaluated and summarized.

This data set has 60 observations of 3 variables (`len`, tooth length in microns; `supp`, either "OJ" for orange juice or "VC" for an ascorbic acid vitamin C supplement; and `dose`, the daily dosage of vitamin C, either 0.5mg, 1mg, or 2mg). Each combination of `supp` and `dose` has 10 observations so there are 30 total observations for each `supp` and 20 for each `dose`. A data summary using the `str` and `summary` functions can be found in Appendix 2.

The data were plotted in a series of panel histograms (one for each `dose` regardless of `supp`, one for each `supp` regardless of `dose`, and one for each combination of `supp` and `dose`) using the `lattice` package's `histogram` function. Code for producing these plots is available in Appendix 3.

Fig. 1: len by dose

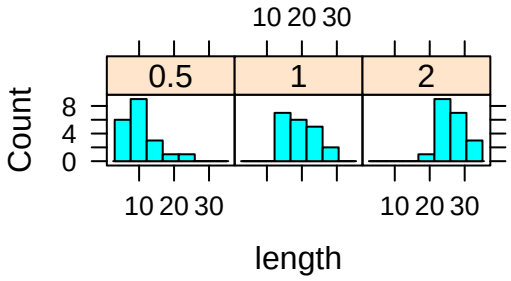


Fig. 2: len by supp

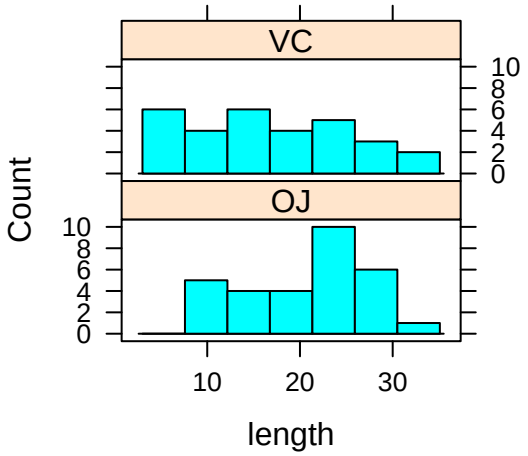
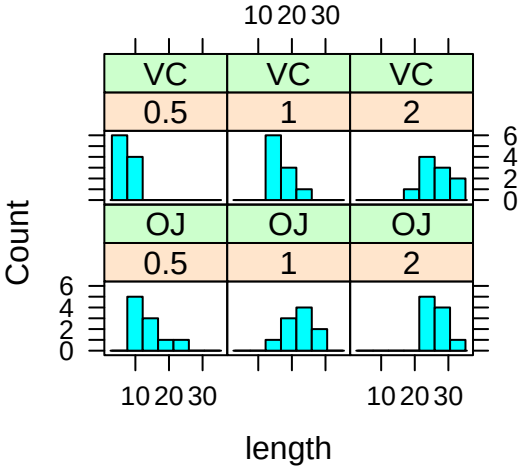


Fig. 3: len by supp and dose



From the plots, it appears that higher dosage yields larger lengths (Fig. 1), regardless of supplement (in Fig. 3, `len` seems to be greater for `dose` = 2 for both `supp` = "VC" and "OJ"). Additionally, in Fig. 3, orange juice ("OJ") seems to be associated with larger lengths than ascorbic acid ("VC"), regardless of dosage (the distribution of length for "OJ" (Fig. 2, as well as for each dosage in Fig. 3) seems to be further to the right than the distribtuion of length for "VC." However, the relationship of `supp` and `len` is less clear in Fig. 2 where the two distributions overlap to a large extent.

The means and standard deviations for each level of `supp` and `dose` and each combination of these factors were calculated using the `dplyr` package's `summarize` function. Code for generating this data frame is available in Appendix 4.

```
## summary data frame:  'total' refers to either dose irrespective of
##         supplement or supplement irrespective of dose
```

```
## Source: local data frame [11 x 5]
##
##     dose     mean       sd  n  supp
## 1    0.5 13.23000 4.459709 10    OJ
## 2    0.5  7.98000 2.746634 10    VC
## 3    0.5 10.60500 4.499763 20 total
## 4      1 22.70000 3.910953 10    OJ
## 5      1 16.77000 2.515309 10    VC
## 6      1 19.73500 4.415436 20 total
## 7      2 26.06000 2.655058 10    OJ
## 8      2 26.14000 4.797731 10    VC
## 9      2 26.10000 3.774150 20 total
## 10 total 20.66333 6.605561 30    OJ
## 11 total 16.96333 8.266029 30    VC
```

As can be seen, the sample means differ. Tests were then performed to determine conclusions about the population means based on these samples.

Hypothesis Tests

The `data` data frame was split into 11 data frames (one corresponding to each row of the above summary) for easy t-testing of each combination of `dose` and `supp`. Seven t-tests with $\alpha = 0.05$ were performed to compare the mean lengths of each group (i.e., comparing the mean `len` for the values of `dose` and `supp` as well as for each of their combinations – a `dose` of 0.5mg was compared to a `dose` of 1mg and a `dose` of 2mg irrespective of `supp`, "VC" was compared to "OJ" irrespective of dose, and "VC" and "OJ" were compared for each level of `dose`). Due to the small sample sizes, t-tests were chosen for comparing the distribution means. Null hypotheses, alternative hypotheses, and p-values are provided below for each hypothesis test and the code used to subset and t-tests the groups is available in Appendix 5.

Hypothesis tests and associated p-values:

1. $H_0 : \mu_{dose=0.5} = \mu_{dose=1}$
   $H_A : \mu_{dose=0.5} \neq \mu_{dose=1}$
   p-value < 0.00001
2. $H_0 : \mu_{dose=0.5} = \mu_{dose=2}$
   $H_A : \mu_{dose=0.5} \neq \mu_{dose=2}$
   p-value < 0.00001
3. $H_0 : \mu_{dose=1} = \mu_{dose=2}$
   $H_A : \mu_{dose=1} \neq \mu_{dose=2}$
   p-value $= 0.00002$
4. $H_0 : \mu_{supp=VC} = \mu_{supp=OJ}$
   $H_A : \mu_{supp=VC} \neq \mu_{supp=OJ}$
   p-value $= 0.06063$
5. $H_0 : \mu_{dose=0.5,supp=VC} = \mu_{dose=0.5,supp=OJ}$
   $H_A : \mu_{dose=0.5,supp=VC} \neq \mu_{dose=0.5,supp=OJ}$
   p-value $= 0.00636$
6. $H_0 : \mu_{dose=1,supp=VC} = \mu_{dose=1,supp=OJ}$
   $H_A : \mu_{dose=1,supp=VC} \neq \mu_{dose=1,supp=OJ}$
   p-value $= 0.00104$
7. $H_0 : \mu_{dose=2,supp=VC} = \mu_{dose=2,supp=OJ}$
   $H_A : \mu_{dose=2,supp=VC} \neq \mu_{dose=2,supp=OJ}$
   p-value $= 0.96385$

Conclusions

Based on the sample data, the null hypotheses (H_0) are accepted for cases 4 (t-test comparing mean `len` of supplements "VC" and "OJ" regardless of `dose`) and 7 (t-test comparing mean `len` of supplements "VC" and "OJ" for `dose` = 2mg) as these t-tests' p-values $> \alpha = 0.05$. H_0 in cases 1-3 and 5-6 must be rejected in favor of the alternative hypotheses (H_A) because their p-values $< \alpha = 0.05$.

In other words, when `supp` is not a factor, the individual values of `dose` are associated with different average odontoblast lengths in microns (as evidenced by hypothesis tests 1-3). Additionally, based on hypothesis test 4, one cannot say that the different supplements (`supp`, orange juice ("OJ") and ascorbic acid ("VC")) are associated with different mean odontoblast lengths. However, taking both `dose` and `supp` into account yields different results: when `dose` = 0.5mg/day and when `dose` = 1mg/day, one cannot say that the mean odontoblast lengths are different for the different supplements (hypothesis tests 5-6) but when `dose` = 2mg/day, one can assume that the mean length is the same regardless of supplement (as H_0 is accepted in that case).

Appendix

1: Loading the data and required packages

```
## effectively turn off scientific notation
options(scipen=999)
## load packages and data, convert dose to factor
library(dplyr, warn.conflicts = FALSE) ## load the dplyr package
library(lattice) ## load the lattice package
data <- ToothGrowth ## assign the ToothGrowth data to a data frame
data$dose <- as.factor(data$dose) ## convert dose to factor
```

2: Structure of the `ToothGrowth` data after assigning to the `data` data frame

```
str(data) ## review the structure of the data frame
```

```
## '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: Factor w/ 3 levels "0.5","1","2": 1 1 1 1 1 1 1 1 1 1 ...
```

```
summary(data)
```

```
##       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
```

3: Code for exploratory data analysis

```
## histogram of length by dose
histogram(~len | dose, data = data, layout = c(3, 1), type = "count",
          main = "Fig. 1: len by dose", xlab = "length")
```

```
## histogram of length by supp
histogram(~len | supp, data = data, layout = c(1, 2), type = "count",
          main = "Fig. 2: len by supp", xlab = "length")
## histogram of length by combination of dose and supplement
histogram(~len | dose * supp, data = data, layout = c(3, 2), type = "count",
          main = "Fig. 3: len by supp and dose", xlab = "length")
```

4: Code for summary statistics

```
## summarize by dose
summDose <- summarize(group_by(data, dose), mean = mean(len), sd = sd(len),
                      n = n())
summDose$supp <- "total"
## summarize by supplement
summSupp <- summarize(group_by(data, supp), mean = mean(len), sd = sd(len),
                      n = n())
summSupp$dose <- "total"
## summarize by dose and supplement
summDoseSupp <- summarize(group_by(data, dose, supp), mean = mean(len),
                          sd = sd(len), n = n())

## combine summaries and sort using dplyr's arrange function
summ <- data.frame()
summ <- rbind(summ, summDose, summSupp, summDoseSupp)
summ <- arrange(summ, dose, supp)

## display the summary data frame
message("summary data frame:  'total' refers to either dose irrespective of
        supplement or supplement irrespective of dose")
summ
```

5: Code for subsetting and t-testing the data

```
## ***** subsetting *****
##
## subset the data frame based on dose, supplement, and combinations of them
## rows where `dose` = "0.5" (irrespective of `supp`)
data0.5 <- data[data$dose=="0.5",]
## rows where `dose` = "1" (irrespective of `supp`)
data1 <- data[data$dose=="1",]
## rows where `dose` = "2" (irrespective of `supp`)
data2 <- data[data$dose=="2",]
## rows where `supp` = "VC" (irrespective of `dose`)
dataVC <- data[data$supp=="VC",]
## rows where `supp` = "OJ" (irrespective of `dose`)
dataOJ <- data[data$supp=="OJ",]
## rows where `dose` = "0.5" and `supp` = "VC"
data0.5VC <- data[data$dose == "0.5" & data$supp=="VC",]
## rows where `dose` = "0.5" and `supp` = "OJ"
data0.5OJ <- data[data$dose == "0.5" & data$supp=="OJ",]
## rows where `dose` = "1" and `supp` = "VC"
data1VC <- data[data$dose == "1" & data$supp=="VC",]
## rows where `dose` = "1" and `supp` = "OJ"
data1OJ <- data[data$dose == "1" & data$supp=="OJ",]
## rows where `dose` = "2" and `supp` = "VC"
data2VC <- data[data$dose == "2" & data$supp=="VC",]
## rows where `dose` = "2" and `supp` = "OJ"
data2OJ <- data[data$dose == "2" & data$supp=="OJ",]
```

```
## ***** t-tests with p-values displayed under hypothesis tests ******
##
## t-test for means of dose = 0.5 and dose = 1
tdose0.5_1 <- t.test(data0.5$len, data1$len)
## determine p-value and display "< 0.00001" if it is or "= p-value" if not
pval0.5_1 <- ifelse(tdose0.5_1$p.value < 0.00001, paste0("< 0.00001"),
                    paste0("= ", round(tdose0.5_1$p.value, digits = 5)))

## t-test for means of dose = 0.5 and dose = 2
tdose0.5_2 <- t.test(data0.5$len, data2$len)
## determine p-value and display "< 0.00001" if it is or "= p-value" if not
pval0.5_2 <- ifelse(tdose0.5_2$p.value < 0.00001, paste0("< 0.00001"),
                    paste0("= ", round(tdose0.5_2$p.value, digits=5)))

## t-test for means of dose = 1 and dose = 2
tdose1_2 <- t.test(data1$len, data2$len)
## determine p-value and display "< 0.00001" if it is or "= p-value" if not
pval1_2 <- ifelse(tdose1_2$p.value < 0.00001, paste0("< 0.00001"),
                    paste0("= ", round(tdose1_2$p.value, digits = 5)))

## t-test for means of supp = VC and supp = OJ
tsupp <- t.test(dataVC$len, dataOJ$len)
## determine p-value and display "< 0.00001" if it is or "= p-value" if not
pvalsupp <- ifelse(tsupp$p.value < 0.00001, paste0("< 0.00001"),
                    paste0("= ", round(tsupp$p.value, digits = 5)))

## t-test for means of dose = 0.5 for supp = VC and supp = OJ
tdose0.5 <- t.test(data0.5VC$len, data0.5OJ$len)
## determine p-value and display "< 0.00001" if it is or "= p-value" if not
pval0.5 <- ifelse(tdose0.5$p.value < 0.00001, paste0("< 0.00001"),
                    paste0("= ", round(tdose0.5$p.value, digits = 5)))

## t-test for means of dose = 1 for supp = VC and supp = OJ
tdose1 <- t.test(data1VC$len, data1OJ$len)
## determine p-value and display "< 0.00001" if it is or "= p-value" if not
pval1 <- ifelse(tdose1$p.value < 0.00001, paste0("< 0.00001"),
                    paste0("= ", round(tdose1$p.value, digits = 5)))

## t-test for means of dose = 2 for supp = VC and supp = OJ
tdose2 <- t.test(data2VC$len, data2OJ$len)
## determine p-value and display "< 0.00001" if it is or "= p-value" if not
pval2 <- ifelse(tdose2$p.value < 0.00001, paste0("< 0.00001"),
                    paste0("= ", round(tdose2$p.value, digits = 5)))
```