Statistical Inference Course Project

Part 2: Basic Inferential Data Analysis

Irena Papst

Overview

We explore the `ToothGrowth` dataset, which addresses the effect of vitamin C on odontoblast growth in guinea pigs. Odontoblasts are cells that contribute to the formation of dentin in vertibrates, a key component of teeth (Wikipedia contributors 2018). We conclude that increasing the dose of either vitamin C supplement tested (orange juice and ascorbic acid) leads to longer odontoblasts. However, the shape of each dose-response curve appears to be different for each supplement.

Data

```
library(tidyverse)
```

```
## Load data
library(datasets)
data <- ToothGrowth

## Display summary of data
str(data)
```

```
## '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 ...
```

The data includes 60 observations of 3 variables:

- `len`: length of odontoblasts (cells responsible for tooth growth);
- `supp`: supplement type, either:
  - `VC`: ascorbic acid (VC);
  - `OJ`: orange juice (OJ);
- `dose`: dose in milligrams per day (0.5, 1, or 2).

The length of ondontoblasts is the response variable, and the supplement and dose variables denote the treatment received by each animal.

We convert the `dose` variable to a categorical (instead of numerical) variable, since we will want to group individual observations by both supplement type and dose. We group the data by supplement type, and then by dose, so that we may examine the effect of each of these treatments on cell length.

```
## Convert dose to factor variable
data$dose <- as.factor(data$dose)

## Group data by supplement, then by dose
data <- data %>% group_by(supp, dose)
```

We check whether any observations are missing data, and then present the sample sizes of each category (grouped by supplement type and dose).

```
## Check for missing values
n_missing <- sum(!(complete.cases(data)))
print(paste0("Number of observations missing data: ",
    n_missing))
```

```
## [1] "Number of observations missing data: 0"
```

```
## Display sample sizes
data %>% summarise(sample_size = n())
```

```
## # A tibble: 6 x 3
## # Groups:   supp [?]
##   supp  dose  sample_size
##   <fct> <fct>       <int>
## 1 OJ    0.5            10
## 2 OJ    1              10
## 3 OJ    2              10
## 4 VC    0.5            10
## 5 VC    1              10
## 6 VC    2              10
```

Exploratory analysis

Preliminary question

It would be interesting to test whether either supplement type or dose affect the length of odontoblasts.

Note that since there is no control group included in this data, we cannot explore whether either supplement/dose can significantly change the length of these cells from baseline (in an untreated animal with all else being equal), but we can at least test whether there are any significant differences in cell length among the supplements (or doses).

Based on this reasoning, we can pose a preliminary analysis question to explore:

Question 1: Are there significant differences in cell length among the different supplements and doses tested?

Box and whiskers plot

Since there is no missing data and the sample size for each category is equal, we can safely compare cell length among groups with no further adjustments.

We start with a basic box plot to get a sense of any potential differences among treatment categories.

```
## Add combined factor variable for both
## supplement and dose
data <- data %>% unite(cat, supp, dose, sep = "-",
    remove = FALSE)

data$cat <- as.factor(data$cat)

## Define custom colour pallettes
col_pallette <- c("orange1", "violetred1")
fill_pallette <- c("grey80", "grey65", "grey50")

## Make boxplot
p <- ggplot(data, aes(x = cat, y = len, col = supp,
    fill = dose)) + ## Display mean in box instead of median
geom_boxplot(aes(middle = mean(len))) + ## Use custom colour palettes
scale_colour_manual(values = col_pallette) +
    scale_fill_manual(values = fill_pallette) +
    ## Add annotations
labs(x = "Category (Supplement-Dose)", y = "Cell length")

print(p)
```

This box plot shows the mean of the data within each category (the horizontal line in each box), as well as the 25th and 75th percentiles (the bottom and top edges of the box, respectively). The whiskers (vertical lines extending from the boxes) span the range of the remaining data beyond the 25th and 75th percentiles, with the exception of outliers (plotted individually as points).

Comparing orange juice and ascorbic acid as delivery methods for vitamin C, it seems that there is not a significant difference in each supplements effect on odontoblast length. There does, however, seem to be a trend in the dosage level; it appears that as dosage increases, so does cell length (independent of supplement). However, we should test this assertion more formally to quantify whether this trend is statistically significant.

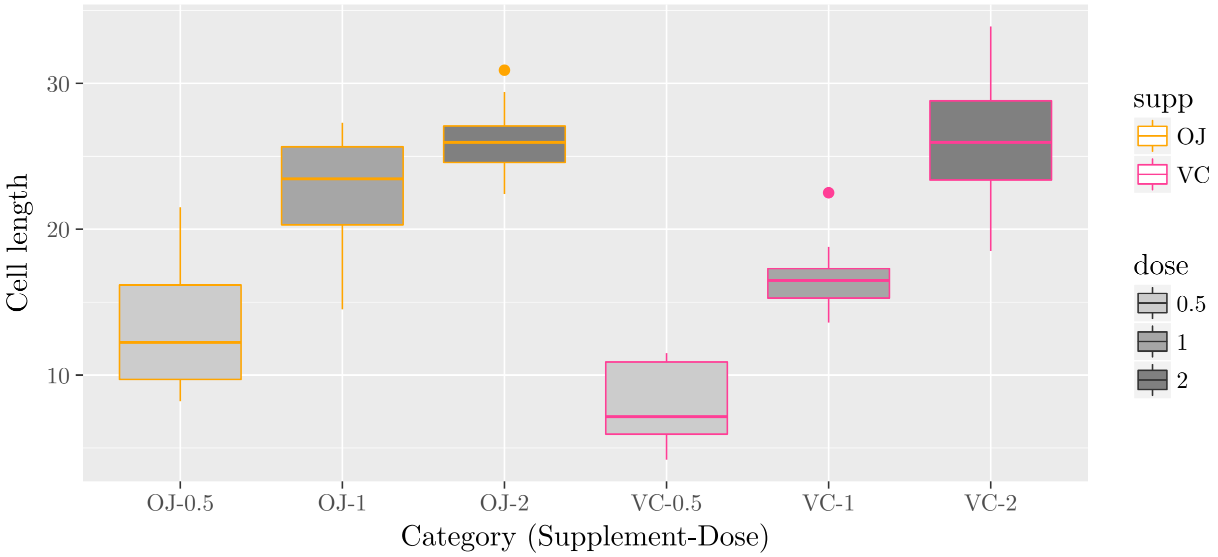

Figure 1: Distributions of cell length by supplement (orange juice and ascorbic acid), and various doses in miligrams per day. Means are marked with a horizontal line in the centre of each box, whose borders are the 25th and 75th percentiles. The whiskers denote the rest of the data, with the exception of outliers, which are plotted individually as points.

T-tests

We perform pairwise T-tests among the 6 different categories of treatment (2 supplements, 3 dosage levels) to test more rigorously whether there are significant differences among groups.

Test assumptions

The assumptions of these tests are:

1. the population was sampled randomly;
2. the observations are independent;
3. the population is approximately normally distributed.

Details of data collection can be found in the original study (Crampton 1947). There is nothing in this study to indicate that any of the above assumptions were violated, so we proceed with the T-tests.

Hypotheses and test parameters

The null hypothesis for these tests is that the population means are equal between any two categories. The alternative hypothesis is that the population means are not equal, and so we use a *two-sided* test. Moreover, each observation comes from a different guinea pig, so we use an *unpaired* test.

Since we are performing a number of pairwise tests, we must adjust our p-values to correct for accumulating error due to multiple comparisons. We will use the Benjamini-Hochberg correction, which controls the false *discovery* rate.

Test results

We perform all 15 t-tests and report a matrix where the different categories label the rows and columns, and the entry denotes whether $p < 0.05$ for the t-test on the pair specified by the row and column labels. For instance, if the entry in row `X` and column `Y` is `TRUE` then $p < 0.05$ for the t-test on the category pair `X-Y`, and so the mean between those groups is significantly different (we reject the null hypothesis that the means of groups `X` and `Y` are equal).

```r
ttest <- with(data, pairwise.t.test(len,
    cat, p.adjust.method = "BH", alternative = "two.sided"))

## Display which pairs are significantly
## different p < 0.05
with(ttest, p.value < 0.05)
```

```
##        OJ-0.5 OJ-1  OJ-2 VC-0.5 VC-1
## OJ-1     TRUE   NA    NA     NA   NA
## OJ-2     TRUE TRUE    NA     NA   NA
## VC-0.5   TRUE TRUE  TRUE     NA   NA
## VC-1     TRUE TRUE  TRUE   TRUE   NA
## VC-2     TRUE TRUE FALSE   TRUE TRUE
```

Conclusion

According to our t-tests, all of the means are significantly different, except for the two supplements at 2 mg/day, where there was not enough evidence to reject the null hypothesis. Paired with the boxplot in figure 1, these results suggest that for each supplement, different doses lead to statistically significant increases in odontoblast length.

Among the two supplements, there seems to be a statistically different dose-response curve, since the means for either supplement at (a) 0.5 mg/day and (b) 1 mg/day are significantly different. If we inspect the boxplot in figure 1, we see that the dose-response curve for orange juice looks sublinear (concave down), while the curve for ascorbic acid looks linear. However, since we could not reject the null hypothesis that the means for both supplement at 2 mg/day were significantly different, it could be the case that both supplements eventually result in the same effect on odontoblast length (at a sufficiently high dosage).

Citations

Crampton, E. W. 1947. “The Growth of the Odontoblasts of the Incisor Tooth as a Criterion of the Vitamin c Intake of the Guinea Pig Five Figures.” *The Journal of Nutrition* 33 (5): 491–504. doi:10.1093/jn/33.5.491.

Wikipedia contributors. 2018. “Odontoblast — Wikipedia, the Free Encyclopedia.” https://en.wikipedia.org/wiki/Odontoblast.