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 tooth growth in guinea pigs.

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;
  - `OJ`: orange juice;
- `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. Then 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

Based on the contents of this data, 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, 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.

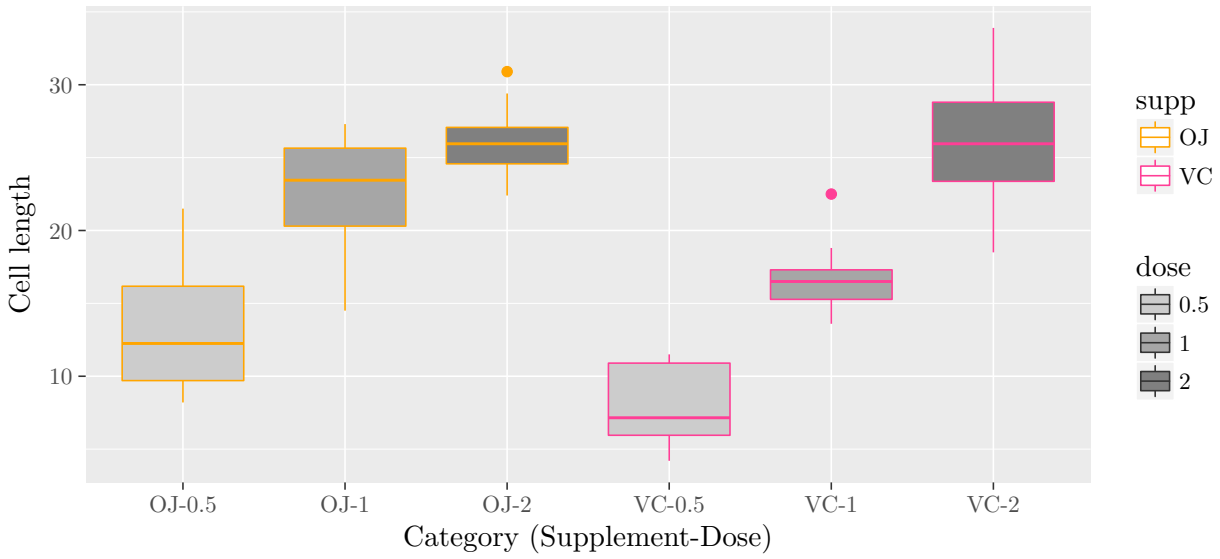
```
## Add combined factor variable for both
## supplement and dose
data <- data %>% unite(cat, supp, dose, sep = "-",
    remove = FALSE)

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)) + geom_boxplot() + ## 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 median of the data within each category (the horizontal line in each box), as well as the 25^{th} and 75^{th} 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 25^{th} and 75^{th} percentiles, with the exeption of outliers (plotted individually as points).

Comparing orange juice (OJ) and ascorbic acid (VC) as delivery methods for vitamin C, it seems that there is not a significant difference between the two supplement types. 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.

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.

The assumptions of these tests are: 1. the population was sampled randomly; 1. the observations are independent; 1. the population is approximately normally distributed.

Details of data collection can be found in (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.

The null hypothesis for these tests is that the population means are equal between any given groups. 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 multiple comparisons. We will use the Bonferorni correction, which is rather strict.

Appendix A: Session Info

```
sessionInfo()
```

```
## R version 3.4.4 (2018-03-15)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS High Sierra 10.13.6
##
## Matrix products: default
## BLAS: /Library/Frameworks/R.framework/Versions/3.4/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.4/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_CA.UTF-8/en_CA.UTF-8/en_CA.UTF-8/C/en_CA.UTF-8/en_CA.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] forcats_0.3.0   stringr_1.3.0   dplyr_0.7.4     purrr_0.2.4
## [5] readr_1.1.1     tidyr_0.8.0     tibble_1.4.2    ggplot2_2.2.1
## [9] tidyverse_1.2.1
##
## loaded via a namespace (and not attached):
##  [1] tinytex_0.5      tidyselect_0.2.4 reshape2_1.4.3   haven_1.1.1
##  [5] lattice_0.20-35  colorspace_1.3-2 htmltools_0.3.6  yaml_2.1.18
##  [9] utf8_1.1.3       rlang_0.2.0      pillar_1.2.1     foreign_0.8-69
## [13] glue_1.2.0       tikzDevice_0.11  modelr_0.1.1     readxl_1.0.0
## [17] bindrcpp_0.2.2   bindr_0.1.1      plyr_1.8.4       munsell_0.4.3
## [21] gtable_0.2.0     cellranger_1.1.0 rvest_0.3.2      codetools_0.2-15
## [25] psych_1.8.4      evaluate_0.10.1  labeling_0.3     knitr_1.20
## [29] parallel_3.4.4   broom_0.4.4      Rcpp_0.12.16     scales_0.5.0
## [33] backports_1.1.2  formatR_1.5      filehash_2.4-1   jsonlite_1.5
## [37] mnormt_1.5-5     hms_0.4.2        digest_0.6.15    stringi_1.1.7
## [41] grid_3.4.4       rprojroot_1.3-2  cli_1.0.0        tools_3.4.4
## [45] magrittr_1.5     lazyeval_0.2.1   crayon_1.3.4     pkgconfig_2.0.1
## [49] xml2_1.2.0       lubridate_1.7.4  assertthat_0.2.0 rmarkdown_1.9
## [53] httr_1.3.1       rstudioapi_0.7   R6_2.2.2         nlme_3.1-131.1
## [57] compiler_3.4.4
```

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