Exploring Effects of Vitamin C on Tooth Growth in Guinea Pigs

Alex Hudson

2019-01-23

Setup

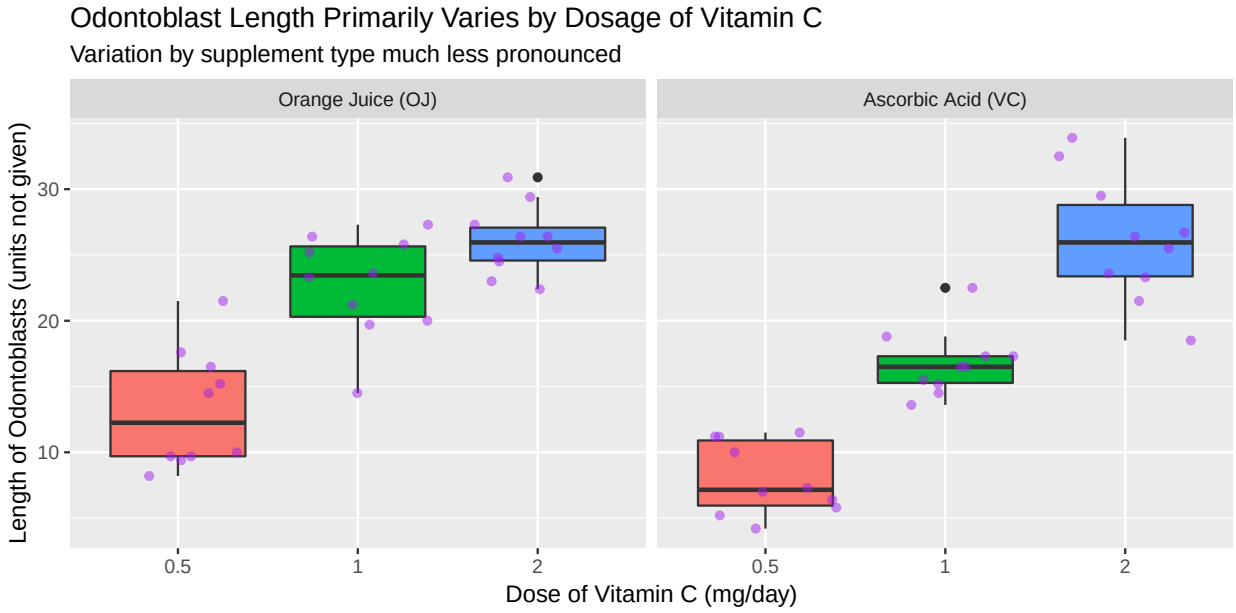
```
library(ggplot2)
library(dplyr)
library(knitr)
library(kableExtra)
library(stringr)
data(ToothGrowth)
```

Overview

For this analysis, we will explore and test for differences in the `ToothGrowth` data set that is found in the R package `datasets`. We will use basic exploratory analysis to theorize about potential differences in subgroups of the data, and we will apply independent two-sample T tests to look for differences in means between different subgroups of the data.

Exploratory Analysis

Initial exploration of the data reveals that there are 60 total measurements: they are split into six subgroups, each of which corresponds to one possible combination of delivery method and vitamin C dosage and contains 10 measurements. In the plot below, we first split the data into two groups: those for which the delivery method is orange juice (left plot), and those for which the delivery method is ascorbic acid (right plot). Within each plot, we further split the data by the vitamin C dosage. For each of the six groups, we construct a boxplot of the group, and we overlay a jitterplot so that we see the data as well as the summary.



Overall, the data vary between about 4 and 35. There appears to be some separation between vitamin C dosage levels in both delivery method groups, with odontoblast length increasing as vitamin C dosage level increases. If we consider the delivery method groups as the only grouping, then it appears that the ascorbic acid group may be slightly more variable than the orange juice group. However, the variability does appear to change across vitamin C dosage levels differently in each delivery method group.

We'll also look at some summary statistics for various groups and subgroups of the data. In the table below, we show the mean, median, standard deviation, and interquartile range for 12 groups: the full data set, two subsets constructed by filtering for delivery method, three subsets constructed by filtering for vitamin C dosage, and six subsets constructed by filtering for both factors. By both interquartile range and standard deviation, the full data set and the subsets created by filtering the delivery method are the most variable groups of data. Due to this variability, it is difficult to determine whether the delivery method groups has significantly different means from each other. Meanwhile, all of the groups that filter on vitamin C dosage have much lesser variability, and there may be significant differences in means for groups with different levels of vitamin C dosage.

Summary Statistics for Odontoblast Length Filtered by Delivery Method and Vitamin C Dosage

Delivery Method	Dosage of Vitamin C	Mean	Median	Standard Deviation	Interquartile Range
No Filter	No Filter	18.81	19.25	7.65	12.20
Orange Juice (OJ)	No Filter	20.66	22.70	6.61	10.20
Ascorbic Acid (VC)	No Filter	16.96	16.50	8.27	11.90
No Filter	0.5	10.61	9.85	4.50	5.03
No Filter	1	19.73	19.25	4.42	7.12
No Filter	2	26.10	25.95	3.77	4.30
Orange Juice (OJ)	0.5	13.23	12.25	4.46	6.48
Orange Juice (OJ)	1	22.70	23.45	3.91	5.35
Orange Juice (OJ)	2	26.06	25.95	2.66	2.50
Ascorbic Acid (VC)	0.5	7.98	7.15	2.75	4.95
Ascorbic Acid (VC)	1	16.77	16.50	2.52	2.03
Ascorbic Acid (VC)	2	26.14	25.95	4.80	5.43

Hypothesis Tests for Comparisons

Next, we'll set up and perform our hypothesis tests. For all of our hypothesis tests, we'll use independent two-sample T tests for differences in means, for which we have the hypotheses $H_0 : \mu_1 - \mu_2 = 0$ and $H_a : \mu_1 - \mu_2 \neq 0$. For an initial significance level, we'll use $\alpha = 0.05$, but since we're conducting 13 hypothesis tests with the same data, we'll use the Bonferroni correction to control the family-wise error rate. Thus, we'll use the adjusted significance level $\alpha_{\text{Bonf.}} = 0.05/13 \approx 0.0038$. The following table shows the P-values and indications of significance for all 13 pairings of subsets that we chose. See the appendix for the naming convention used to identify subsets of data.

Significance of Independent Two-Sample T Tests for 13 Pairs of Subsets of Data Using Bonferroni Correction for Multiple Tests

Data Subsets	P-value	Significant (Bonferroni)
TGOJ and TGVC	0.0606	FALSE
TGOJ05 and TGVC05	0.0064	FALSE
TGOJ10 and TGVC10	0.0010	TRUE
TGOJ20 and TGVC20	0.9639	FALSE
TG05 and TG10	0.0000	TRUE
TG05 and TG20	0.0000	TRUE
TG10 and TG20	0.0000	TRUE
TGOJ05 and TGOJ10	0.0001	TRUE
TGOJ05 and TGOJ20	0.0000	TRUE
TGOJ10 and TGOJ20	0.0392	FALSE
TGVC05 and TGVC10	0.0000	TRUE
TGVC05 and TGVC20	0.0000	TRUE
TGVC10 and TGVC20	0.0001	TRUE

We have four tests in which the main difference between groups is the delivery method, and we have nine tests in which the main difference between groups is the vitamin C dosage. Of the four delivery method tests, only one generates a P-value that is considered significant after applying the Bonferroni correction. Meanwhile, of the nine vitamin C dosage tests, eight generate a P-value that is considered significant after applying the Bonferroni correction. Further, six of those nine tests generate P-values that are less than 0.0001, which is well below the significance threshold.

Required Assumptions

The independent two-sample T test requires us to make some assumptions about the population that produced our data. First, the data must be independent, both within their own groups and from the other group. Second, the data within each group must be identically distributed, and that distribution must be the normal distribution. If these assumptions are met (or are reasonable), then we can use Gosset's T distribution as a method for testing our hypotheses. In our case, we have no reason to suspect that any of these assumptions are significantly violated.

Conclusions

From the results of the hypothesis tests, it appears reasonable to conclude that there is not a major difference in mean odontoblast lengths between delivery methods. Only one of the four hypothesis tests for which the delivery method was the primary difference was statistically significant. Meanwhile, it also appears reasonable to conclude that there is a significant difference in mean odontoblast lengths across different vitamin C dosage levels. Eight of the nine hypothesis tests for which vitamin C dosage was the primary difference were statistically significant, and six of those nine tests were highly significant.

Appendix

Data

The data are stored in the `ToothGrowth` data set in R. From the documentation for the data, "The response is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, orange juice or ascorbic acid (a form of vitamin C and coded as `VC`)." In the data, the numeric variable `len` records the length of the odontoblasts, the factor variable `supp` denotes the delivery method with `OJ` for orange juice and `VC` for ascorbic acid, and the numeric variable `dose` records the dosage of vitamin C in milligrams per day.

Data Subset Naming

For the table of hypothesis test results, the naming convention for each subset of data is as follows: if the name contains "OJ" or "VC," then the subset contains only observations for which the delivery method is orange juice or ascorbic acid, respectively; if the name contains "05," "10," or "20,", then the subset contains only observations for which the vitamin C dosage is 0.5, 1.0, or 2.0 mg/day, respectively. For example, the subset denoted "TGOJ" is the subset of data for which the delivery method is orange juice, and the subset denoted "TGVC10" is the subset of data for which the delivery method is ascorbic acid and the vitamin C dosage is 1.0 mg/day.

Code

What follows is the code required to produce each of the figures seen in the report.

The following chunk of code generates the side-by-side boxplot with overlaid jitter plot that show the original data split into different groups.

```
ggplot(data = ToothGrowth) +
    geom_boxplot(mapping = aes(x = factor(dose), y = len, fill = factor(dose))) +
    geom_jitter(mapping = aes(x = factor(dose), y = len),
                color = "purple", alpha = 0.5, height = 0) +
    facet_wrap(facets = vars(supp), nrow = 1,
               labeller = as_labeller(c(OJ = "Orange Juice (OJ)",
                                        VC = "Ascorbic Acid (VC)"))) +
    guides(fill = "none") +
    labs(x = "Dose of Vitamin C (mg/day)",
         y = "Length of Odontoblasts (units not given)",
         title = "Odontoblast Length Primarily Varies by Dosage of Vitamin C",
         subtitle = "Variation by supplement type much less pronounced")
```

The following chunk of code generates the first table, which shows summary statistics for each of the different groups and subgroups of the original data. Note that this requires the `knitr` chunk option `results = "asis"` to be rendered as seen above.

```
tab1 <- ToothGrowth %>% group_by(supp, dose) %>%
    summarize(mean = mean(len), median = median(len),
              sd = sd(len), iqr = IQR(len))
tab2 <- ToothGrowth %>% group_by(dose) %>%
    summarize(mean = mean(len), median = median(len),
              sd = sd(len), iqr = IQR(len))
tab3 <- ToothGrowth %>% group_by(supp) %>%
    summarize(mean = mean(len), median = median(len),
              sd = sd(len), iqr = IQR(len))
tab4 <- ToothGrowth %>%
    summarize(mean = mean(len), median = median(len),
              sd = sd(len), iqr = IQR(len))
```

```
tab5 <- bind_rows(tab4, tab3, tab2, tab1) %>%
    select(5:6, 1:4) %>%
    mutate(supp = str_replace_na(supp, "No Filter"),
           supp = str_replace(supp, "OJ", "Orange Juice (OJ)"),
           supp = str_replace(supp, "VC", "Ascorbic Acid (VC)"),
           dose = str_replace_na(dose, "No Filter"))

kab1 <- kable(tab5,
              digits = 2,
              col.names = c("Delivery Method",
                            "Dosage of Vitamin C",
                            "Mean",
                            "Median",
                            "Standard Deviation",
                            "Interquartile Range"),
              caption = paste("Summary Statistics for Odontoblast Length",
                              " Filtered by Delivery Method and Vitamin C",
                              " Dosage",
                              sep = "", collapse = ""))
kable_styling(kab1, latex_options = "HOLD_position")
```

The following chunk of code generates the second table, which shows P-values and significance identifiers for each of the 13 hypothesis tests conducted. Note that this requires the `knitr` chunk option `results = "asis"` to be rendered as seen above.

```
## set initial alpha
alpha <- 0.05
## construct subsetting vectors and subsetted data
dose <- ToothGrowth$dose
supp <- ToothGrowth$supp
TGOJ <- ToothGrowth[supp == "OJ", ]$len
TGVC <- ToothGrowth[supp == "VC", ]$len
TG05 <- ToothGrowth[dose == 0.5, ]$len
TG10 <- ToothGrowth[dose == 1.0, ]$len
TG20 <- ToothGrowth[dose == 2.0, ]$len
TGOJ05 <- ToothGrowth[supp == "OJ" & dose == 0.5, ]$len
TGOJ10 <- ToothGrowth[supp == "OJ" & dose == 1.0, ]$len
TGOJ20 <- ToothGrowth[supp == "OJ" & dose == 2.0, ]$len
TGVC05 <- ToothGrowth[supp == "VC" & dose == 0.5, ]$len
TGVC10 <- ToothGrowth[supp == "VC" & dose == 1.0, ]$len
TGVC20 <- ToothGrowth[supp == "VC" & dose == 2.0, ]$len

## overall: supp OJ vs supp VC
t01 <- t.test(TGOJ, TGVC, alternative = "two.sided", mu = 0, paired = FALSE,
              var.equal = FALSE, conf.level = 1 - alpha)
## each dose: supp OJ vs supp VC
t02 <- t.test(TGOJ05, TGVC05, alternative = "two.sided", mu = 0, paired = FALSE,
              var.equal = FALSE, conf.level = 1 - alpha)
t03 <- t.test(TGOJ10, TGVC10, alternative = "two.sided", mu = 0, paired = FALSE,
              var.equal = FALSE, conf.level = 1 - alpha)
t04 <- t.test(TGOJ20, TGVC20, alternative = "two.sided", mu = 0, paired = FALSE,
              var.equal = FALSE, conf.level = 1 - alpha)
## overall: each pair of doses
t05 <- t.test(TG05, TG10, alternative = "two.sided", mu = 0, paired = FALSE,
```

```
              var.equal = FALSE, conf.level = 1 - alpha)
t06 <- t.test(TG05, TG20, alternative = "two.sided", mu = 0, paired = FALSE,
              var.equal = FALSE, conf.level = 1 - alpha)
t07 <- t.test(TG10, TG20, alternative = "two.sided", mu = 0, paired = FALSE,
              var.equal = FALSE, conf.level = 1 - alpha)
## supp OJ: each pair of doses
t08 <- t.test(TGOJ05, TGOJ10, alternative = "two.sided", mu = 0, paired = FALSE,
              var.equal = FALSE, conf.level = 1 - alpha)
t09 <- t.test(TGOJ05, TGOJ20, alternative = "two.sided", mu = 0, paired = FALSE,
              var.equal = FALSE, conf.level = 1 - alpha)
t10 <- t.test(TGOJ10, TGOJ20, alternative = "two.sided", mu = 0, paired = FALSE,
              var.equal = FALSE, conf.level = 1 - alpha)
## supp VC: each pair of doses
t11 <- t.test(TGVC05, TGVC10, alternative = "two.sided", mu = 0, paired = FALSE,
              var.equal = FALSE, conf.level = 1 - alpha)
t12 <- t.test(TGVC05, TGVC20, alternative = "two.sided", mu = 0, paired = FALSE,
              var.equal = FALSE, conf.level = 1 - alpha)
t13 <- t.test(TGVC10, TGVC20, alternative = "two.sided", mu = 0, paired = FALSE,
              var.equal = FALSE, conf.level = 1 - alpha)

# full results with test statistics, confidence intervals and three
# significance identifiers
ttests <- mget(paste("t", c(rep("0", 9), rep("", 4)), 1:13, sep = ""))
ttestresults <- data.frame(data = sapply(ttests, function(x) x$data.name),
                           tstat = sapply(ttests, function(x) x$statistic),
                           tdf = sapply(ttests, function(x) x$parameter),
                           pval = sapply(ttests, function(x) x$p.value),
                           cilow = sapply(ttests, function(x) x$conf.int[1]),
                           cihigh = sapply(ttests, function(x) x$conf.int[2]),
                           alpha = alpha)
ttestresults <- ttestresults %>%
    mutate(alpha_fwer = alpha / length(pval),
           alpha_fdr = alpha * rank(pval)/length(pval),
           sig_alpha = pval < alpha,
           sig_fwer = pval < alpha_fwer,
           sig_fdr = pval < alpha_fdr)
# filtered results with Bonferroni corrected values
ttestresults2 <- ttestresults %>%
    select(data, pval, sig_fwer)

kab2 <- kable(ttestresults2,
              digits = 4,
              col.names = c("Data Subsets",
                            "P-value",
                            "Significant (Bonferroni)"),
              caption = paste("Significance of Independent Two-Sample T Tests",
                              " for 13 Pairs of Subsets of Data Using",
                              " Bonferroni Correction for Multiple Tests",
                              sep = "", collapse = ""))
kable_styling(kab2, latex_options = "HOLD_position")
```