An Analysis of the ToothGrowth data in R

Overview

In this report, we perform simple analyses of the ToothGrowth data in the R datasets package. After exploring the data, we propose two hypothesis about differences in growth levels between different supplement-dose combinations, and then design and perform appropriate tests to validate these claims. We've found consistent improvement in growth with increasing dosage with both supplements and much more effectiveness of vitamin C via orange juice at lower dose levels.

Loading and Processing the Data

First, the dataaset is made available for processing and basic information is obtained:

```
library(datasets); data("ToothGrowth"); str(ToothGrowth)
```

```
## '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 dataset has 60 observations of 3 variables. *len* is the response, and reports 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, reported by the *dose* variable) by one of two delivery methods (variable *supp*): orange juice, coded as "OJ" and ascorbic acid, coded as "VC". For the purpose of the following anlysis, it is convenient to tranform the *dose* variable from numerical to factor, as it encodes 3 levels:

```
ToothGrowth$dose <- factor(ToothGrowth$dose)
```

There's no missing value so there's no need for some imputation strategy to further transform the data.

```
colSums(is.na(ToothGrowth))
```

```
##  len supp dose
##    0    0    0
```

Exploring the Data

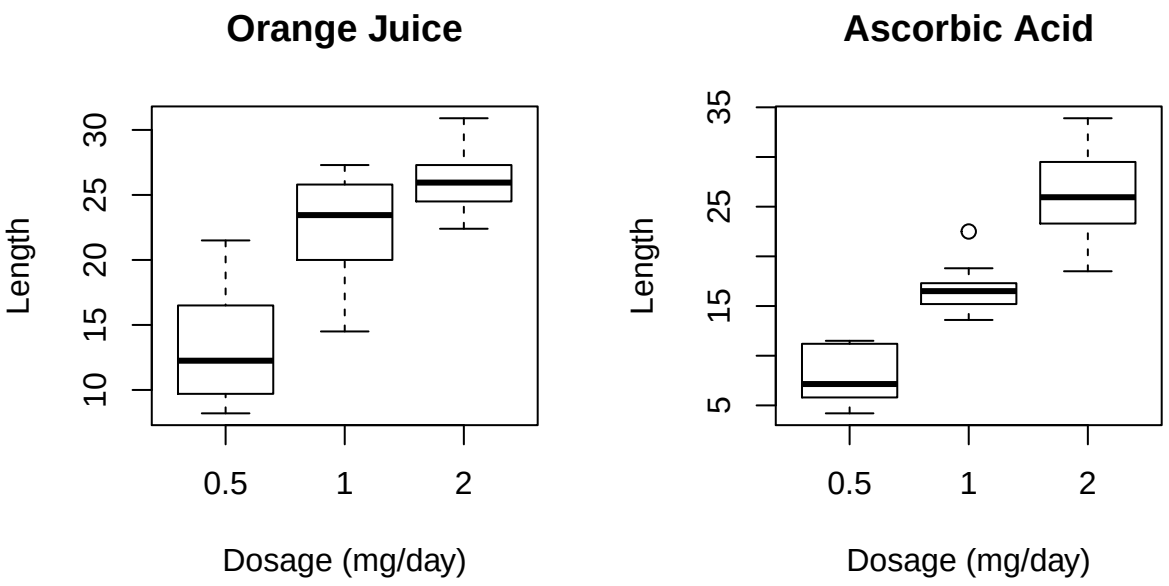
The subjects are split into equally sized supplement-dosage subgroups:

```
table(ToothGrowth$supp, ToothGrowth$dose)
```

```
##
##      0.5  1  2
##   OJ  10 10 10
##   VC  10 10 10
```

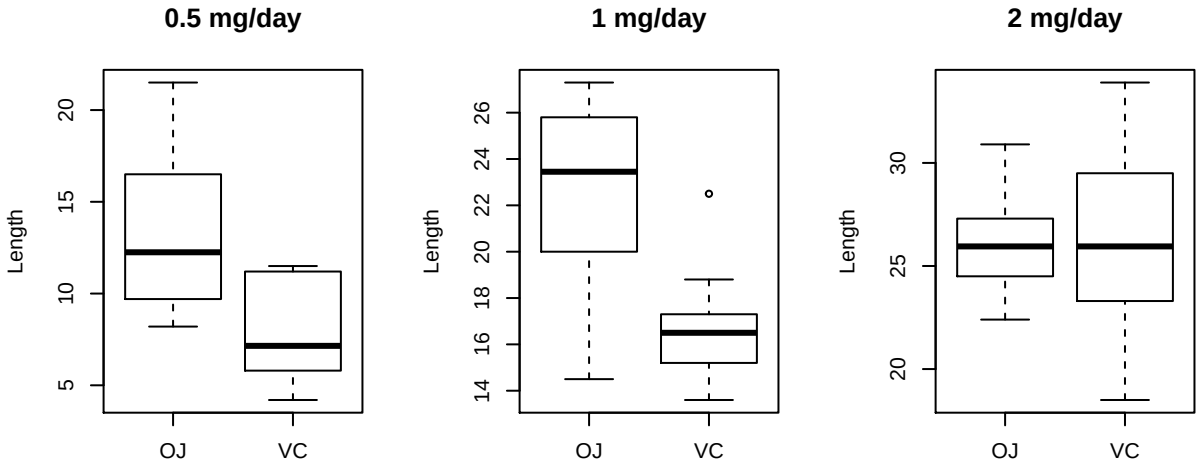
Let's first compare the response levels in terms of dosage in each supplement group:

```
par(mfrow = c(1,2))
with(subset(ToothGrowth, supp=="OJ"),
     boxplot(len ~ dose, main="Orange Juice", xlab="Dosage (mg/day)", ylab="Length"))
with(subset(ToothGrowth, supp == "VC"),
     boxplot(len ~ dose, main="Ascorbic Acid", xlab="Dosage (mg/day)", ylab="Length"))
```



The subjects given orange juice show noticeable growth improvements by doubling the 0.5 mg/day dosage, but not the 1 mg/day one. There's overlap in the range of data between each pair of successive levels. On the other hand, growth in the subjects under ascorbic acid seems to consistently improve by increasing the dosage. Secondly, we compare the two delivery methods at the same levels of dosage:

```
par(mfrow = c(1,3))
with(subset(ToothGrowth,dose=="0.5"),boxplot(len~supp, main="0.5 mg/day", ylab="Length"))
with(subset(ToothGrowth, dose=="1"), boxplot(len ~ supp, main="1 mg/day", ylab="Length"))
with(subset(ToothGrowth, dose=="2"), boxplot(len ~ supp, main="2 mg/day", ylab="Length"))
```



From the above plots, it appears there's no significant difference between orange juice and ascorbic acid at the highest dosage. At the lower levels, orange juice subjects show more noticeable growth.

Data Analyses

Based on the on the exploratory analyses done before, we tentatively make two statements/hypotheses:

1. increasing dosage consistenly brings more growth, but not for orange juice;
2. lower levels of dosage are more effective if the subjects are given orange juice.

We validate the above statements with two groups of t-tests (sample sizes are small), for which we need to test the assumption of normality of the data. We perform a shapiro test on each *supplement-dosage* subgroup, as shown in the Appendix A.1, and fail to reject the hypothesis of normality in each case.

To validate the above statements, we perform two-sample (independent groups) t-test assuming different variances. For the first statement, we compare all possible combinations of dosages for each supplement:

```
tests_by_supp <- list()
for (supplement in levels(ToothGrowth$supp)) {
  supp_subset <- ToothGrowth %>% filter(supp == supplement) %>%
    mutate(id = rep(1:10, 3)) %>% spread(id, len) %>% select(-supp)
  combs <- combn(1:nrow(supp_subset),2)
  tests_results <- data.frame()
  for(j in 1:ncol(combs)) {
    t_test <- t.test(supp_subset[combs[,j][1],2:11], supp_subset[combs[,j][2],2:11],
                     paired = FALSE, var.equal = FALSE, alternative = "less")
    tests_results <- rbind(tests_results, data.frame(doses =
                                                       paste(supp_subset[combs[,j][1],1],
                                                             supp_subset[combs[,j][2],1],
                                                             sep = "-"),
                                                     t = t_test$statistic,
                                                     p.value = t_test$p.value))
  }
  tests_by_supp[[supplement]] = tests_results
}
```

Details of the test results are shown in Appendix A.2. Although less significant, there's still some notable difference between the 1.0 and 2.0 doses with orange juice, so that the first statement cannot be supported.

In order to validate the second statement, we compare Orange Juice and Ascorbic Acid subgroups by dosage:

```
tests_by_dose <- data.frame()
for (d in levels(ToothGrowth$dose)) {
  t_test <- t.test(len ~ I(relevel(supp, "OJ")), paired = FALSE, var.equal = FALSE,
                   data = subset(ToothGrowth, dose == d), alternative = "greater")
  tests_by_dose <- rbind(tests_by_dose, data.frame(dose = d, t = t_test$statistic,
                                                   p.value = t_test$p.value))
}
```

Details of these test results can be found in Appendix A.3. As expected, lower level of dosage with orange juice are conducive of much more growth than with ascorbic acid, and there's no real difference at the 2.0 mg/day dose.

Results

We've consistently detected more growth by doubling dosages for both orange juice and ascorbic acid, although results seems to be less supported for orange juice. More tests with larger sample size are probably needed. At lower levels of dosage (i.e. $[0.5 - 1.0]$ mg/day), vitamin C administered via orange juice is much more effective than Ascorbic Acid and produces more growth. At the highest level (i.e. 2 mg/day), vitamin C can either be administered through orange juice or ascorbic acid.

Appendix

A.1. Testing the assumptions for the t-tests

Based on the nature of the tests, we split the data into the six possible *supplement-dosage* subgroups and perform a shapiro test of normality in each one of them and collecting the p-values:

```
df <- ToothGrowth %>% mutate(id = rep(1:10, 6)) %>% spread(id, len)
df
```

```
##   supp dose    1    2    3    4    5    6    7    8    9   10
## 1   OJ  0.5 15.2 21.5 17.6  9.7 14.5 10.0  8.2  9.4 16.5  9.7
## 2   OJ    1 19.7 23.3 23.6 26.4 20.0 25.2 25.8 21.2 14.5 27.3
## 3   OJ    2 25.5 26.4 22.4 24.5 24.8 30.9 26.4 27.3 29.4 23.0
## 4   VC  0.5  4.2 11.5  7.3  5.8  6.4 10.0 11.2 11.2  5.2  7.0
## 5   VC    1 16.5 16.5 15.2 17.3 22.5 17.3 13.6 14.5 18.8 15.5
## 6   VC    2 23.6 18.5 33.9 25.5 26.4 32.5 26.7 21.5 23.3 29.5
```

```
norm_tests <- sapply(apply(df[,3:12], 1, shapiro.test), function(elt) elt$p.value)
names(norm_tests) <- c("OJ-0.5","OJ-1","OJ-2","VC-0.5","VC-1","VC-2")
norm_tests
```

```
##    OJ-0.5      OJ-1      OJ-2    VC-0.5      VC-1      VC-2
## 0.1820408 0.4152983 0.8147908 0.1695627 0.2697855 0.9194497
```

A.2. Comparing dosages for each supplement

For each supplement, we report the results of comparing growth in all possible dose subgroups. In each casse, we test the alternative hypothesis that the mean of the lower level dose group is less the higher one. For each combination, we report the test statistic and the p-value:

```
## $OJ
##    doses         t      p.value
## t  0.5-1 -5.048635 4.392460e-05
## t1 0.5-2 -7.817021 6.618919e-07
## t2   1-2 -2.247761 1.959757e-02
##
## $VC
##    doses          t      p.value
## t  0.5-1  -7.463430 3.405509e-07
## t1 0.5-2 -10.387795 2.340789e-08
## t2   1-2  -5.469814 4.577802e-05
```

A.3. Comparing supplements for each dose level

For each dose level, we report the results of comparing growth between the Orange Juice and Ascorbic Acid subgroups. In each case, we test the alternative hypothesis that vitamin C through Orange Juice bring more growth than Ascorbic Acid:

```
##    dose          t      p.value
## t   0.5  3.1697328 0.0031793034
## t1    1  4.0327696 0.0005191879
## t2    2 -0.0461361 0.5180742056
```