Exploring ToothGrowth data with R

adatum

2015-06-21

Overview

The `ToothGrowth` dataset in R contains results from experiments on the effects of varying doses (0.5 mg, 1 mg, 2 mg) of Vitamin C delivered by two different methods, orange juice (OJ) or ascorbic acid (VC), on the length of odontoblast cells (in micrometers) of guinea pigs' teeth.[1] Each of the six cases includes 10 specimens, for a total of 60 guinea pigs. We will examine the dataset and conduct hypothesis tests on the effectiveness of the delivery methods.

Exploratory Data Analysis

First we load the libraries needed for the analysis and take a look at the structure and summary information of the dataset, confirming that there are data for 60 guinea pigs and that the ranges of values make sense for the variables .

```
library(dplyr)
library(ggplot2)
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 ...
```

```
summary(ToothGrowth)
```

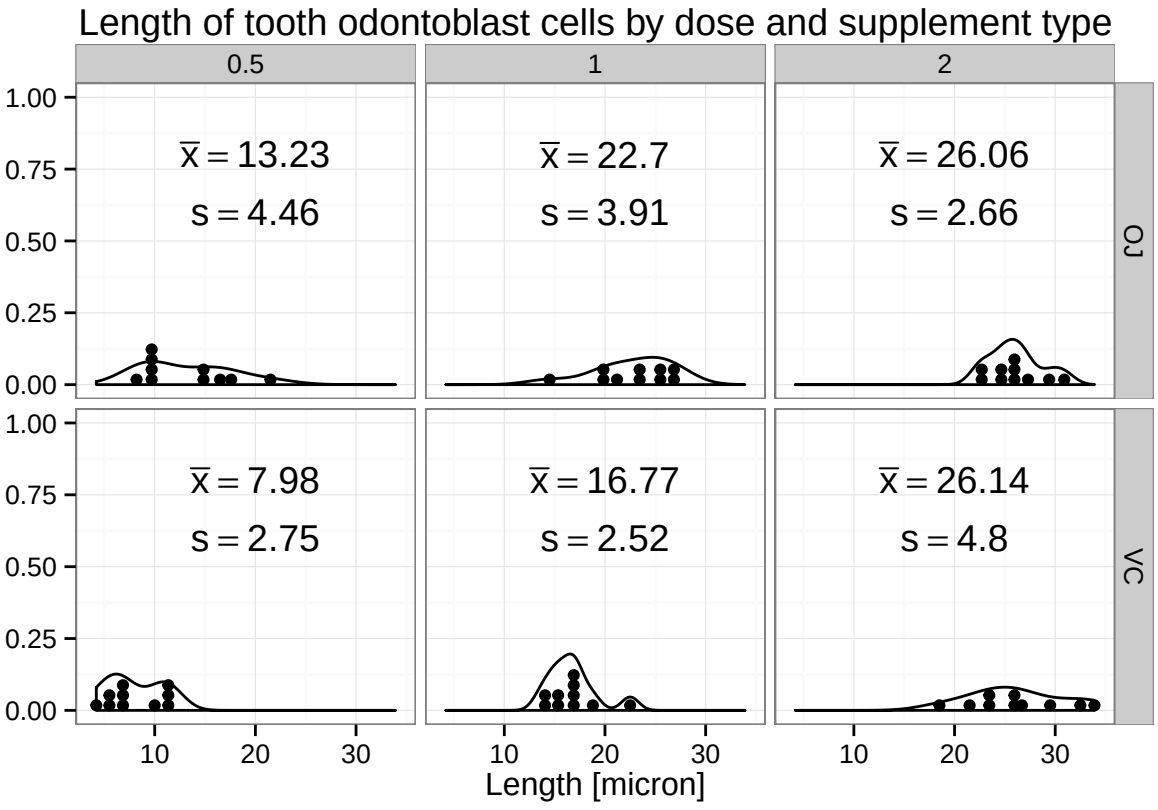
```
##       len        supp         dose
##  Min.   : 4.20   OJ:30   Min.   :0.500
##  1st Qu.:13.07   VC:30   1st Qu.:0.500
##  Median :19.25           Median :1.000
##  Mean   :18.81           Mean   :1.167
##  3rd Qu.:25.27           3rd Qu.:2.000
##  Max.   :33.90           Max.   :2.000
```

Before we can do hypothesis testing, we must verify whether the data meet criteria for the tests. By plotting histograms of each of the six cases in the dataset we can visually inspect the spread and calculate some summary statistics.

[1] Crampton, E. W., "The Growth of the Odontoblasts of the Incisor Tooth as a Criterion of the Vitamin C Intake of the Guinea Pig," The Journal of Nutrition, 33 (5): 491-504, 1947.

```r
tg <- ToothGrowth %>%
        group_by(dose, supp) %>%
        summarise(len.sd = round(sd(len), 2), len.mean = round(mean(len), 2))

ggplot(data = ToothGrowth, aes(x = len)) +
        geom_dotplot(binwidth = 1) +
        geom_density() +
        facet_grid(supp ~ dose) +
        geom_text(x = 20, y = 0.8,
                  data = tg,
                  aes(label = paste0("bar(x) == ", len.mean)),
                  parse = T
                  ) +
        geom_text(x = 20, y = 0.6,
                  data = tg,
                  aes(label = paste0("s == ", len.sd)),
                  parse = T
                  ) +
        labs(title = "Length of tooth odontoblast cells by dose and supplement type",
             x = "Length [micron]",
             y = ""
             ) +
        theme_bw()
```



With only 10 data points per case, it is difficult to establish whether the data are normally distributed. However, we can see that the standard deviation, and thus variance, is not necessarily similar in all cases. Therefore, we cannot assume constant variance in our tests.

Hypothesis Testing

We will use an independent group t-test, with variance not assumed to be similar, to test the hypothesis of whether orange juice is more effective than ascorbic acid supplements. The groups are independent, since each case of the experiment was randomly assigned to a different group of 10 guinea pigs. The t-test is used, since the number of data points is small.

```
with(ToothGrowth, t.test(ToothGrowth[supp == "OJ" & dose == 0.5, "len"], ToothGrowth[supp ==
    "VC" & dose == 0.5, "len"], alternative = "greater")[c("conf.int", "p.value")])
```

```
## $conf.int
## [1] 2.34604     Inf
## attr(,"conf.level")
## [1] 0.95
##
## $p.value
## [1] 0.003179303
```

```
with(ToothGrowth, t.test(ToothGrowth[supp == "OJ" & dose == 1, "len"], ToothGrowth[supp ==
    "VC" & dose == 1, "len"], alternative = "greater")[c("conf.int", "p.value")])
```

```
## $conf.int
## [1] 3.356158      Inf
## attr(,"conf.level")
## [1] 0.95
##
## $p.value
## [1] 0.0005191879
```

```
with(ToothGrowth, t.test(ToothGrowth[supp == "OJ" & dose == 2, "len"], ToothGrowth[supp ==
    "VC" & dose == 2, "len"], alternative = "greater")[c("conf.int", "p.value")])
```

```
## $conf.int
## [1] -3.1335     Inf
## attr(,"conf.level")
## [1] 0.95
##
## $p.value
## [1] 0.5180742
```

For the 0.5 mg and 1 mg doses, we reject the null hypothesis that the difference between the mean cell lengths for OJ and VC cases is zero. The p-values are both < 0.05 and the 95% confidence intervals do not contain zero, therefore there is a statistically significant increase in effectiveness in Vitamin C delivery by orange juice compared to ascorbic acid at these doses.

However, for a dose of 2 mg, we cannot reject the null hypothesis since the p-value is > 0.05 and the confidence interval includes zero, hence there is not a statistically significant difference between the mean cell lengths for the different Vitamin C delivery methods at the 95% confidence level.

In performing these t-tests, we have assumed that the data are reasonably normal. However, we cannot be confident of this due to the small number of data points.