Exploratory analysis of the ToothGrowth data in R

Mykyta Zharov

11/1/2019

Overview

In this project I am going to analyze the ToothGrowth data from the R datasets package, which has information about the effect of vitamin C on tooth growth for Guinea pigs. I am going to perform basic exploratory data analysis, provide a basic summary of the data, use confidence intervals and hypothesis tests to compare tooth growth by supplement type and dosage level and state the conclusions.

Exploratory data analysis

Let us start by loading the data and printing out a few first rows.

```
# load data
data(ToothGrowth)
head(ToothGrowth)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

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). Let us explore the structure of the data set.

```
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 ...
```

From the above output we see that the data set contains 60 observations of 3 variables. The variables `len` and `dose` are numerical, a variable `supp` is categorical with two factors "OJ" and "VC", which stand for orange juice and ascorbic acid respectively. To better understand the dataset let us look at the data summary object.

```
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
```

From the summary we observe that 30 guinea pigs were given orange juice and 30 guinea pigs were given ascorbic acid. Also we see that the length of odontoblasts has minimal value 4.20, maximal value 33.90 and

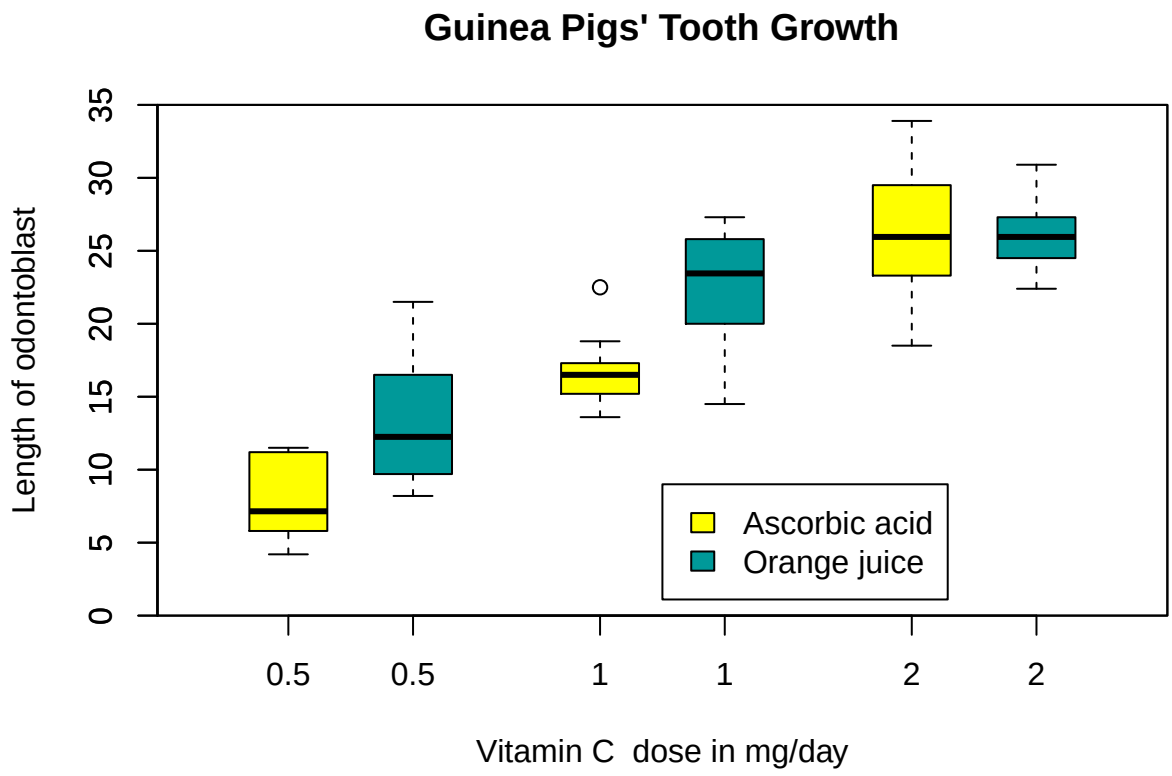
mean value is 18.81, and median 19.25. The dose variable has minimal value of 0.5, maximal value of 2.0 and mean value 1.167.

```
unique(ToothGrowth$dose)
```

```
## [1] 0.5 1.0 2.0
```

We also observe tht there are only three possible values for the dose variable, 0.5, 1.0 and 2.0. Let us visualise the data using a boxplot. We plot the length of a odontoblast for different levels of supplement type and values of dose.

```
boxplot(len ~ dose, data = ToothGrowth,
        boxwex = 0.25, at = 1:3 - 0.2,
        subset = supp == "VC", col = "yellow",
        main = "Guinea Pigs' Tooth Growth",
        xlab = "Vitamin C  dose in mg/day",
        ylab = "Length of odontoblast",
        xlim = c(0.5, 3.5), ylim = c(0, 35), yaxs = "i")
boxplot(len ~ dose, data = ToothGrowth, add = TRUE,
        boxwex = 0.25, at = 1:3 + 0.2,
        subset = supp == "OJ", col = "#009999")
legend(2, 9, c("Ascorbic acid", "Orange juice"),
       fill = c("yellow", "#009999"))
```



From the boxplot above we clearly see that the increase of the dosage of vitamin C contribute to increase of the length of odontoblasts. We observe that the trend is linear for the case of ascorbic acid. We also see that for dosage 0.5mg and 1.0mg of vitamin C the values of length of odontoblasts for pigues taking orange juice are higher than for pigues taking ascorbic acid and for dosage of 2.0mg the values of the length of odontoblasts for orange juice and for ascorbic acid are similar. The plot also shows that the higher dosage of 2.0mg has less improvement in the length of odontoblast when orange juice supplement is used.

Hypothesis tests

In this section we are going to perform hypothesis tests to compare tooth length by supplement and dosage.

Compare tooth length by supplement type

We are going to compare the tooth length by supplement using the Welch Two Sample t-test. Welch's t-test, or unequal variances t-test, is a two-sample location test which is used to test the hypothesis that two populations have equal means. The test has following assumptions:

- The variables are independent and identically distributed.
- Both populations have normal distributions with unequal variances.

Let the null hypothesis be H_0: the mean lengths of odontoblast for both supplements (ascorbic acid and orange juice) are the same. The alternative hypothesis will be H_1: the mean length of odontoblast is greater for supplement orange juice than for ascorbic acid.

Let us split the data into two samples by supplement type.

```
OJ = ToothGrowth$len[ToothGrowth$supp == 'OJ']
VC = ToothGrowth$len[ToothGrowth$supp == 'VC']
```

Now we perform a one-tailed Welch Two Sample t-test with a confidence level 95% as follows:

```
t.test(x=OJ, y=VC, alternative = "greater", paired = FALSE, var.equal = FALSE,
       conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  OJ and VC
## t = 1.9153, df = 55.309, p-value = 0.03032
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.4682687       Inf
## sample estimates:
## mean of x mean of y
##  20.66333  16.96333
```

We see from the output that the p-value of the test 0.03032 is lower than 0.05, and we reject the null hypothesis H_0. We can conclude that it is very likely that the orange juice supplement has greater effect on the length of odontoblast than the ascorbic acid supplement.

Compare tooth length by dosage level

Let the null hypothesis be H_0: the mean lengths of odontoblast for both dosage levels (0.5 and 1.0) are the same. The alternative hypothesis will be H_1: the mean length of odontoblast is less for dosage level 0.5 than for dosage level 1.0.

Let us split the data into three samples by dosage level.

```
Dose05 = ToothGrowth$len[ToothGrowth$dose == 0.5]
Dose10 = ToothGrowth$len[ToothGrowth$dose == 1.0]
Dose20 = ToothGrowth$len[ToothGrowth$dose == 2.0]
```

Now we perform a one-tailed Welch Two Sample t-test with a confidence level 95% as follows:

```
t.test(x=Dose05, y=Dose10, alternative = "less", paired = FALSE, var.equal = FALSE,
       conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  Dose05 and Dose10
## t = -6.4766, df = 37.986, p-value = 6.342e-08
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -6.753323
## sample estimates:
## mean of x mean of y
##    10.605    19.735
```

We see from the output that the p-value of the test 6.342e-08 is lower than 0.05, and we reject the null hypothesis H_0. We can conclude that it is very likely that the dosage 1.0 has greater effect on the length of odontoblast than the dosage 0.5.

Let the null hypothesis be H_0: the mean lengths of odontoblast for both dosage levels (1.0 and 2.0) are the same. The alternative hypothesis will be H_1: the mean length of odontoblast is less for dosage level 1.0 than for dosage level 2.0.

```
t.test(x=Dose10, y=Dose20, alternative = "less", paired = FALSE, var.equal = FALSE,
       conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  Dose10 and Dose20
## t = -4.9005, df = 37.101, p-value = 9.532e-06
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf -4.17387
## sample estimates:
## mean of x mean of y
##    19.735    26.100
```

We see from the output that the p-value of the test 9.532e-06 is lower than 0.05, and we reject the null hypothesis H_0. We can conclude that it is very likely that the dosage 2.0 has greater effect on the length of odontoblast than the dosage 1.0.

Compare tooth length by supplement for dosage level 2.0

In previous section we have seen with the boxplot that the mean values for the tooth length lie close to each other for both supplement types for dosage level 2.0. Using a hypothesis test, we want to check if the mean values for both samples are identical.

Let the null hypothesis be H_0: the mean lengths of odontoblast for both supplement types (ascorbic acid and orange juice) for dosage level 2.0 are the same. The alternative hypothesis will be H_1: the mean length of odontoblast for both supplement types (ascorbic acid and orange juice) for dosage level 2.0 are not the same.

Define the samples.

```
Dose20OJ=ToothGrowth$len[ToothGrowth$dose == 2.0 & ToothGrowth$supp == 'OJ']
Dose20VC=ToothGrowth$len[ToothGrowth$dose == 2.0 & ToothGrowth$supp == 'VC']
```

```r
t.test(x=Dose20OJ, y=Dose20VC, alternative = "two.sided", paired = FALSE,
       var.equal = FALSE, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  Dose20OJ and Dose20VC
## t = -0.046136, df = 14.04, p-value = 0.9639
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.79807  3.63807
## sample estimates:
## mean of x mean of y
##     26.06     26.14
```

We see from the output that the p-value of the test 0.9639 is greater than 0.05, and we do not reject the null hypothesis H_0. We can conclude that there is not sufficient evidence to show that there is a difference in length of odontoblast when using supplement “OJ” and “VC” at dosage level 2.0mg.