Statistical Inference Course Project Part 2

Katya Demidova

26 December 2015

Overview

In this document we're going to use the ToothGrowth data from the R `datasets` package. We will analyse if the response is the length of odontoblasts (teeth) in each group of 10 guinea pigs is different:

- at each of three dose levels of Vitamin C (0.5, 1, and 2 mg),
- with each of two delivery methods (orange juice or ascorbic acid).

Load the Data and Basic Analysis

First, we will load the data and perform basic exploration:

```
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
```

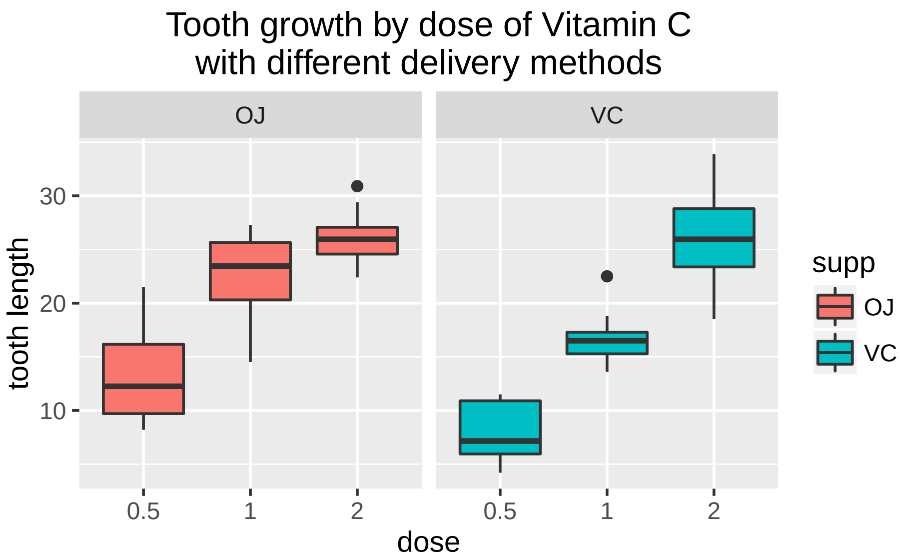
```
summary(ToothGrowth)
```

```
##       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
```

```
ToothGrowth$dose <- as.factor(ToothGrowth$dose) # coerce 'dose' to factor
```

The ToothGrowth data has 60 observations of 3 variables:

- `len`: tooth length (numeric),
- `supp`: supplement type / delivery method (factor with 2 levels `OJ` and `VC`, orange juice and ascorbic acid respectively),
- `dose`: dose of Vitamin C in miligrams. This variable is factorised for convenience.



From the plot above, it is clear that there is a positive correlation between dose level and tooth length in guinea pigs. We might also make an assumption that orange juice causes larger tooth growth than ascorbic acid.

Statistical Inference

Tooth Growth by Dose Level

As mentioned above, there seems to be a positive correlation between dose level and tooth length in guinea pigs. To compare tooth growth by `dose` and check if there is a statistically significant difference between the means, we will use t-tests (full results can be seen in Appendix).

```
# make three subsets
dose05or1 <- subset(ToothGrowth, dose == "0.5" | dose == "1")
dose1or2 <- subset(ToothGrowth, dose == "1" | dose == "2")
dose05or2 <- subset(ToothGrowth, dose == "0.5" | dose == "2")
```

```
t.test(len ~ dose, data = dose05or2)
t.test(len ~ dose, data = dose1or2)
t.test(len ~ dose, data = dose05or2)
```

P-values and confidence intervals of these t-tests:

```
##   compared.doses     p.values conf.int.lower conf.int.higher
## 1      0.5 and 1 4.397525e-14        -18.156         -12.834
## 2        1 and 2 1.906430e-05         -8.996          -3.734
## 3      0.5 and 2 4.397525e-14        -18.156         -12.834
```

From these numbers, we can reject all of three null hypotheses and make a conclusion that **dosage of Vitamin C affects tooth growth in guinea pigs**.

Tooth growth by supplement type

We have made an assumption that that orange juice causes larger tooth growth than ascorbic acid. To compare tooth growth by `supp` and check if there is a statistically significant difference between the means, we will run a t-test with the null hypothesis *H0*: mean tooth growth in guinea pigs getting Vitamin C from orange juice is the same as mean tooth growth in guinea pigs getting Vitamin C from ascorbic acid:

```
t.test(len ~ supp, data = ToothGrowth)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1710156  7.5710156
## sample estimates:
## mean in group OJ mean in group VC
##         20.66333         16.96333
```

The resulting p-value is `0.0606345`. Though it is very close to `0.05` (most common alpha-level), the difference between the means is **not** significant. Furthermore, the 95% confidence interval contains 0. We cannot reject the null hypothesis and can conclude that **delivery method doesn't affect tooth growth in guinea pigs**.

Tooth growth by supplement type and dosage

We will compute means and standard deviations for 6 groups:

```
ToothGrowth %>% group_by(supp, dose) %>% summarise(mean(len), sd(len))
```

```
## Source: local data frame [6 x 4]
## Groups: supp [?]
##
##     supp   dose mean(len)  sd(len)
##   (fctr) (fctr)     (dbl)    (dbl)
## 1     OJ    0.5     13.23 4.459709
## 2     OJ      1     22.70 3.910953
## 3     OJ      2     26.06 2.655058
## 4     VC    0.5      7.98 2.746634
## 5     VC      1     16.77 2.515309
## 6     VC      2     26.14 4.797731
```

Next, we will compare tooth growth by delivery method (`supp`) in groups with different dosage:

```
# make three subsets
dose05 <- subset(ToothGrowth, dose == "0.5")
dose1 <- subset(ToothGrowth, dose == "1")
dose2 <- subset(ToothGrowth, dose == "2")
```

```r
t.test(len ~ supp, data = dose05)
t.test(len ~ supp, data = dose1)
t.test(len ~ supp, data = dose2)
```

P-values and confidence intervals of these t-tests:

```
##   doses p.values conf.int.lower conf.int.higher
## 1   0.5  0.00636          1.719           8.781
## 2     1  0.00104          2.802           9.058
## 3     2  0.96385         -3.798           3.638
```

As we can see, delivery method **does affect** tooth growth in guinea pigs when dosage is 0.5 or 1, with **orange juice causing more tooth growth**. However, **when dosage is 2 miligrams, orange juice does not significantly differ from ascorbic acid**.

Assumptions and Conclusions

Assuming that (1) delivery method and dosage are randomly assigned, (2) population data is normally distributed, and (3) sample group is representative for the population, we can make several conclusions:

- higher doses of Vitamin C lead to larger tooth length in guinea pigs;
- orange juice is a more successful delivery method (comparing to ascorbic acid) when dosage os either 0.5 or 1 mg;
- when dosage is 2 mg, there is no significant difference between orange juice and ascorbic acid.

In other words, if we needed our guinea pig to have long teeth, we could choose any delivery method and dosage of 2 mg.

Appendix

Hidden Code Chunks

```r
library(ggplot2)
library(dplyr)
options(scipen=2)
```

```r
library(datasets)
data(ToothGrowth)
```

```r
ggplot(data = ToothGrowth, aes(dose, len)) + geom_boxplot(aes(fill = supp)) + ggtitle("Tooth growth by 
```

Section Tooth Growth by Dose Level

```r
t.test(len ~ dose, data = dose05or2)
t.test(len ~ dose, data = dose1or2)
t.test(len ~ dose, data = dose05or2)
```

```r
compared.doses <- c("0.5 and 1", "1 and 2", "0.5 and 2")

p.values <- c(t.test(len ~ dose, data = dose05or2)$p.value,
              t.test(len ~ dose, data = dose1or2)$p.value,
              t.test(len ~ dose, data = dose05or2)$p.value)

conf.int.lower <- round(c(t.test(len ~ dose, data = dose05or2)$conf.int[1],
                    t.test(len ~ dose, data = dose1or2)$conf.int[1],
                    t.test(len ~ dose, data = dose05or2)$conf.int[1]), 3)

conf.int.higher <- round(c(t.test(len ~ dose, data = dose05or2)$conf.int[2],
                    t.test(len ~ dose, data = dose1or2)$conf.int[2],
                    t.test(len ~ dose, data = dose05or2)$conf.int[2]), 3)

data.frame(compared.doses, p.values, conf.int.lower, conf.int.higher)
```

Section Tooth Growth by Dose Level and Supplement Type

```r
t.test(len ~ supp, data = dose05)
t.test(len ~ supp, data = dose1)
t.test(len ~ supp, data = dose2)
```

```r
doses <- c("0.5", "1", "2")

p.values <- round(c(t.test(len ~ supp, data = dose05)$p.value,
              t.test(len ~ supp, data = dose1)$p.value,
              t.test(len ~ supp, data = dose2)$p.value), 5)

conf.int.lower <- round(c(t.test(len ~ supp, data = dose05)$conf.int[1],
                    t.test(len ~ supp, data = dose1)$conf.int[1],
                    t.test(len ~ supp, data = dose2)$conf.int[1]), 3)

conf.int.higher <- round(c(t.test(len ~ supp, data = dose05)$conf.int[2],
                    t.test(len ~ supp, data = dose1)$conf.int[2],
                    t.test(len ~ supp, data = dose2)$conf.int[2]), 3)

data.frame(doses, p.values, conf.int.lower, conf.int.higher)
```