Sample-based analysis of exponential distribution

Michael Krämer

25.10.2015

Abstract

The report shows the result of a basic analysis of the ToothGrowth dataset from the library 'datasets. It has the title "The Effect of Vitamin C on Tooth Growth in Guinea Pigs".

General data description

The data covers 3 variables per observation:

- len (numeric Tooth length)
- supp (Supplement type (VC or OJ) meaning ascorbic acid or orange juice)
- dose (vitamin dose in milligrams)

```
summary(tooth)
```

```
##       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
```

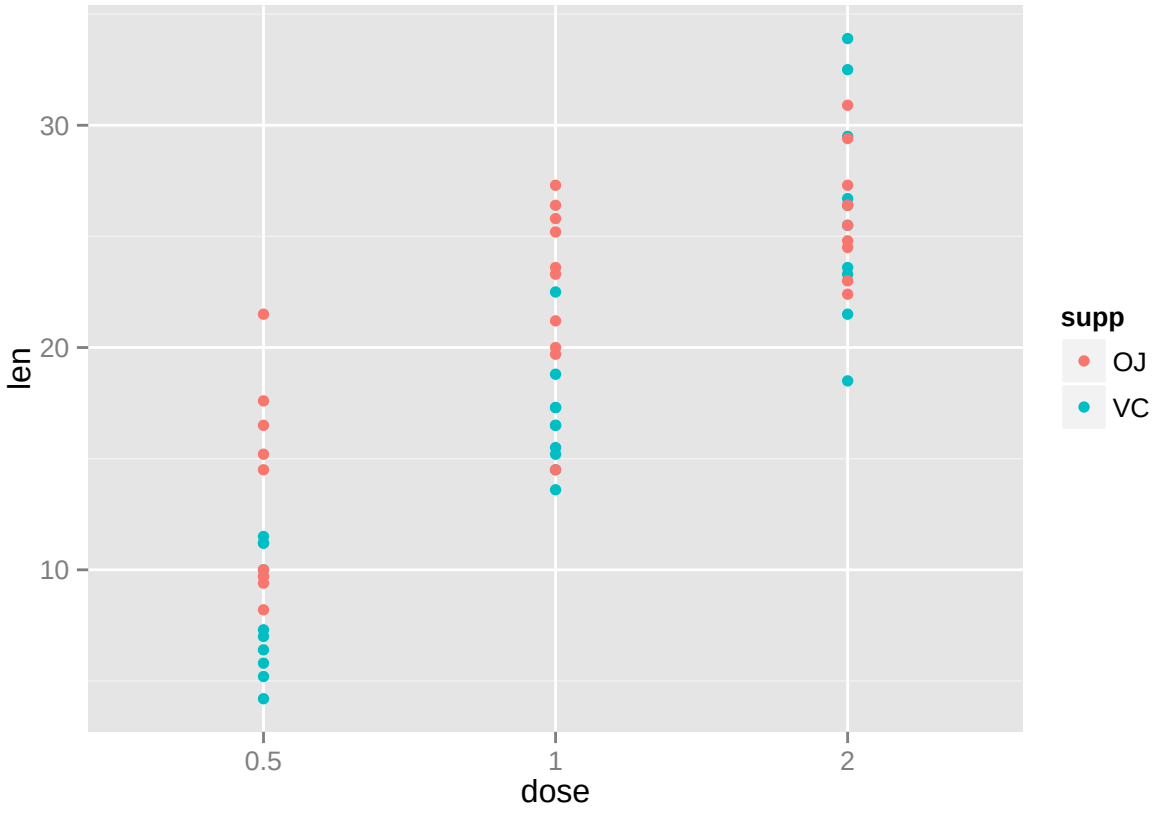
```
summary(factor(tooth$dose))
```

```
## 0.5   1   2
##  20  20  20
```

The dataset covers 60 observations, where 30 cover each supplement type. It were 3 doses tested (0.5, 1 and 2 mg) and of each dose exist 10 observations per supplement type which makes up the 60 overall.

Since there is neither a time information given nor identification for individuals in the test group, the observations will not be interpreted as paired. To give a rough idea about the spread of the data, the following plot is provided.

Exploratory analysis



To get a better numerical overview of the existing 6 groups, let's have a look at the mean and the standard deviation of the tooth length in each group.

```
data.frame(
    tooth %>%
    group_by(supp, dose) %>%
    summarise(mu = mean(len), sigma = sd(len)))
```

```
##   supp dose    mu    sigma
## 1   OJ  0.5 13.23 4.459709
## 2   OJ  1.0 22.70 3.910953
## 3   OJ  2.0 26.06 2.655058
## 4   VC  0.5  7.98 2.746634
## 5   VC  1.0 16.77 2.515309
## 6   VC  2.0 26.14 4.797731
```

Basically the mean of the tooth length seems to be higher with higher doses for both supplement types. Interestingly, the tables shows that the standard deviation decreases with higher doses for the OJ observations while it increases for the VC observations.

Hypothesis testing

To examine the assumptions from the summary and to figure out if the increase in tooth length by the dose has statistical significance, a t test is used.

```r
oj05 <- tooth %>% filter(supp == "OJ", dose == 0.5) %>% select(len)
oj2 <- tooth %>% filter(supp == "OJ", dose == 2.0) %>% select(len)
t.test(oj2, oj05, paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  oj2 and oj05
## t = 7.817, df = 14.668, p-value = 1.324e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   9.324759 16.335241
## sample estimates:
## mean of x mean of y
##     26.06     13.23
```

Under the hypotheis H_0 that the dose has no effect on the tooth length it would be very unlikely to encounter such data which the very low p-value shows as well as the high t value and the confidence interval which does not include 0.

Further let's try to figure out if there is statistical significance between the 2 supplement types.

```r
vc2 <- tooth %>% filter(supp == "VC", dose == 2.0) %>% select(len)
t.test(oj2, vc2, paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  oj2 and vc2
## 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
```

In contrast to the consideration of the dose, the supplement type has clearly no significant effect on the tooth growth. The p-value of >96% quantifies the probability to observe data like is under the hypothesis H_0 (that the supplement type has no effect). In this case, we would clearly not reject H_0.

Sources

The sources for the report can be found on Github at https://github.com/mkraemerx/datasciencecoursera/tree/master/06StatisticalInference