Statistical Inference Part #2 Basic inferential data analysis

Valentina Scipione

24 May 2015

Basic Inferential Data Analysis

In the second part of this report we analyze the ToothGrowth data in the R datasets package.

Summary of the data:

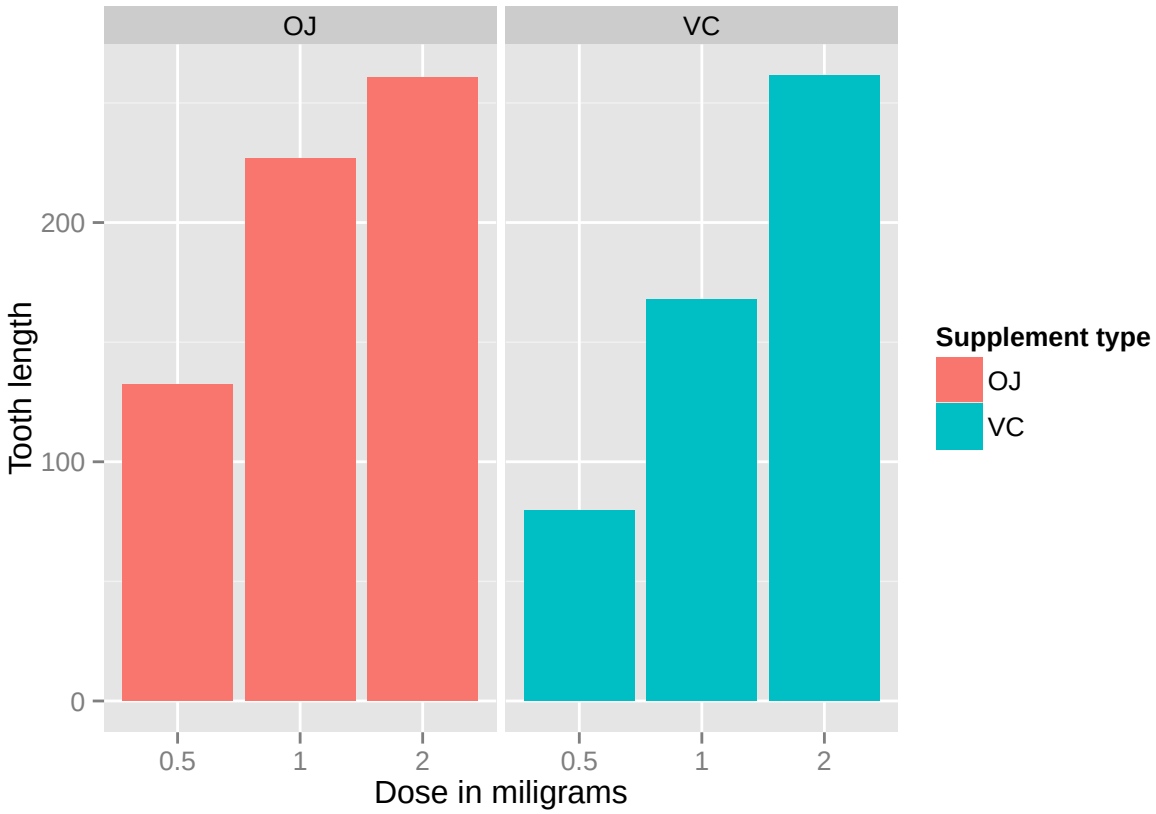
```
# load the dataset
library(datasets)
data(ToothGrowth)

# convert variable dose from numeric to factor
ToothGrowth$dose <- as.factor(ToothGrowth$dose)

# summary statistics for all variables
summary(ToothGrowth)
```

```
##       len        supp     dose
##  Min.   : 4.20   OJ:30   0.5:20
##  1st Qu.:13.07   VC:30   1  :20
##  Median :19.25           2  :20
##  Mean   :18.81
##  3rd Qu.:25.27
##  Max.   :33.90
```

```
# split of cases between different dose levels and delivery methods
table(ToothGrowth$dose, ToothGrowth$supp)
```

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

```
# Plot data set
library(ggplot2)
ggplot(data=ToothGrowth, aes(x=as.factor(dose), y=len, fill=supp)) +
    geom_bar(stat="identity",) +
    facet_grid(. ~ supp) +
    xlab("Dose in miligrams") +
    ylab("Tooth length") +
    guides(fill=guide_legend(title="Supplement type"))
```



We use confidence intervals to compare tooth growth by supp and dose, assuming that the two groups are independent and unequal variances.

95% confidence intervals for two variables and the intercept are as follows:

```
fit <- lm(len ~ dose + supp, data=ToothGrowth)
confint(fit)
```

```
##                 2.5 %    97.5 %
## (Intercept) 10.475238 14.434762
## dose1        6.705297 11.554703
## dose2       13.070297 17.919703
## suppVC      -5.679762 -1.720238
```

Conclusions

Assuming the observations are not paired and the two groups have unequal variances, the 95% confidence intervals between the two supplement type groups mean that if we collect a different set of data and estimate parameters of the linear model many times, 95% of the time, the coefficient estimations will be in these ranges. For each coefficient (i.e. intercept, dose and suppVC), the null hypothesis is that the coefficients are zero, meaning that no tooth length variation is explained by that variable. All p-values are less than 0.05, rejecting the null hypothesis and suggesting that each variable explains a significant portion of variability in tooth length, assuming the significance level is 5%.