Course Project Part 2_Basic Inferential Analysis

Casey645

March 6, 2016

Part#2 Basic Inferential Data Analysis

This report is produced as part of the "Statistical Inference" course project by Coursera which is a part of specialization "Data Science" by John Hopkins University. In this first part, we perform basic inferential data analysis using the ToothGrowth data in the R datasets package.

1.Load the ToothGrowth data and perform some basic exploratory data analysis

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

#look at the dataset variables
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
```

```
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 ...
```

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

#review dataset variables after conversion
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: Factor w/ 3 levels "0.5","1","2": 1 1 1 1 1 1 1 1 1 1 ...

#number of rows of dataset
nrow(ToothGrowth)

##[1] 60
```

2.Provide a basic summary of the data.

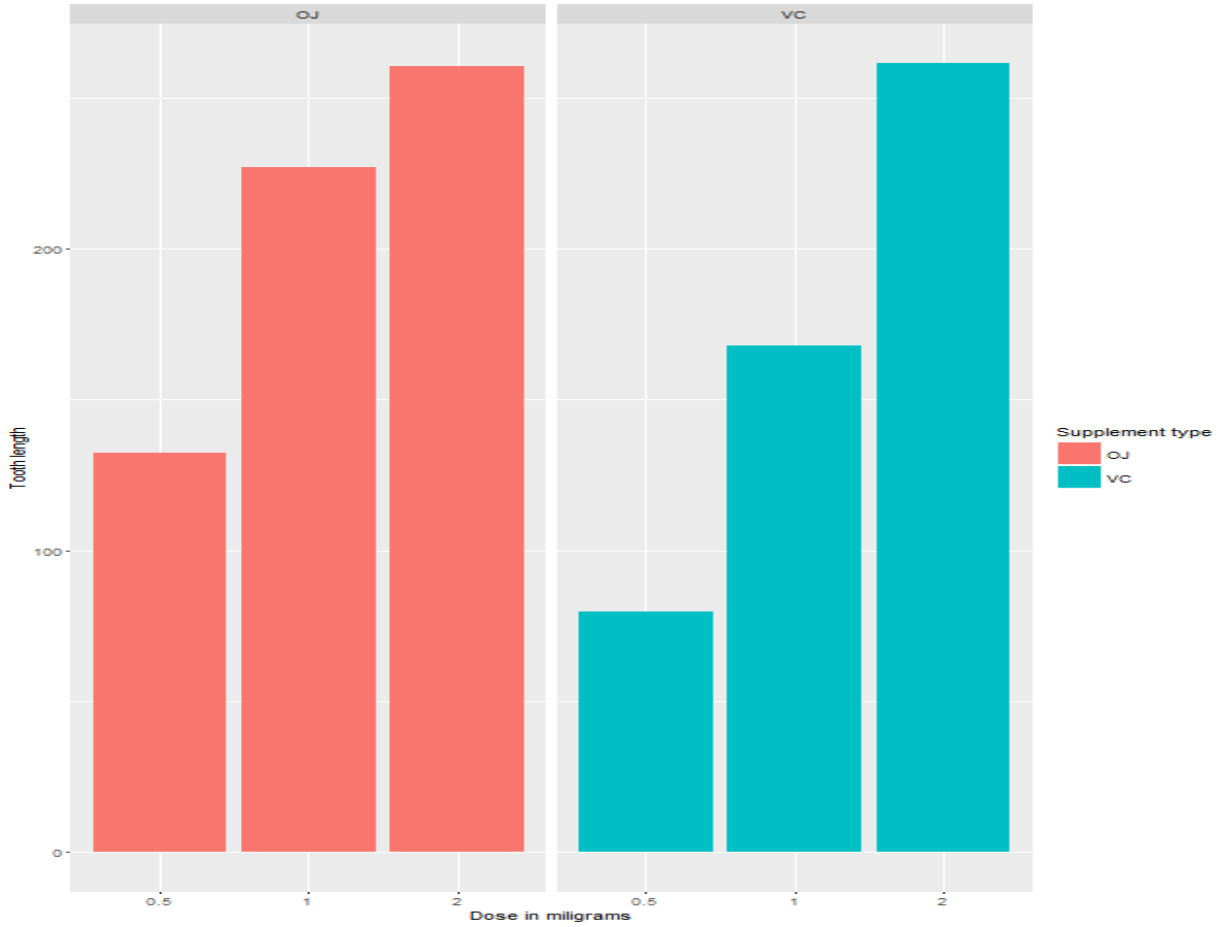
```
#summary statistics for all variables
summary(ToothGrowth)

##       len       supp     dose
##  Min.   : 4.2   OJ:30   0.5:20
##  1st Qu.:13.1   VC:30   1  :20
##  Median :19.2           2  :20
##  Mean   :18.8
##  3rd Qu.:25.3
##  Max.   :33.9

#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

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"))
```



As can be seen above, there is a clear positive correlation between the tooth length and the dose levels of Vitamin C, for both delivery methods.

3.Use confidence intervals and hypothesis tests to compare tooth growth by supp and dose.

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) 10.475  14.43
##dose1        6.705  11.55
##dose2       13.070  17.92
##suppVC      -5.680  -1.72
```

The confidence intervals 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%.

The effect of the dose can also be identified using regression analysis. One interesting question that can also be addressed is whether the supplement type (i.e. orange juice or ascorbic acid) has any effect on the tooth length.

```
summary(fit)
```

The model explains 70% of the variance in the data. The intercept is r
fitcoefficients[[1]], meaningthatwithnosupplementofVitaminC, theaveragetoothlengthisrfitc
oefficients[[1]] units. The coefficient of dose is r
fit
coefficients[[2]]. Itcanbeinterpretedasincreasingthedelievereddose1mg, allelseequal(i. e. nochangeinthesu
coefficients[[2]] units. The last coefficient is for the supplement type. Since the supplement type is a categorical variable, dummy variables are used. The computed coefficient is for suppVC and the value is r
fit
coefficients[[3]]meaningthatdeliveringagivendoseasascorbicacid, withoutchangingthedose, wouldresultin
coefficients[[3]]) units of decrease in the tooth length. Since there are only two categories, we can also conclude that on average, delivering the dosage as orange juice would increase the tooth length by r abs(fit$coefficients[[3]]) units.