Statistical Inference-Project2-Inferential Data Analysis

English Garden

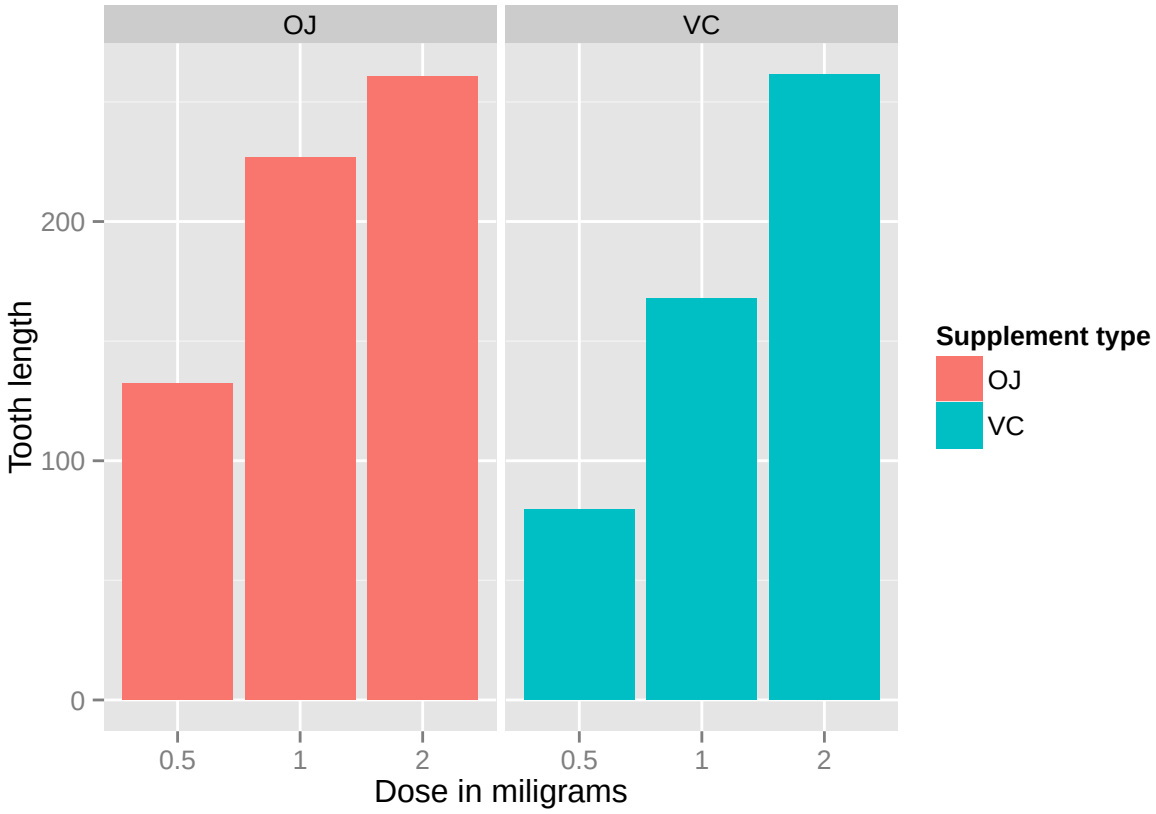
June 19, 2015

In the second part of the project of the Statistical Inference course, we analyze the `ToothGrowth` data in the R datasets package.

```
library(datasets)
data<-ToothGrowth
dim(ToothGrowth)
head(ToothGrowth)
?ToothGrowth
```

This data set consists of 60 observations of three variables: tooth length, supplement type (VC or OJ), dose in milligrams. The reponse is the length of odontoblasts (teeth) in each of 10 guinea pigs 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).

```
library(datasets)
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"))
```



The figure above seperates the observations to two groups based on the supplement type VC or OJ, plots the tooth length variation with respect to the amount of dose. It shows a clear positive correlation between the tooth length and the dose levels of Vitamin C, for both delivery methods.

The effect of the dose and supplement can be identified using regression analysis. This helps us to answer whether the supplement type has an effect on the tooth length. As we seen the figure above, the tooth length is clearly correlated with the dose, here we can find out how much variation of the tooth length can be explained by the supplement type.

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

```
##
## Call:
## lm(formula = len ~ dose + supp, data = ToothGrowth)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.600 -3.700  0.373  2.116  8.800
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.2725     1.2824   7.231 1.31e-09 ***
## dose          9.7636     0.8768  11.135 6.31e-16 ***
## suppVC       -3.7000     1.0936  -3.383   0.0013 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.236 on 57 degrees of freedom
## Multiple R-squared:  0.7038, Adjusted R-squared:  0.6934
## F-statistic: 67.72 on 2 and 57 DF,  p-value: 8.716e-16
```

The model explains approximately 70% of the variance in the data. The intercept 9.2725 means that without supplement of Vitamin C the average tooth length is 9.2725 units. The coefficient of `dose` is 9.7635714. We can interpret it as increasing the dose by 1 mg, all other variables equal (here supplement type), would increase the tooth length by 9.7635714 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 -3.7 means that delivering a given dose as ascorbic acid without changing the dose would result in 3.7 units of decrease in the tooth length. Since there are only two categories, we can infer from this that on average delivering the dosage as orange juice would increase the tooth length by 3.7 units.

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

```
confint(fit)
```

```
##                 2.5 %    97.5 %
## (Intercept)  6.704608 11.840392
## dose         8.007741 11.519402
## suppVC      -5.889905 -1.510095
```

The confidence intervals mean that if we collect different sets of data and estimate parameters of this linear model many times, 95% of the time, the coefficient estimations will fall into these ranges. For each of the three coefficients, intercept, `dose` and `suppVC`, the null hypothesis states that there is no tooth length variation

explained by that variable, the coefficients themselves should be zero. Since all p-values for these three coefficients are less than 0.05, we reject the null hypothesis and accept that each variable explains a significant part of viariation in tooth length, assuming the significance level is 5%.