Statistical Inference Course Project - Part II

Eugene Fallon

July 26, 2015

Project instruction

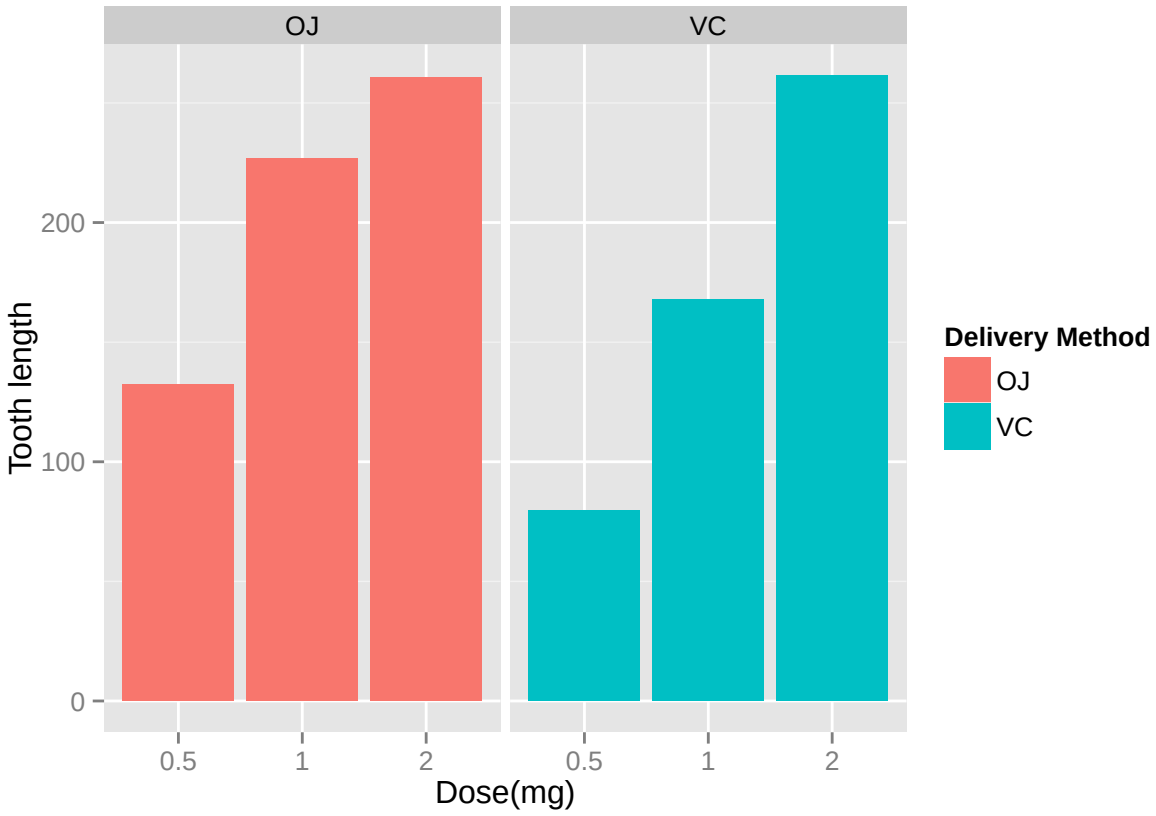
Now in the second portion of the class, we're going to analyze the ToothGrowth data in the R datasets package.

1) Load the ToothGrowth data and perform some basic exploratory data analyses
2) Provide a basic summary of the data.
3) Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose. (Only use the techniques from class, even if there's other approaches worth considering)
4) State your conclusions and the assumptions needed for your conclusions.

Initial data analysis:

Our initial analysis of the data shows that we have a set of 60 observations of the length of teeth for each of 10 guinea pigs. The observations are for three dose levels of Vitamin C (0.5, 1 and 2 mg) each with 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(mg)") +
    ylab("Tooth length") +
    guides(fill=guide_legend(title="Delivery Method"))
```



Data summary

The plot above shows a clear positive correlation between the tooth length and the dose levels of Vitamin C, for both delivery methods.

Further analysis

The effect of the dose can also be identified using regression analysis. One interesting question that can also be addressed is whether the delivery method (i.e. orange juice or ascorbic acid) has any effect on the tooth length. In other words, how much of the variance in tooth length, if any, can be explained by the delivery method?

```
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
```

Inference

Based on our analysis above, we can conclude that the model explains 70% of the variance in the data. The intercept is 9.2725, meaning that with no supplement of Vitamin C, the average tooth length is 9.2725 units. The coefficient of dose is 9.7635714. It can be interpreted as increasing the delievered dose 1 mg, all else equal (i.e. no change in the delivery method), would increase the tooth length 9.7635714 units. The last coefficient is for the delivery method. The computed coefficient is for suppVC and the value is -3.7 meaning that delivering a given dose as ascorbic acid, would result in 3.7 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 3.7 units.

Confidence levels

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 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%.