Data Science Statistical Inference - ToothGrowth

Dr Paul Fergus

Introduction

In this second report some basic inferential data will be carried out on the ToothGrowth dataset. The dataset contains 60 observations. There are 3 features, the first is the length of the teeth of guinea pigs, the second is teh supplement type, and the third gives the dose.

Methodology

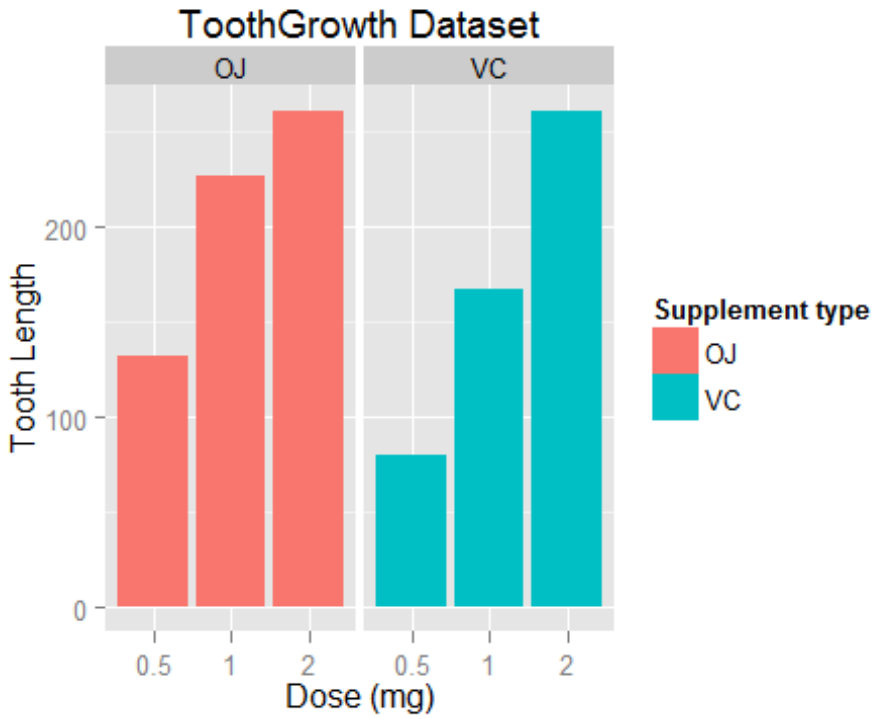
The table below provides a basic summary of the data.

```
summary(ToothGrowth)
```

```
##       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
```

The diagram below shows dose levels against tooth length and is grouped into supplement type.

```
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") +
  ggtitle("ToothGrowth Dataset") +
  guides(fill=guide_legend(title="Supplement type"))
```



The figures shows us that their are significant differences in the length of guinea pigs teeth when different dosages are used. The supplement type has as an affect of the outcome with OJ supplements improving the length of teeth when dosages 0.5 adn 1 are used, which suggests it is better than using VC. However, when dosage 2 is used there are no marked differences between tooth length.

Below we find the confidence intervals to compare tooth growth by supplement nad dose.

```
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
```

Using a linear model is shows that 70% of the variance can be explained by the model. The intercept is approx 9.27 which states that the average length of a tooth is 9.27 with any Vitamin C supplement. The dose coefficient is approx. 9.76 meaning that increasing the delivery dose by 1mg will increase teh tooth length by 9.76 units. The suppVC value is -3.7 which means that giving a dose with VC will result in a decrease in tooth growth by 3.7. Given that their are only two supplment types, this means that the alternative, OJ, will result in a 3.7 increase.

95% confidence intervals and the intercept are shown below.

```
confint(fit)
```

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

If we collect different sample data and estimate parameters of the linear model over a number of simulations, 95% of the time, the calculated confidence interval from some future experiement will emcompass the true value of the population parameter.