Statistical Inference Project, Part 2: Basic Inferential Data Analysis

Tehty

Jul 23, 2015

Overview

In this second portion of the assignment, we're going to analyze the ToothGrowth data in the R datasets package.

Steps:

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.
4. Conclusions and the assumptions

The `ToothGrowth` data set consists of 60 observations of 3 variables:

- `len` (numeric)
- `supp` (VC or OJ)
- `dose` (numeric)

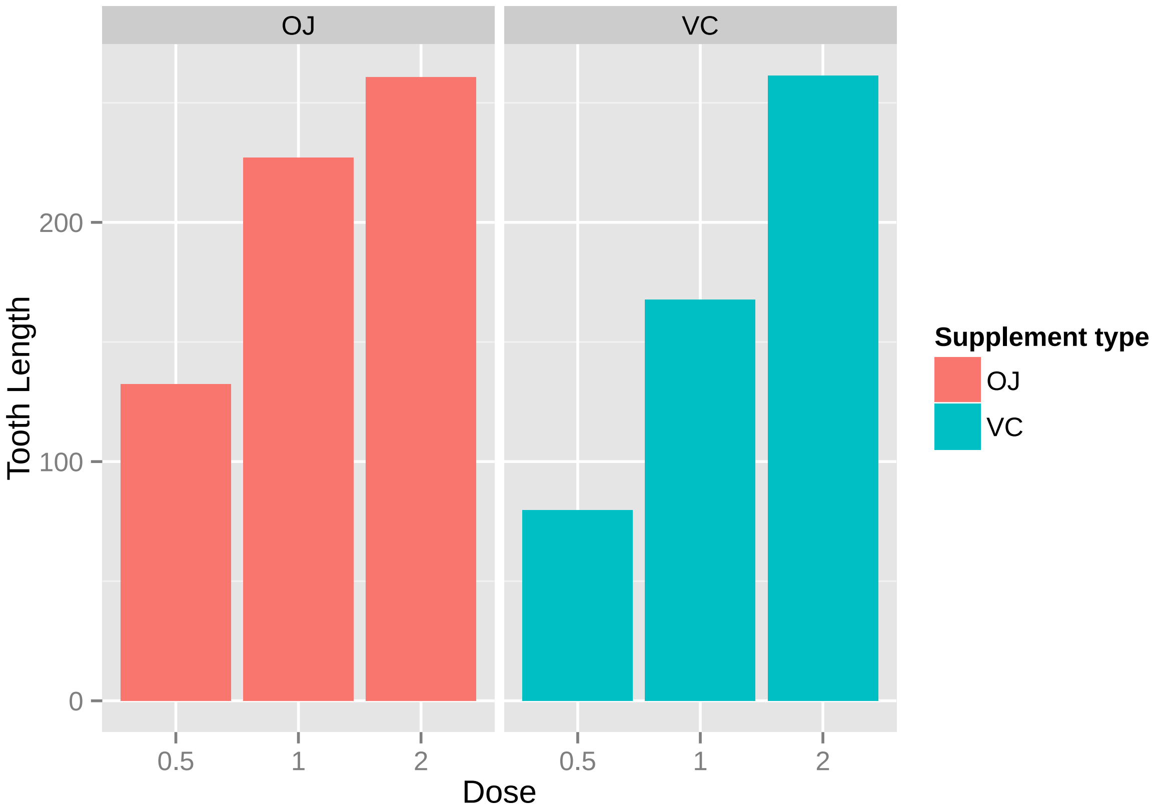
```
library(datasets)
head(ToothGrowth)
```

```
##    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
```

```
summary (ToothGrowth)
```

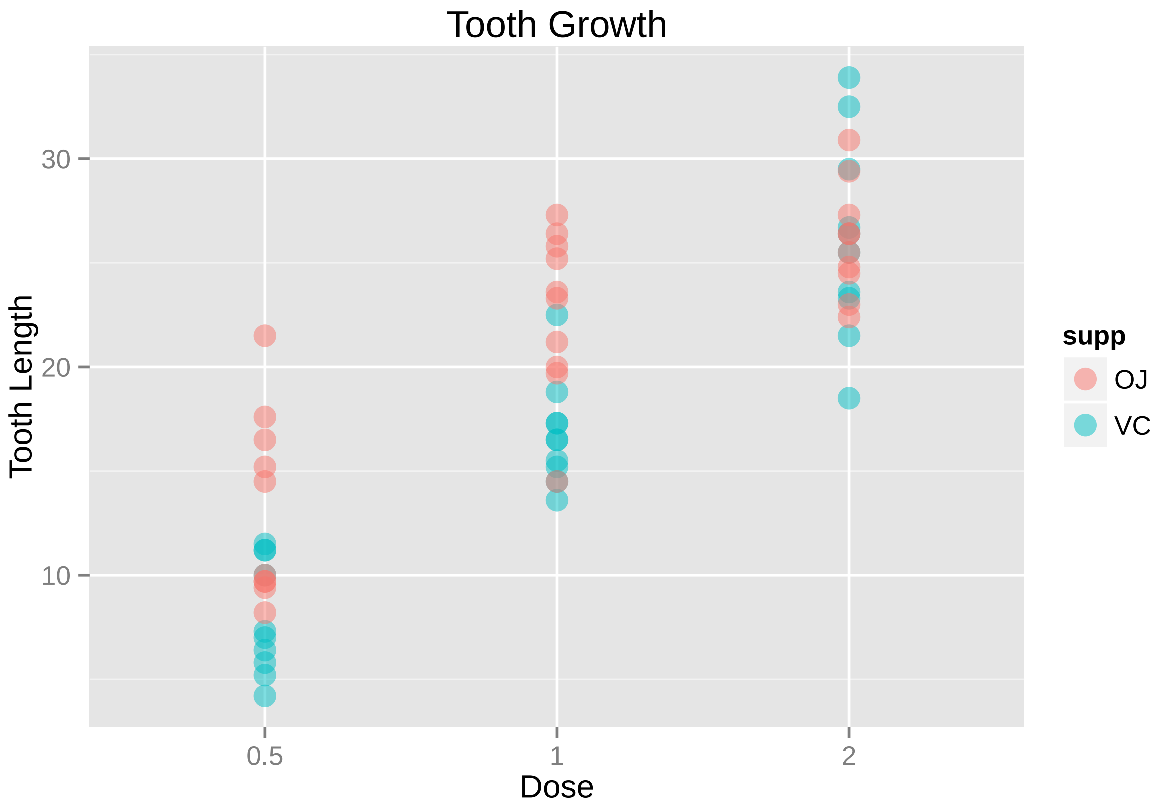
```
##       len        supp         dose      
##  Min.   : 4.20   OJ:30   Min.   :0.500  
##  1st Qu.:13.07   VC:30   1st Qu.:0.500  
##  Median :19.25           Median :1.000  
##  Mean   :18.81           Mean   :1.167  
##  3rd Qu.:25.27           3rd Qu.:2.000  
##  Max.   :33.90           Max.   :2.000  
```

```
library(ggplot2)
ggplot(ToothGrowth, aes(x=as.factor(dose), y=len, fill=supp)) +
  geom_bar(stat="identity",) +
  facet_grid(. ~ supp) +
  labs(x ="Dose", y="Tooth Length")+
  guides(fill=guide_legend(title="Supplement type"))
```



As per the histogram above, it is a clear positive correlation between the tooth length and the dose levels.

```
graph <- ggplot(ToothGrowth, aes(x=as.factor(dose), y=len))
graph + geom_point(aes(color=supp), size = 4, alpha = 1/2) + labs(title = "Tooth Growth") + labs(x ="Do
```



From above results, we can identify below

1. Tooth length with VC supplement has wider distribution than those with OJ supplement
2. Teeth are longer with OJ than those with VC at the dose 0.5 and 1.0 level
3. The larger the dose, the longer the tooth.

Now, let us examing the variance in tooth length in relation of the supplement type.

```
fitting <- lm(len ~ dose + supp, ToothGrowth)
summary(fitting)
```

```
##
## 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
```

Assumption: All else equal

The intercept is 9.2725, means that with no supplement of Vitamin C, the average tooth length is 9.2725.

The coefficient of `dose` is 9.7635714. When increasing dose 1 mg, would increase the tooth length of 9.7635714

The last coefficient `suppVC` with the value is -3.7 explains that a given dose of VC, would result in decrease of 3.7 in the tooth length.

Confidence intervals for these two variables and the intercept as below

```
confint(fitting)
```

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

For each coefficient (the intercept, `dose` and `suppVC`), the null hypothesis is where the coefficent is 'zero' - means 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, with the assumption of 5% of significance level.

Hypothesis Test

Cosidering there are three levels of dose (0.5, 1.0, 2.0), we perform t test in following orders: (1) dose 0.5 vs. dose 1.0; (2) dose 1.0 vs. dose 2.0; and (3) dose 0.5 vs. dose 2.0.

```
t.test(ToothGrowth$len[ToothGrowth$dose == 1.0], ToothGrowth$len[ToothGrowth$dose == 0.5], paired = FALS
```

```
## [1]  6.276252 11.983748
## attr(,"conf.level")
## [1] 0.95
```

```
t.test(ToothGrowth$len[ToothGrowth$dose == 2.0], ToothGrowth$len[ToothGrowth$dose == 1.0], paired = FALS
```

```
## [1] 3.735613 8.994387
## attr(,"conf.level")
## [1] 0.95
```

```
t.test(ToothGrowth$len[ToothGrowth$dose == 2.0], ToothGrowth$len[ToothGrowth$dose == 0.5], paired = FALS
```

```
## [1] 12.83648 18.15352
## attr(,"conf.level")
## [1] 0.95
```

Conclusion

From the result, the confidence intervals for three comparisons are all above zero, therefore, we can conclude that the larger the supp dose, the longer the tooth.