Inferential Data Analysis of the R `ToothGrowth` Data Set

Hiro Miyake

September 21, 2016

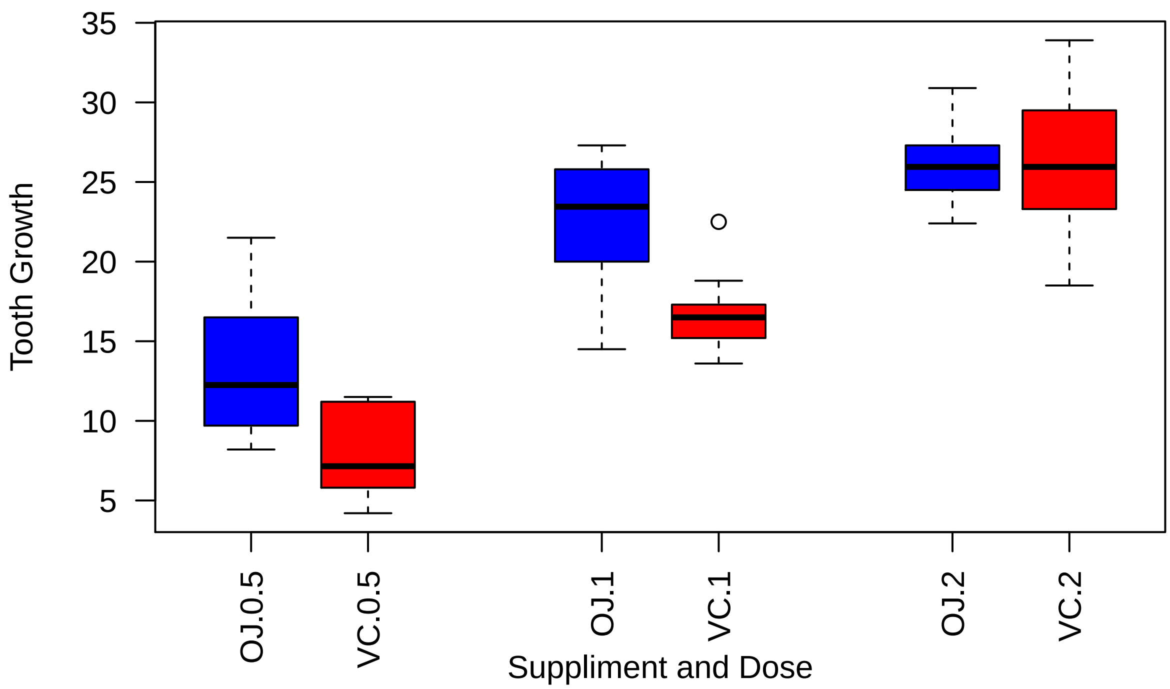
Overview

In this report I perform exploratory analysis and inferential data analysis on the R `ToothGrowth` data set. Briefly, this data set includes the length of odontoblasts, which are cells responsible for tooth growth, for 60 guinea pigs as a function of vitamin C doses (0.5, 1, and 2 mg/day) and delivery method (equivalently, suppliment) [orange juice (OJ) or ascorbic acid (VC)]. Further details on this data set can be found here.

Exploratory analysis

Following the examples of box plots here and here, the data can be summarized with the box plot below.

```
res = boxplot(len~supp*dose, data=ToothGrowth, col=(c("blue","red")),
        xlab="Suppliment and Dose", ylab = "Tooth Growth", las = 2, at = c(1,2,4,5,7,8))
```



Summary statistics

The box plot shows the median (thick black line within the box) and approximately the first (bottom of box) and third (top of box) quartile of the data (for further detail on what the box plot shows, see here). The table below is a summary of these values for all six instances considered.

```r
tgsummary = res$stats[c(2,3,4),]
colnames(tgsummary) = c("OJ.0.5", "VC.0.5", "OJ.1",
                        "VC.1", "OJ.2", "VC.2")
rownames(tgsummary) = c("1st Quartile", "Median", "3rd Quartile")
tgsummary
```

```
##              OJ.0.5 VC.0.5  OJ.1 VC.1  OJ.2  VC.2
## 1st Quartile   9.70   5.80 20.00 15.2 24.50 23.30
## Median        12.25   7.15 23.45 16.5 25.95 25.95
## 3rd Quartile  16.50  11.20 25.80 17.3 27.30 29.50
```

Tooth growth and dose

From the box plot, it looks like higher the dose, bigger the tooth growth. I will perform a t-test for delivery method VC for the two doses 0.5 and 2 mg/day to test if this difference is statistically significant. I assume unequal variance.

```r
growthdoseVC = subset(ToothGrowth, supp == 'VC' & (dose == 0.5 | dose == 2))
t.test(len~dose, paired = FALSE, var.equal = FALSE, data = growthdoseVC)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -10.3878, df = 14.327, p-value = 4.682e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -21.90151 -14.41849
## sample estimates:
## mean in group 0.5   mean in group 2
##              7.98             26.14
```

The p-value is 4.682×10^{-8} and the 95% confidence interval does not contain zero, which means it is likely that the dose affects tooth growth.

Tooth growth and delivery method

From the box plot, it appears there is no statistically significant correlation in tooth growth between different delivery methods for the same doses. I will perform a t-test between the two delivery methods for dose 2 mg/day. I assume unequal variance.

```r
growthsuppd2 = subset(ToothGrowth, dose == 2)
t.test(len~supp, paired = FALSE, var.equal = FALSE, data = growthsuppd2)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = -0.0461, df = 14.04, p-value = 0.9639
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
##  -3.79807  3.63807
## sample estimates:
## mean in group OJ mean in group VC
##            26.06            26.14
```

The p-value is 0.9639 and the 95% confidence interval contains zero, which means it is unlikely there is a correlation between tooth growth and delivery method.

Conclusion

Through inferential analyses, I conclude that higher doses lead to bigger tooth growth, but the delivery method does not matter. I assume the data are independently identically distributed Gaussian and that it is symmetrically distributed and not skewed.