Statistical Inference Project - Part 2

Alan C Bonnici

Tuesday, September 16, 2014

Synopsis

This project answers the following question:

> 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 hypothesis tests to compare tooth growth by supp and dose. (Use the techniques from class even if there's other approaches worth considering) 4 State your conclusions and the assumptions needed for your conclusions.

The Solution

Analysis of the data

According to the R Documentation, the ToothGrowth dataset is

> *The response 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).*

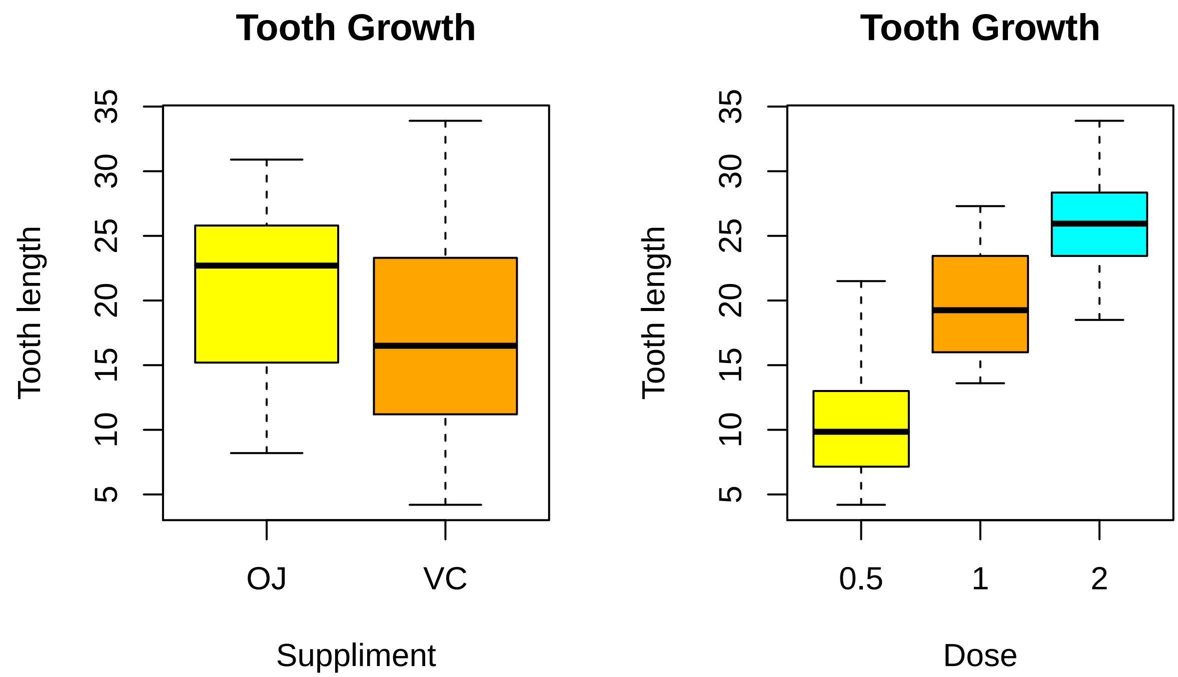
The dataset is a data frame with 60 observations on 3 variables.

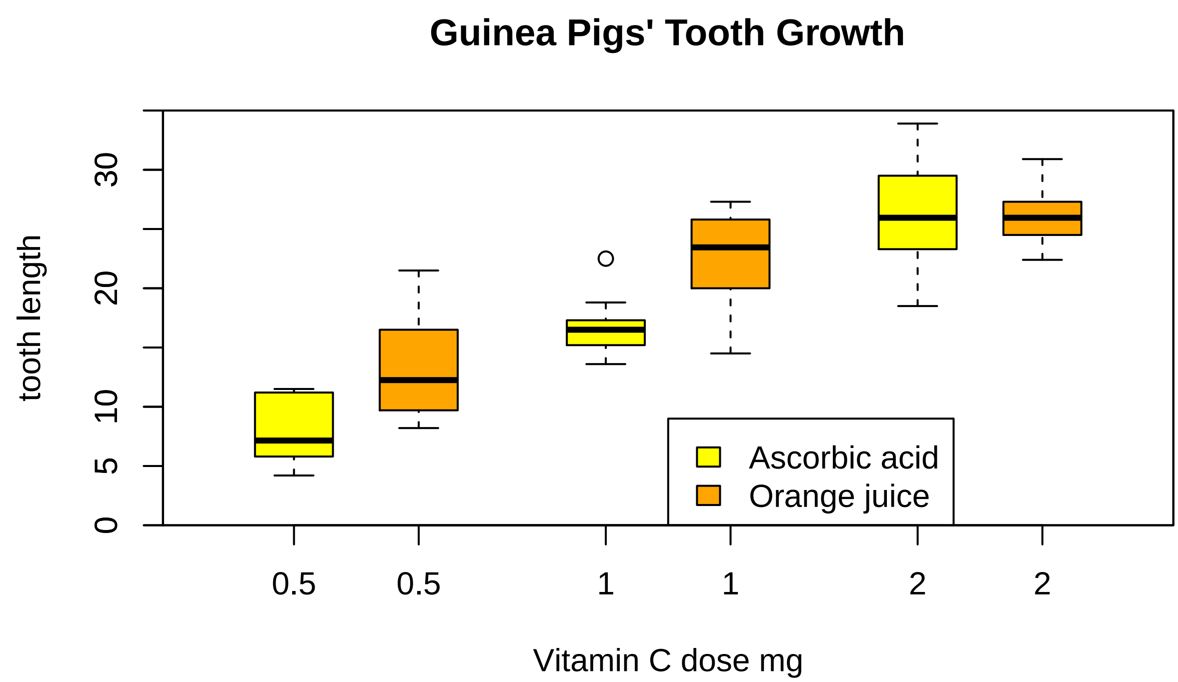
len	numeric	Tooth length
supp	factor	Supplement type (VC or OJ).
dose	numeric	Dose in milligrams.

The first few rows of the data are shown below:

```
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
```





The graph indicates a positive correlation between dose levels and tooth growth for both delivery methods although the correlation for ascorbic acid is greater than that for orange juice.

Summary information for Ascorbic acid

```
summary(ToothGrowth[ToothGrowth$supp == "VC", "len"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     4.2    11.2    16.5    17.0    23.1    33.9
```

Summary information for Orange juice

```
summary(ToothGrowth[ToothGrowth$supp == "OJ", "len"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     8.2    15.5    22.7    20.7    25.7    30.9
```

OJ has a greater mean than VC as shown hereunder.

```
round(sapply(with(ToothGrowth, split(len, supp), mean), mean), 2)
```

```
##    OJ    VC
## 20.66 16.96
```

The group mean increases as the dosage increases.

```
round(with(ToothGrowth, aggregate(len, list(dose), mean)), 2)
```

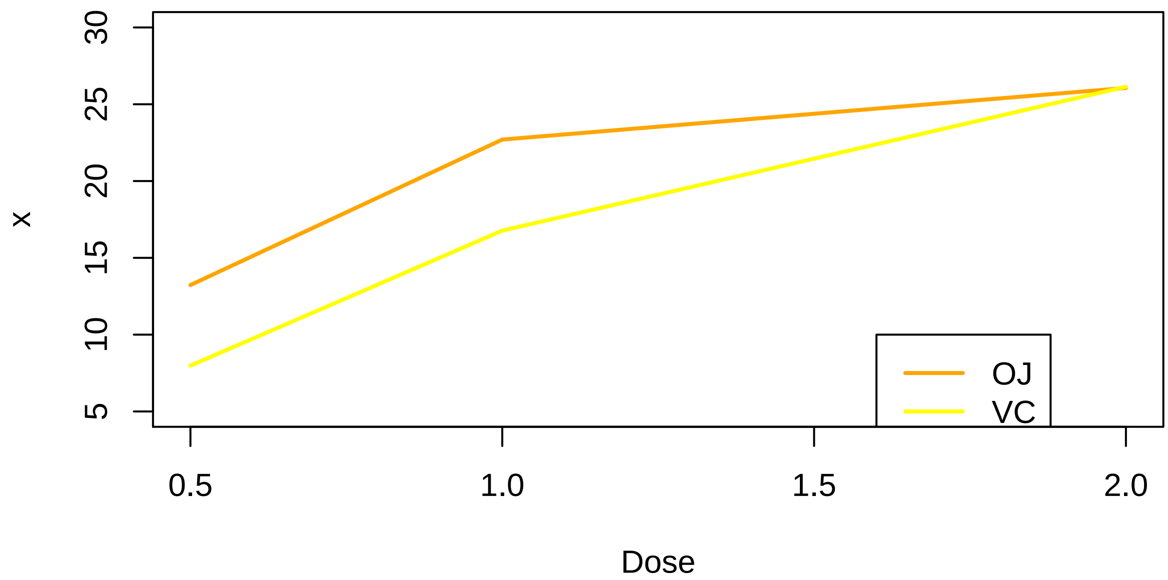
```
##   Group.1     x
## 1     0.5 10.61
## 2     1.0 19.73
## 3     2.0 26.10
```

From the table below it is clear that

```
data <- with(ToothGrowth, aggregate(len, list(dose, supp), mean))
dataOJ <- data[data$Group.2 == "OJ",]
dataVC <- data[data$Group.2 == "VC",]
with (dataOJ, plot (Group.1, x, type="l", col="orange", lwd=2, pch=23,
                    xlab="Dose", ylan="mean", ylim=c(5, 30)))
```

```
## Warning: "ylan" is not a graphical parameter
## Warning: "ylan" is not a graphical parameter
## Warning: "ylan" is not a graphical parameter
## Warning: "ylan" is not a graphical parameter
## Warning: "ylan" is not a graphical parameter
## Warning: "ylan" is not a graphical parameter
```

```r
with (dataVC, lines (Group.1, x, type="l", col="yellow", lwd=2, pch=22))
legend (1.6, 10, c("OJ", "VC"), lty=c(1,1), lwd=2, col=c("orange", "yellow"))
```



T-Tests on the data: Supliment Type

Looking at the data as a whole,

```r
t.test(len ~ supp, data=ToothGrowth)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 1.915, df = 55.31, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.171  7.571
## sample estimates:
## mean in group OJ mean in group VC
##            20.66            16.96
```

From the above data, one cannot state at the 95% confidence interval that one supplement is more effective than the other. This is because the confidence interval comprises 0 within it and so the null hypothesis cannot be rejected.

T-Tests on the data: Suppliment by Dosage

The purpose of these tests is to see whether the two supplements are more effective at the different doses.

```
suppByDose <- list(d0.5 = subset(ToothGrowth, dose == 0.5),
                   d1.0 = subset(ToothGrowth, dose == 1.0),
                   d2.0 = subset(ToothGrowth, dose == 2.0))

lapply(suppByDose, function(d) t.test(len ~ supp, data = d))
```

```
## $d0.5
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 3.17, df = 14.97, p-value = 0.006359
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.719 8.781
## sample estimates:
## mean in group OJ mean in group VC
##            13.23             7.98
##
##
## $d1.0
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 4.033, df = 15.36, p-value = 0.001038
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.802 9.058
## sample estimates:
## mean in group OJ mean in group VC
##            22.70            16.77
##
##
## $d2.0
##
##  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:
##  -3.798  3.638
## sample estimates:
## mean in group OJ mean in group VC
##            26.06            26.14
```

From the output of the t-test the dose levels of 0.5 and 1.0 we can at a confidence interval of 95% accept the Alternative hypothesis $H\ a$ that the effect of OJ is more effective than VC. The same statement cannot be made for the dose of 2.0.

T-Tests on the data: Dosage levels

The purpose of these tests is to compare the dosage levels irrespective of supplement.

```
doseLevels <- list(d0.5_v_1.0 = subset(ToothGrowth, dose %in% c(0.5, 1.0)),
                   d0.5_v_2.0 = subset(ToothGrowth, dose %in% c(0.5, 2.0)),
                   d1.0_v_2.0 = subset(ToothGrowth, dose %in% c(1.0, 2.0)))

lapply(doseLevels, function(d) t.test(len ~ dose, data = d))
```

```
## $d0.5_v_1.0
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -6.477, df = 37.99, p-value = 1.268e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.984  -6.276
## sample estimates:
## mean in group 0.5   mean in group 1
##             10.61             19.73
##
##
## $d0.5_v_2.0
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -11.8, df = 36.88, p-value = 4.398e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.16 -12.83
## sample estimates:
## mean in group 0.5   mean in group 2
##             10.61             26.10
##
##
## $d1.0_v_2.0
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -4.901, df = 37.1, p-value = 1.906e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.996 -3.734
## sample estimates:
## mean in group 1 mean in group 2
##           19.73           26.10
```

At the 95% confidence interval the very small p-value and the very high absolute value of t indicate that for all three comparisons, the alternative hypothesis that the "*true difference in means is not equal to 0*" seems to prevail.

Code Reproducability

This project follows the reproducible research methodology that allows others to view the code and execute it if they so desire. The code for this project can be downloaded from https://github.com/chribonn/statinference-005.