Statistical Inference, Coursera

Course Project Part 2: Inferential data analysis

Chris Rupley, October 24, 2014

The ToothGrowth Dataset

We wish to explore the ToothGrowth dataset in R. This is a set of data recorded from an experiment on guinea pigs where the length of their teeth was measured as a function of different amounts of different supplements given to them. Further information on the dataset can be found in R from `?ToothGrowth`.

First, let's load the dataset and see how it is structured.

```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

The `len` or "length" variable is our output so let us take a look at how each of the 60 observations is distributed over the different supplements (`supp`) and dosages (`dose`).

```
##     dose
## supp 0.5  1  2
##   OJ  10 10 10
##   VC  10 10 10
```

So all the observations are nicely evenly distributed over each of the 6 possible experimental conditions.

Comparison of the two supplements

Let's see what, if any, effect our choice of supplement has on guinea pig tooth growth. We have two different supplements used; vitamin C and orange juice. In this case, we can choose to make the following our null hypothesis:

> Orange juice is no more effective at increasing tooth growth than vitamin C supplements.

First, we will take a look at the average tooth length for each supplement group.

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

It certainly appears that the orange juice supplement does have an effect on tooth growth with the average tooth length being about 22% longer for OJ. But is this difference enough to be considered statistically significant?

To find out, I will perform a t-test with a 95% confidence interval threshold. That is to say, if we can say with at least 95% certainty that tooth length will be longer in one condition than the other, then we may reject our null hypothesis in favor of our statistically significant result. Let's take a look at the range of values for the 95% confidence interval.

```
##      OJ range      VC range
##  "18.2-23.13" "13.88-20.05"
```

So it appears that the ranges of the intervals do, in fact, overlap. Therefore, in spite of the fact that there appeared to be rather large difference in tooth length for the two supplements, we find that this difference is not enough to be considered statistically significant (with 95% confidence) and we must accept our null hypothesis.

Comparison of dose

We can perform a similar analysis of the effect of dose size on guinea pig tooth growth. Proceeding as before, our null hypothesis states,

> The supplement dosage amount has no effect on increasing tooth growth.

Looking at the mean tooth length by dose,

```
##   0.5     1     2
## 10.61 19.73 26.10
```

Again, it appears that the tooth lengths are increasing with dosage, but let's check the confidence intervals to be sure.

```
##      0.5 dose      1.0 dose      2.0 dose
##   "8.5-12.71"  "17.67-21.8" "24.33-27.87"
```

In this case, we have three 95% confidence intervals that do not overlap. We can therefore reject our null hypotheses in favor of the alternative conclusion.

> **An increase in supplement dose leads to an increase in tooth growth.**

Conclusion

From this evaluation of the `ToothGrowth` dataset in R, two conclusions were reached.

- Orange juice is no more effective at increasing tooth growth than is vitamin C.
- Increasing supplement dose does lead to an increase in tooth growth.

In the first case, the null hypothesis was accepted since it failed a t-test with its $\alpha > 0.05$ (confidence interval $< 95\%$) and in the second case, the null hypothesis was rejected since it passed the t-test.

To reach these conclusions, the following assumptions were made.

- The various observations were independend identically distributed (i.i.d) for each experimental condition.
- There was a sufficient number of observations for the Central Limit Theorem to be valid.
- A t-test success threshold of $\alpha < 0.05$ is adequate.

Appendix

Figures

Let's take a look at a visualization of the ToothGrowth data.

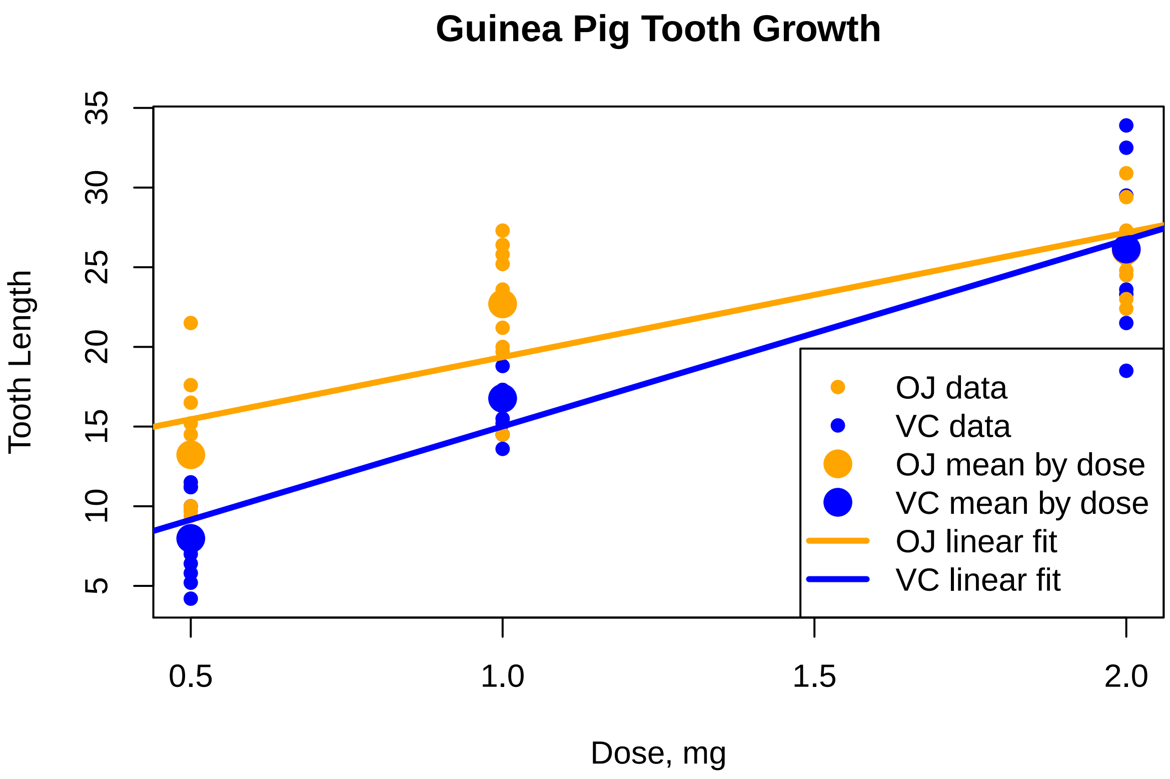
```r
palette(c("orange", "blue"))
with(ToothGrowth, plot(dose, len, col = supp, pch = 16,
                       main = "Guinea Pig Tooth Growth",
                       xlab = "Dose, mg", ylab = "Tooth Length"))

Tmeans <- tapply(ToothGrowth$len, list(ToothGrowth$dose, ToothGrowth$supp), mean)
Tsd <- tapply(ToothGrowth$len, list(ToothGrowth$dose, ToothGrowth$supp), sd)
x <- as.numeric(dimnames(Tmeans)[[1]])

points(x, Tmeans[,1], pch = 16, cex = 2, col = 1)
points(x, Tmeans[,2], pch = 16, cex = 2, col = 2)

fitOJ <- lm(len ~ dose, ToothGrowth, supp == "OJ")
fitVC <- lm(len ~ dose, ToothGrowth, supp == "VC")
abline(fitOJ, col = 1, lwd = 3)
abline(fitVC, col = 2, lwd = 3)

legend("bottomright",
       c("OJ data", "VC data",
         "OJ mean by dose", "VC mean by dose",
         "OJ linear fit", "VC linear fit"),
       pch = c(16,16,16,16,NA,NA),
       pt.cex = c(1,1,2,2),
       lty = c("blank","blank","blank","blank","solid","solid"),
       col = rep(c(1,2),3),
       lwd = c(NA,NA,NA,NA,3,3))
```



Code

The following is all of the unechoed R code used in creating this report.

```
# Load the data and display the structure
data(ToothGrowth)
str(ToothGrowth)

# Output table of counts of observations by experimental conditions
table(ToothGrowth[,2:3])

# Compute the mean tooth length by supplement
suppmean <- with(ToothGrowth, tapply(len, supp, mean))
print(suppmean)

# inline percent difference calculation
round(abs(suppmean[1]-suppmean[2])/suppmean[2]*100,0)

# Compute the t confidence interval by supplement
nOJ <- sum(ToothGrowth$supp == "OJ")
OJrange <- round(suppmean[1] + c(-1,1) * qt(0.975, nOJ - 1) *
    sd(ToothGrowth$len[ToothGrowth$supp == "OJ"])/sqrt(nOJ),2)

nVC <- sum(ToothGrowth$supp == "VC")
VCrange <- round(suppmean[2] + c(-1,1) * qt(0.975, nVC - 1) *
    sd(ToothGrowth$len[ToothGrowth$supp == "VC"])/sqrt(nVC),2)

suppresult <- c(paste0(OJrange[1],"-",OJrange[2]),
```

```r
                paste0(VCrange[1],"-",VCrange[2]))
names(suppresult) <- c("OJ range", "VC range")
print(suppresult)

# Compute mean tooth length by dose
dosemean <- with(ToothGrowth, tapply(len, dose, mean))
print(dosemean)

# Compute t intervals by dose
np5 <- sum(ToothGrowth$dose == 0.5)
dp5range <- round(dosemean[1] + c(-1,1) * qt(0.975, np5 - 1) *
    sd(ToothGrowth$len[ToothGrowth$dose == 0.5])/sqrt(np5),2)

n1 <- sum(ToothGrowth$dose == 1)
d1range <- round(dosemean[2] + c(-1,1) * qt(0.975, n1 - 1) *
    sd(ToothGrowth$len[ToothGrowth$dose == 1])/sqrt(n1),2)

n2 <- sum(ToothGrowth$dose == 2)
d2range <- round(dosemean[3] + c(-1,1) * qt(0.975, n2 - 1) *
    sd(ToothGrowth$len[ToothGrowth$dose == 2])/sqrt(n2),2)

doseresult <- c(paste0(dp5range[1],"-",dp5range[2]),
                paste0(d1range[1],"-",d1range[2]),
                paste0(d2range[1],"-",d2range[2]))
names(doseresult) <- c("0.5 dose", "1.0 dose", "2.0 dose")
print(doseresult)
```