Statistical Inference Project Part 2

Matt Turner

Monday, January 12, 2015

The Question

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/or hypothesis tests to compare tooth growth by supp and dose.
4. State your conclusions and the assumptions needed for your conclusions.

The Solution

1) Basic Exploratory Analysis

A quick search for R and ToothGrowth will reveal a web page describing what the data is representing. It indicates that the ToothGrowth dataset contains a data frame with 60 observations on 3 variables. The 60 observations are are taken from 10 Guinea pigs and describe the length of their teeth following 3 different dosage levels of Vitamin C (0.5, 1 and 2mg) and from 2 different delivery mechanisms (orange juice and ascorbic acid).

The first thing we need to do is load the data into R and have a quick look at ToothGrowth using the *describe* command from the **Hmisc** package. Doing so yields the following:

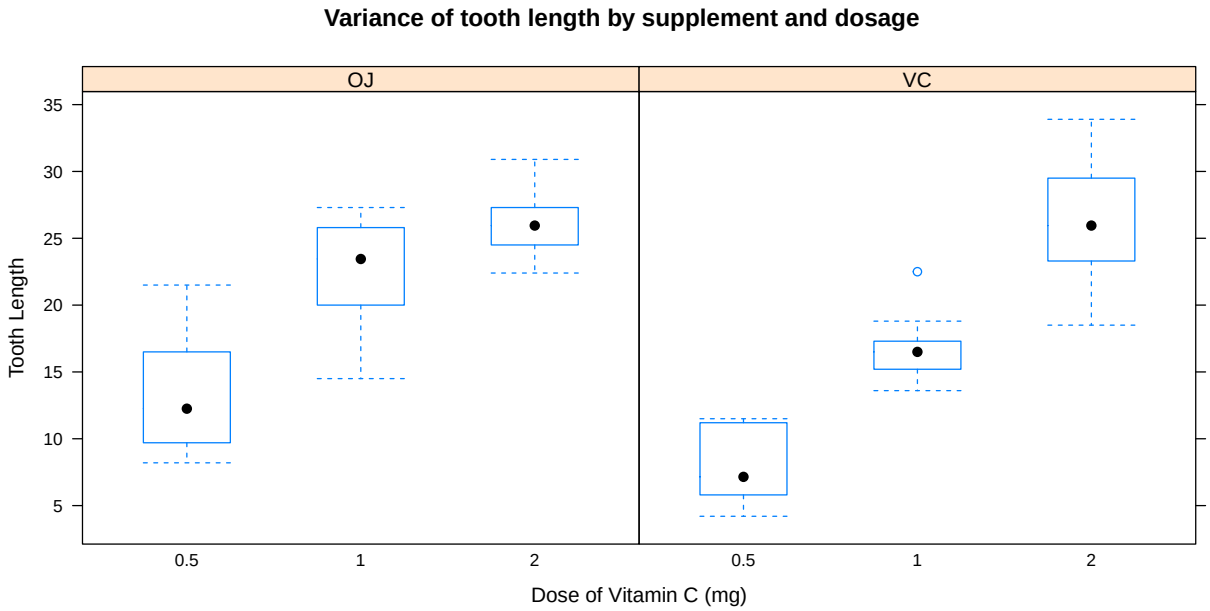
```
describe(ToothGrowth)
```

```
## ToothGrowth
##
##  3  Variables      60  Observations
## ---------------------------------------------------------------------------
## len
##       n missing  unique    Info    Mean     .05     .10     .25     .50
##      60       0      43       1   18.81    6.37    8.11   13.07   19.25
##     .75     .90     .95
##   25.27   27.30   29.57
##
## lowest :  4.2  5.2  5.8  6.4  7.0, highest: 29.4 29.5 30.9 32.5 33.9
## ---------------------------------------------------------------------------
## supp
##       n missing  unique
##      60       0       2
##
## OJ (30, 50%), VC (30, 50%)
## ---------------------------------------------------------------------------
## dose
##       n missing  unique    Info    Mean
##      60       0       3    0.89   1.167
##
## 0.5 (20, 33%), 1 (20, 33%), 2 (20, 33%)
## ---------------------------------------------------------------------------
```

Here, we can see our 3 variables, *len*, *supp* and *dose*, which correspond to tooth length, supplement (or delivery) type, and dosage respectively.
It confirms that there are 2 possible values for *supp* with 30 observations for each, and 3 for *dose* with 20 observations each as expected. We can also see there are no missing values for any variable, which is obviously a good thing and saves us from having to do some interpolation.
The most useful information shown here is in the *len* variable. We can see the observed tooth lengths range from 4.2 to 33.9, with a mean of 18.81 and a median of 19.25.

2) Brief summary of data

Box plots are a good way to quickly summarise a set of data, particularly a set such as ours that has multiple factors:



Code shown in Appendix A

We can see here the clear pattern that a higher dosage of Vitamin C implies a larger tooth length. It seems to also be apparent that for lower doses, using orange juice (OJ) as the supplement correlates to longer teeth than if you were to use the ascorbic acid (VC). However, for the highest dose of 2mg, it is clear that both supplements result in a roughly equal median tooth length, although orange juice is less varied in these results.

3) Comparing tooth growth by supplement and dosage

We now need to do more detailed analysis to determine whether the conclusions drawn from the above boxplot are correct.
We will use hypothesis testing and confidence intervals to determine whether OJ or VC are better linked to tooth length.
There are a few **assumptions** that we need to make for our tests: Firstly, we assume the *populations are independent from one another*. We also assume that they are *not paired*, and that they have *unequal variances*.

The first test will be to directly compare OJ to VC regardless of which dosage has been administered. This is done as follows: *Note the null hypothesis is that there is no relation between the test variables and tooth length.*

```
OjVc <- t.test(len~supp, paired=F, var.equal=F, data=ToothGrowth)
OjVc
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1710156  7.5710156
## sample estimates:
## mean in group OJ mean in group VC
##         20.66333         16.96333
```

The output here indicates a confidence interval between -0.1710156 and 7.5710156. Since this interval contains zero, we cannot reject the null hypothesis. The P-value is also greater (just) than the 0.05 that we are looking for for a 95% confidence, and so again indicates the relationship is not significant. Therefore, we must assume that neither OJ or VC are better for tooth length when dosage is ignored.

The next stage here will be to do the same test but at each dosage level. If this is to follow the conclusions drawn from the boxplot earlier then we will expect to see OJ better than VC for dose level of 0.5 and 1, but no difference for dose at level 2.
To save some unnecessary output, the 3 tests have been run and the relevent statistics extraced and put into a table. *Code shown in Appendix B.*

	P-Value	LowerConfInt	UpperConfInt
0.5 Dosage	0.00636	1.71906	8.78094
1.0 Dosage	0.00104	2.80215	9.05785
2.0 Dosage	0.96385	-3.79807	3.63807

We can see in the table that for doses of 0.5 and 1.0 we achieved a 95% confidence interval that does **not** contain zero. This is a good indicator that one supplement type is better that the other for tooth growth.
In addition we can also see that the P-Value for doses of 0.5 and 1.0 are below the required 0.05 required to be statistically significant. This again indicates that one type of supplement is better than the other for tooth growth. We must therefore reject the null hypothesis.
For doses of 2mg however, we see that the confidence interval contains zero and the P-Value is much too high to be statistically significant. We can therefore accept the null hypothesis and conclude that for doses of 2mg the supplement type has little to no impact on tooth growth.

4) Conclusions

As stated, our conclusions have relied on the assumptions that the populations are independent from one another, they are not paired and that they have unequal variances.
As a result we have seen from the hypothesis tests that at a dosage of 2mg neither type of supplement is better than the other for prompting tooth growth.
However, for lower doses there is definite evidence that one of the supplement types has a stronger impact on tooth growth than the other. By looking at our boxplots, it is clear that orange juice is the supplement more strongly related to tooth growth. Therefore, out tests have fully supported the inferences gained from the boxplot, indicating that the analysis has been performed correctly.
In conclusion, if you are only able to administer small doses of Vitamin C to encourage tooth growth in Guinea Pigs, use orange juice to get the best results.

Appendix

All chunks set with "eval=FALSE" to prevent unnecessary re-running of code

Preliminary code:

```
library(lattice)
library(xtable)
library(Hmisc)
library(datasets)
data(ToothGrowth)
```

Appendix A Code:

```
bwplot(len ~ as.factor(dose)|supp, data=ToothGrowth,
       xlab="Dose of Vitamin C (mg)", ylab="Tooth Length",
       main="Variance of tooth length by supplement and dosage")
```

Appendix B Code:

```
Dose0.5 <- t.test(len~supp, paired=F, var.equal=F, data=ToothGrowth[ToothGrowth$dose==0.5,])
Dose1.0 <- t.test(len~supp, paired=F, var.equal=F, data=ToothGrowth[ToothGrowth$dose==1.0,])
Dose2.0 <- t.test(len~supp, paired=F, var.equal=F, data=ToothGrowth[ToothGrowth$dose==2.0,])

Dose0.5pval <- Dose0.5$p.value
Dose0.5lowconf <- Dose0.5$conf[1]
Dose0.5highconf <- Dose0.5$conf[2]

Dose1.0pval <- Dose1.0$p.value
Dose1.0lowconf <- Dose1.0$conf[1]
Dose1.0highconf <- Dose1.0$conf[2]

Dose2.0pval <- Dose2.0$p.value
Dose2.0lowconf <- Dose2.0$conf[1]
Dose2.0highconf <- Dose2.0$conf[2]

TestResults <- cbind(c(Dose0.5pval, Dose1.0pval, Dose2.0pval),
                     c(Dose0.5lowconf, Dose1.0lowconf, Dose2.0lowconf),
                     c(Dose0.5highconf, Dose1.0highconf, Dose2.0highconf))
rownames(TestResults) <- c("0.5 Dosage","1.0 Dosage","2.0 Dosage")
colnames(TestResults) <- c("P-Value","LowerConfInt","UpperConfInt")

print.xtable(xtable(TestResults, digits=5), comment=FALSE)
```