ToothGrowth Exploratory Analysis

thecapacity

January 24, 2015

Overview

This is the project file for my Peer Assignment of the Statistical Inference class.

I will seek to analyze the ToothGrowth data in the R datasets package, by:

1. Loading the ToothGrowth data
2. Performing some basic exploratory data analyses
3. Providing a basic summary of the data.
4. Using confidence intervals and/or hypothesis tests to compare tooth growth by `supp` and `dose`. *Note I will only use techniques from class*
5. Stating my conclusions and the assumptions needed for those conclusions.

I will create a report to answer the questions. Given the nature of the series, ideally knitr will be used to create the reports and convert to a pdf. The pdf report will be **no more than 6 pages total including supporting material** if needed (code, figures, etcetera).

This will also be published at: http://rpubs.com/thecapacity/StatInf_Proj2 (http://rpubs.com/thecapacity/StatInf_Proj2)

Approach

I will plan to ensure I cover the following activities:

- Perform an exploratory data analysis with at least a single plot or table highlighting basic features of the data?
- Perform some relevant confidence intervals and/or tests?
- Review the results of the tests and/or intervals and interpret them in the context of the problem?
- Describe the assumptions needed for conclusions?

Load Data

Here we focus on loading the ToothGrowth data from the R datasets package, and provide some initial information on the data size and structure.

According to this page (https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/ToothGrowth.html) the `ToothGrowth` dataset records the effect of Vitamin C on Tooth Growth in Guinea Pigs.

We load the data via:

```
# Load the Data
data("ToothGrowth")
str(ToothGrowth)
```

```
## '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 ...
```

```
# attach(ToothGrowth)
```

The data frame correctly has `60` observations on `3` variables, where:

- `len` indicates overall *Tooth length*
- `supp` indicates *Supplement type (`VC` or `OJ`)*
- `dose` indicates the *Dose in milligrams*

```
# A small sample
head(ToothGrowth, 5)
```

```
##    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
```

Exploratory Analysis

Using confidence intervals and/or hypothesis tests to compare tooth growth by `supp` and `dose`, I will perform some basic exploratory data analyses that will be collected and visualized, to be summarized in the next section.

This will include:

- At least a single plot or table highlighting basic features of the data
- A relevant confidence intervals and/or other potential tests

Note I will only use techniques from class.

First let us collect and visualize some basic statistical measures:

Overall

```
mn_o <- mean(ToothGrowth$len)
sd(ToothGrowth$len)
```

```
## [1] 7.649315
```

By Supp

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

```
## [1] 20.66333
```

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

```
## [1] 6.605561
```

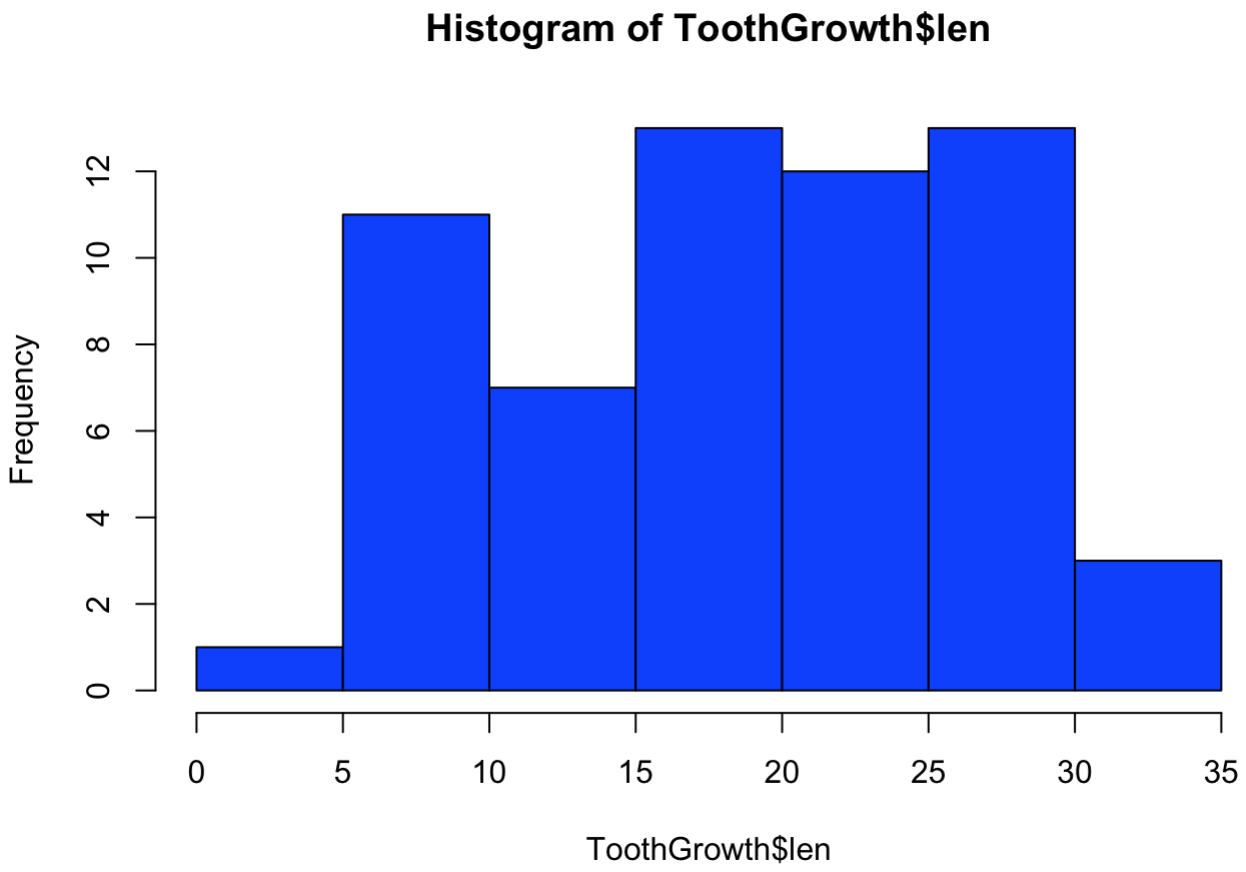
```
mean(ToothGrowth[ToothGrowth$supp=="VC",]$len)
```

```
## [1] 16.96333
```

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

```
## [1] 8.266029
```

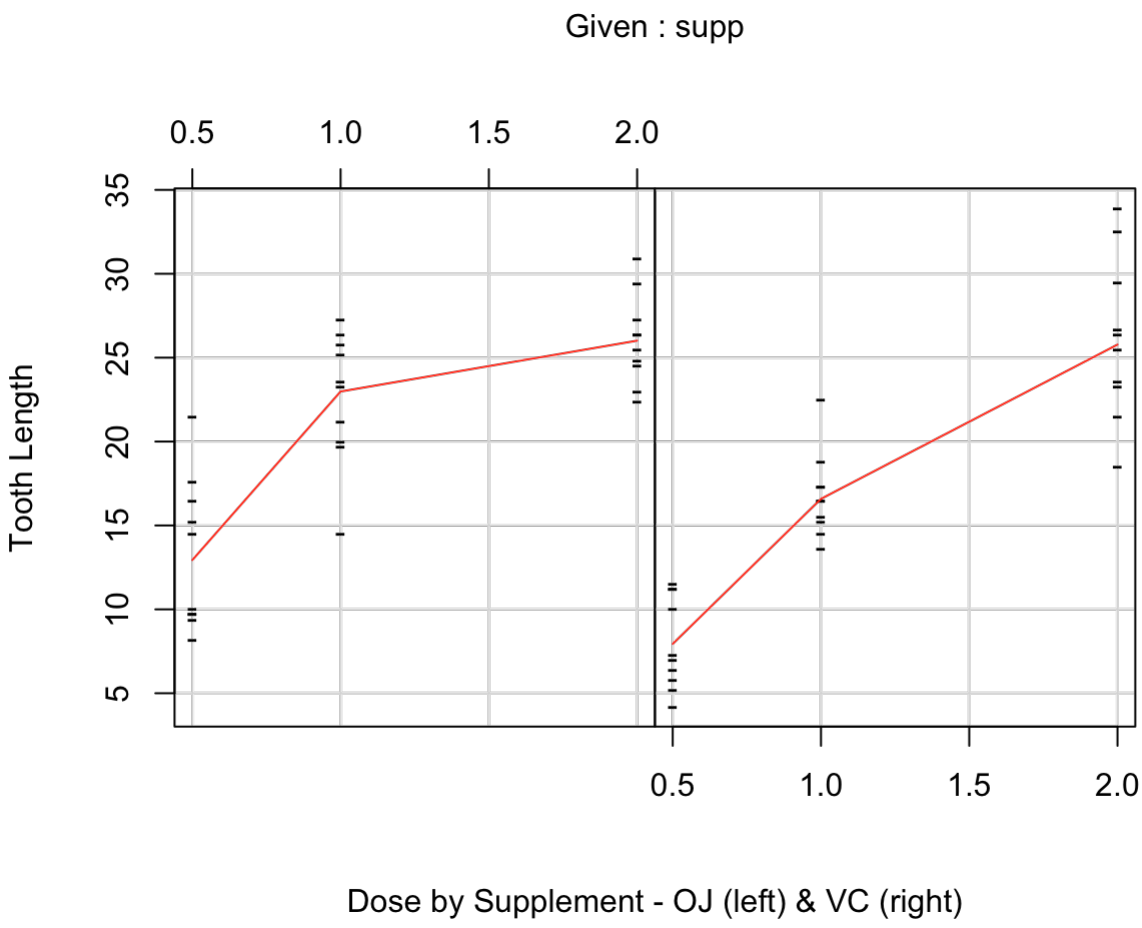
```
hist(ToothGrowth$len, col="blue")
```



The first thing we notice is that the is something less than a normal distribution to the tooth lengths, however even with the seeming deviation we will consider this acceptable.

The next item to explore is whether or not the particular supplement, and dosage, has a discernible effect:

```
coplot(len ~ dose | supp, data = ToothGrowth, ylab="Tooth Length", xlab="Dose by Supplement - OJ (l
eft) & VC (right)", panel=panel.smooth, show.given=FALSE, pch="-")
```



The co-plot suggests that there is little difference between either supplement at high doses, but at lower levels the `OJ` group (*left pannel*) seems to have had longer teeth.

This may suggest two elements worth further experimental research:

- Is there an upper limit of dosage effectiveness, as seems to be indicated on our graph?
- Does OJ include additional supplements that combine with the intrinsic vitamin C to enhance effective growth promotion?

However, before we conclude our exploratory analysis with any type of 'eureka moment' we should evaluate our confidence intervals.

This may be particularly important and a **paired interval** will be used because each of 10 guinea pigs was given all three dose levels of Vitamin C (0.5, 1, and 2 mg) via each of two delivery methods (orange juice or ascorbic acid), and should not be treated as an independent group.

```
x1 <- ToothGrowth[ToothGrowth$supp=="OJ",]$len
x2 <- ToothGrowth[ToothGrowth$supp=="VC",]$len

res <- t.test(x2, x1, paired=TRUE)
res
```

```
##
##  Paired t-test
##
## data:  x2 and x1
## t = -3.3026, df = 29, p-value = 0.00255
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.991341 -1.408659
## sample estimates:
## mean of the differences
##                    -3.7
```

Given the 'reuse' of Guinea Pig 'patents' it seems reasonable (particularly based on lecture discussions) to assume equal variance for both groups.

Note, this is tested across all three dosage levels, which again seems reasonable given the consistency/re-use of subjects.

From the paired T-test we can see that the `95%` confidence interval is `-5.9913414, -1.4086586` and our P-value is `0.0025498`, which does not include `0`. This implies that we have enough evidence to reject the null hypothesis and, given the order of comparison, the `VC` is gauged to be less effective than the `OJ` supplement.

Summary

The following section provides a basic summary of the data, including the tests and/or intervals used, along with contextual interpretation.

In summary, we find willing evidence to suggest Orange Juice (`OJ`) supplementation has greater efficacy than supplementing with Vitamin C (`VC`) alone, for low dosages.

Conclusions

The following discussion focused on restating my conclusions and the assumptions needed for those conclusions and further describing assumptions underlying the analysis and summary.

Our conclusions are built on some key assumptions that are worth reviewing. These include:

- Assuming a relatively normal distribution to the tooth lenghts
- A visual assessment (based on the `coplot(...)`) of differences between supplement at varrying doses
- A paired interval should be used
- The assumptions of equal variance for both groups
- A single confidence test across all three dosage levels
- Effective independence of dosage levels across the experiment.

The last point is worth considering further: None of the data included time-observations so it is assumed that each dosage test was given a reasonable 'washout' period. However, if there were delayed 'uptake' that affected growth conditions then this implicit assumption could have significant negative affects on the reasonableness of interpretation. However, given Vitamin C is relatively soluble, digestible and easily processed & excreted by biological organisms it seems reasonable.