Statistical Inference part 2

Alun Meredith

26 September 2015

Introduction

The dataset ToothGrowth documents the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. Three seperate dose levels of vitamin C (0.5, 1, 2 mg/day) and two different deliver methods (Orange juice and ascorbic acid) were used. There were 6 different groups of 10 guinea pigs for each combination of method and dose, with no control group. We shall be looking at the results of this experiment and try to establish the confidence for which we can say that vitamin C dosages and its different delivery mechanisms have on the length of odontoblasts.

Exploratory Analysis

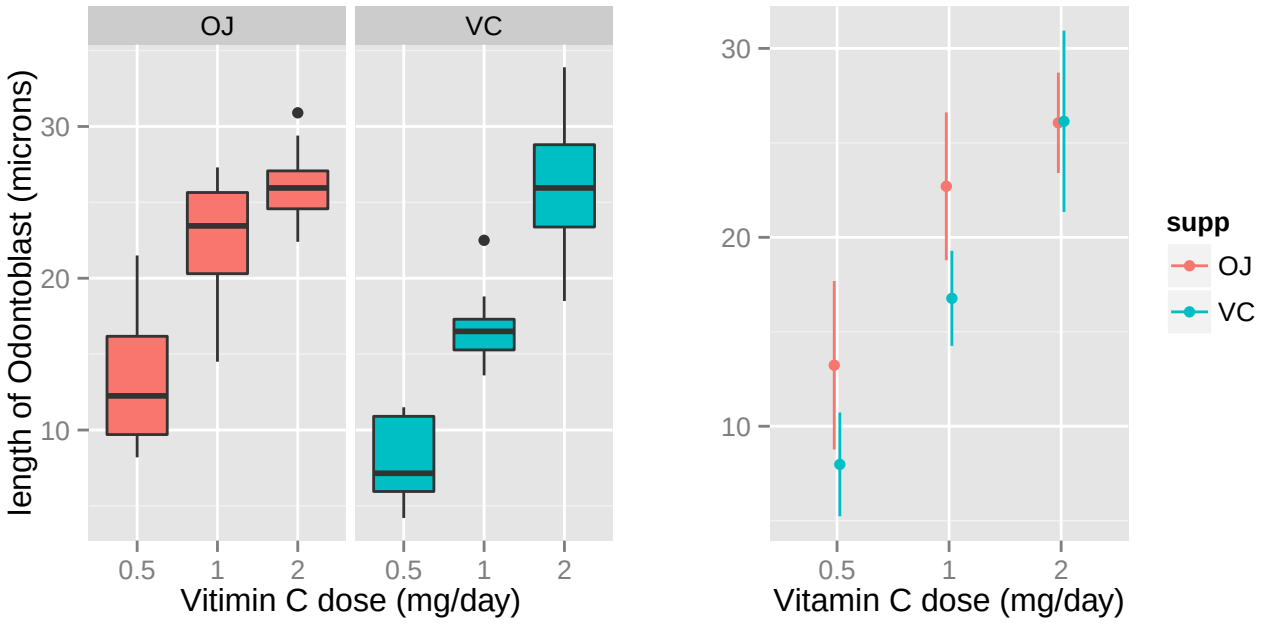
For each animal is fed in addition to its diet of hay one of three levels of vitamin C intake by either an aqueous solution of ascorbic acid or orange juice from an age of 28 +/- 3 days for a duration of 42 days. At this point the length of odontoblasts is measured. The table below shows the mean and variance for each of the groups of animals based on the diet they were fed.

```
ToothGrowth$dose = as.factor(ToothGrowth$dose)
grouped <- group_by(ToothGrowth, dose, supp) %>%
  summarise(Mean = mean(len), Variance = var(len))
```

Table 1: Mean and variance for length of odontoblasts (mm) in groups of 10 Guinea pigs based on 3 dosages (mg/day) and two supply methods (VC: ascorbic acid and OJ: Orange Juice)

dose	supp	Mean	Variance
0.5	OJ	13.23	19.89
0.5	VC	7.98	7.54
1	OJ	22.70	15.30
1	VC	16.77	6.33
2	OJ	26.06	7.05
2	VC	26.14	23.02

We can visualise this information by plotting the Mean +/- standard deviation for each group. We have also constructed a box and wisker plot of this information showing the InterQuartile ranges.



The plot above on the right shows the recorded mean for length for each subset of does and delivery mechanism. The vertical bars show an error 1 standard deviation from the mean. This represents a 68% confidence from a normal population distribution. The left plot shows the quartiles of each of these sample groups and seperating some outliers.

Comparing the interquartile ranges of different doses while controlling for delivery method we see that the only overlap is between 1 and 2 mg/day dose for Orange Juice. This observation is mirrored by the standard deviation ranges and suggests with reasonable confidence the relationship between Vitamin C dose and Odontoblast length.

However comparing both the box and wisker interquartile and standard deviation ranges across the two delivery methods while keeping the dosages constant we see an overlap of these ranges universally with the exception of the 2mg/day dose which doesn't overlap its interquartile range between orange juice and ascorbic acid. This suggests that although the orange juice is generally seen with higher Odontoblast lengths it is difficult to assign confidence to this observation.

Hypothesis testing

One measurement of the mean was taken from our sample of 10, this low degree of freedom suggests a student's t distribution would be more suitable than a normal distribution to assess the hypothesis.

To compare the two delivery methods we will use the independent Group t confidence interval approximation for unequal variances, where x and y are the two groups, orange juice and ascorbic acid:

$$\bar{Y} - \bar{X} \pm t_{df} \times (\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y})^{\frac{1}{2}}$$

$$df = \frac{(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y})^2}{\frac{(\frac{s_x^2}{n_x})^2}{n_x - 1} + \frac{(\frac{s_y^2}{n_y})^2}{n_y - 1}}$$

While considering a t distribution we can detirmine the p values of the main groupings of data. The delivery mechanisms and the doses.

First calculating the p value sampling all data split into two delivery mechanism groups.

```
t.test(len~supp, data=ToothGrowth)$p.value
```

```
## [1] 0.06063451
```

This shows a p value slightly weaker than the 95% confidence interval generally accepted as significant. It suggests that, while optimisitic the delivery mechanism may have an effect on the length of dodontoblasts, this data alone is not enough to establish this link.

```
t.test(ToothGrowth[ which(ToothGrowth$dose == 0.5),"len"],
       ToothGrowth[ which(ToothGrowth$dose == 1),"len"])$p.value

t.test(ToothGrowth[ which(ToothGrowth$dose == 1),"len"],
       ToothGrowth[ which(ToothGrowth$dose == 2),"len"])$p.value
```

```
## [1] 1.268301e-07
## [1] 1.90643e-05
```

These two tests, the independent group t confidence intervals for unequal variance, show very small p-values for the difference between doses [0.5, 1] mg/day (first value) and [1, 2] mg/day (second value). The small values represent the very small likelihood that these values are seperated by chance, strongly supporting increased doses of vitamin C in the diet to odontoblast length.

Going a little more in depth we can compare the t confidence intervals for the smaller groups where we vary only one parameter. Below is the 3 p-values comparing Orange Juice to ascorbic acid for each of the 3 dose levels. Followed by the p values comparing 0.5 and 1 mg/day doses for given delivery methods (orange juice then acid) and again comparing 1mg/day to 2mg/day.

```
ToothGrowth$id = rep(1:10,3)
ttest = ToothGrowth %>%
  spread(dose, len) %>%
  select(-id)
ttest2 = ToothGrowth %>%
  spread(supp, len) %>%
  select(-id)

ttest %>% summarise_each(funs(t.test(.[supp=="VC"], .[supp == "OJ"])$p.value), dose = 2:4)
ttest2 %>% summarise_each(funs(t.test(.[dose == 0.5], .[dose == 1])$p.value), supp = 2:3)
ttest2 %>% summarise_each(funs(t.test(.[dose == 1], .[dose == 2])$p.value), supp = 2:3)
```

```
##         dose1       dose2     dose3
## 1 0.006358607 0.001038376 0.9638516
##          supp1        supp2
## 1 8.784919e-05 6.811018e-07
##        supp1        supp2
## 1 0.03919514 9.155603e-05
```

The values above afirm our previous calculations that dose is strongly significant (p-values < 10^-5) and delivery method is weakly significant (p-values ~ 10^-2) but show two outliers; the dose 2mg comparison of delivery method and orange juice delivery comparing 1mg/day to 2mg/day. These two low confidence estimates match the datapoints with intervals overlapping in our exploratory analysis earlier.

Both of the outlying group p-values include data from Orange Juice given at 2mg/day data so it is plausible that this set of observations has some outlying data which is strongly effecting the results. Another explanation could be that for orange juice there is a ceiling to the dosage intake which is causing the length to fall off with respect to what was expected.

While the t tests show confidence in the effect of dose and hope for the effect of delivery mechanisms due to a borderline confidence interval, the small number of observations and presence of some likely outlying data suggests further observations are required to evaluate this effect fully.