Statistical Inference: Assignment 2

ND

17/09/2014

Introduction

This document contains a brief analysis of the ToothGrowth dataset provided with R. The dataset contains the results of an experiment to investigate the effect of vitamin C on tooth growth in guinea pigs. The R markdown source for this document is available on github.

Exploratory data analysis

The first step is to load the datasets package using the command `library(datasets)`. This package contains the ToothGrowth dataset and many more besides. The full list can be found using the command `library(help = "datasets")`.

```
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 ...
```

Using `str` to display the structure of the dataset reveals a dataframe of three variables: `len`, the measured tooth length; `supp`, a factor indicating supplement type (orange juice or vitamin C); and `dose`, the dose in milligrams.

```
with(ToothGrowth, table(dose, supp))
```

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

The table of dose values and supplement types confirms that there are 10 observations for each of the six combinations of dose level and supplement type.

Plotting the data is a good way to get a feel for its range and distribution. The following plots are created using ggplot2, loaded with the command `library(ggplot2)`.

```
ggplot(ToothGrowth, aes(y = len, x = supp)) + geom_boxplot() +
  xlab("Supplement Type") + ylab("Tooth Length") +
  theme_bw()
```

```
ggplot(ToothGrowth, aes(y = len, x = factor(dose))) + geom_boxplot() +
  xlab("Dose Level") + ylab("Tooth Length") +
  theme_bw()
```

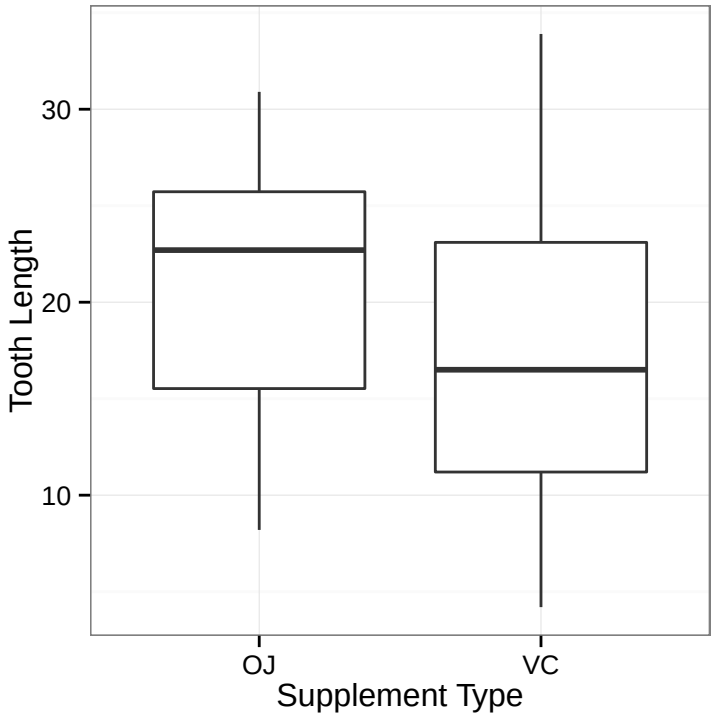

Figure 1: Relationship between tooth length and supplement type

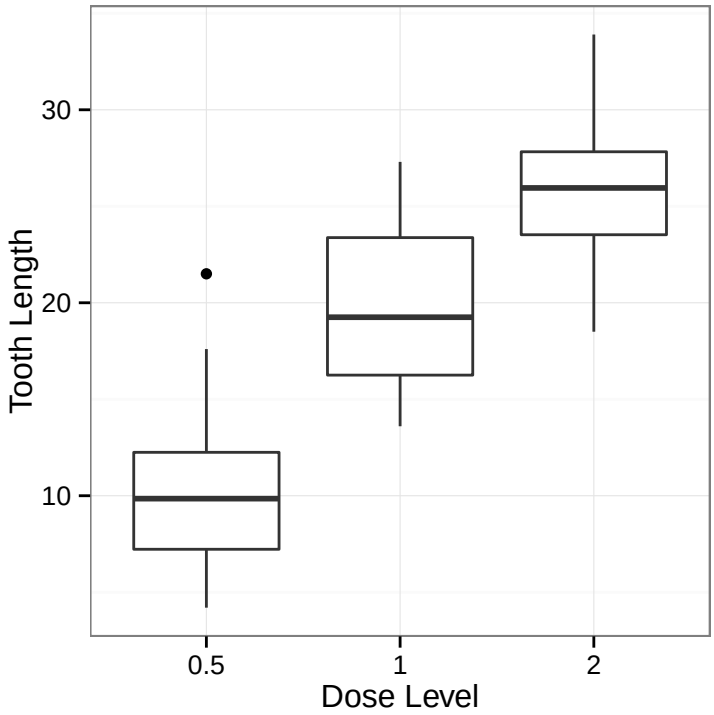

Figure 2: Relationship between tooth length and dose level

Figure 1 suggests there is a small difference between the two supplement types in terms of measured tooth length. The orange juice group has a greater median tooth length than the vitamin C group. Figure 2 shows a clear relationship between dose level and measured tooth length. As dose level increase, so does the measured tooth length.

Since there are two groups in the dataset it makes sense to use a facet plot that shows the differences between group as well as the differences between dose. The red lines in Figure 3 connect the mean measured tooth length of each dose group. A difference in mean between the two supplement types is evident at the lower dosages. At a dose of 2 mg the means are virtually indistinguishable.

```
ggplot(ToothGrowth, aes(y = len, x = dose)) + geom_point() +
  xlab("Dose") + ylab("Tooth Length") + ggtitle("Supplement Type") +
  stat_summary(fun.y=mean, colour="red", geom="line", aes(group = 1)) +
  facet_grid(. ~ supp) +
  theme_bw()
```

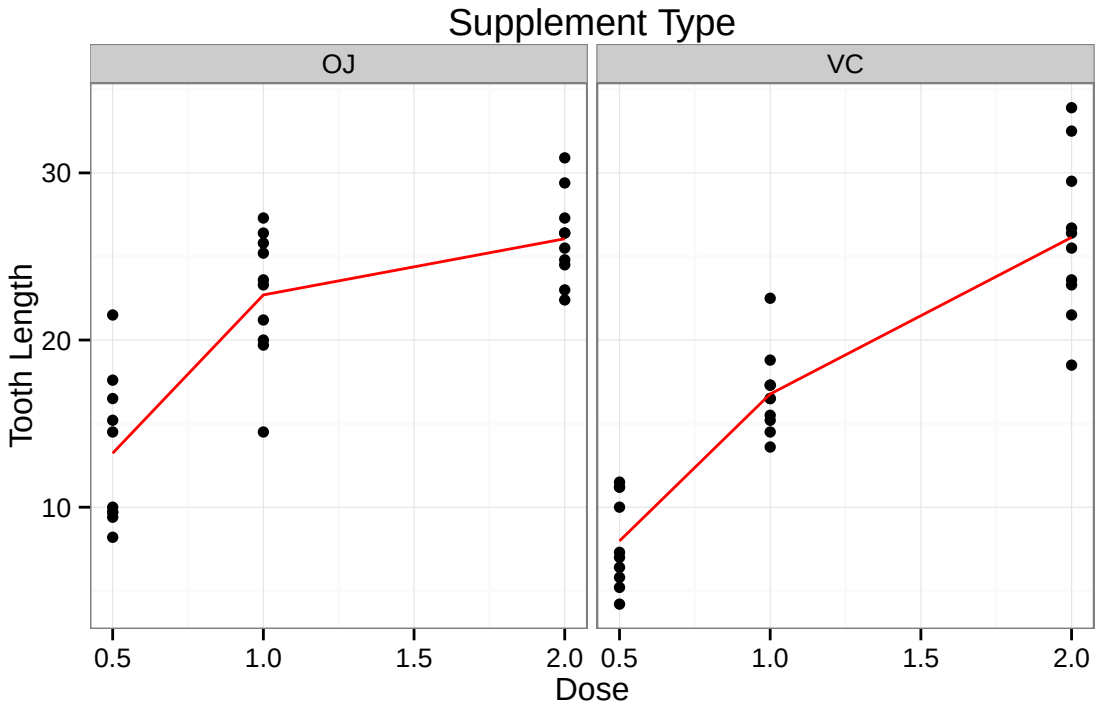

Figure 3: Relationship between tooth length and dose level by supplement type

Further testing is required to investigate the statistical significance of the differences between supplement group and dose level.

Data summary

The initial data analysis showed the two supplement factors and three dose levels. The mean measured tooth length across all the whole dataset was 18.8133, with a standard deviation of 7.6493. The means and standard deviations of the length variable for different combinations of supplement and dose are shown in the following summary tables.

```
library(data.table)
data <- data.table(ToothGrowth)

library(xtable)
options(xtable.comment = FALSE)
```

```
print(xtable(data[, list(mean=mean(len), sd=sd(len)), by=c("supp")],
             caption="Mean and standard deviation by supplement type"),
      include.rownames=FALSE)
```

supp	mean	sd
VC	16.96	8.27
OJ	20.66	6.61

Table 1: Mean and standard deviation by supplement type

The `xtable` package is used to create tidier looking alternatives to the standard output. The `data.table` package is used here simply for its convenient ability to summarise multiple groups by multiple functions. The following tables are generated in the same way as Table 1 (the code may be viewed in the source document).

dose	mean	sd
0.50	10.61	4.50
1.00	19.73	4.42
2.00	26.10	3.77

Table 2: Mean and standard deviation by dose level

dose	supp	mean	sd
0.50	VC	7.98	2.75
1.00	VC	16.77	2.52
2.00	VC	26.14	4.80
0.50	OJ	13.23	4.46
1.00	OJ	22.70	3.91
2.00	OJ	26.06	2.66

Table 3: Mean and standard deviation by dose and supplement type

Statistical testing

Different supplement at all dose levels

The following t.test considers the difference between supplement type in terms of its relationship with mean measured tooth length across all the dose levels (this is the comparison shown in Figure 1).

```
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
```

Since the confidence interval includes a small negative region, there is a small chance that the true difference in means is negative, i.e. the VC mean could be larger than the OJ mean, so it is not possible

to conclude at the 95% confidence level that that OJ is associated with greater tooth length than VC across all dose levels.

Different supplement at the same dose level

A set of t-tests were performed to consider the difference between supplement type at each of the three dose levels (this comparison can be made using Figure 3). To avoid cluttering the report with the full output from the t-test function its return values are first assigned to an object and then the required values (lower and upper confidence limits, means and p-values) are copied into the dataframe `t.dose` for output with `xtable` (this code can be seen in the document's source). The test conditions are the same as before: Welch Two Sample t-test with the alternative hypothesis that the true difference in means is not equal to 0.

comparison	dose	lower	upper	mean.x	mean.y	p.value
OJ and VC	0.50	1.72	8.78	13.23	7.98	0.01
OJ and VC	1.00	2.80	9.06	22.70	16.77	0.00
OJ and VC	2.00	-3.80	3.64	26.06	26.14	0.96

Table 4: t-test results comparing supplement types by dose level

The alternative hypothesis effectively states that the means are different. Table 4 shows the results of the t-tests on the mean tooth length with orange juice (x) and vitamin C (y). At the first two dose levels, the alternative hypothesis can be accepted at the 95% confidence level as the interval is completely above zero. At the 2.0 mg dose level the alternative hypothesis is rejected at the 95% confidence level because the interval includes zero (i.e. there could be no difference in means). The size of the p-values also reflect this.

Different dose with the same supplement

A final set of t-tests with the same conditions were performed to consider the difference between dose levels for each of the two supplement types (this comparison can also be made using Figure 3).

comparison	supp	lower	upper	mean.x	mean.y	p.value
1.0 mg and 0.5 mg	OJ	5.52	13.42	22.70	13.23	0.00
2.0 mg and 1.0 mg	OJ	0.19	6.53	26.06	22.70	0.04
1.0 mg and 0.5 mg	VC	6.31	11.27	16.77	7.98	0.00
2.0 mg and 1.0 mg	VC	5.69	13.05	26.14	16.77	0.00

Table 5: t-test results comparing dose levels by supplement type

The results show that in all the cases considered the alternative hypothesis can be accepted; there is a significant difference between the mean tooth length at the two dose levels. The larger of each dose level is associated with the greater tooth length.

Conclusions and assumptions

Tooth length appears to be associated with both supplement type and dose level. At dosage levels of 0.5 and 1.0 mg, orange juice is associated with greater tooth length than vitamin C. At a dosage level of 2.0 mg there is little difference between orange juice and vitamin C. Within the experiment's range, higher dosage levels are associated with greater tooth length for both supplement types.

To use the t-test is must be assumed that the populations are normally distributed. Due to the central limit theorem the distribution of sample means will approach a Normal distribution as the sample size increases. The shape of the box plots suggests the data are approximately normally distributed but this could be further investigated, for example using q-q plots.