Inferential Data Analysis

Kevin Tham

April 25, 2018

Inferential Data Analysis

In this section, we will analyse the dataset named ‘ToothGrowth’ in the R datasets package. We will undertake a statistical study of the data in order to test the effect of Vitamin C on the tooth growth of guinea pigs.

Exploratory Analysis

First we begin by loading the dataset and calling the `head()` and `str()` functions to give us a brief overview of the data.

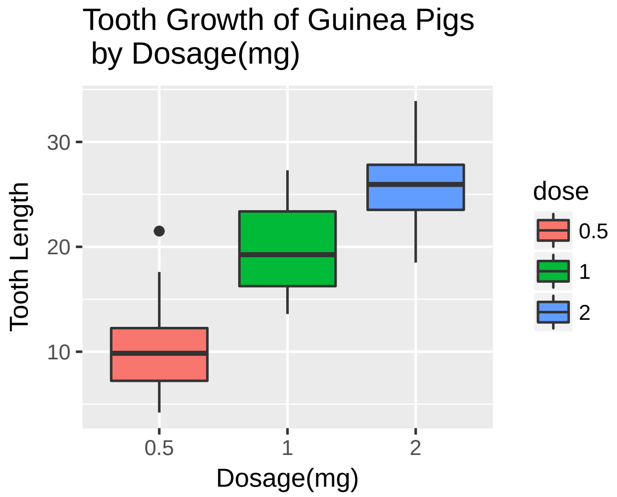
```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

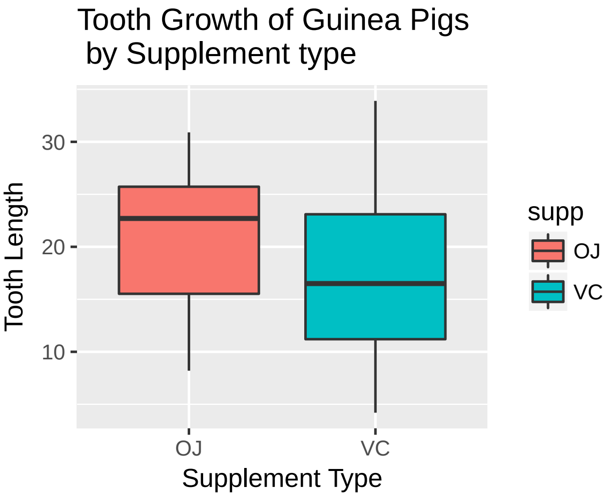
```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

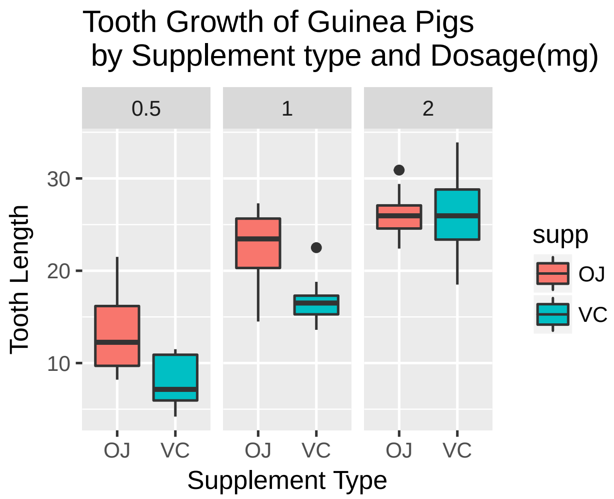
The dataset contains three variables and 60 observations. The variables correspond to:

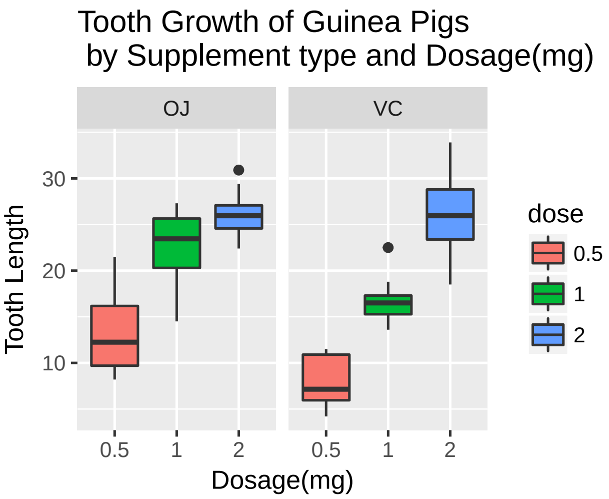
1. `len`: Tooth length
2. `supp`: Supplement type (VC:ascorbic acid, OJ:orange juice)
3. `dose`: Dose in milligrams/day

Since we want to compare numeric variables (`len`) against categorical data (`supp` and `dose`), we can utilise boxplots to display our dataset.









The boxplot highlights certain features of the data:

1. Increasing dosage appears to be correlated with increasing tooth length.
2. For 0.5mg and 1.0mg dosage, orange juice appears to result in larger toothlength than ascorbic acid, however for 2.0mg dosage it appears that the medians of the distributions are very similar.

Hypothesis Testing

In order to determine if the supplement type or dosage has an effect on the tooth growth of guinea pigs, and if possible, the magnitude of the effect. Therefore we will utilise hypothesis testing in order to achieve this.

In hypothesis testing, we have a null hypothesis (H_0, representing status quo or independence from variable being tested) and an alternative hypothesis (H_1, representing a positive effect). The aim is employ statistics to decide whether the null hypothesis H_0 should be rejected or not.

This is done by taking the difference in the means of each population being tested (for eg. splitting the data based on supplement type). The difference in means is treated as a random variable and the probability of it being zero (variable has no effect), the p-value, is calculated. If it is below a certain level, which we specify to be 0.05, then we reject the null hypothesis as the probability of the null hypothesis being true is low.

The following assumptions are made:

- Tooth growth follows a normal distribution (may or may not have different means and variance for different variables)
- Within each variable, tooth growth is independent and identically distributed (i.i.d.)

Hypothesis Testing of Dosage Levels

We begin by conducting hypothesis testing on the effect of dosage levels of Vitamin C on tooth growth. This is done by testing each possible dosage level pairwise. The null hypothesis here is that the dosage level does not affect tooth growth as a whole.

	t.statistic	DoF	p.value	Lower.CL	Upper.CL	Group.1.mean	Group.2.mean
Dose=0.5 vs Dose=1	-6.48	38.0	1.00e-07	-12.0	-6.28	10.6	19.7
Dose=1 vs Dose=2	-4.90	37.1	1.91e-05	-9.0	-3.73	19.7	26.1
Dose=0.5 vs Dose=2	-11.80	36.9	0.00e+00	-18.2	-12.80	10.6	26.1

From the above we can see that for all three combinations, the p-value is much smaller than 0.05. The confidence intervals do not contain 0 as well. Thus, there is a less than 5% probability of obtaining the results in the data if the null hypothesis is true and we reject it. We conclude that

Hypothesis Testing of Supplement Type

Next we test the effect of supplement type on tooth growth.

	t.statistic	DoF	p.value	Lower.CL	Upper.CL	Group.1.mean	Group.2.mean
Orange Juice vs Ascorbic Acid	1.92	55.3	0.0606	-0.171	7.57	20.7	17

This time the p-value of 0.0606 is larger than the specified value of 0.05. Therefore we fail to reject the null hypothesis and conclude that there is not enough evidence that supplement type affects tooth growth as a whole. However we would like to look deeper into this as grouping the data by dosage levels might review a different picture, as is suggested by the boxplots above.

Hypothesis Testing of Supplement Type grouped by Dosage Levels

Here, the data is grouped by the three different dosage levels first, and hypothesis tests are performed on each group to test the effect of supplement type.

t.statistic	DoF	p.value	Lower.CL	Upper.CL	OJ.mean	VC.mean	Dose
3.1700	15.0	0.00636	1.72	8.78	13.2	7.98	0.5
4.0300	15.4	0.00104	2.80	9.06	22.7	16.80	1.0
-0.0461	14.0	0.96400	-3.80	3.64	26.1	26.10	2.0

For dosage levels of 0.5mg and 1mg, the p-values are both below 0.05 with confidence intervals that do not contain zero, while for 2mg the p-value is much larger than 0.05. Therefore we reject the null hypothesis for 0.5mg and 1mg, and conclude with 95% confidence level that there is a difference in mean between the two supplement types for 0.5mg and 1mg, while there is no difference in mean for 2mg.

Appendix (R Script)

```r
rm(list=ls())
knitr::opts_chunk$set(echo = FALSE)
if (!require("pacman"))
  install.packages("pacman", repos = "http://cran.us.r-project.org")
pacman::p_load(knitr, dplyr, ggplot2, tidyr, grid, gridExtra)

data(ToothGrowth)
head(ToothGrowth)
str(ToothGrowth)

ToothGrowth$dose <- as.factor(ToothGrowth$dose)
g1 <- ggplot(ToothGrowth, aes(x=dose,y=len)) +
  geom_boxplot(aes(fill=dose)) +
  labs(title="Tooth Growth of Guinea Pigs \n by Dosage(mg)",
       x="Dosage(mg)", y="Tooth Length")

g2 <- ggplot(ToothGrowth, aes(x=supp,y=len)) +
  geom_boxplot(aes(fill=supp)) +
  labs(title="Tooth Growth of Guinea Pigs \n by Supplement type",
       x="Supplement Type", y="Tooth Length")

g3 <- ggplot(ToothGrowth, aes(x=supp,y=len)) +
  geom_boxplot(aes(fill=supp)) +
  facet_grid(.~ dose) +
  labs(title="Tooth Growth of Guinea Pigs \n by Supplement type and Dosage(mg)",
       x="Supplement Type", y="Tooth Length")

g4 <- ggplot(ToothGrowth, aes(x=dose,y=len)) +
  geom_boxplot(aes(fill=dose)) +
  facet_grid(.~ supp) +
  labs(title="Tooth Growth of Guinea Pigs \n by Supplement type and Dosage(mg)",
       x="Dosage(mg)", y="Tooth Length")

grid.arrange(g1, g2, ncol = 2)
grid.arrange(g3, g4, ncol = 2)

data0.5_1 <- filter(ToothGrowth, dose == 0.5 | dose == 1)
test0.5_1 <- t.test(len ~ dose, data = data0.5_1, paired = FALSE,
                    var.equal = FALSE, conf.level = 0.95)
res0.5_1 <- with(test0.5_1, data.frame(
  't-statistic' = statistic,
  'DoF' = parameter,
  'p-value' = p.value,
  'Lower CL' = conf.int[1],
  'Upper CL' = conf.int[2],
  'Group 1 mean' = estimate[1],
  'Group 2 mean' = estimate[2],
  row.names = 'Dose=0.5 vs Dose=1'
))

data1_2 <- filter(ToothGrowth, dose == 1 | dose == 2)
```

```r
test1_2 <- t.test(len ~ dose, data = data1_2, paired = FALSE,
                  var.equal = FALSE, conf.level = 0.95)
res1_2 <- with(test1_2, data.frame(
  't-statistic' = statistic,
  'DoF' = parameter,
  'p-value' = p.value,
  'Lower CL' = conf.int[1],
  'Upper CL' = conf.int[2],
  'Group 1 mean' = estimate[1],
  'Group 2 mean' = estimate[2],
  row.names = 'Dose=1 vs Dose=2'
))

data0.5_2 <- filter(ToothGrowth, dose == 0.5 | dose == 2)
test0.5_2 <- t.test(len ~ dose, data = data0.5_2, paired = FALSE,
                    var.equal = FALSE, conf.level = 0.95)
res0.5_2 <- with(test0.5_2, data.frame(
  't-statistic' = statistic,
  'DoF' = parameter,
  'p-value' = p.value,
  'Lower CL' = conf.int[1],
  'Upper CL' = conf.int[2],
  'Group 1 mean' = estimate[1],
  'Group 2 mean' = estimate[2],
  row.names = 'Dose=0.5 vs Dose=2'
))

res_dose <- res0.5_1 %>% bind_rows(res1_2) %>% bind_rows(res0.5_2)
row.names(res_dose) <- c('Dose=0.5 vs Dose=1', 'Dose=1  vs Dose=2',
                        'Dose=0.5 vs Dose=2')
kable(x = signif(res_dose, 3))

testsup <- t.test(len ~ supp, data = ToothGrowth, paired = FALSE,
                  var.equal = FALSE, conf.level = 0.95)

ressup <- with(testsup, data.frame(
  't-statistic' = statistic,
  'DoF' = parameter,
  'p-value' = p.value,
  'Lower CL' = conf.int[1],
  'Upper CL' = conf.int[2],
  'Group 1 mean' = estimate[1],
  'Group 2 mean' = estimate[2],
  row.names = 'Orange Juice vs Ascorbic Acid'
))

kable(x = signif(ressup, 3))

data0.5 <- filter(ToothGrowth, dose == 0.5)
test0.5 <- t.test(len ~ supp, data = data0.5, paired = FALSE,
                  var.equal = FALSE, conf.level = 0.95)

res0.5 <- with(test0.5, data.frame(
  't-statistic' = statistic,
  'DoF' = parameter,
  'p-value' = p.value,
  'Lower CL' = conf.int[1],
  'Upper CL' = conf.int[2],
  'OJ mean' = estimate[1],
  'VC mean' = estimate[2],
  'Dose' = 0.5
))

data1 <- filter(ToothGrowth, dose == 1)
test1 <- t.test(len ~ supp, data = data1, paired = FALSE,
                var.equal = FALSE, conf.level = 0.95)

res1 <- with(test1, data.frame(
  't-statistic' = statistic,
  'DoF' = parameter,
  'p-value' = p.value,
  'Lower CL' = conf.int[1],
  'Upper CL' = conf.int[2],
  'OJ mean' = estimate[1],
  'VC mean' = estimate[2],
  'Dose' = 1
))

data2 <- filter(ToothGrowth, dose == 2)
test2 <- t.test(len ~ supp, data = data2, paired = FALSE,
                var.equal = FALSE, conf.level = 0.95)

res2 <- with(test2, data.frame(
  't-statistic' = statistic,
  'DoF' = parameter,
  'p-value' = p.value,
  'Lower CL' = conf.int[1],
  'Upper CL' = conf.int[2],
  'OJ mean' = estimate[1],
  'VC mean' = estimate[2],
  'Dose' = 2
))

res_supdose <- res0.5 %>% bind_rows(res1) %>% bind_rows(res2)
kable(x = signif(res_supdose, 3))
```