Statistical Inference - Course Project - Part 2 - Basic Inferential Data Analysis

Ioannis Moros

01/07/2017

R Setup

```
require(ggplot2)
```

Introduction

Now in the second portion of the project, we're going to analyze the ToothGrowth data in the R datasets package.

1. Load the data and perform basic exploratory data analysis.

```
my_data <- datasets::ToothGrowth
```

```
str(my_data)
```

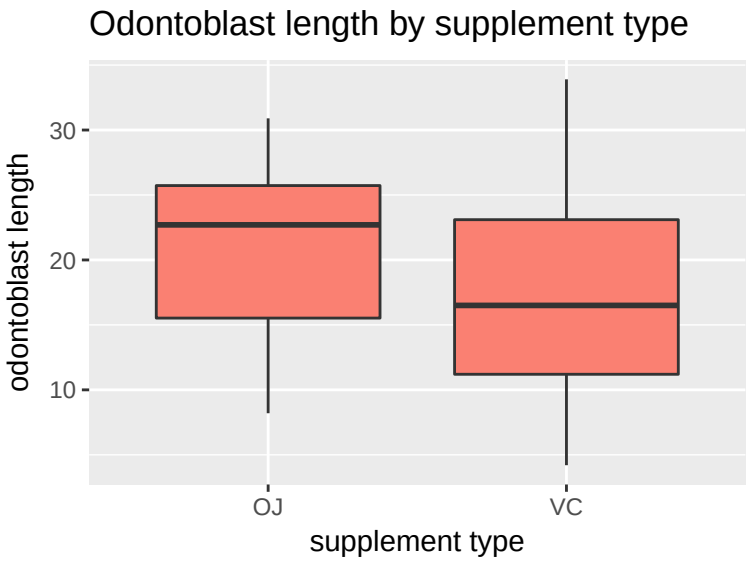
```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

```
summary(my_data)
```

```
##       len        supp         dose
##  Min.   : 4.20   OJ:30   Min.   :0.500
##  1st Qu.:13.07   VC:30   1st Qu.:0.500
##  Median :19.25           Median :1.000
##  Mean   :18.81           Mean   :1.167
##  3rd Qu.:25.27           3rd Qu.:2.000
##  Max.   :33.90           Max.   :2.000
```

The boxplot of odontoblast length by supplement type

```
g <- ggplot(my_data, aes(x = supp, y = len))
g + geom_boxplot(aes(group = supp), fill = "salmon") +
        xlab("supplement type") +
        ylab("odontoblast length") +
        ggtitle("Odontoblast length by supplement type")
```



2. Hypothesis tests and conclusions.

We want to investigate if there is a significant difference in odontoblast length between the 2 delivery methods, orange juice and ascorbic acid. We will perform a two-sided test using the paired t-test:

H_0: there is no difference in growth between the delivery methods

H_1: there is a difference

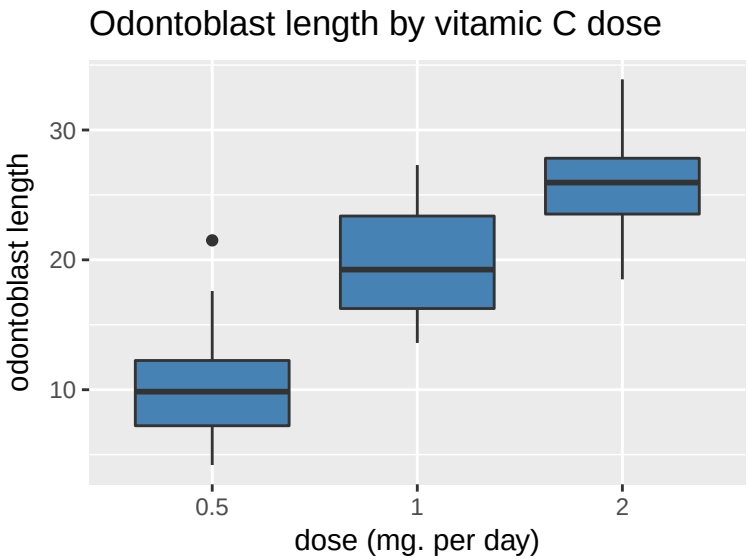
```
t.test(len ~ supp, data = ToothGrowth, paired = FALSE)
```

```
##
##  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 p-value of the t-test is $0.06 > 0.05$, so we retain the null hypothesis at the 95% level, i.e. we conclude that odontoblast growth is not significantly different between the two delivery methods, with 95% confidence. We observe that the confidence interval (-0.17, 7.57) contains 0, which is enough evidence to retain the null.

Now we investigate the effect of different vitamin C dosages.

```
p <- ggplot(my_data, aes(x = as.factor(dose), y = len))
p + geom_boxplot(aes(group = dose), fill = "steel blue") +
        xlab("dose (mg. per day)") +
        ylab("odontoblast length") +
        ggtitle("Odontoblast length by vitamic C dose")
```



It looks like higher dosages of vitamic C are associated with longer odontoblasts. Let's test that hypothesis for dose levels 0.5 and 1 mg/day. Again, we use the paired t-test, this time for a one-sided test

H_0: odontoblast is the same for vitamin C dose of 0.5 mg/day and 1 mg/day

H_1: odontoblasts are longer for dose of 1 mg/day

```
t.test(len ~ dose, data = subset(my_data, dose %in% c(1, 0.5)), alt = "less", paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -6.4766, df = 37.986, p-value = 6.342e-08
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -6.753323
## sample estimates:
## mean in group 0.5   mean in group 1
##            10.605            19.735
```

The p-value is practically 0 , so we reject the null for any reasonable significance level and we conclude that a dosage of 1mg/day is associated with longer odontoblasts than a dosage of 0.5 mg/day.

3. Assumptions:

1. Independence of observations - each guinea pig was given exactly one dosage via exactly one delivery method.
2. Normality - odontoblast length is approximately normally distributed for each group of the independent variables (supp, dose)
3. Homogeneity of variances.