Coursera Statistical Inference Project part 2

Ben de Haan

22 Aug 2015

Introduction and methods

This report discusses the R `ToothGrowth` dataset. The dataset is composed of results on the length of odontoblasts in guinea pigs based on different doses of vitamin C intake. The guinea pigs received either 0.5, 1, or 2 mg/day of vitamin C through orange juice (OJ) or asorbic acid (VC).

```
## '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 questions that this report aims to answer are: * What is the 95% confidence interval for odontoblast length for each dose and intake method? * Is there a difference in odontoblast length between the orange juice (OJ) and asorbic acid (VC) group? ** H_0 : there is no significant difference in odontoblast length between OJ and VC group * Is there a difference in odontoblas length between the groups for each different dose? ** H_0 : there is no significant difference in odontoblast length between groups for each dose

Questions will be answered using t-tests and confidence interval calculations by the T-distribution.

To test the difference between groups, the T-test is used. However, this assumes lengths are normally distributed. If the data is not normally distributed, the following solutions are available: 1. Confidence intervals can be retrieved through bootstrapping. 2. A log transformation of length may solve the abnormality

Exploratory Data Analysis

Summary statistics

Confidence intervals for the mean were retrieved as follows:

```
# Calculte the errors
oj.error <- qt(0.975, df = 29) * oj.sd / sqrt(30)
vc.error <- qt(0.975, df = 29) * vc.sd / sqrt(30)
tot.error <- qt(0.975, df = 58) * tot.sd / sqrt(60)

# Calculate the confidence intervals
oj.ci <- oj.mean + c(-1,1) * oj.error
vc.ci <- vc.mean + c(-1,1) * vc.error
tot.ci <- tot.mean + c(-1,1) * tot.error

# Add to summary statistics table
cis.lower <- c(oj.ci[1], vc.ci[1], tot.ci[1])
cis.upper <- c(oj.ci[2], vc.ci[2], tot.ci[2])
summary.stats$cis.lower <- cis.lower
summary.stats$cis.upper <- cis.upper
```

```
##                Group  Mean 95% CI lower 95% CI upper Variance
## 1 Orange juice (OJ) 20.66        18.20        23.13    43.63
## 2 Asorbic acid (VC) 16.96        13.88        20.05    68.33
## 3             Total 18.81        16.84        20.79    58.51
##   Standard deviation
## 1              6.606
## 2              8.266
## 3              7.649
```

The groups are of equal size:

```r
table(ToothGrowth$supp)
```

```
## 
## OJ VC 
## 30 30
```

As are the doses within groups:

```r
table(ToothGrowth[ToothGrowth$supp=="OJ",]$dose)
```

```
## 
## 0.5   1   2 
##  10  10  10
```

```r
table(ToothGrowth[ToothGrowth$supp=="VC",]$dose)
```

```
## 
## 0.5   1   2 
##  10  10  10
```

Figure 1: Odontoblast length by supplement and dose

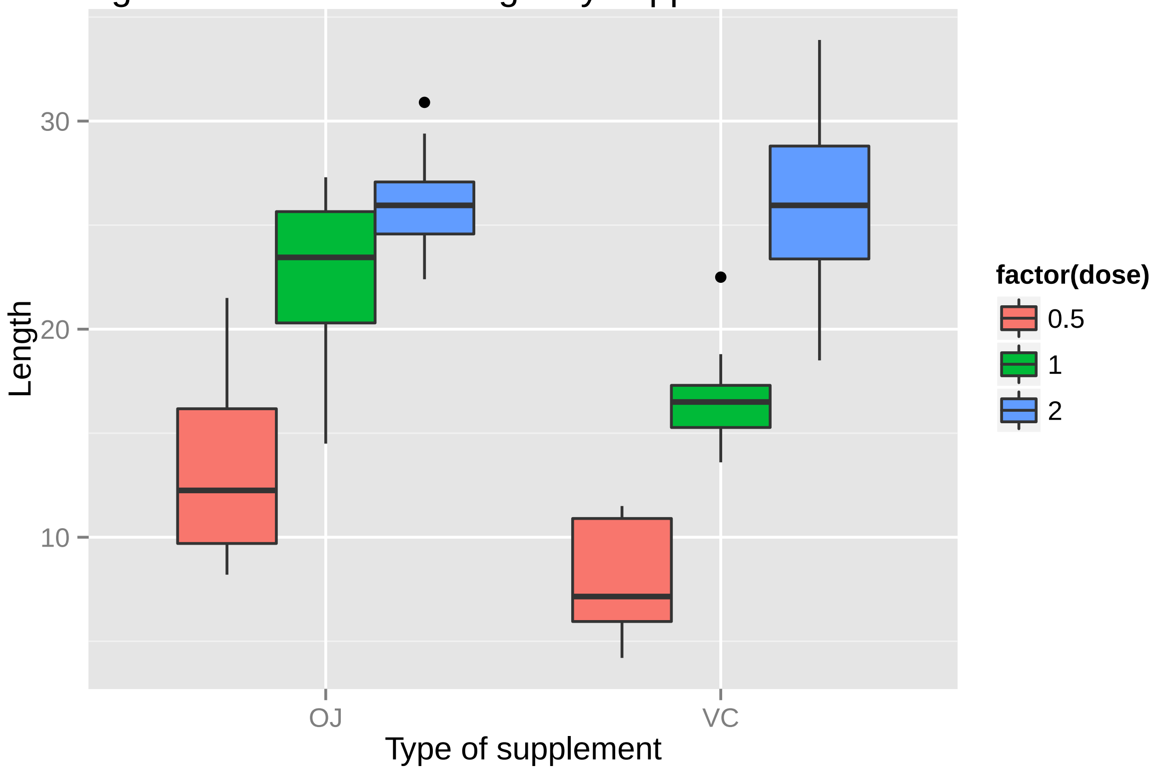
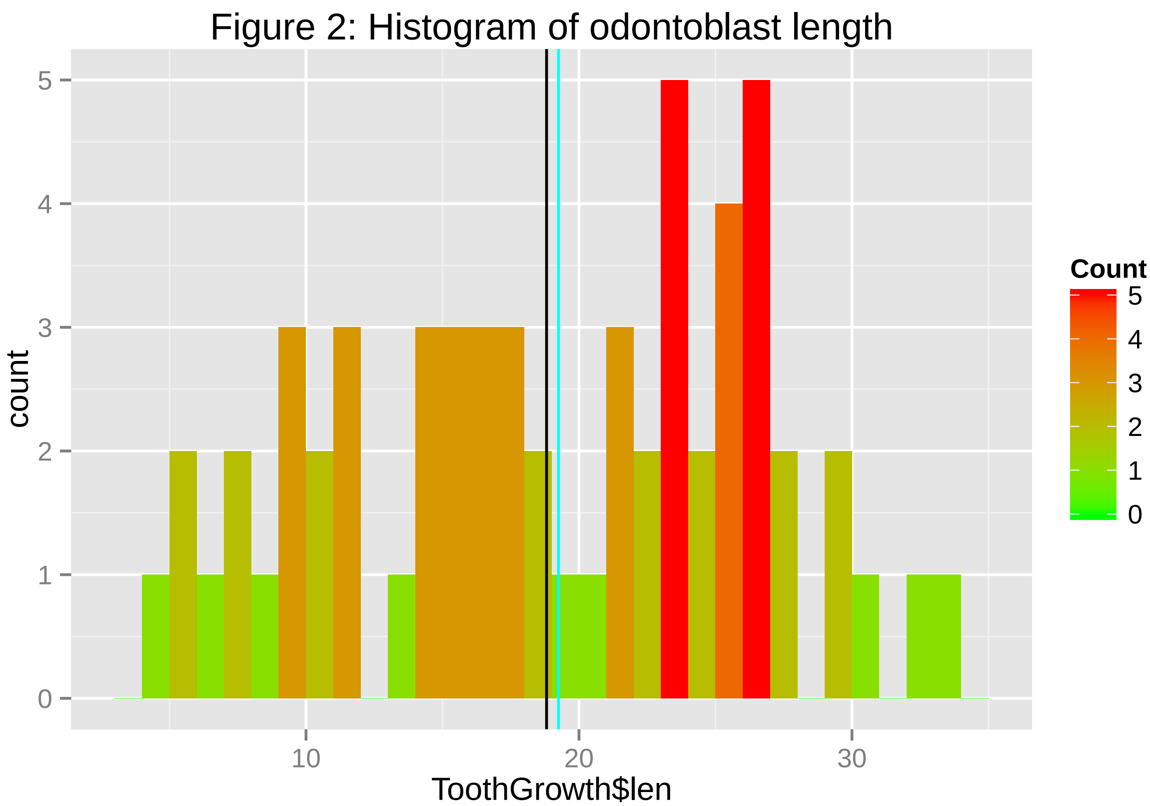
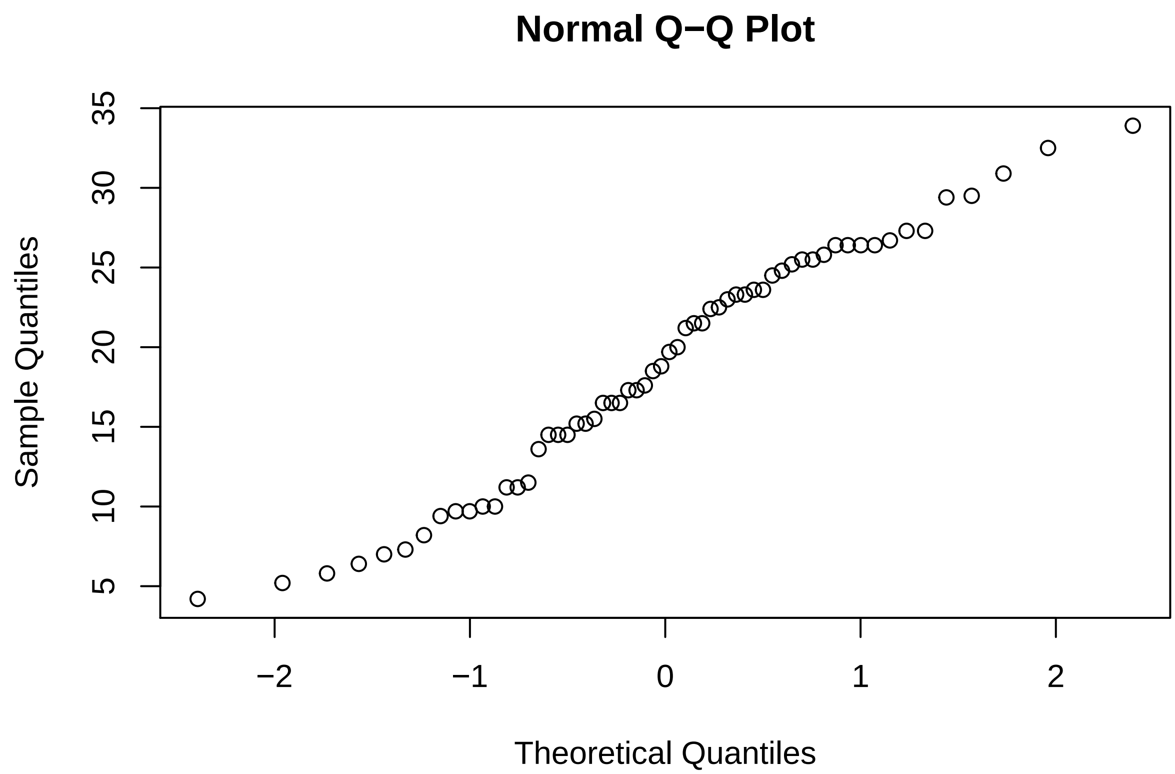


Figure 1 suggests some issues that might require further investigation: * Orange juice seems to be more effective at a lower dose, but evens out at the higher doses. * Higher doses seem to correspond with a higher odontoblast length

Normality

Odontoblast lengths seem to be approximately normally distributed, with the mean (black line) and median (cyan) quite close together (see figure 2).





The Shapiro-Wilk test of normality reveals that we cannot reject H_0 that the lengths are normally distributed ($p > 0.05$).

```
shapiro.test(ToothGrowth$len)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ToothGrowth$len
## W = 0.9674, p-value = 0.1091
```

Results

When we compare the two groups we see that p > 0.05. Therefore, we cannot reject H_0 that there is no significant difference between the groups.

```
t.test(ToothGrowth[ToothGrowth$supp=="OJ",]$len,
       ToothGrowth[ToothGrowth$supp=="VC",]$len)
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth[ToothGrowth$supp == "OJ", ]$len and ToothGrowth[ToothGrowth$supp == "VC", ]$len
## 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 of x mean of y
##     20.66     16.96
```

If we split up and compare between doses, we have different results.

For a dose of 0.5 mg/day:

```
t.test(ToothGrowth[ToothGrowth$supp=="OJ" & ToothGrowth$dose==0.5,]$len,
       ToothGrowth[ToothGrowth$supp=="VC" & ToothGrowth$dose==0.5,]$len)
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth[ToothGrowth$supp == "OJ" & ToothGrowth$dose == 0.5,  and ToothGrowth[ToothGrowth$s
## t = 3.17, df = 14.97, p-value = 0.006359
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.719 8.781
## sample estimates:
## mean of x mean of y
##     13.23      7.98
```

Since $p < 0.05$, we can reject H_0. This means that there is a difference between the supplement types for a dose of 0.5 mg/day.

For a dose of 1 mg/day:

```r
t.test(ToothGrowth[ToothGrowth$supp=="OJ" & ToothGrowth$dose==1,]$len,
       ToothGrowth[ToothGrowth$supp=="VC" & ToothGrowth$dose==1,]$len)
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth[ToothGrowth$supp == "OJ" & ToothGrowth$dose == 1,  and ToothGrowth[ToothGrowth$su
## t = 4.033, df = 15.36, p-value = 0.001038
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.802 9.058
## sample estimates:
## mean of x mean of y
##     22.70     16.77
```

Again, $p < 0.05$, we can reject H_0. This means that there is a difference between the supplement types for a dose of 1 mg/day as well.

For a dose of 2 mg/day:

```r
t.test(ToothGrowth[ToothGrowth$supp=="OJ" & ToothGrowth$dose==2,]$len,
       ToothGrowth[ToothGrowth$supp=="VC" & ToothGrowth$dose==2,]$len)
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth[ToothGrowth$supp == "OJ" & ToothGrowth$dose == 2,  and ToothGrowth[ToothGrowth$su
## t = -0.0461, df = 14.04, p-value = 0.9639
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.798  3.638
## sample estimates:
## mean of x mean of y
##     26.06     26.14
```

Here, $p > 0.05$, so we cannot reject H_0. This means that there is no significant difference between the supplement types for a dose of 2 mg/day.

Discussion

Assuming that odontoblas length is normally distributed, the results show that there is no significant difference between different intake methods for a dose of 2 mg/day. For a dose of 0.5 or 1 mg/day, orange juice was significantly more effective than asorbic acid.