Statistical inference project. Part 2.

Anton

Tuesday, September 16, 2014

Basic inferential data analysis

In this part of the project, we analyze the ToothGrowth data in the R datasets package. Load the data and look on its structure.

```
data("ToothGrowth")
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 ...
```

The variable *dose* should be a factor. Encode *dose* as a factor and summarize the data.

```
ToothGrowth$dose <- factor(ToothGrowth$dose)
summary(ToothGrowth)
```

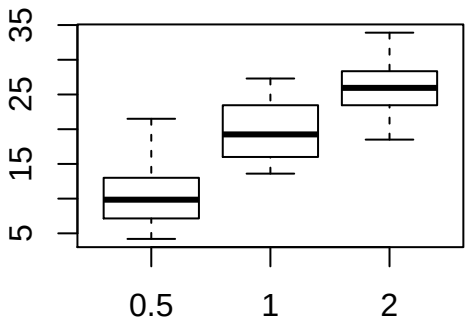
```
##       len       supp     dose
##  Min.   : 4.2   OJ:30   0.5:20
##  1st Qu.:13.1   VC:30   1  :20
##  Median :19.2           2  :20
##  Mean   :18.8
##  3rd Qu.:25.3
##  Max.   :33.9
```

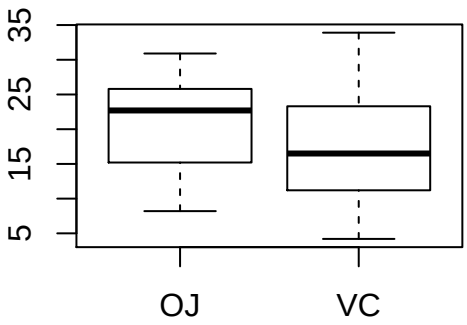
```
with(ToothGrowth, table(dose, supp))
```

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

There is no NA data and for each of ten pigs we have an observation of each dose level and each delivery method (in assumption that there is one group of 10 pigs).
Let's get started and look on data briefly:

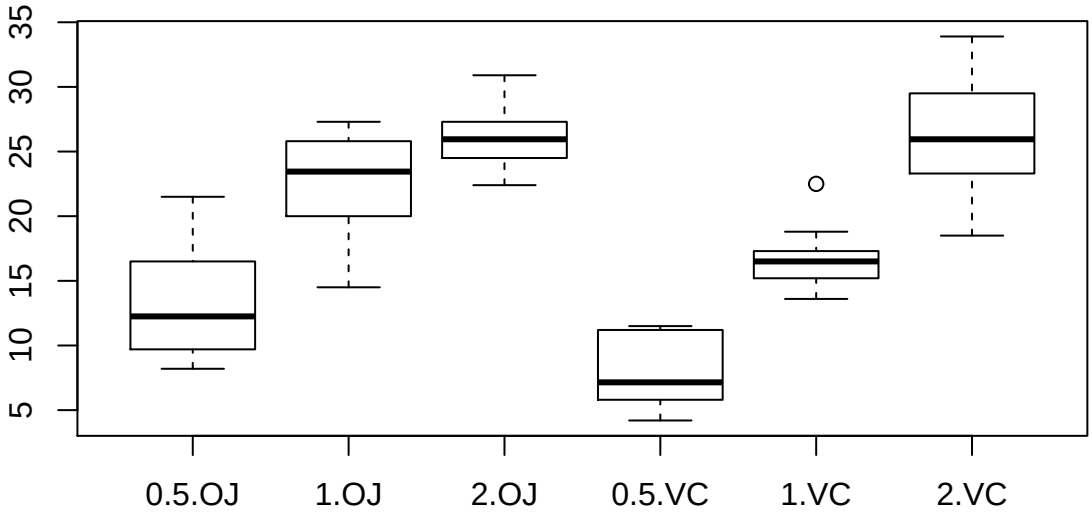
```
par(mfrow=c(1,2))
boxplot(len ~ dose, data = ToothGrowth)
boxplot(len ~ supp, data = ToothGrowth)
```





We can see that:
- the higher dose, the faster growth of teeth
- orange juice looks like more effective delivery method.
Now look at interaction of doses and delivery methods.

```
par(mfrow=c(1,1))
boxplot(len ~ dose*supp, data = ToothGrowth)
```



We can see that on small doses (0.5, 1.0) orange juice looks like more efficient.
But what about 2mg dose? Let's look more precisely:

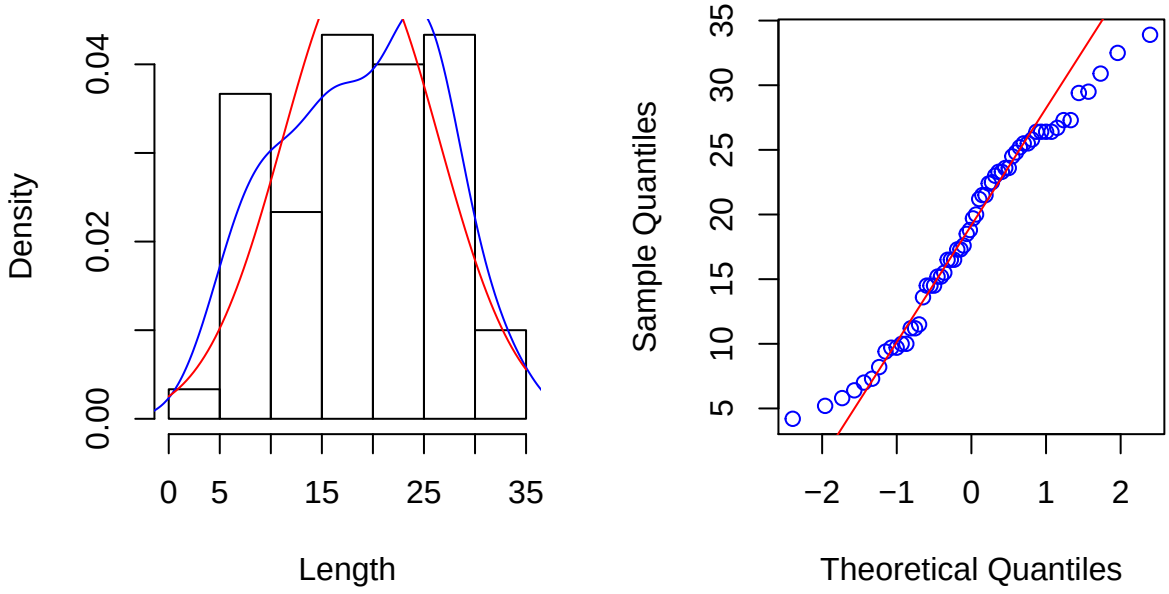
```r
agmean <- aggregate(len ~ ., data = ToothGrowth,  mean)
names(agmean)[3] <- "lmean"
agmedian <- aggregate(len ~ ., data = ToothGrowth, median)
names(agmedian)[3] <- "lmedian"
cbind(agmean, agmedian["lmedian"])
```

```
##   supp dose lmean lmedian
## 1   OJ  0.5 13.23   12.25
## 2   VC  0.5  7.98    7.15
## 3   OJ    1 22.70   23.45
## 4   VC    1 16.77   16.50
## 5   OJ    2 26.06   25.95
## 6   VC    2 26.14   25.95
```

For 2mg dose: the mean of length just a little bit higher for ascorbic acid while there is no difference in medians. Suppose there is no difference in delivery methods for this dose.

To prove our hypotheses let's make some tests.
We have small samples of same group so use paired t-test. Nevertheless, do we have a distribution close to normal?
The data looks relatively normal:

```r
par(mfrow=c(1,2))
hist(ToothGrowth$len, freq = FALSE, xlab = "Length", main = NULL)
lines(density(ToothGrowth$len), col = "blue")
curve(dnorm(x, mean(ToothGrowth$len), sd(ToothGrowth$len)), col="red", add = TRUE)
qqnorm(ToothGrowth$len, main="", col = "blue")
qqline(ToothGrowth$len, col="red")
```



Let's perform tests and look on confidence intervals and p-values:

- H_0 0.5mg dose, 1mg dose, 2mg dose have same effectiveness.

```
dose1 <- ToothGrowth[ToothGrowth$dose == 0.5, ]
dose2 <- ToothGrowth[ToothGrowth$dose == 1, ]
dose3 <- ToothGrowth[ToothGrowth$dose == 2, ]
t.test(dose2$len, dose1$len, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  dose2$len and dose1$len
## t = 6.967, df = 19, p-value = 1.225e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   6.387 11.873
## sample estimates:
## mean of the differences
##                    9.13
```

```
t.test(dose3$len, dose2$len, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  dose3$len and dose2$len
## t = 4.605, df = 19, p-value = 0.0001934
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.472 9.258
## sample estimates:
## mean of the differences
##                   6.365
```

```
t.test(dose3$len, dose1$len, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  dose3$len and dose1$len
## t = 11.29, df = 19, p-value = 7.19e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  12.62 18.37
## sample estimates:
## mean of the differences
##                   15.49
```

In all cases, we have small p-values and confidence intervals don't have a zero. Correctly reject null.

- H_0 - it's no matter orange juice or ascorbic acid.

```
oj <- ToothGrowth[ToothGrowth$supp == "OJ", ]
vc <- ToothGrowth[ToothGrowth$supp == "VC", ]
t.test(oj$len, vc$len, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  oj$len and vc$len
## t = 3.303, df = 29, p-value = 0.00255
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.409 5.991
## sample estimates:
## mean of the differences
##                     3.7
```

The p-value small enough and confidence interval don't have a zero. Correctly reject null.
However, we remember *2mg dose* case:

```
oj <- ToothGrowth[ToothGrowth$supp == "OJ" & ToothGrowth$dose==2, ]
vc <- ToothGrowth[ToothGrowth$supp == "VC" & ToothGrowth$dose==2, ]
t.test(oj$len, vc$len, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  oj$len and vc$len
## t = -0.0426, df = 9, p-value = 0.967
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.329  4.169
## sample estimates:
## mean of the differences
##                   -0.08
```

In this case we fail to reject null.