Analysis of the ToothGrowth dataset

Michal Siwek

Overview

This is part 2 of the Project Assignement for Statistical Inference course by JHU at Coursera. It analyzes the ToothGrowth data in the R datasets package. In particular it shows:

- basic data explorations
- basic summary of data
- comparison of tooth growth by supp and dose using t-tests
- assumptions needed for performing the t-tests

Exploratory Data Analysis and basic summaries of the data

Let's have a quick view of the data:

```
head(ToothGrowth)
```

```
##    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
```

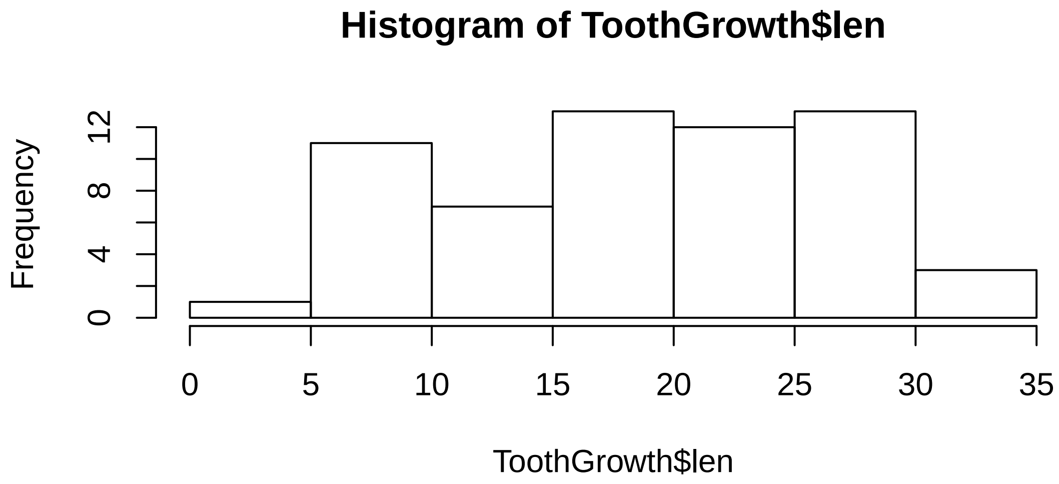
```
str(ToothGrowth)
```

```
## '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 ...
```

```
summary(ToothGrowth)
```

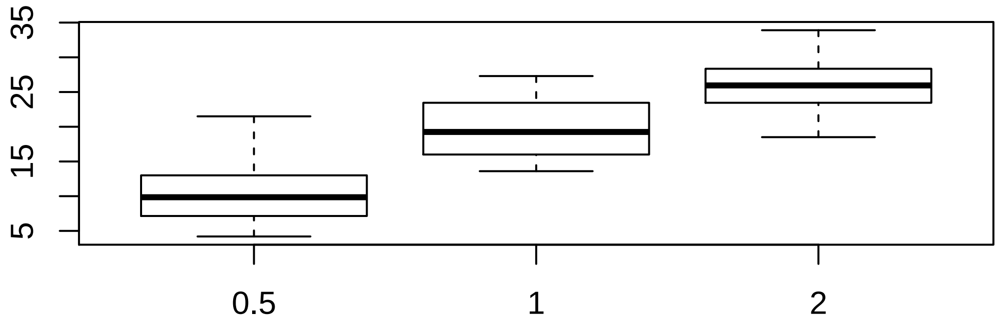
```
##       len        supp         dose      
##  Min.   : 4.20   OJ:30   Min.   :0.500  
##  1st Qu.:13.07   VC:30   1st Qu.:0.500  
##  Median :19.25           Median :1.000  
##  Mean   :18.81           Mean   :1.167  
##  3rd Qu.:25.27           3rd Qu.:2.000  
##  Max.   :33.90           Max.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

```
hist(ToothGrowth$len)
```

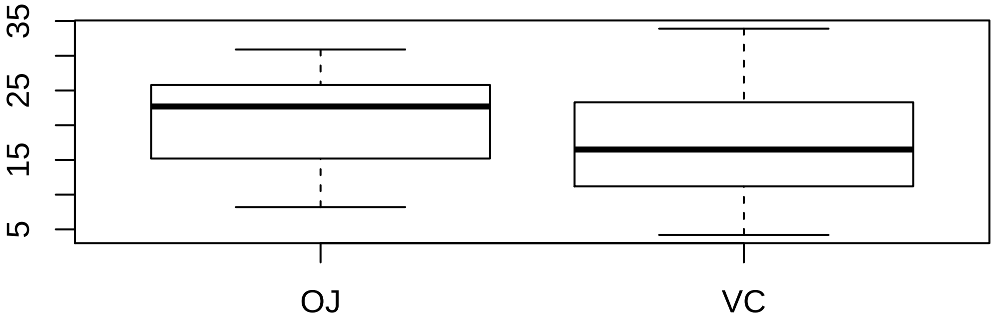


Some exploratory plots:

```
boxplot(tapply(ToothGrowth$len, ToothGrowth$dose, as.vector))
```

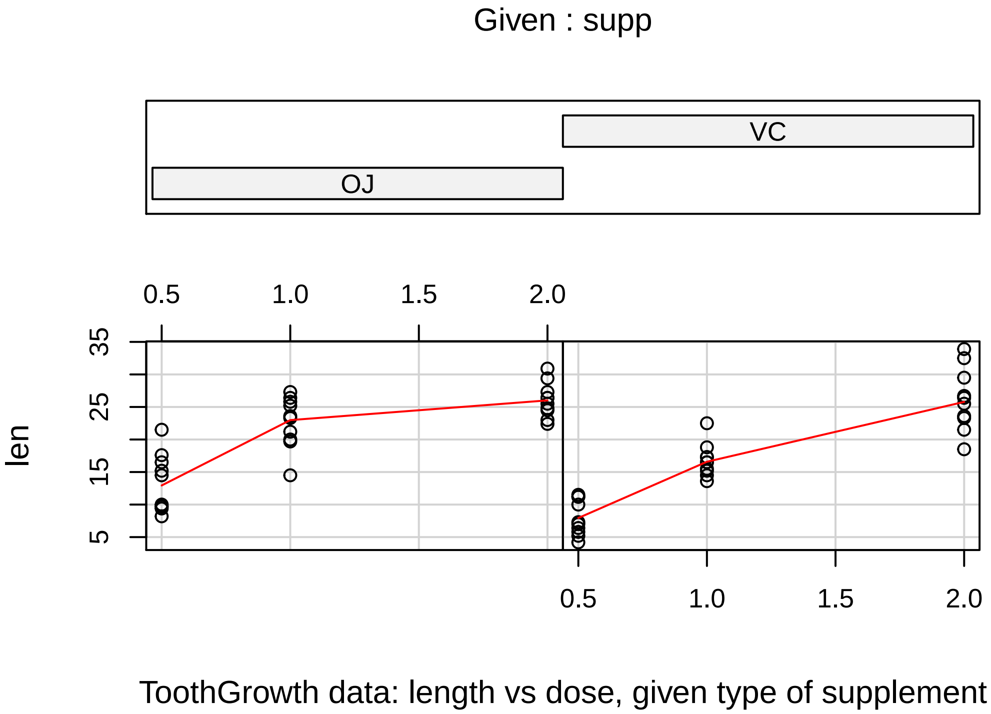


```
plot(ToothGrowth$supp, ToothGrowth$len)
```



Let's take a look at the exploratory plot offered for the ToothGrowth dataset in the help page of the dataset package.

```
require(graphics)
coplot(len ~ dose | supp, data = ToothGrowth, panel = panel.smooth,
       xlab = "ToothGrowth data: length vs dose, given type of supplement")
```



All these plots give a general suggestion that: * the dose size has a positive impact on the tooth length * orange juice gives higher tooth growth than ascorbid acid

Comparing Tooth Growth by Supp and Dose

Comparison will be made using both **95% t confidence intervals** and **hypotheses tests at $\alpha = .05$**. I will show the calculations for supp in more detail, but I will skip most details for Dose, as the calculations run along the same lines in both cases.

Comparing by Supp

Calculating auxiliary variables.

```
g1 <- ToothGrowth[ToothGrowth$supp=='OJ',]$len
g2 <- ToothGrowth[ToothGrowth$supp=='VC',]$len
n1 <- length(g1); n2 <- length(g2)
s1 <- var(g1); s2 <- var(g2)
```

Using Confidence Intervals

Let's first assume **paired observations**. The confidence interval is:

```
diff <- g1-g2
mn <- mean(diff); s <- sd(diff)
mn + c(-1,1)*qt(.975,n1-1)*s/sqrt(n1)
```

```
## [1] 1.408659 5.991341
```

The same can be achived using t.test function:

```
t.test(diff)$conf.int
```

```
## [1] 1.408659 5.991341
## attr(,"conf.level")
## [1] 0.95
```

We see that **orange juce has a significantly more positive impact on the tooth growth**, as the interval is entirely above 0.

Now let's assume **independent groups with equal variance**. The confidence interval is:

```
mn <- mean(g1) - mean(g2)
pooled_var = ((n1-1)*s1 + (n2-1)*s2)/(n1+n2-2)
mn + c(-1,1)*qt(.975,n1+n2-2)*sqrt(pooled_var * (1/n1+1/n2))
```

```
## [1] -0.1670064  7.5670064
```

The same can be achived using t.test function:

```
t.test(g1, g2, paired=F, var.equal=T)$conf.int
```

```
## [1] -0.1670064  7.5670064
## attr(,"conf.level")
## [1] 0.95
```

We could also asssume **independent groups with unequal variance**:

```
t.test(g1, g2, paired=F, var.equal=F)$conf.int
```

```
## [1] -0.1710156  7.5710156
## attr(,"conf.level")
## [1] 0.95
```

In both cases of independent group tests there is **no significant difference** between the groups.

Using Hypotheses Tests

Let's reframe the problem in terms of hypotheses testing. Let's first assume **paired observations**. Test settings are:

- $H_0 : diff = 0$
- $H - a : diff \neq 0$
- $\alpha = .05$

Critical value for two sided test:

```
qt(0.975, n1-1)
```

```
## [1] 2.04523
```

Test statistic $TS = \frac{\bar{X}-\mu_0}{S/\sqrt{n}}$:

```
(mn - 0) / (s/sqrt(n1))
```

```
## [1] 3.302585
```

so we **reject** H_0 as the test statistic is greater than the critical value.

The same using t.test function:

```
t.test(diff)
```

```
## 
##  One Sample t-test
## 
## data:  diff
## t = 3.3026, df = 29, p-value = 0.00255
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  1.408659 5.991341
## sample estimates:
## mean of x 
##       3.7
```

Now let's assume **independent groups with equal or unequal variance**. The test settings are:

- $H_0 : mn1 = mn2$
- $H_a : mn1 \neq mn2$
- $\alpha = .05$

Assuming **equal variance**:

```
t.test(g1, g2, paired=F, var.equal=T)
```

```
##
##  Two Sample t-test
##
## data:  g1 and g2
## t = 1.9153, df = 58, p-value = 0.06039
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1670064  7.5670064
## sample estimates:
## mean of x mean of y
##  20.66333  16.96333
```

Assuming **unequal variance**:

```
t.test(g1, g2, paired=F, var.equal=F)
```

```
##
##  Welch Two Sample t-test
##
## data:  g1 and g2
## 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 of x mean of y
##  20.66333  16.96333
```

I both cases of independent group tests we **can't reject** H_0 as the p-values are above .05.

Comparing by Dose

Comparing tooth length by dose runs along the same lines. Here I perform the tests only for paired observations. I compare each pair of doses, deducting sample means of bigger doses from those of smaller doses.

```
doses <- combn(c(.5,1,2),2)
for(i in seq_len(ncol(doses))) {
    data <- subset(ToothGrowth, dose==doses[1,i])$len -
        subset(ToothGrowth, dose==doses[2,i])$len
    cat(">>>>> H0: dose of", doses[1,i], "vs. dose of", doses[2,i])
    print(t.test(data))}
```

```
## >>>>> H0: dose of 0.5 vs. dose of 1
##  One Sample t-test
##
## data:  data
## t = -6.9669, df = 19, p-value = 1.225e-06
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -11.872879  -6.387121
## sample estimates:
## mean of x
##     -9.13
##
## >>>>> H0: dose of 0.5 vs. dose of 2
##  One Sample t-test
##
## data:  data
## t = -11.2915, df = 19, p-value = 7.19e-10
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -18.3672 -12.6228
## sample estimates:
## mean of x
##   -15.495
##
## >>>>> H0: dose of 1 vs. dose of 2
##  One Sample t-test
##
## data:  data
## t = -4.6046, df = 19, p-value = 0.0001934
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -9.258186 -3.471814
## sample estimates:
## mean of x
##    -6.365
```

We see that in each case that:

- the 95% t confidence intervals are entirely below 0
- estimated means of differences are below 0 with p-values below .5

so we conclude that **the bigger the dose of a supplement, the higher the tooth growth**.

Conclusions and Assumptions

Conclusions:

- the dose size has a positive impact on the tooth length
- orange juice gives higher tooth growth than ascorbid acid

Assumptions required to hold for these conclusions:

- the sample of guinea pigs is a proper **random draw** (no confounding factors)
- the **observations are paired -** i.e. the same 10 guinea pigs were treated with every possible comination of supplement and dose
- **the order of pigs is preserved** in the data, i.e. the i-th row pig is the same pig under each treatment
- t-statistic assumes that **the underlying population have a normal distribution**, we should assume that it is close to normal (at least symmetric and mound shaped)