Analyzing the Tooth Growth dataset in R

Manish Mittal

Overview:

In this report I will analyze the ToothGrowth data in the R datasets package and compare tooth growth by supp and dose

Summary of data:

The data contains 3 columns: len, supp, dose. Len is continous data ranging from 4.2 to 33.9. Supp is a factor with 2 levels OJ and VC. Finally, dose is ordinal ranging from 0.5 to 2.0. Below is a plot the visualizes the entire dataset across these 3 dimensions.

```
library(ggplot2)
ToothGrowth <- ToothGrowth
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
```

```
ggplot(data=ToothGrowth, aes(x=dose,y=len)) + geom_point(aes(colour=supp))
```
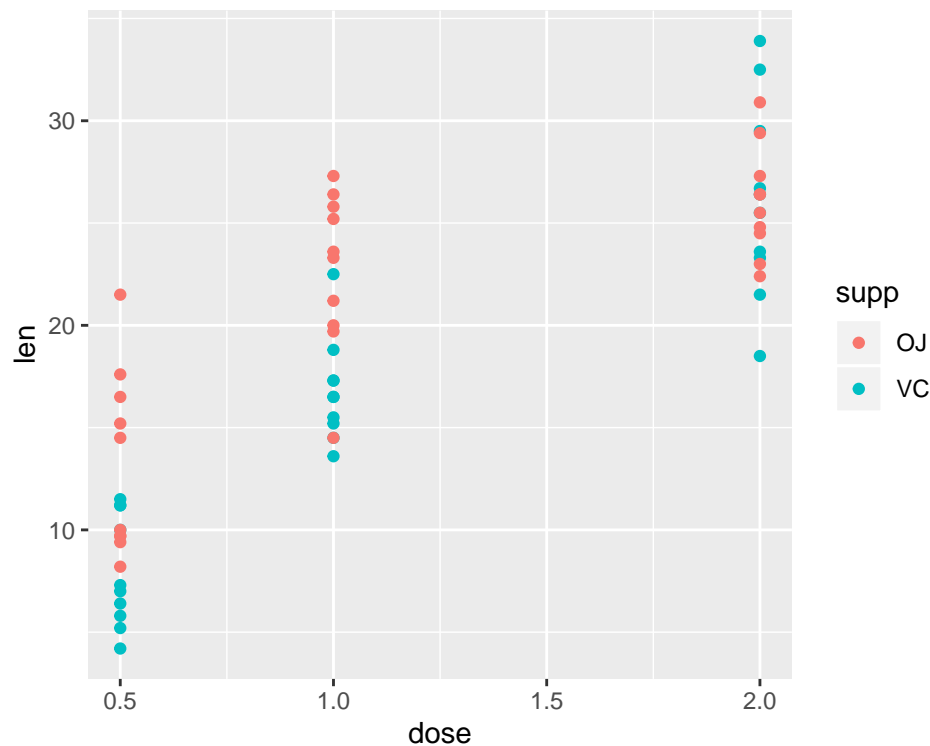
Testing the effect of supp

Null hypothesis is that there is no effect of treatment supp on len. To test for this hypotheis we peform a two sample t-tets comparing the means for the two groups. The assumptions in this test is that the variance is equal and data is normally distributed. The ratio of the two variances is ~1.5 which is considered reasonable for this assumption, however, we will run a test for unequal variances. The VC subset is normally distributed but the OJ subset shows some deviations from normal distribution. The t-test for comparing the means has a p=value of 0.06 indicating that we cannot reject the null hypothesis that the means for the two groups is different. The difference between the two group has a 95% CI of (-0.17,7.57) which includes 0.

```
var(subset(ToothGrowth,supp=='OJ',select='len')$len)
```
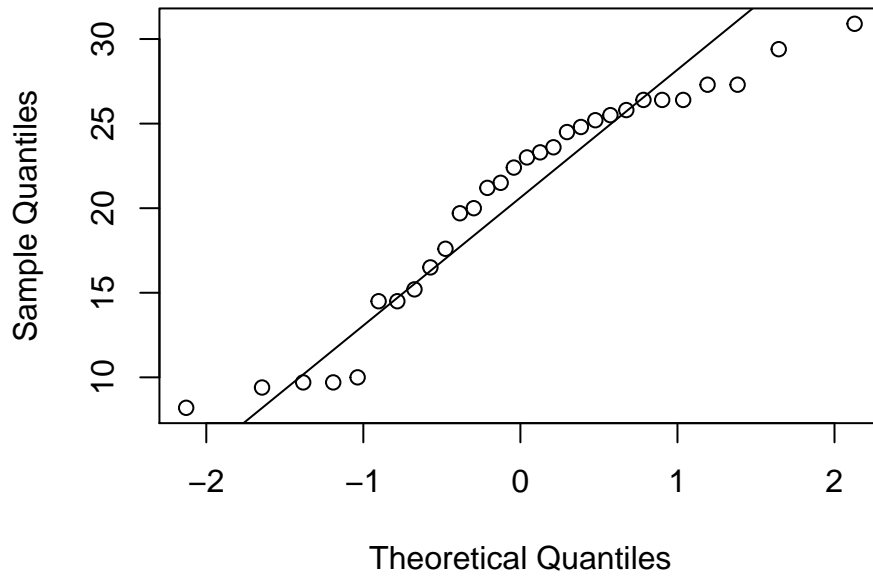
```
## [1] 43.63344
```

```
var(subset(ToothGrowth,supp=='VC',select='len')$len)
```
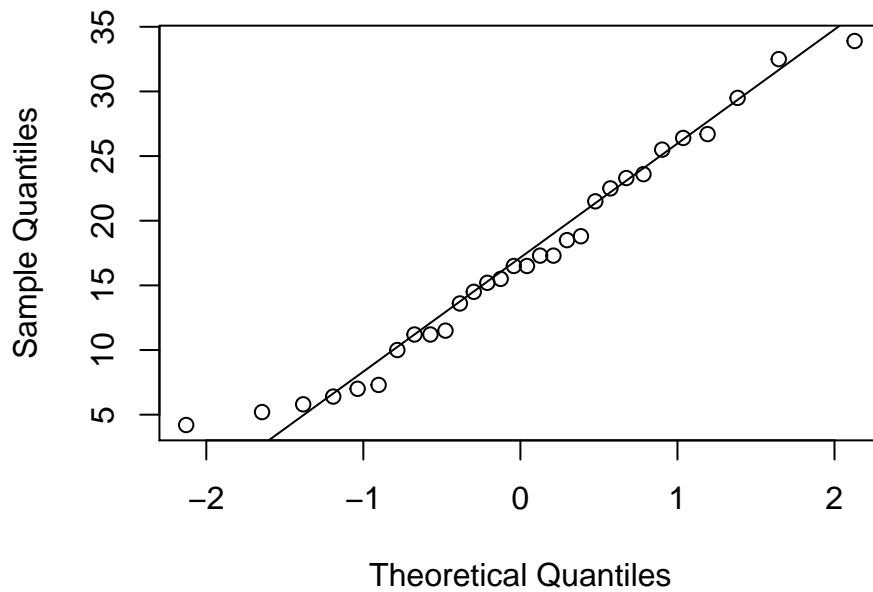
```
## [1] 68.32723
```

```
qqnorm(subset(ToothGrowth,supp=='OJ',select='len')$len,main='OJ supp qqplot')
qqline(subset(ToothGrowth,supp=='OJ',select='len')$len)
```

OJ supp qqplot


```r
qqnorm(subset(ToothGrowth,supp=='VC',select='len')$len,main='VC supp qqplot')
qqline(subset(ToothGrowth,supp=='VC',select='len')$len)
```

VC supp qqplot

```
t.test(len ~ supp, data=ToothGrowth, alternative = "two",var.equal=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
```

Testing the effect of dose

Similarly, I have also run t-tests for each of dose levels by running 3 t-tests. First I compare the variances and find that a dose of 0.5 and 1.0 have simialr variance but a dose of 2.0 has a lower varaince. Normal test by making a qqplot indicates that normality is a reasonable assumoption. Finally the t-test for all 3 cases has a p-values < 0.001 indicating that all three dose levels result in a significant difference.

```
dose1 = subset(ToothGrowth,dose==0.5,select='len')$len
dose2 = subset(ToothGrowth,dose==1.0,select='len')$len
dose3 = subset(ToothGrowth,dose==2.0,select='len')$len
var(dose1)
```

```
## [1] 20.24787
```
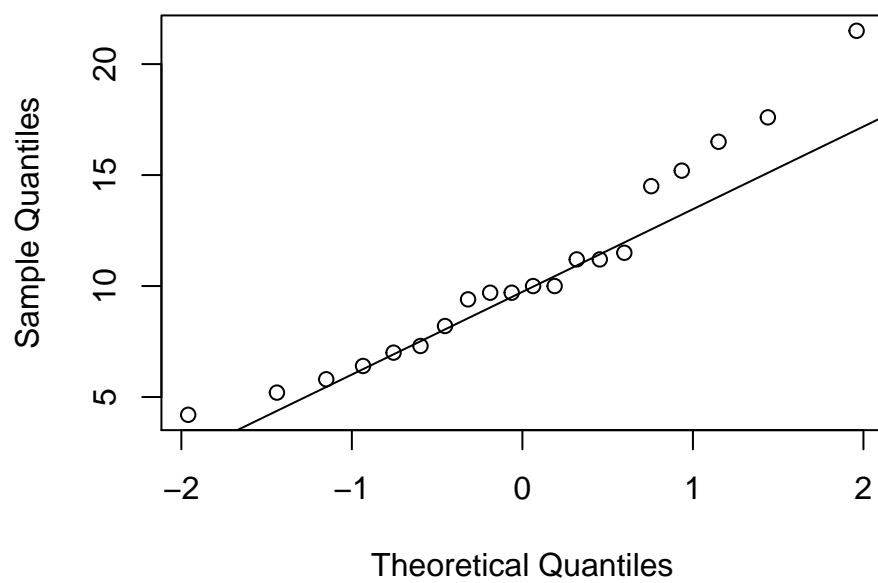
```
var(dose2)
```

```
## [1] 19.49608
```
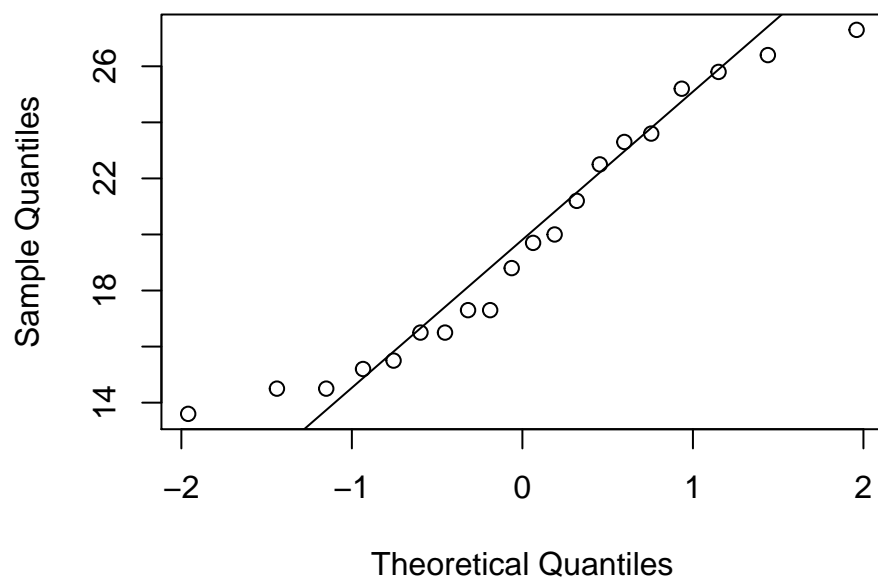
```
var(dose3)
```

```
## [1] 14.24421
```

```
qqnorm(dose1, main='dose = 0.5')
qqline(dose1)
```
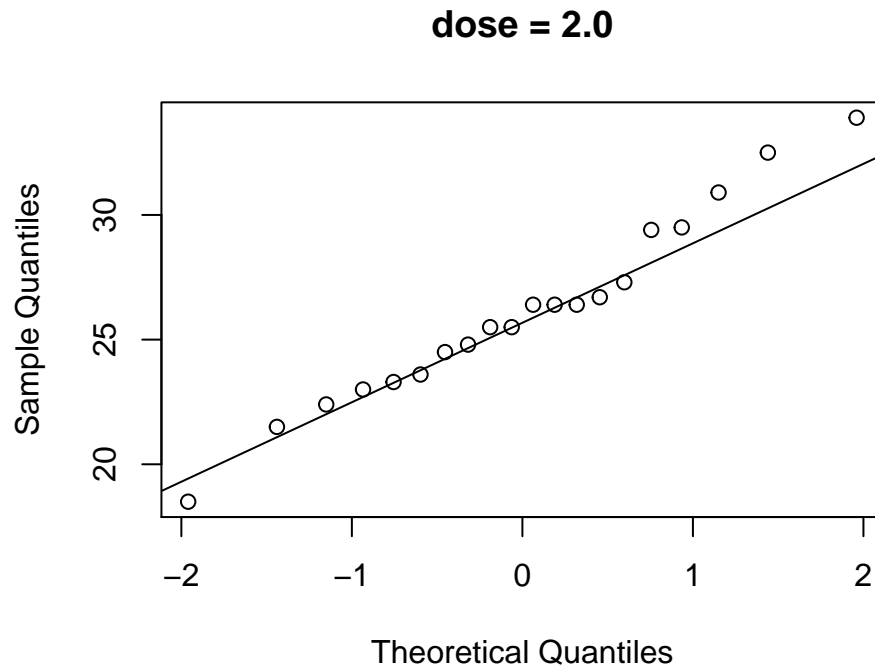
dose = 0.5


```
qqnorm(dose2, main='dose = 1.0')
qqline(dose2)
```

dose = 1.0

```
qqnorm(dose3, main='dose = 2.0')
qqline(dose3)
```

dose = 2.0


```
t.test(dose1,dose2,alternative = "two",var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  dose1 and dose2
## t = -6.4766, df = 38, p-value = 1.266e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.983748  -6.276252
## sample estimates:
## mean of x mean of y
##    10.605    19.735
```

```
t.test(dose2,dose3,alternative = "two",var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  dose2 and dose3
## t = -4.9005, df = 37.101, p-value = 1.906e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.996481 -3.733519
```

```
## sample estimates:
## mean of x mean of y
##    19.735    26.100
```

```r
t.test(dose1,dose3,alternative = "two",var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  dose1 and dose3
## t = -11.799, df = 36.883, p-value = 4.398e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.15617 -12.83383
## sample estimates:
## mean of x mean of y
##    10.605    26.100
```