Basic Inferential Data Analysis - Neeraj Ahire

Overview

In this assignment we will investigate the ToothGrowth dataset in the R datasets package. Let's load the dataset and check the summary.

```
library(datasets)
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
```

```
str(ToothGrowth)
```

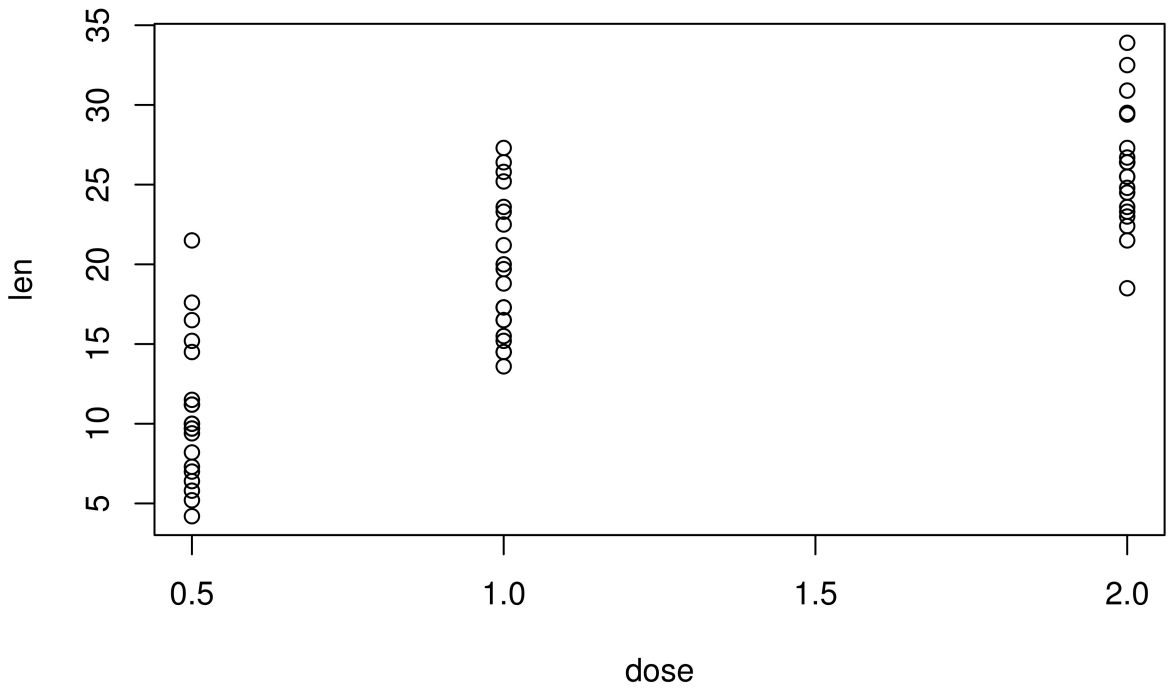
```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

So we have three columns in our dataframe with 'len' indicating the length of a special type of tooth cell, 'supp' indicating delivery method of supplement, and 'dose' indication dosage. 'supp' has two factor levels "vc" and "OJ" and 'dose' and three value 0.5, 1, 2 mg/day. Complete details can be found from documentation for the dataset.

Basic EDA

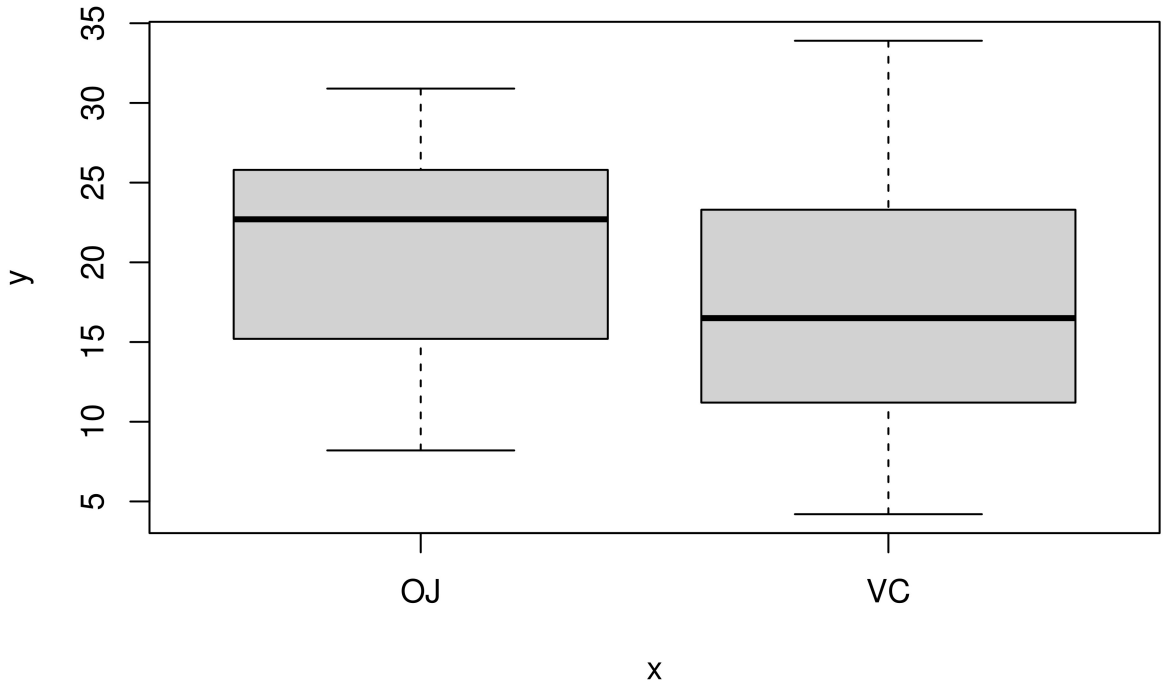
Lets plot the tooth length by supp

```
with(ToothGrowth, plot(dose, len))
```



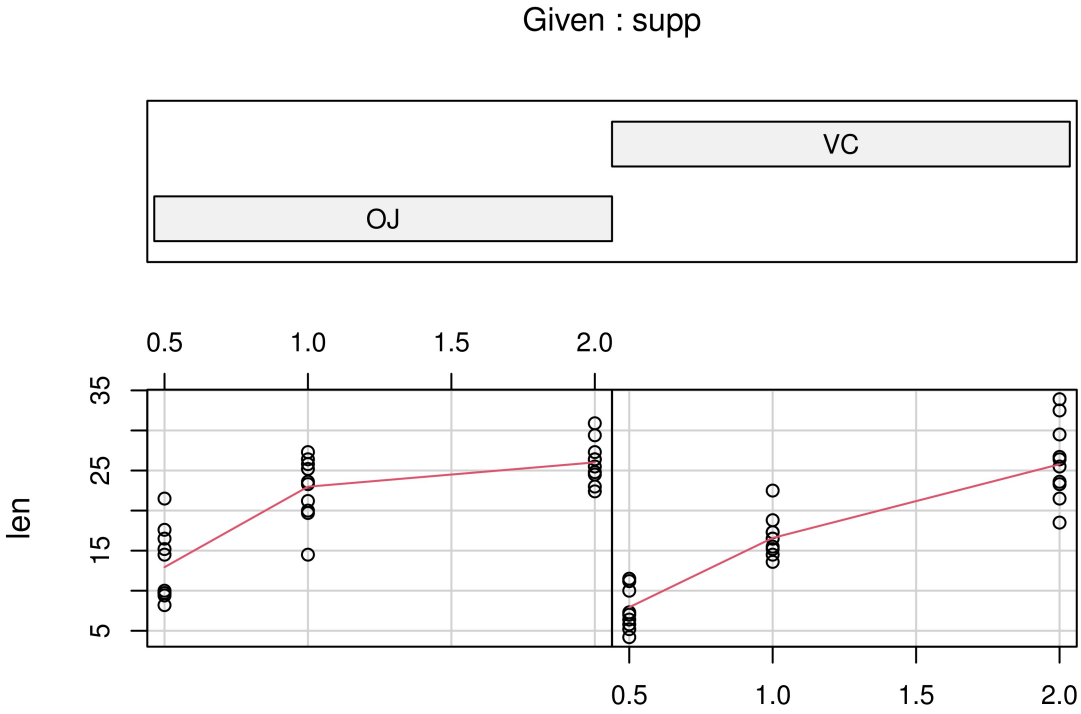
And now by supp

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



Lets see a combined panel plot.

```
coplot(len ~ dose | supp, data = ToothGrowth, panel = panel.smooth,xlab = "ToothGrowth data: length vs
```



ToothGrowth data: length vs dose, given type of supplement

So far we can see that the tooth length seems to be less for "VC" than "OJ" and that the length seems to increase by dosage for each supplement.

Hypothesis tests and CI's

Now we shall do hypothesis tests to check whether means are different and calculate p-values and confidence intervals. We shall use t-distribution and t tests.

First let's see if 'len' means for the different 'supp' values are same or different indicating if there is any difference in length for the two supplement delivery methods and check the statistical significance. The T test yields the following result. Our null hypothesis is that the difference between means is zero.

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

```
##
##  Welch Two Sample t-test
##
## data:  len[supp == "VC"] and len[supp == "OJ"]
## 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:
##  -7.5710156  0.1710156
## sample estimates:
## mean of x mean of y
##  16.96333  20.66333
```

The T- test shows us the confidence intervals and p value. Since, p value > 0.05 threshold we fail to reject the null hypothesis. Hence, supplement delivery method has no affect on tooth length.

Now, lets do the hypothesis tests for comparing the effect of dosage on length. Lets do the T tests for each dosage value (0,5, 1, 2). Null hypothesis being mean of 'len' is zero.

```
with(ToothGrowth, t.test(len[dose==0.5]))
```

```
##
##  One Sample t-test
##
## data:  len[dose == 0.5]
## t = 10.54, df = 19, p-value = 2.241e-09
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   8.499046 12.710954
## sample estimates:
## mean of x
##    10.605
```

```
with(ToothGrowth, t.test(len[dose== 1]))
```

```
##
##  One Sample t-test
##
## data:  len[dose == 1]
## t = 19.988, df = 19, p-value = 3.218e-14
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  17.66851 21.80149
## sample estimates:
## mean of x
##    19.735
```

```
with(ToothGrowth, t.test(len[dose== 2]))
```

```
##
##  One Sample t-test
##
## data:  len[dose == 2]
## t = 30.927, df = 19, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  24.33364 27.86636
## sample estimates:
## mean of x
##      26.1
```

In each case $p < 0.05$ hence we reject the null hypothesis that mean is zero indicating that dosage has a positive effect on the tooth length.

Now well shall compare means across each dosage i.e three such groups (0.5, 1), (0.5, 2), (1, 2). The T tests yield the following result. Null hypothesis being difference between means of each group is zero.

```
with(ToothGrowth, t.test(len[dose==0.5], len[dose == 1]))
```

```
##
##  Welch Two Sample t-test
##
## data:  len[dose == 0.5] and len[dose == 1]
## t = -6.4766, df = 37.986, p-value = 1.268e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.983781  -6.276219
## sample estimates:
## mean of x mean of y
##    10.605    19.735
```

```
with(ToothGrowth, t.test(len[dose==0.5], len[dose == 2]))
```

```
##
##  Welch Two Sample t-test
##
## data:  len[dose == 0.5] and len[dose == 2]
## 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
```

```
with(ToothGrowth, t.test(len[dose== 1], len[dose == 2]))
```

```
##
##  Welch Two Sample t-test
##
## data:  len[dose == 1] and len[dose == 2]
## 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
```

We can see that in every case $p < 0.05$, therefore we can reject the null hypothesis and conclude that the means are different and also that length increases with dosage.

Conclusion

We conclude that supplement delivery method has no effect on tooth length and that length increases with increase in dosage. The assumption in our analysis is that we assume out data is Gaussian which it may or may not be true.