Analysis of the ToothGrowth Data in R

Kumar Chandrakant

December 23, 2015

Synopsis

Through this snalysis we want to establish patterns in the data set ToothGrowth available in R. We would be using hypothesis testing and confidence intervals to establish this.

Exploratory Analysis

Let us begin by loading the data set into session.

```
library(plyr)
```

```
## Warning: package 'plyr' was built under R version 3.2.2
```

```
data(ToothGrowth)
```

Let us now explore this dataset through str function.

```
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 ...
```

We can see that this is a data frame consisting of 60 observations and 3 variables. The variable supp has two values “OJ” and “VC”. The variable dose has three values 0.5, 1.0 and 2.0. The variable len provides the tooth growth for the supp and dose combination.

We would now summarise the data set to yield the mean, standard deviation and sample size for every combination of supp and dose variables. We would use the ddply function in R.

```
data <- ddply(ToothGrowth, c("supp", "dose"), summarise,
              mean = mean(len), sd = sd(len), count=length(len))
```

Let us look at the summarised data now.

```
data
```

```
##   supp dose  mean       sd count
## 1   OJ  0.5 13.23 4.459709    10
## 2   OJ  1.0 22.70 3.910953    10
## 3   OJ  2.0 26.06 2.655058    10
## 4   VC  0.5  7.98 2.746634    10
## 5   VC  1.0 16.77 2.515309    10
## 6   VC  2.0 26.14 4.797731    10
```

Statistical Inference

Assumptions

Before we get onto the statistical inference of this data set, let us postulate some key assumptions of the analysis:

- Populations from where sample have been drawn have the same variance.
- Populations from where sample have been drawn are normally distributed.
- Each value is sampled independently from each other value.
- The values in samples does not contain any outliers

Hypothesis Testing

Let us perform the two sample t-test for all the doses 0.5, 1.0 and 2.0 to see if the difference in mean is significant to conclude anything. The details of the test are as following:

- Null hypothesis: True difference in means is equal to 0
- Alternate hypothesis: True difference in means is not equal to 0
- Significance level of 5%, i.e. 0.05.

```
t.test(subset(ToothGrowth, supp=="OJ" & dose==0.5)$len,
       subset(ToothGrowth, supp=="VC" & dose==0.5)$len, var.equal=T)
```

```
##
##  Two Sample t-test
##
## data:  subset(ToothGrowth, supp == "OJ" & dose == 0.5)$len and subset(ToothGrowth, supp
== "VC" & dose == 0.5)$len
## t = 3.1697, df = 18, p-value = 0.005304
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.770262 8.729738
## sample estimates:
## mean of x mean of y
##     13.23      7.98
```

```
t.test(subset(ToothGrowth, supp=="OJ" & dose==1.0)$len,
       subset(ToothGrowth, supp=="VC" & dose==1.0)$len, var.equal=T)
```

```
##
##  Two Sample t-test
##
## data:  subset(ToothGrowth, supp == "OJ" & dose == 1)$len and subset(ToothGrowth, supp ==
"VC" & dose == 1)$len
## t = 4.0328, df = 18, p-value = 0.0007807
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.840692 9.019308
## sample estimates:
## mean of x mean of y
##     22.70     16.77
```

```
t.test(subset(ToothGrowth, supp=="OJ" & dose==2.0)$len,
       subset(ToothGrowth, supp=="VC" & dose==2.0)$len, var.equal=T)
```

```
##
##  Two Sample t-test
##
## data:  subset(ToothGrowth, supp == "OJ" & dose == 2)$len and subset(ToothGrowth, supp ==
"VC" & dose == 2)$len
## t = -0.046136, df = 18, p-value = 0.9637
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.722999  3.562999
## sample estimates:
## mean of x mean of y
##     26.06     26.14
```

Conclusions

- Since the t-statistic of the tests for doses 0.5 and 1.0 have come below the significance level we can conclude that the null hypothesis is incorrect. Hence for these doses supp “OJ” performs better than “VC”. However the t-statistic in case of dose 2.0 comes out to be more than the significance level and hence we can conclude that the null hypothesis is true which says that the performance of both supp “OJ” and “VC” are similar for dose 2.0.
- The same conclusion can be drawn based on the confidence intervals calculated above. The first two tests for doses 0.5 and 1.0 have lower and upper confidence intervals above zero which shows that the population means are indeed different at 95% confidence. However the upper and lower confidence intervals of the third tests for dose 2.0 contains zero suggesting that the populations means can be similar.