ANLY\_510-50\_HW2

Dean D'souza

2016-06-11

## HomeWork Questions : Part 1

* **Use theCO2 data file in R**
* To get definitions of the columns type help(CO2)

# Initializing the CO2 dataset  
data("CO2")  
# Double-checking as global environment was not describing the dataset  
head(CO2)

## Plant Type Treatment conc uptake  
## 1 Qn1 Quebec nonchilled 95 16.0  
## 2 Qn1 Quebec nonchilled 175 30.4  
## 3 Qn1 Quebec nonchilled 250 34.8  
## 4 Qn1 Quebec nonchilled 350 37.2  
## 5 Qn1 Quebec nonchilled 500 35.3  
## 6 Qn1 Quebec nonchilled 675 39.2

# Assigning to a new variable for ease of use  
dat1<-CO2  
# Clearing memory  
remove(CO2)  
# Summary to get an idea of the data  
summary(dat1)

## Plant Type Treatment conc   
## Qn1 : 7 Quebec :42 nonchilled:42 Min. : 95   
## Qn2 : 7 Mississippi:42 chilled :42 1st Qu.: 175   
## Qn3 : 7 Median : 350   
## Qc1 : 7 Mean : 435   
## Qc3 : 7 3rd Qu.: 675   
## Qc2 : 7 Max. :1000   
## (Other):42   
## uptake   
## Min. : 7.70   
## 1st Qu.:17.90   
## Median :28.30   
## Mean :27.21   
## 3rd Qu.:37.12   
## Max. :45.50   
##

* Calculate means & standard deviations for 4 groups broken down by Type and Treatment

As there are two numerical variables we try to find the mean and standard deviations across Type and Treatment for both resulting as follows:

# Mean and Standard Deviation based on Type and conc  
with(dat1, tapply(conc, Type, mean))

## Quebec Mississippi   
## 435 435

with(dat1, tapply(conc, Type, sd))

## Quebec Mississippi   
## 297.7231 297.7231

# Mean and Standard Deviation based on Type and uptake  
with(dat1, tapply(uptake, Type, mean))

## Quebec Mississippi   
## 33.54286 20.88333

with(dat1, tapply(uptake, Type, sd))

## Quebec Mississippi   
## 9.673830 7.815773

# Mean and Standard Deviation based on Treatment and conc  
with(dat1, tapply(conc, Treatment, mean))

## nonchilled chilled   
## 435 435

with(dat1, tapply(conc, Treatment, sd))

## nonchilled chilled   
## 297.7231 297.7231

# Mean and Standard Deviation based on Treatment and uptake  
with(dat1, tapply(uptake, Treatment, mean))

## nonchilled chilled   
## 30.64286 23.78333

with(dat1, tapply(uptake, Treatment, sd))

## nonchilled chilled   
## 9.704994 10.884312

* Perform one-way tests twice: once for Type and once for Treatment

# One way test based on Type and conc  
oneway.test(conc ~ Type, dat1)

##   
## One-way analysis of means (not assuming equal variances)  
##   
## data: conc and Type  
## F = 3.0621e-30, num df = 1, denom df = 82, p-value = 1

# One way test based on Type and uptake  
oneway.test(uptake ~ Type, dat1)

##   
## One-way analysis of means (not assuming equal variances)  
##   
## data: uptake and Type  
## F = 43.519, num df = 1.000, denom df = 78.533, p-value = 4.451e-09

# One way test based on Treatment and conc  
oneway.test(conc ~ Treatment, dat1)

##   
## One-way analysis of means (not assuming equal variances)  
##   
## data: conc and Treatment  
## F = 3.0621e-30, num df = 1, denom df = 82, p-value = 1

# One way test based on Treatment and uptake  
oneway.test(uptake ~ Treatment, dat1)

##   
## One-way analysis of means (not assuming equal variances)  
##   
## data: uptake and Treatment  
## F = 9.2931, num df = 1.000, denom df = 80.945, p-value = 0.003107

It was interesting to see that when tested with conc the p-values were 1 indicating no relationship and that conc was more of a categorical variable. With regarrds to uptake as the p-values were significantly less, it indicated that this variable has a relationship with Type and Treatment.

* Perform a two-way test for Type and Treatment

# Two way test using Type and Treatment with uptake as the response variable  
co2\_lm2<-lm(uptake ~ Type + Treatment + Type\*Treatment, data=dat1)  
anova(co2\_lm2)

## Analysis of Variance Table  
##   
## Response: uptake  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Type 1 3365.5 3365.5 52.5086 2.378e-10 \*\*\*  
## Treatment 1 988.1 988.1 15.4164 0.0001817 \*\*\*  
## Type:Treatment 1 225.7 225.7 3.5218 0.0642128 .   
## Residuals 80 5127.6 64.1   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

From the above results we can see that both Type and Treatment do have a relationship with uptake, however the interaction between the two had less of an impact on uptake. This would mean that individually, the relationships between the variables was important but when we consider interactions, it would not be as important and useful.

## HomeWork Questions : Part 2 - mtcars

* **Use the table() function with the following combinations**

#Following the same steps as with the previous question  
library(car)  
data(mtcars)  
head(mtcars)

## mpg cyl disp hp drat wt qsec vs am gear carb  
## Mazda RX4 21.0 6 160 110 3.90 2.620 16.46 0 1 4 4  
## Mazda RX4 Wag 21.0 6 160 110 3.90 2.875 17.02 0 1 4 4  
## Datsun 710 22.8 4 108 93 3.85 2.320 18.61 1 1 4 1  
## Hornet 4 Drive 21.4 6 258 110 3.08 3.215 19.44 1 0 3 1  
## Hornet Sportabout 18.7 8 360 175 3.15 3.440 17.02 0 0 3 2  
## Valiant 18.1 6 225 105 2.76 3.460 20.22 1 0 3 1

dat2<-mtcars  
remove(mtcars)

* The variables vs and am

# Using table() function with vs and am  
vsam<-table(dat2$vs,dat2$am)

* The variables gear and carb

# Using table() function with gear and carb  
gearcarb<-table(dat2$gear,dat2$carb)

* The variables cyl and gear

# Using table() function with cyl and gear  
cylgear<-table(dat2$cyl,dat2$gear)

* For each of the three cases above guess what the results of a Chi-Squared analysis will be
* Ignore warnings for low values in the cells

# viewing the tables individually  
vsam

##   
## 0 1  
## 0 12 6  
## 1 7 7

gearcarb

##   
## 1 2 3 4 6 8  
## 3 3 4 3 5 0 0  
## 4 4 4 0 4 0 0  
## 5 0 2 0 1 1 1

cylgear

##   
## 3 4 5  
## 4 1 8 2  
## 6 2 4 1  
## 8 12 0 2

From the above data we can draw the following conclusions:  
1. For the first set of variables (vs and am) we can expect the p-value to be close to 0.05 as there may be a relationship between the two.  
2. For the second set of variables (gear and carb) we can expect the p-value to be close to 0.05 as there may be a relationship between the two, though we can also expect it to be higher due to the variation across stacks.  
3. For the third set of variables (cyl and gear) we can expect the p-value to be close to 0.05 with a greater chance of it being under 0.05 as there should be a good relationship between the two variables.

* **Perform a Chi-Squared analysis on the mtcars data set for each of the three cases above**

# Chisquare test on the first pair  
chisq.test(vsam)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: vsam  
## X-squared = 0.34754, df = 1, p-value = 0.5555

# Chisquare test on the first pair  
chisq.test(gearcarb)

##   
## Pearson's Chi-squared test  
##   
## data: gearcarb  
## X-squared = 16.518, df = 10, p-value = 0.08573

# Chisquare test on the first pair  
chisq.test(cylgear)

##   
## Pearson's Chi-squared test  
##   
## data: cylgear  
## X-squared = 18.036, df = 4, p-value = 0.001214

The above results confirm our assumptions for the most part with the exception of the first case, as we observe a p-value much higher than assumed. This indicates that the counts are more due to chance.