output: word\_document: default

title: "Assignment 1" author: "Dineo Nono" date: "21 February 2017" output: md\_document

## Hello Octocat

I love Octocat. She's the coolest cat in town.

data("anscombe")  
dim.data.frame(anscombe)

## [1] 11 8

colnames(anscombe)

## [1] "x1" "x2" "x3" "x4" "y1" "y2" "y3" "y4"

head(anscombe)

## x1 x2 x3 x4 y1 y2 y3 y4  
## 1 10 10 10 8 8.04 9.14 7.46 6.58  
## 2 8 8 8 8 6.95 8.14 6.77 5.76  
## 3 13 13 13 8 7.58 8.74 12.74 7.71  
## 4 9 9 9 8 8.81 8.77 7.11 8.84  
## 5 11 11 11 8 8.33 9.26 7.81 8.47  
## 6 14 14 14 8 9.96 8.10 8.84 7.04

tail(anscombe)

## x1 x2 x3 x4 y1 y2 y3 y4  
## 6 14 14 14 8 9.96 8.10 8.84 7.04  
## 7 6 6 6 8 7.24 6.13 6.08 5.25  
## 8 4 4 4 19 4.26 3.10 5.39 12.50  
## 9 12 12 12 8 10.84 9.13 8.15 5.56  
## 10 7 7 7 8 4.82 7.26 6.42 7.91  
## 11 5 5 5 8 5.68 4.74 5.73 6.89

summary(anscombe)

## x1 x2 x3 x4   
## Min. : 4.0 Min. : 4.0 Min. : 4.0 Min. : 8   
## 1st Qu.: 6.5 1st Qu.: 6.5 1st Qu.: 6.5 1st Qu.: 8   
## Median : 9.0 Median : 9.0 Median : 9.0 Median : 8   
## Mean : 9.0 Mean : 9.0 Mean : 9.0 Mean : 9   
## 3rd Qu.:11.5 3rd Qu.:11.5 3rd Qu.:11.5 3rd Qu.: 8   
## Max. :14.0 Max. :14.0 Max. :14.0 Max. :19   
## y1 y2 y3 y4   
## Min. : 4.260 Min. :3.100 Min. : 5.39 Min. : 5.250   
## 1st Qu.: 6.315 1st Qu.:6.695 1st Qu.: 6.25 1st Qu.: 6.170   
## Median : 7.580 Median :8.140 Median : 7.11 Median : 7.040   
## Mean : 7.501 Mean :7.501 Mean : 7.50 Mean : 7.501   
## 3rd Qu.: 8.570 3rd Qu.:8.950 3rd Qu.: 7.98 3rd Qu.: 8.190   
## Max. :10.840 Max. :9.260 Max. :12.74 Max. :12.500

## x1 x2 x3 x4 y1 y2 y3 y4  
## 1 10 10 10 8 8.04 9.14 7.46 6.58  
## 2 8 8 8 8 6.95 8.14 6.77 5.76  
## 3 13 13 13 8 7.58 8.74 12.74 7.71  
## 4 9 9 9 8 8.81 8.77 7.11 8.84  
## 5 11 11 11 8 8.33 9.26 7.81 8.47  
## 6 14 14 14 8 9.96 8.10 8.84 7.04  
## 7 6 6 6 8 7.24 6.13 6.08 5.25  
## 8 4 4 4 19 4.26 3.10 5.39 12.50  
## 9 12 12 12 8 10.84 9.13 8.15 5.56  
## 10 7 7 7 8 4.82 7.26 6.42 7.91  
## 11 5 5 5 8 5.68 4.74 5.73 6.89



library(readr)  
df<-data.frame(read.csv("analgesic.csv",header=T))  
df

## ID Group Measurement\_1 Measurement\_2 Measurement\_3  
## 1 1 Analgesic 26 26 21  
## 2 2 Analgesic 29 26 23  
## 3 3 Analgesic 24 28 22  
## 4 4 Analgesic 25 22 24  
## 5 5 Analgesic 24 28 23  
## 6 6 Analgesic 22 23 26  
## 7 7 Analgesic 25 25 30  
## 8 8 Analgesic 28 21 21  
## 9 9 Analgesic 22 26 20  
## 10 10 Analgesic 18 25 29  
## 11 11 Analgesic 25 29 28  
## 12 12 Analgesic 26 25 23  
## 13 13 Analgesic 26 25 26  
## 14 14 Analgesic 19 30 27  
## 15 15 Analgesic 24 20 24  
## 16 16 Analgesic 23 24 27  
## 17 17 Analgesic 24 32 28  
## 18 18 Analgesic 24 17 25  
## 19 19 Analgesic 23 25 23  
## 20 20 Analgesic 30 18 25  
## 21 21 Placebo 19 12 18  
## 22 22 Placebo 10 16 18  
## 23 23 Placebo 12 11 20  
## 24 24 Placebo 17 17 18  
## 25 25 Placebo 18 18 20  
## 26 26 Placebo 12 16 16  
## 27 27 Placebo 14 17 17  
## 28 28 Placebo 20 19 18  
## 29 29 Placebo 16 19 15  
## 30 30 Placebo 17 15 13  
## 31 31 Placebo 18 21 14  
## 32 32 Placebo 20 13 16  
## 33 33 Placebo 12 8 21  
## 34 34 Placebo 20 17 16  
## 35 35 Placebo 17 21 15  
## 36 36 Placebo 19 17 15  
## 37 37 Placebo 14 19 13  
## 38 38 Placebo 17 19 13  
## 39 39 Placebo 11 20 18  
## 40 40 Placebo 15 18 12

dim(df)

## [1] 40 5

colnames(df)

## [1] "ID" "Group" "Measurement\_1" "Measurement\_2"  
## [5] "Measurement\_3"

head(df)

## ID Group Measurement\_1 Measurement\_2 Measurement\_3  
## 1 1 Analgesic 26 26 21  
## 2 2 Analgesic 29 26 23  
## 3 3 Analgesic 24 28 22  
## 4 4 Analgesic 25 22 24  
## 5 5 Analgesic 24 28 23  
## 6 6 Analgesic 22 23 26

tail(df)

## ID Group Measurement\_1 Measurement\_2 Measurement\_3  
## 35 35 Placebo 17 21 15  
## 36 36 Placebo 19 17 15  
## 37 37 Placebo 14 19 13  
## 38 38 Placebo 17 19 13  
## 39 39 Placebo 11 20 18  
## 40 40 Placebo 15 18 12

summary(df)

## ID Group Measurement\_1 Measurement\_2   
## Min. : 1.00 Analgesic:20 Min. :10.00 Min. : 8.0   
## 1st Qu.:10.75 Placebo :20 1st Qu.:17.00 1st Qu.:17.0   
## Median :20.50 Median :20.00 Median :20.0   
## Mean :20.50 Mean :20.12 Mean :20.7   
## 3rd Qu.:30.25 3rd Qu.:24.00 3rd Qu.:25.0   
## Max. :40.00 Max. :30.00 Max. :32.0   
## Measurement\_3   
## Min. :12.00   
## 1st Qu.:16.00   
## Median :20.50   
## Mean :20.52   
## 3rd Qu.:24.25   
## Max. :30.00

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tidyr)  
#changing data from wide to long format  
df\_1 <- gather(df, key=key, value=value, Measurement\_1:Measurement\_3, -ID)  
#Group data according to means across the measurements made on each individual  
df\_2 <- group\_by(df\_1, ID, Group)  
#Find descriptive stats for data df\_2  
summarise(df\_2, mean = mean(value))

## Source: local data frame [40 x 3]  
## Groups: ID [?]  
##   
## ID Group mean  
## <int> <fctr> <dbl>  
## 1 1 Analgesic 24.33333  
## 2 2 Analgesic 26.00000  
## 3 3 Analgesic 24.66667  
## 4 4 Analgesic 23.66667  
## 5 5 Analgesic 25.00000  
## 6 6 Analgesic 23.66667  
## 7 7 Analgesic 26.66667  
## 8 8 Analgesic 23.33333  
## 9 9 Analgesic 22.66667  
## 10 10 Analgesic 24.00000  
## # ... with 30 more rows

##chunk 1  
# Null hypothesis  
#HO: The chicks weights is dependent on the feed  
#Alternative hypothesis   
#H1: the chicks weight is not dependent on the feed  
# read chick weight  
x<-read.csv("chick-weights (4).csv")  
# tidy the data  
model.frame.default(formula = x$weight ~ x$feed)

## x$weight x$feed  
## 1 179 horsebean  
## 2 160 horsebean  
## 3 136 horsebean  
## 4 227 horsebean  
## 5 217 horsebean  
## 6 168 horsebean  
## 7 108 horsebean  
## 8 124 horsebean  
## 9 143 horsebean  
## 10 140 horsebean  
## 11 309 linseed  
## 12 229 linseed  
## 13 181 linseed  
## 14 141 linseed  
## 15 260 linseed  
## 16 203 linseed  
## 17 148 linseed  
## 18 169 linseed  
## 19 213 linseed  
## 20 257 linseed  
## 21 244 linseed  
## 22 271 linseed  
## 23 243 soybean  
## 24 230 soybean  
## 25 248 soybean  
## 26 327 soybean  
## 27 329 soybean  
## 28 250 soybean  
## 29 193 soybean  
## 30 271 soybean  
## 31 316 soybean  
## 32 267 soybean  
## 33 199 soybean  
## 34 171 soybean  
## 35 158 soybean  
## 36 248 soybean  
## 37 423 sunflower  
## 38 340 sunflower  
## 39 392 sunflower  
## 40 339 sunflower  
## 41 341 sunflower  
## 42 226 sunflower  
## 43 320 sunflower  
## 44 295 sunflower  
## 45 334 sunflower  
## 46 322 sunflower  
## 47 297 sunflower  
## 48 318 sunflower  
## 49 325 meatmeal  
## 50 257 meatmeal  
## 51 303 meatmeal  
## 52 315 meatmeal  
## 53 380 meatmeal  
## 54 153 meatmeal  
## 55 263 meatmeal  
## 56 242 meatmeal  
## 57 206 meatmeal  
## 58 344 meatmeal  
## 59 258 meatmeal  
## 60 368 casein  
## 61 390 casein  
## 62 379 casein  
## 63 260 casein  
## 64 404 casein  
## 65 318 casein  
## 66 352 casein  
## 67 359 casein  
## 68 216 casein  
## 69 222 casein  
## 70 283 casein  
## 71 332 casein

# state statistical test  
chickanova <-aov(weight~feed, data = x)  
summary(chickanova)

## Df Sum Sq Mean Sq F value Pr(>F)   
## feed 5 231129 46226 15.37 5.94e-10 \*\*\*  
## Residuals 65 195556 3009   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#underlying assumption  
"compares means or samples of more than 3 groups"

## [1] "compares means or samples of more than 3 groups"

# degrees of freedom and p value  
df = 5  
"p-value < 0.05"

## [1] "p-value < 0.05"

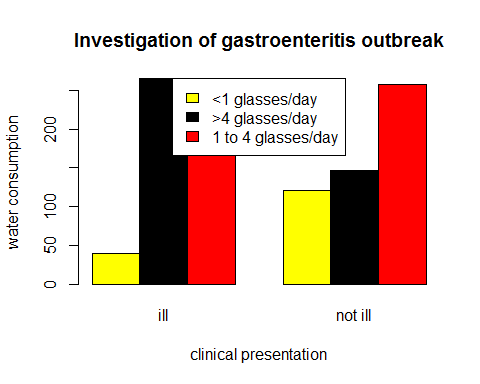
"accept null hypothesis"

## [1] "accept null hypothesis"

### chunk 2  
##Null hypothesis  
#Drinking contaminated water does not cause gastroenteritis  
#Alternative hypothesis  
#drinking contaminated water does cause gastroenteritis  
library(knitr)  
# read outbreak of severe gastroenteritis  
x <- read.csv("gastroenteritis.csv")  
y <- xtabs(~Consumption + Outcome, data = x)  
y

## Outcome  
## Consumption ill not ill  
## < 1 glasses/day 39 121  
## > 4 glasses/day 265 146  
## 1 to 4 glasses/day 265 258

barplot(y, beside = TRUE, ylab = "water consumption", xlab = "clinical presentation",main = "Investigation of gastroenteritis outbreak", col = c("yellow", "black", "red"))  
legend("top", c("<1 glasses/day", ">4 glasses/day", "1 to 4 glasses/day"), fill = c("yellow", "black", "red"))

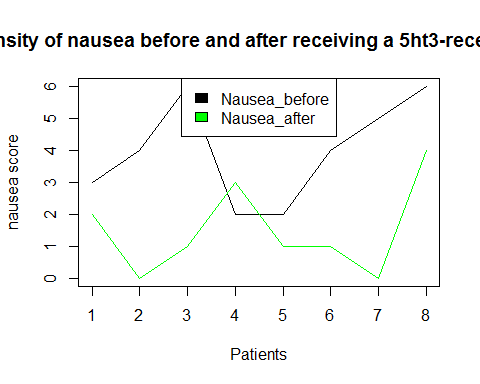


# statistical test  
Z <- chisq.test(y, correct = TRUE)  
  
##Underlying assumptions  
#two categorical variables from a single population  
#used to determine whether there is an association between 2 variable names   
## interpretation of outcomes  
#p< 0.05, therefor reject the null hypothesis and accept the alternative hypothesis

###chunk 3  
##Null hypothesis  
#Recieving chemotherapy does not cause nausea  
##Alternative hypothesis  
#Recieving chemotherapy does cause nausea  
# read the intensity of nausea  
A <- read.csv("nausea.csv")  
  
#tidy data  
# rating scale is anchored at 0(no nausea) to 6(severe nausea and vomiting) therefore change row 8 coloumn 3.  
A[8,3] = 4  
A

## Patient Nausea\_before Nausea\_after  
## 1 1 3 2  
## 2 2 4 0  
## 3 3 6 1  
## 4 4 2 3  
## 5 5 2 1  
## 6 6 4 1  
## 7 7 5 0  
## 8 8 6 4

##plotting of data  
plot(A$Nausea\_before~A$Patient, type = "l", ylim = c(0,6), xlab = "Patients", ylab = "nausea score", main = "The intensity of nausea before and after receiving a 5ht3-receptor blocker")  
lines(A$Nausea\_after~A$Patient, col= "green")  
legend ("top", c("Nausea\_before", "Nausea\_after"),fill = c("black", "green"))



# statistical test  
wilcox.test(A$Nausea\_before, A$Nausea\_after, paired = TRUE)

## Warning in wilcox.test.default(A$Nausea\_before, A$Nausea\_after, paired =  
## TRUE): cannot compute exact p-value with ties

##   
## Wilcoxon signed rank test with continuity correction  
##   
## data: A$Nausea\_before and A$Nausea\_after  
## V = 34, p-value = 0.02897  
## alternative hypothesis: true location shift is not equal to 0

## Test assumption  
#The data is paired and non-parametric  
#two measurements were taken from the same sample group.   
#P< 0.05 Therefore reject the null hypothesis and accept the alternative hypothesis

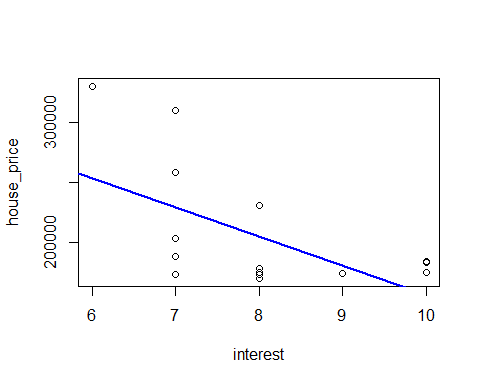
##Null hypothesis  
#The housing price is not dependent on the interest rate  
##Alternative hypothesis  
#The housing price is dependent on the interest rate  
# import data  
  
G <- read.csv("housing-prices.csv")  
G

## interest\_rate median\_house\_price\_USD  
## 1 10 183800  
## 2 10 183200  
## 3 10 174900  
## 4 9 173500  
## 5 8 172900  
## 6 7 173200  
## 7 8 173200  
## 8 8 169700  
## 9 8 174500  
## 10 8 177900  
## 11 7 188100  
## 12 7 203200  
## 13 8 230200  
## 14 7 258200  
## 15 7 309800  
## 16 6 329800

# Tidy data  
interest= G$interest\_rate  
house\_price = G$median\_house\_price\_USD  
head(cbind(interest, house\_price))

## interest house\_price  
## [1,] 10 183800  
## [2,] 10 183200  
## [3,] 10 174900  
## [4,] 9 173500  
## [5,] 8 172900  
## [6,] 7 173200

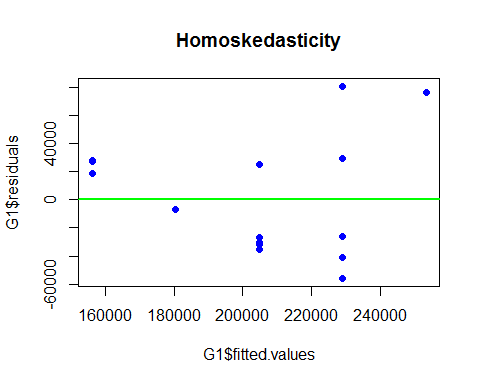
# scatter plot  
plot( interest, house\_price, xlab= "interest", ylab = "house\_price")  
abline(lm(G$median\_house\_price\_USD ~ G$interest\_rate, data = G), col= "blue", lwd= 2)



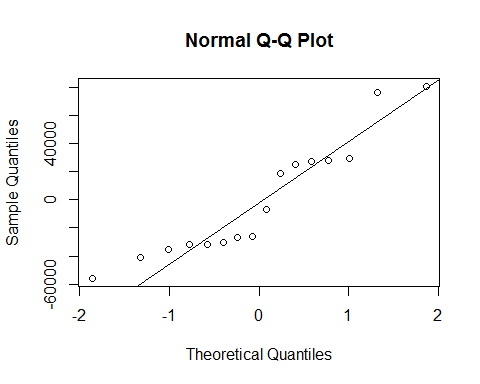
# linear regression  
G1 <- lm(G$median\_house\_price\_USD ~ G$interest\_rate, data = G)  
summary(G1)

##   
## Call:  
## lm(formula = G$median\_house\_price\_USD ~ G$interest\_rate, data = G)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -55865 -31631 -16406 27212 80735   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 399229 74427 5.364 9.99e-05 \*\*\*  
## G$interest\_rate -24309 9205 -2.641 0.0194 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 43180 on 14 degrees of freedom  
## Multiple R-squared: 0.3325, Adjusted R-squared: 0.2848   
## F-statistic: 6.974 on 1 and 14 DF, p-value: 0.01937

# diagnostic plot 1  
plot( x= G1$fitted.values, y=G1$residuals, main = "Homoskedasticity", pch = 19, col= "blue")  
abline(h=0, col= "green", lwd= 2)



#diagnostic plot 2 : Gaussian residual distribution  
qqnorm(G1$residuals)  
qqline(G1$residuals)



# Binary outcome variable: the variables have non-Gaussian error distribution  
glm (G$median\_house\_price\_USD ~ G$interest\_rate, data = G)

##   
## Call: glm(formula = G$median\_house\_price\_USD ~ G$interest\_rate, data = G)  
##   
## Coefficients:  
## (Intercept) G$interest\_rate   
## 399229 -24309   
##   
## Degrees of Freedom: 15 Total (i.e. Null); 14 Residual  
## Null Deviance: 3.91e+10   
## Residual Deviance: 2.61e+10 AIC: 390.8

## test assumptions  
 #I did the scatter plot to determine the trend or the relationship  
 # disgnostics for linear regression  
 # I did qq plot to determine if the residuals are normally distributed   
 # I did the Gaussian residual distribution to determine the variance of the fitted values  
 # I did the generalized linear model because the diagnostics I ran were not normally distributed.  
 ## test interpretation  
 # p-value = 0.01937 therefore reject the null hypotehsis and accept the alternative hypothesis.  
 # degrees of freedom : 15 Total (i.e. Null); 14 Residual  
 # test statistics  
 # F-statistic: 6.974 on 1  
  
##KOLMOGOROV-SMIRNOV AND ANDERSON DARLING TESTS WERE USED TO TEST FOR PARAMETRIC AND NON-PARAMETRIC DATA