Module 5

Michael Pearson

1/17/2020

## Set up a temporary file name for the downloaded zip file  
  
temp <- tempfile()  
  
download.file("http://archive.ics.uci.edu/ml/machine-learning-databases/00503/HCV-Egy-Data.zip", temp)  
  
## unzip the downloaded file, see what the name of the csv is  
tried <- unzip(temp)  
  
tried

## [1] "./HCV-Egy-Data.csv" "./Discretization-Criteria.csv"

## Read the csv file that has been unzipped  
  
catz <- read.csv("./HCV-Egy-Data.csv", header = TRUE)  
  
  
## show the head of the data  
  
head(catz)

## Age Gender BMI Fever Nausea.Vomting Headache Diarrhea  
## 1 56 1 35 2 1 1 1  
## 2 46 1 29 1 2 2 1  
## 3 57 1 33 2 2 2 2  
## 4 49 2 33 1 2 1 2  
## 5 59 1 32 1 1 2 1  
## 6 58 2 22 2 2 2 1  
## Fatigue...generalized.bone.ache Jaundice Epigastric.pain WBC RBC  
## 1 2 2 2 7425 4248807  
## 2 2 2 1 12101 4429425  
## 3 1 1 1 4178 4621191  
## 4 1 2 1 6490 4794631  
## 5 2 2 2 3661 4606375  
## 6 2 2 1 11785 3882456  
## HGB Plat AST.1 ALT.1 ALT4 ALT.12 ALT.24 ALT.36 ALT.48 ALT.after.24.w  
## 1 14 112132 99 84 52 109 81 5 5 5  
## 2 10 129367 91 123 95 75 113 57 123 44  
## 3 12 151522 113 49 95 107 116 5 5 5  
## 4 10 146457 43 64 109 80 88 48 77 33  
## 5 11 187684 99 104 67 48 120 94 90 30  
## 6 15 131228 66 104 121 96 65 73 114 29  
## RNA.Base RNA.4 RNA.12 RNA.EOT RNA.EF Baseline.histological.Grading  
## 1 655330 634536 288194 5 5 13  
## 2 40620 538635 637056 336804 31085 4  
## 3 571148 661346 5 735945 558829 4  
## 4 1041941 449939 585688 744463 582301 10  
## 5 660410 738756 3731527 338946 242861 11  
## 6 1157452 1086852 5 5 5 4  
## Baselinehistological.staging  
## 1 2  
## 2 2  
## 3 4  
## 4 3  
## 5 1  
## 6 4

## Now convert to a data table so that the genders can be broken out into separate tables for analysis  
  
library(data.table)  
catz <- data.table(catz)  
setkey(catz, Gender)  
males <- catz[Gender == 1]  
females <- catz[Gender == 2]  
  
  
## Compare the means of the samples for age, BMI, red blood count, hemoglobin, and AST1 and ALT1 by gender  
  
  
## Age Comparison  
age\_test <- t.test(males$Age, females$Age, conf.level = 0.95)  
  
print(age\_test)

##   
## Welch Two Sample t-test  
##   
## data: males$Age and females$Age  
## t = 0.3693, df = 1378.2, p-value = 0.712  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.7521527 1.1010281  
## sample estimates:  
## mean of x mean of y   
## 46.40453 46.23009

## BMI comparison  
  
  
BMI\_test <- t.test(males$BMI, females$BMI, conf.level = 0.95)  
  
print(BMI\_test)

##   
## Welch Two Sample t-test  
##   
## data: males$BMI and females$BMI  
## t = -0.24141, df = 1375.4, p-value = 0.8093  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.4832104 0.3773119  
## sample estimates:  
## mean of x mean of y   
## 28.58274 28.63569

## red blood cell count comparison  
  
  
RBC\_test <- t.test(males$RBC, females$RBC, conf.level = 0.95)  
  
print(RBC\_test)

##   
## Welch Two Sample t-test  
##   
## data: males$RBC and females$RBC  
## t = 0.035644, df = 1375.2, p-value = 0.9716  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -35896.46 37225.08  
## sample estimates:  
## mean of x mean of y   
## 4422455 4421791

## hemoglobin comparison  
  
  
HGB\_test <- t.test(males$HGB, females$HGB, conf.level = 0.95)  
  
print(HGB\_test)

##   
## Welch Two Sample t-test  
##   
## data: males$HGB and females$HGB  
## t = 0.046329, df = 1378, p-value = 0.9631  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.1765473 0.1850881  
## sample estimates:  
## mean of x mean of y   
## 12.58982 12.58555

## the aspartate transaminase ratio on the first day of observation comparison  
  
  
AST1\_test <- t.test(males$AST.1, females$AST.1, conf.level = 0.95)  
  
print(AST1\_test)

##   
## Welch Two Sample t-test  
##   
## data: males$AST.1 and females$AST.1  
## t = 0.48664, df = 1382.6, p-value = 0.6266  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -2.060442 3.419978  
## sample estimates:  
## mean of x mean of y   
## 83.10750 82.42773

## The alanine transaminase ratio on the first day of observation comparison  
  
  
AST1\_test <- t.test(males$ALT.1, females$ALT.1, conf.level = 0.95)  
  
  
print(AST1\_test)

##   
## Welch Two Sample t-test  
##   
## data: males$ALT.1 and females$ALT.1  
## t = -0.83501, df = 1379.9, p-value = 0.4039  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -3.897862 1.570283  
## sample estimates:  
## mean of x mean of y   
## 83.34653 84.51032

## And let's see if there was a gender difference in the opening histology  
  
  
whiz <- table(catz$Gender, catz$Baselinehistological.staging)  
  
print(whiz)

##   
## 1 2 3 4  
## 1 172 183 162 190  
## 2 164 149 193 172

whizzy <- chisq.test(whiz)  
  
print(whizzy)

##   
## Pearson's Chi-squared test  
##   
## data: whiz  
## X-squared = 6.6702, df = 3, p-value = 0.08319