Lab_3

Natarajan Shankar

July 27, 2016

```
# Load required libraries
library(gmodels)
library(ggplot2)
library(car)
library(lsr)
```

```
# Load the base 1993 data provided along with assignment handout
load("GSS.Rdata")
```

DATA ANALYSIS AND SHORT ANSWER SECTION

==========

Problem 14: Conduct a Chi-Square test to determine if there is association between marital status (marital) and political orientation (politics)

==========

```
# extract just the two fields of interest - marital and politics - so that we can
# deal with a smaller data frame. Call the data frame - dataDF
dataDF <- data.frame(GSS$marital, GSS$politics)
names(dataDF) <- c("marital", "politics")

# ============================================================================
# View the data to understand its makeup and quirks
# ============================================================================
# Check class type of "marital", verify as "factor" for applying chi square test
# Check class type of "politics", verify as "factor" for applying chi square test
class(dataDF$marital); class(dataDF$politics)
```

```
## [1] "factor"
```

```
## [1] "factor"
```

```
# Check levels of "marital" and "politics"
levels(dataDF$marital) ; levels(dataDF$politics)
```

```
## [1] "married"       "widowed"       "divorced"      "separated"
## [5] "never married" "NA"
```

```
## [1] "Liberal"      "Tend Lib"     "Moderate"     "Tend Cons"
## [5] "Conservative"
```

```
# Also look at summary to see NAs
summary(dataDF$marital); summary(dataDF$politics)
```

```
##       married       widowed      divorced     separated never married
##           795           165           213            40           286
##            NA
##             1
```

```
##      Liberal     Tend Lib     Moderate    Tend Cons Conservative
##          193          193          527          248          282
##         NA's
##           57
```

```
# ============================================================================
# Clean the data
# ============================================================================
# Per the codebook, marital answers of "no answer" are included. Remove them
dataDF <- dataDF[!dataDF$marital == "no answer",]

# "marital" attribute has NA as factor, clean it up prior to processing
# "politics" has <NA>, these are true NOT available, remove them.
#dataDF <- dataDF[dataDF$marital != "NA",]
dataDF <- dataDF[-which(dataDF$marital == "NA"), ]
dataDF <- dataDF[!is.na(dataDF$politics),]

# Drop the "NA"factor from marital, those rows have been removed
dataDF$marital = droplevels(dataDF$marital)

# Check how many data points are in the base data frame
nrow(dataDF)
```

```
## [1] 1442
```

```
# ============================================================================
# Run the Chisquare test, also display CrossTable to verify expected cell count
# ============================================================================
# Run the Chisquare test
CrossTable(dataDF$marital, dataDF$politics, chisq=TRUE, expected=TRUE, sresid=TRUE, format="SPSS")
```

```
##
##    Cell Contents
## |-------------------------|
## |                   Count |
## |         Expected Values |
## | Chi-square contribution |
## |             Row Percent |
## |          Column Percent |
## |           Total Percent |
## |            Std Residual |
## |-------------------------|
##
## Total Observations in Table:  1442
##
##                | dataDF$politics
## dataDF$marital |      Liberal  |     Tend Lib  |     Moderate  |    Tend Cons  | Conservative  |
## ---------------|---------------|---------------|---------------|---------------|---------------|---------
##        married |           93  |           92  |          271  |          140  |          173  |
##                |      102.391  |      102.924  |      281.042  |      132.255  |      150.387  |
##                |        0.861  |        1.160  |        0.359  |        0.454  |        3.400  |
##                |       12.094% |       11.964% |       35.241% |       18.205% |       22.497% |      53.
##                |       48.438% |       47.668% |       51.423% |       56.452% |       61.348% |
##                |        6.449% |        6.380% |       18.793% |        9.709% |       11.997% |
##                |       -0.928  |       -1.077  |       -0.599  |        0.673  |        1.844  |
## ---------------|---------------|---------------|---------------|---------------|---------------|---------
##        widowed |           15  |           16  |           57  |           24  |           37  |
##                |       19.839  |       19.942  |       54.454  |       25.626  |       29.139  |
```

```
##                 |       1.180  |       0.779  |       0.119  |       0.103  |       2.121  |
##                 |      10.067% |      10.738% |      38.255% |      16.107% |      24.832% |      10.3
##                 |       7.812% |       8.290% |      10.816% |       9.677% |      13.121% |
##                 |       1.040% |       1.110% |       3.953% |       1.664% |       2.566% |
##                 |      -1.086  |      -0.883  |       0.345  |      -0.321  |       1.456  |
## ----------------|--------------|--------------|--------------|--------------|--------------|---------
##        divorced |          22  |          36  |          79  |          38  |          29  |         1
##                 |      27.162  |      27.304  |      74.555  |      35.085  |      39.895  |
##                 |       0.981  |       2.770  |       0.265  |       0.242  |       2.975  |
##                 |      10.784% |      17.647% |      38.725% |      18.627% |      14.216% |      14.
##                 |      11.458% |      18.653% |      14.991% |      15.323% |      10.284% |
##                 |       1.526% |       2.497% |       5.479% |       2.635% |       2.011% |
##                 |      -0.991  |       1.664  |       0.515  |       0.492  |      -1.725  |
## ----------------|--------------|--------------|--------------|--------------|--------------|---------
##       separated |           7  |           3  |          22  |           6  |           1  |
##                 |       5.193  |       5.220  |      14.253  |       6.707  |       7.627  |
##                 |       0.629  |       0.944  |       4.211  |       0.075  |       5.758  |
##                 |      17.949% |       7.692% |      56.410% |      15.385% |       2.564% |       2.7
##                 |       3.646% |       1.554% |       4.175% |       2.419% |       0.355% |
##                 |       0.485% |       0.208% |       1.526% |       0.416% |       0.069% |
##                 |       0.793  |      -0.972  |       2.052  |      -0.273  |      -2.400  |
## ----------------|--------------|--------------|--------------|--------------|--------------|---------
##  never married  |          55  |          46  |          98  |          40  |          42  |         1
##                 |      37.415  |      37.610  |     102.696  |      48.327  |      54.953  |
##                 |       8.265  |       1.872  |       0.215  |       1.435  |       3.053  |
##                 |      19.573% |      16.370% |      34.875% |      14.235% |      14.947% |      19.4
##                 |      28.646% |      23.834% |      18.596% |      16.129% |      14.894% |
##                 |       3.814% |       3.190% |       6.796% |       2.774% |       2.913% |
##                 |       2.875  |       1.368  |      -0.463  |      -1.198  |      -1.747  |
## ----------------|--------------|--------------|--------------|--------------|--------------|---------
##   Column Total  |         192  |         193  |         527  |         248  |         282  |        14
##                 |      13.315% |      13.384% |      36.546% |      17.198% |      19.556% |
## ----------------|--------------|--------------|--------------|--------------|--------------|---------
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  44.2255     d.f. =  16     p =  0.0001822704
##
##
##
##        Minimum expected frequency: 5.192788
```

- All cells show expected value above 5
- There is not a need to run the Fisher test, Chi square test is appropriate

```
# Conduct an effect size calculation
effect_size <- sqrt(44.2255/(nrow(dataDF) * 4))
effect_size
```

```
## [1] 0.08756363
```

==========

Answers to questions : Problem 14

==========

Answer to Question 14 - A

- Null Hypothesis : “Marital” and “Politics” are independent
- Alternative Hypothesis : Gain confidence that “Marital” and “Politics” are related in some way

Answer to Question 14 - B

```
* Value of the test statistic is : 44.2255
* p value is : 0.0001822704
```

Answer to Question 14 - C

```
* The effect size is : ``0.0875636``
```

- Verify using lsr package also: `0.0875636`

Answer to Question 14 - D

- Conclusion:
- There was an significant association between “marital” and “politics” with the test statistic of 44.225 at $p < 0.05$
- This seems to represent that politics is impacted by marital
- Based upon this data, the NULL hypopthesis is rejected and confidence is gained that the alternative hypothesis is true.
- Based upon the computed effect size of 0.08756363, the effect size is “small”, leaning to “medium”, for 16 degrees of freedom

Problem 15: Conduct a Pearson Correlation Analysis to examine the association between age when married (agewed) and hours of TV watched (tvhours)

```
# extract just the two fields of interest
dataDF <- data.frame(GSS$agewed, GSS$tvhours)
names(dataDF) <- c("agewed", "tvhours")

# ============================================================================
# View the data to understand its makeup and quirks
# ============================================================================
# Check class type of "agewed"
# Check class type of "tvhours"
class(dataDF$agewed); class(dataDF$tvhours)
```

```
## [1] "numeric"
```

```
## [1] "numeric"
```

```
# Both variables verified to be numeric

# Check data types in "agewed" and "tvhours"
summary(dataDF$agewed); summary(dataDF$tvhours)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   18.00   21.00   19.06   24.00   99.00
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.000   2.000   3.605   4.000  99.000
```

```
# ============================================================================
# Clean the data
# ============================================================================
# The "no answer" for agewed is coded as 99, remove these
# the "no answer" for tvhours is also coded as 99, remove these
dataDF <- dataDF[!dataDF$agewed == 99,]
dataDF <- dataDF[!dataDF$tvhours == 99,]

# agewed value of 0 appears to be erroneous, remove them
dataDF <- dataDF[!dataDF$agewed == 0,]

# Re-check data types in "agewed" and "tvhours"
summary(dataDF$agewed); summary(dataDF$tvhours)
```

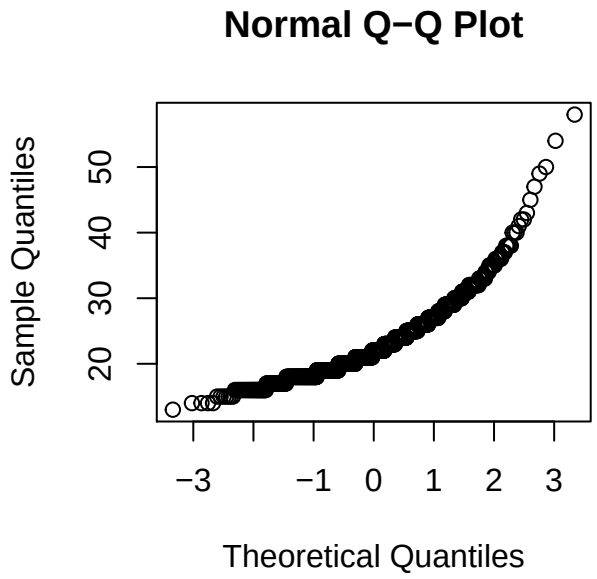
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.00   19.00   22.00   22.77   25.00   58.00
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.000   2.000   2.902   4.000  24.000
```

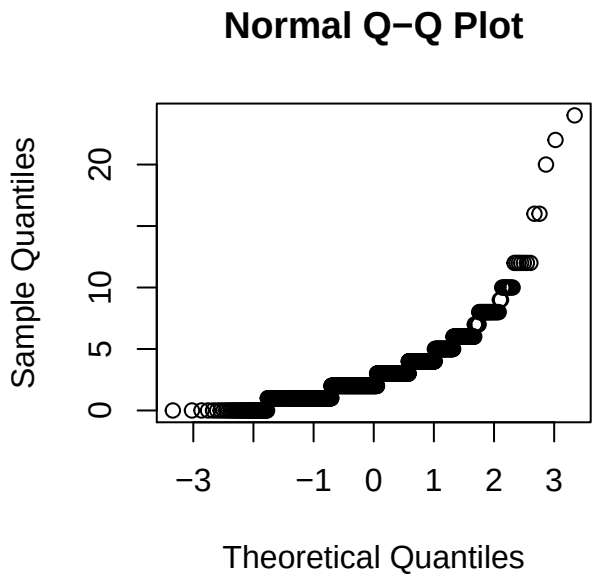
```r
# Check how many data points are in the base data frame
nrow(dataDF)
```

```
## [1] 1194
```

```r
# ============================================================================
# Get a feel for Normality of data
# ============================================================================
qqnorm(dataDF$agewed)
```




```r
qqnorm(dataDF$tvhours)
```



```r
# ============================================================================
# Run the Pearson test, assume that sample size is large enough to not
# require the running od Spearman test
# ============================================================================
pearson_result <- cor.test(dataDF$agewed, dataDF$tvhours, method="pearson", conf.level=0.95, na.rm=TRUE)
pearson_result
```

```
##
##  Pearson's product-moment correlation
##
## data:  dataDF$agewed and dataDF$tvhours
## t = -1.0349, df = 1192, p-value = 0.3009
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.08654554  0.02681630
## sample estimates:
##         cor
## -0.02996096
```

==========

Answers to questions : Problem 15

==========

Answer to Question 15 - A

- Null hypothesis is that correlation is 0, i.e no correlation between agewed and tvhours
- Alternative hypothesis is that the correlaton between agewed and tvhours is different from 0

Answer to Question 15 - B

- Test statistic is : `-1.03`
- p-value is : `0.3009361`
- Correlation coefficient is `-.03`

Answer to Question 15 - C

- Conclusion:
- Pearson's correlation between agewed and tv hours is `-0.029961` and p-value is `0.3009361`
- Based upon the correlation statistic and the statistically non-significant p value it can be conlcuded that agewed and tvhours are weakly correlated and that the Null hypotheses of zero correlation cannot be rejected

==========

Problem 16: Create a new binary variable "Married" that denotes whether an individual is currently married or not married. Then consider just the 23 year olds in this sample. Conduct a Wilcox ran-sum test to determine whether the new married variable is associated with children of 23 year olds

==========

```
# Look only for 23 year olds
dataDF <- GSS[GSS$age == 23,]

# For convenience, extract just the two fields of interest
dataDF <- data.frame(dataDF$marital, dataDF$childs)
names(dataDF) <- c("marital", "childs")

# ============================================================================
# View the data to understand its makeup and quirks
# ============================================================================
# Check class type of "agewed" - should be numeric
# Check class type of "tvhours" - should be factor
class(dataDF$marital); class(dataDF$childs)
```

```
## [1] "factor"
```

```
## [1] "numeric"
```

```
# Check data types in "marital" and "childs"
summary(dataDF$marital); summary(dataDF$childs)
```

```
##       married       widowed      divorced     separated never married
##             8             0             0             0            20
##            NA
##             0
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.4643  1.0000  3.0000
```

```
# ============================================================================
# Clean the data
# ============================================================================
# "marital" attribute has NA, clean it up prior to processing
dataDF <- dataDF[dataDF$marital != "NA",]

# Drop the "NA"factor from marital, those rows have been removed
dataDF$marital = droplevels(dataDF$marital)
```

```
## Create a new variable "married"
dataDF$married <- factor(ifelse(dataDF$marital == "married", as.numeric(1), as.numeric(0)))

# From the codebook, childs = 9 is a "no answer". Remove these rows
dataDF <- dataDF[!dataDF$childs == 99,]

# For convenience, extract just the two fields of interest
dataDF <- data.frame(dataDF$married, dataDF$childs)
names(dataDF) <- c("married", "childs")

# Check data types in "marital" and "childs"
class(dataDF$married); class(dataDF$childs)
```

```
## [1] "factor"
```

```
## [1] "numeric"
```

```
# Check homogeniety of variance
leveneTest(dataDF$childs, dataDF$married, center=mean)
```

```
## Levene's Test for Homogeneity of Variance (center = mean)
##       Df F value  Pr(>F)
## group  1  5.6117 0.02555 *
##       26
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# p-value is statistically significant, The Homogeneity of variance requirement is violated
# However, the sample size is large, let's continue with Wilcox Rank Sum test

# ================================================================================
# Run the Wilcox Rank Sum test, ignore the warning message
# ================================================================================
# run the wilcox test
wilcoxModel <- wilcox.test(dataDF$childs, as.numeric(dataDF$married), paired=FALSE)
```

```
## Warning in wilcox.test.default(dataDF$childs, as.numeric(dataDF$married),
## : cannot compute exact p-value with ties
```

```
wilcoxModel
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  dataDF$childs and as.numeric(dataDF$married)
## W = 132, p-value = 3.392e-06
## alternative hypothesis: true location shift is not equal to 0
```

```r
# What is the mean of your new "married" variable, proportion of cases coded 1, out of total
mean_married <- sum(dataDF[dataDF$married == "1",]$childs)/length(dataDF[dataDF$married == "1",]$childs)

# Proportion of married 23 year olds with children, amongst all with children
m_23_children <- nrow(dataDF[((dataDF$married == 1) & (dataDF$childs != 0)), ])/
                        nrow(dataDF[dataDF$married == 1, ])
```

==========

Answers to questions : Problem 16

==========

Answer to Question 16 - A

- Mean of the new "married" variable is: `1.25` children per married 23 year old
- Proportion of married 23 year olds with children : `0.875`

Answer to Question 16 - B

- Null hypothesis : The two groups (married and childs) are not different
- Alternative hypothesis : The two groups (married and childs) are different

Answer to Question 16 - C

- Test statistic is `132`
- p-value is `3.3919121\times 10^{-6}`

Answer to Question 16 - D : effect size calculation

```r
rFromWilcox <- function (wilcoxModel, N) {
  z <- qnorm(wilcoxModel$p.value/2)
  r <- z/sqrt(N)
  cat(wilcoxModel$data.name, "Effect size, r = ", r)
}

rFromWilcox(wilcoxModel, nrow(dataDF))
```

```
## dataDF$childs and as.numeric(dataDF$married) Effect size, r =  -0.8779242
```

Answer to Question 16 - E : Conclusion

- Conclusion: Based upon the p-value and the statistic above, the two groups are different
- and the NULL hypothesis can be rejected and confidence in alternate is gained
- Married status is associated with number of children for respondents who are 23 years old.
- The computed effect size is subjectively "huge".

==========

Problem 17: Conduct an ANOVA to determine if there is an association between religious affiliation (relig) and age when married (agewed)

==========

```
# extract just the fields of interest, "relig", "age", and "agewed"
# put into a convenience dataframe, dataDF
dataDF <- data.frame(GSS$relig, GSS$age, GSS$agewed)
names(dataDF) <- c("relig", "age", "agewed")

# ============================================================================
# View the important data to understand its makeup and quirks
# ============================================================================
# Check class type of "relig" - should be factor
# Check class type of "agewed" - should be factor
class(dataDF$relig); class(dataDF$agewed)
```

```
## [1] "factor"
```

```
## [1] "numeric"
```

```
# Check data types in "relig" and "agewed"
summary(dataDF$relig); summary(dataDF$agewed)
```

```
## Protestant   Catholic     Jewish       None      Other         DK
##        953        333         31        140         35          1
##         NA
##          7
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   18.00   21.00   19.06   24.00   99.00
```

```
# ============================================================================
# Clean the data
# ============================================================================
#Remove instances where agewed is greater than age
dataDF <- dataDF[!(GSS$agewed > GSS$age), ]

# remove the age column now, we are done with it
dataDF <- data.frame(dataDF$relig, dataDF$agewed)
names(dataDF) <- c("relig", "agewed")

# The "no answer" for agewed is coded as 99, remove these as well
dataDF <- dataDF[!is.na(dataDF$relig),]
dataDF <- dataDF[!dataDF$agewed == 99,]
```

```r
# agewed value of 0 appears to be erroneous, remove them
dataDF <- dataDF[!dataDF$agewed == 0,]

# get rid of the NAs in the relig column
dataDF <- dataDF[-which(dataDF$relig == "NA"), ]

# get rid of the DKs in the relig column
dataDF <- dataDF[-which(dataDF$relig == "DK"), ]

# Drop the "NA"factor from relig, those rows have been removed
dataDF$relig = droplevels(dataDF$relig)

# ============================================================================
# Run Levene's test, Homogenity of variance is a requirement
# ============================================================================
leveneTest(dataDF$agewed, dataDF$relig, center = mean)
```

```
## Levene's Test for Homogeneity of Variance (center = mean)
##         Df F value Pr(>F)
## group    4  0.8367 0.5018
##       1189
```

```r
# The Levene Test p-value is not statistically significant, the homogenity of variance assumption
# is not violated

# ============================================================================
# now run ANOVA and gather the output
# ============================================================================
aovData <- aov(agewed ~ relig, data = dataDF)
summary(aovData)
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## relig          4    804  200.99    8.12 1.85e-06 ***
## Residuals   1189  29430   24.75
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Manually compute the effect size
effectSize <- (804 - (2* 24.75))/(804 + 29430 + 24.75)
effectSize ^ 2
```

```
## [1] 0.000621751
```

==========

Answers to questions : Problem 17

==========

Answer to Question 17 - A

- Null hypothesis : All group means are equal
- Alternative hypothesis : At least one group mean is different from another

Answer to Question 17 - B

- The test statistic is an F value of : 8.12
- The p value is : 1.85e-06

Answer to Question 17 - C

- There is statistically significant difference between individual pairs of groups
- This is indicated by the p-value of < 0.05

Answer to Question 17 - D

- Conclusions:
- The F ratio has the value 8.12, degrees of freedom (4, 1189), F(4, 1189) = 8.12
- The F ratio value > 1 is indicating that all group means are not equal, This along
- the p-value indicates that the Null hypothesis
- can be rejected and confidence in the alternative is gained. However, the effect
- size will determine actual practical significance
- The effect size is `6.2175104\times 10^{-4}`
- In spite of statistically significanct results, the effect size is “low”"