MIDS W203 Lab 3

Matthew Potts

April 5, 2016

Part 1. Multiple Choice

1. C
2. B, C
3. C
4. A
5. D
6. C
7. B
8. B

Part 2. Test Selection

9. E
10. D
11. A
12. B
13. D

Part 3. Data Analysis and Short Answer

14. Task 1: Conduct a chi-square test to determine if there is an association between marital status (marital) and political orientation (politics).

First, I want to look at the variables used in the test and do some cleaning/recoding if necessary.

```
# Load libraries, load GSS data.
library(ggplot2)
library(car)
load("GSS.Rdata")

# Look at a summary of the marital and politics variable
summary(GSS$marital)
```

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

```
summary(GSS$politics)
```

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

```
# Look at the table of marital and politics together.
table(GSS$politics, GSS$marital)
```

```
##
##                married widowed divorced separated never married  NA
##   Liberal           93      15       22         7            55   1
##   Tend Lib          92      16       36         3            46   0
##   Moderate         271      57       79        22            98   0
##   Tend Cons        140      24       38         6            40   0
##   Conservative     173      37       29         1            42   0
```

```
# I noticed that there was an actual NA category in the marital column. I did not
#notice when I used the summary function on marital. If it was a real NA it
#probably would have said NA's like the politics summary did.

# The NA level should be removed. I subset to remove NA, drop the empty NA level
# and re-table the data.
GSS$marital[GSS$marital == "NA"] <- NA
GSS$marital <- droplevels(GSS$marital)
table(GSS$politics, GSS$marital)
```

```
##
##                married widowed divorced separated never married
##   Liberal           93      15       22         7            55
##   Tend Lib          92      16       36         3            46
##   Moderate         271      57       79        22            98
##   Tend Cons        140      24       38         6            40
##   Conservative     173      37       29         1            42
```

14a. The null hypothesis is that there is no association between marital status and political orientation. The alternative hypothesis is that there is an association bewteen these two variables. The alpha level is 0.05.

```
# Run the chi-square test to get a test statistic and p-value.
cs <- chisq.test(GSS$marital, GSS$politics)
cs
```

```
##
##  Pearson's Chi-squared test
##
## data:  GSS$marital and GSS$politics
## X-squared = 44.225, df = 16, p-value = 0.0001823
```

14b. The test statistic is 44.225 and the p-value is 0.0001823.

```r
# Use the Cramer's V calculation of effect size and the function from a sync
# material.
cramers_v = function(cs)
{
    cv = sqrt(cs$statistic / (sum(cs$observed) * (min(dim(cs$observed))-1)))
    print.noquote("Cramer's V:")
    return(as.numeric(cv))
}

cramers_v(cs)
```

```
## [1] Cramer's V:
```

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

14c. The effect size is 0.08756

14d. The p-value indicates that there may be an association between marital status and political orientation and we can reject the null hypothesis that there is no association between the two variables.

While we can reject the null hypothesis, the practical significance of the association may be small because of our small effect size calculation.

15. Task 2: Conduct a Pearson correlation analysis to examine the association between age when married (agewed) and hours of tv watched (tvhours).

```r
# Look at a summary of the agewed and tvhours variable
summary(GSS$agewed)
```

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

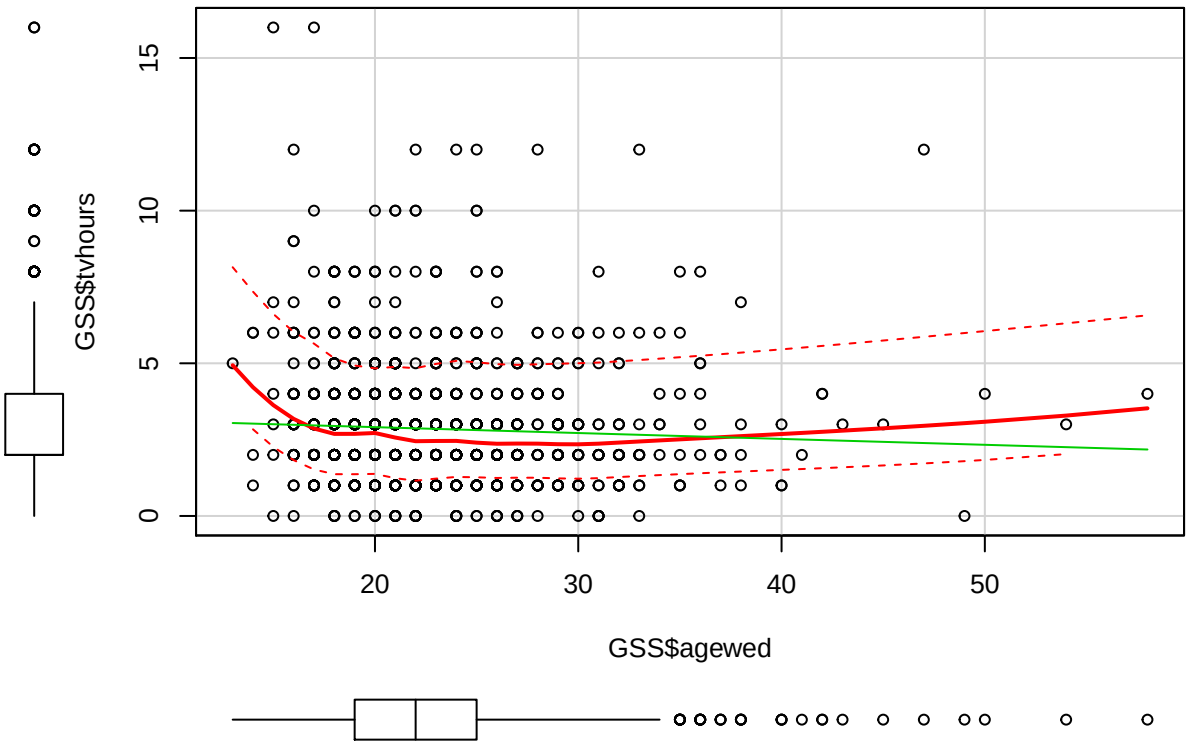
```r
summary(GSS$tvhours)
```

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

```r
# Agewed has some problems that we spotted in the last lab. Let's use that
# code to clean up that variable.

# Find weird values where agewed is greater than current age convert to NA.
weird <- GSS$id[GSS$agewed > GSS$age]
weird
```

```
##  [1]  215  359  361  565  595  609  848  853 1013 1222 1339 1581 1594
```

```r
# There are some 0 values and some 99 values that do not make sense. There are
#also 13 values (mostly agewed = 99, and a few agewed = 26) where the agewed
# is greater than the age of the person.

# Take the values of agewed that are greater than age and convert to NA.
for(i in weird) {
  GSS$agewed[GSS$id == i] <- NA
}

# Turn agewed = 0 to NAs
GSS$agewed[GSS$agewed == 0] <- NA

# Turn agewed = 99 to NAs
GSS$agewed[GSS$agewed == 99] <- NA

# tvhours is the average hours of tv watched daily. Some of the values are over
# 24 hrs. In fact, some of the values are really high, which is pretty suspect.
# I'm going to assume there are 8 hrs a day that you cannot physically watch tv
# so anything over 16 hrs I'm coding as NA. Even the amount of people who watch
# less than an hour a day on average seems off, but I'll let that slide for now.

# Turn tvhours > 16 to NAs
GSS$tvhours[GSS$tvhours > 16] <- NA

# I would just like to look at the data before running a correlation test.
scatterplot(GSS$agewed, GSS$tvhours)
```



15a. The null hypothesis is that there is no correlation between agewed and tvhours. The alternative is that there is a correlation between agewed and tvhours. The alpha level is 0.05.

```
cor.test(GSS$agewed, GSS$tvhours)
```

```
##
##  Pearson's product-moment correlation
##
## data:  GSS$agewed and GSS$tvhours
## t = -1.6656, df = 1187, p-value = 0.09607
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.104849596  0.008587917
## sample estimates:
##         cor
## -0.04828654
```

15b. The test statistic is -1.6656 and the p-value is 0.09607.

15c. Both the statistical significance and the effect size point towards no correlation between these two variables. With a p-value of 0.09607 we cannont reject the null hypothesis. Visual inspection of this data backs up this test statistic.

16. Task 3: Create a new binary/dummy variable, "married", that denotes whether an individual is currently married or not currently married. Next, we want to consider just the subpopulation of 23-year olds in this sample. Conduct a Wilcox rank-sum test to determine whether your new "married" variable is associated with the number of children (childs) for respondents who are 23 years old.

```
#Create dummy variables for all of the levels in marital. Use dummy_married.
for(level in unique(GSS$marital)){
  GSS[paste("dummy", level, sep = "_")] <- ifelse(GSS$marital == level, 1, 0)
}

#subset the GSS data frame to just 23 year olds.
GSS_23 <- subset(GSS, age == 23)

#Find the mean of the dummy_married variable
mean(GSS_23$dummy_married)
```

```
## [1] 0.2857143
```

16a. The mean of the dummy_married variable is 0.2857.

16b. The null hypothesis is that the mean amount of children currently married and not currently married is the same. The alternative is that the means for the dummy variable is not the same. The alpha level is 0.05.

```r
wilcox.test(GSS_23$childs ~ GSS_23$dummy_married)
```

```
## Warning in wilcox.test.default(x = c(0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, :
## cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  GSS_23$childs by GSS_23$dummy_married
## W = 19, p-value = 0.0002656
## alternative hypothesis: true location shift is not equal to 0
```

16c. The Wilcox rank-sum test statistic is 19 and the p-value is 0.0002656.

```r
# Use cohens d effect size calculation function from async class.
cohens_d <- function(x, y) {
    # this function takes two vectors as inputs, and compares
    # their means

    # first, compute the pooled standard error
  lx = length(subset(x,!is.na(x)))
  ly = length(subset(y,!is.na(y)))
    # numerator of the pooled variance:
    num = (lx-1)*var(x, na.rm=T) + (ly-1)*var(y, na.rm=T)
    pooled_var = num / (lx + ly - 2) # variance
    pooled_sd = sqrt(pooled_var)

    # finally, compute cohen's d
    cd = abs(mean(x, na.rm=T) - mean(y, na.rm=T)) / pooled_sd
    return(cd)
}

# Split childs into 2 vectors based on married and not married
childs_married <- GSS_23$childs[GSS_23$dummy_married == 1]
childs_notmarried <- GSS_23$childs[GSS_23$dummy_married == 0]
cohens_d(childs_notmarried, childs_married)
```

```
## [1] 1.976885
```

16d. The Cohen's D effect size calculation yields a value of 1.976885.

16e. Using the Wilcox rank-sum test we can reject the null hypothesis at the 0.05 alpha level. The effect size test also revealed a strong practical significance.

17. Task 4: Conduct an analysis of variance to determine if there is an association between religious affiliation (relig) and age when married (agewed).

```r
# agewed has been cleaned up, so let's check relig
summary(GSS$relig)
```

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

```r
# relig has a NA factor level that should not be there. Lets clean that up
# like we did with marital status.
GSS$relig[GSS$relig == "NA"] <- NA
GSS$relig <- droplevels(GSS$relig)
summary(GSS$relig)
```

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

15a. The null hypothesis is the mean of agewed is the same across all of the groups in relig. The alternative hpothesis is the mean of agewed is not the same across all of the groups in relig. The alpha level is 0.05.

```r
# perform an analysis of variance test
aovm <- aov(agewed ~ relig, GSS)
summary(aovm)
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## relig          5    805  161.09   6.508 5.56e-06 ***
## Residuals   1189  29430   24.75
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 305 observations deleted due to missingness
```

15b. The test statistic is 6.508 and the p-value is 5.56x10^-6.

```r
# use a pairwise t test to explore the statistical significance of differences
# between pairs of groups.
tt <- pairwise.t.test(GSS$agewed, GSS$relig, p.adjust.method = "bonferroni")
tt
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  GSS$agewed and GSS$relig
##
##          Protestant Catholic Jewish None Other
```

```
## Catholic -          -        -      -    -
## Jewish   -          -        -      -    -
## None     -          -        -      -    -
## Other    -          -        -      -    -
## DK       -          -        -      -    -
##
## P value adjustment method: bonferroni
```

15c. There are no significant differences between the individual pairs of groups.

```
# Look at the effect size of the test statistic
library(lsr)
etaSquared(aovm, type = 2, anova = TRUE)
```

```
##               eta.sq eta.sq.part         SS   df        MS        F
## relig     0.02663962  0.02663962   805.4632    5 161.09265 6.508279
## Residuals 0.97336038          NA 29430.0799 1189  24.75196       NA
##                      p
## relig     5.562517e-06
## Residuals           NA
```

15d. At the 0.05 alpha level we can reject the null hypothesis that the mean of agewed is the same across all groups. However the pairwise t test showed no significant pairs of groups. The effect size of the test statistic was relatively small. This leads me to believe that there is not much pratical significance in the differences of variability between means.