W203 Lab 3

Mohammad Jawad Habib

April 3, 2016

Part 1: Multiple Choice

- Q1 : b
- Q2 : b, c
- Q3 : b
- Q4 : c
- Q5 : d
- Q6 : a
- Q7 : b
- Q8 : b

Part 2: Test Selection

- Q9 : e, Chi-square test
- Q10: d, ANOVA
- Q11: c, Wilcoxon Rank-Sum Test
- Q12: b, Pearson correlation
- Q13: d, Chi-square test

Part 3: Data Analysis and Short Answer

```r
# setwd("W203 Week 12/Lab 3")
load("GSS.Rdata")
```

14. Task 1: Chi-Square Test on Marital Status and Political Orientation

A. Null and Alternative Hypothesis

H0: marital status and political orientation are independent H1: marital status and poliical orientation are not independent (that is, knowning about marital status can help predict the political orientation).

We can assume that the various marital statuses are independent of each other i.e. one person is not married and widowed at the same time. We can assume independence for political categorization as well.

We will also see below that none of the expected frequencies are below 5.

```r
mp <- data.frame(GSS$marital, GSS$politics)

# remove the value coded as "NA" from "marital"
mp$GSS.marital[mp$GSS.marital == "NA"] <- NA
```

```r
# remove "NA" as a factor from marital
mp$GSS.marital <- factor(mp$GSS.marital,
                         c("married", "widowed", "divorced", "separated",
                           "never married"))

# remove all rows that have NA for "marital" or "politics"
mp <- mp[complete.cases(mp),]

# run the Chi Square test
mpcs <- chisq.test(table(mp))

# see that no expected frequencies are below 5
mpcs$expected
```

```
##                  GSS.politics
## GSS.marital         Liberal   Tend Lib  Moderate   Tend Cons Conservative
##   married        102.391123 102.924411 281.04230 132.255201   150.386963
##   widowed         19.839112  19.942441  54.45423  25.625520    29.138696
##   divorced        27.162275  27.303745  74.55479  35.084605    39.894591
##   separated        5.192788   5.219834  14.25312   6.707351     7.626907
##   never married   37.414702  37.609570 102.69556  48.327323    54.952843
```

B. Test Statistics and p-value

We get a test statistic and p-value as follows:

```r
# test statistic
mpcs$statistic
```

```
## X-squared
##   44.2255
```

```r
# p-value
mpcs$p.value
```

```
## [1] 0.0001822704
```

Given the p-value above we can reject the null hypothesis. We can say that marital status does seem to be related to political orientation.

C. Effect Size Calculation

We will use Cramer's V for the effect size.

```r
cv <- sqrt(mpcs$statistic / (length(mp$GSS.marital)*min(nrow(mpcs$observed) - 1, ncol(mpcs$observed) - 
names(cv) <- "Cramer's V"
cv
```

```
## Cramer's V
## 0.08756363
```

D. Interpretation

Our Chi-Square test reveals that "politics" is significantly related to "marital" status:

$$\chi^2(16) = 44.225, p < 0.01$$

. The contingency table is shown below.

```
mpcs$observed
```

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

We can also see from the standardized residuals which ones are significant outside of +/- 1.96 (p < 0.05).

```
mpcs$stdres > 1.96 | mpcs$stdres < -1.96
```

```
##                 GSS.politics
## GSS.marital     Liberal Tend Lib Moderate Tend Cons Conservative
##   married         FALSE    FALSE    FALSE     FALSE         TRUE
##   widowed         FALSE    FALSE    FALSE     FALSE        FALSE
##   divorced        FALSE    FALSE    FALSE     FALSE         TRUE
##   separated       FALSE    FALSE     TRUE     FALSE         TRUE
##   never married    TRUE    FALSE    FALSE     FALSE         TRUE
```

Being "married" is significantly related with "Conservative" political view. "widowed" is not significantly related with any political view. "divorced" is significantly related with "Conservative". "separated" is significantly related with "Moderate" and "Conservative". "never married" is significantly related with "Liberal" and "Conservative".

15. Task 2: Pearson Correlation on Age when Married and Hours of TV watched

A. Null and Alternative Hypothesis

H0: there is no relationship between agewed and tvhours (r = 0) H1: there is a positive or negative relationship between agewed and tvhours

B. Test Statistics and p-value

GSS website is not very clear on how missing values in `agewed` and `tvhours` are coded. For `agewed` let's assume that 0 and 99 are dummy values. for `tvhours` we can assume that anything greater than 24 is a dummy value because there are only 24 hours in the day.

Note: our sample size is large so we can assume normality of our sampling distribution. This assumption is required for establishing whether correlation coefficient is significant.

```r
at <- data.frame(GSS$agewed, GSS$tvhours)
at <- at[!(at$GSS.agewed %in% c(0, 99)) & at$GSS.tvhours <= 24,]

library(Hmisc)

atpc <- rcorr(as.matrix(at))
atpc
```

```
##             GSS.agewed GSS.tvhours
## GSS.agewed        1.00       -0.03
## GSS.tvhours      -0.03        1.00
##
## n= 1194
##
##
## P
##             GSS.agewed GSS.tvhours
## GSS.agewed             0.3009
## GSS.tvhours 0.3009
```

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

```
##
##  Pearson's product-moment correlation
##
## data:  at$GSS.agewed and at$GSS.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
```

From the above, we get a test statistic (r) of -0.03 and a p-value of 0.3009. And we get a 95% confidence interval that passes through zero (-0.087, 0.027)

C. Interpretation

The results indicate that agewed is not correlated with tvhours. We also get

$$r^2 = 9e - 04$$

which means that only 0.09% of variability is shared between agewed and tvhours.

I ran `shapiro.wilk` on `agewed` and `tvhours` and the two are not normally distributed. We can also check this by `hist`. Therefore, I'm going to run a Kendall's Tau. Note: we relied on a large sample size before to assume normality but we don't need that assumption with the non-parametric Kendall's Tau. I did not use Spearman because of ties in ranked data and I also did not use bootstrapping (which would be trivial anyways).

```
cor.test(at$GSS.agewed, at$GSS.tvhours, method = "kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  at$GSS.agewed and at$GSS.tvhours
## z = -2.9978, p-value = 0.002719
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##         tau
## -0.06486956
```

Even though $p < 0.01$, the tau is -0.06 which shows that there may be a weak negative relationship between agewed and tvhours.

16. Task 3: Wilcox Rank-Sum on Marital Status and Number of Children for 23 year olds

We will remove the observation where marital value is set to "NA" and use the rest of the data for our analysis. We will also subset the data for age==23 aftewards (we could do this before too).

```
mc <- GSS[, c("age", "marital", "childs")]
mc$marital[mc$marital == "NA"] <- NA
mc$marital <- factor(mc$marital,
                     c("married", "widowed", "divorced", "separated",
                       "never married"))
mc <- mc[complete.cases(mc),]

mc$married <- ifelse(mc$marital == "married", 1, 0)
# mc$married <- factor(mc$married, c("married", "not married"))

# just keep those who are 23
mc <- mc[mc$age == 23,]
```

A. Mean of Married Variable

```
sum(mc$married) / length(mc$married)
```

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

The proportion of observations coded married = 1 is 0.286 in our subset.

B. Null and Alternative Hypothesis

H0: median number of children is the same for married and non-married 23-year olds H1: median number of children is not equal for married and non-married 23-year olds

C. Test Statistic and p-value

```r
# mc$married <- factor(mc$married, c("married", "not married"))
mcwrs <- wilcox.test(childs ~ as.factor(married), data = mc, exact = FALSE)
mcwrs
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  childs by as.factor(married)
## W = 19, p-value = 0.0002656
## alternative hypothesis: true location shift is not equal to 0
```

From the above test, we see that W = 19 and p = 0.000266 ($p < 0.01$).

D. Calculate Effect Size

```r
z <- qnorm(mcwrs$p.value/2)
r <- z/sqrt(length(mc$married))
r
```

```
## [1] -0.6891632
```

We see from the above that the effect size r is -0.689 which is conventionally considered a large effect (above 0.5).

E. Interpretation

Our results show that number of children in married 23-year olds (Mdn = 1) differed significantly from unmarried 23-year olds (Mdn = 0), W=19, p = 0.000266, r = -0.689. That is married 23-year olds had significantly more children than unmarried ones.

17. Task 4: ANOVA on Religious Affiliation and Age When Married

A. Null and Alternative Hypothesis

H0: mean age when married is the same across all religious affiliation H1: mean age when married is NOT the same across all religious affiliation

B. Test Statistic and p-value

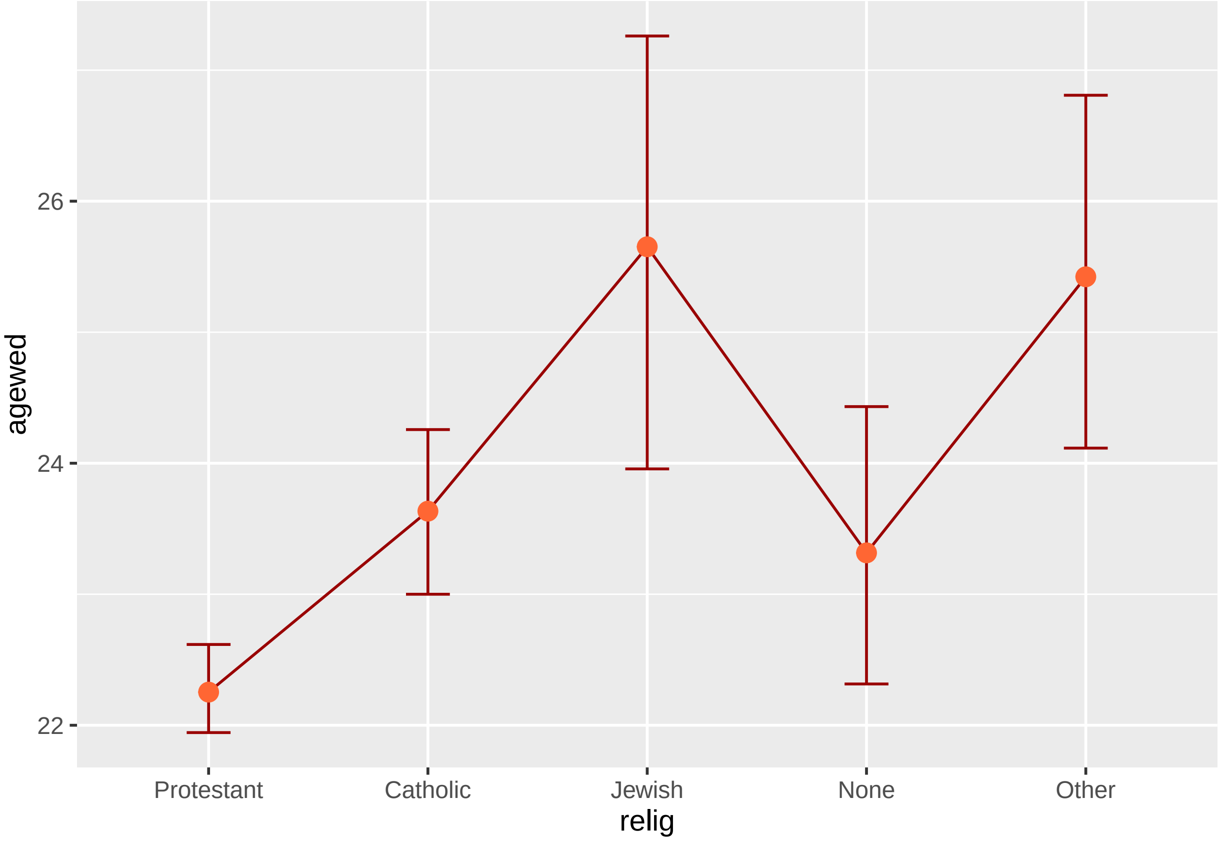
We will remove the observations where agewed equals 0 or 99. We will also remove the observations where relig is NA or DK (per the GSS website).

```
ar <- data.frame(GSS$agewed, GSS$relig)
ar <- ar[!(ar$GSS.agewed %in% c(0, 99)) & !(ar$GSS.relig %in% c("NA", "DK")),]

ar$GSS.relig <- factor(ar$GSS.relig,
                       c("Protestant", "Catholic", "Jewish", "None", "Other"))
names(ar) <- c("agewed", "relig")

require(ggplot2)

arp <- ggplot(ar, aes(relig, agewed)) +
  stat_summary(fun.y = mean, geom = "line", size = 0.5, aes(group = 1),
               colour = "#990000") +
  stat_summary(fun.data = mean_cl_boot, geom = "errorbar",
               width = 0.2, size = 0.5, colour = "#990000") +
  stat_summary(fun.y = mean, geom = "point", size = 3, colour = "#FF6633")
arp
```



The errorbars for protestants do not overlap with errorbars of catholic, jewish and other.

```
require(pastecs)
by(ar$agewed, ar$relig, stat.desc)
```

```
## ar$relig: Protestant
##      nbr.val     nbr.null       nbr.na          min          max
## 7.870000e+02 0.000000e+00 0.000000e+00 1.300000e+01 5.800000e+01
##        range          sum       median         mean      SE.mean
## 4.500000e+01 1.751300e+04 2.100000e+01 2.225286e+01 1.787230e-01
## CI.mean.0.95          var      std.dev     coef.var
## 3.508308e-01 2.513827e+01 5.013808e+00 2.253107e-01
## --------------------------------------------------------
## ar$relig: Catholic
##      nbr.val     nbr.null       nbr.na          min          max
##  265.0000000    0.0000000    0.0000000   14.0000000   49.0000000
##        range          sum       median         mean      SE.mean
##   35.0000000 6263.0000000   23.0000000   23.6339623    0.3073684
## CI.mean.0.95          var      std.dev     coef.var
##    0.6052055   25.0359634    5.0035950    0.2117121
## --------------------------------------------------------
## ar$relig: Jewish
##      nbr.val     nbr.null       nbr.na          min          max
##   23.0000000    0.0000000    0.0000000   20.0000000   37.0000000
##        range          sum       median         mean      SE.mean
##   17.0000000  590.0000000   26.0000000   25.6521739    0.8634170
## CI.mean.0.95          var      std.dev     coef.var
##    1.7906173   17.1462451    4.1408025    0.1614211
## --------------------------------------------------------
## ar$relig: None
##      nbr.val     nbr.null       nbr.na          min          max
##   95.0000000    0.0000000    0.0000000   14.0000000   38.0000000
##        range          sum       median         mean      SE.mean
##   24.0000000 2215.0000000   22.0000000   23.3157895    0.5145722
## CI.mean.0.95          var      std.dev     coef.var
##    1.0216952   25.1545353    5.0154297    0.2151087
## --------------------------------------------------------
## ar$relig: Other
##      nbr.val     nbr.null       nbr.na          min          max
##   26.0000000    0.0000000    0.0000000   19.0000000   32.0000000
##        range          sum       median         mean      SE.mean
##   13.0000000  661.0000000   26.0000000   25.4230769    0.7172020
## CI.mean.0.95          var      std.dev     coef.var
##    1.4771052   13.3738462    3.6570270    0.1438467
```

```r
require(car)
leveneTest(ar$agewed ~ ar$relig, center = median)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value Pr(>F)
## group    4  0.8521 0.4922
##       1191
```

```r
by(ar$agewed, ar$relig, shapiro.test)
```

```
## ar$relig: Protestant
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.85224, p-value < 2.2e-16
##
```

```
## --------------------------------------------------------
## ar$relig: Catholic
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.91753, p-value = 6.316e-11
##
## --------------------------------------------------------
## ar$relig: Jewish
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.935, p-value = 0.1402
##
## --------------------------------------------------------
## ar$relig: None
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.94581, p-value = 0.0006409
##
## --------------------------------------------------------
## ar$relig: Other
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.96314, p-value = 0.4571
```

The non-significant Levene Test above shows that we have homogeneity of variance. We do not have normally distributed groups for Protestant, Catholic and None. We do have normally distributed groups for Jewish and Other.

Section 10.3.2 of our text book says that F-statistic controls Type I error well under conditions of non-normality. Power of F-statistic is also relatively unaffected by non-normality per our text book when group sizes are equal. Unfortunately, that is not the case so we will stick with a robust ANOVA as described in section 10.6.6.3 of our text book.

```r
require(WRS)

arwide <- unstack(ar, agewed ~ relig)

# resample to make groups sizes equal
arwide <- data.frame(lapply(arwide, sample, size  = 30, replace = TRUE))

# t1way
levels(arwide)
```

```
## NULL
```

```
WRS::med1way(arwide)
```

```
## [1] "NOTE: This function was modified in Dec. 2004"
## [1] "A new approximate critical value is used if crit=NA"
## [1] "This might improve type I error probabilities substantially"
## [1] "For discrete data with ties, this function is NOT recommended."
## [1] "Use the function medpb; it is best for general use"
## [1] "WARNING: tied values detected."
## [1] "Estimate of standard error might be highly inaccurate, even with n large"
## [1] "WARNING: tied values detected."
## [1] "Estimate of standard error might be highly inaccurate, even with n large"
## [1] "WARNING: tied values detected."
## [1] "Estimate of standard error might be highly inaccurate, even with n large"
## [1] "WARNING: tied values detected."
## [1] "Estimate of standard error might be highly inaccurate, even with n large"
## [1] "WARNING: tied values detected."
## [1] "Estimate of standard error might be highly inaccurate, even with n large"
```

```
## $TEST
## [1] 2.60868
##
## $crit.val
## [1] 2.126715
##
## $p.value
## [1] 0.021
```

```
WRS::t1waybt(arwide, nboot = 2000)
```

```
## [1] "Taking bootstrap samples. Please wait."
## [1] "Working on group  1"
## [1] "Working on group  2"
## [1] "Working on group  3"
## [1] "Working on group  4"
## [1] "Working on group  5"
```

```
## $test
## [1] 3.449421
##
## $p.value
## [1] 0.0215
```

We have a non-significant result for median age by religion and a marginally significant result for mean.

Using the results from t1waybt above, we get a test statistic (F) of 3.380 and p.value of 0.042. From the med1way we get a test statistic (F) of 1.69 and p.value of 0.09.

C. Statistical Differences Between Individual Pairs

Since we did not have any directional hypothesis beforehand, we will run post-hoc tests.

```r
arpt <- pairwise.t.test(ar$agewed, ar$relig, paired = FALSE, p.adjust.method = "bonferroni")
arpt
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  ar$agewed and ar$relig
##
##          Protestant Catholic Jewish  None
## Catholic 0.00097    -        -       -
## Jewish   0.01263    0.62106  -       -
## None     0.49274    1.00000  0.43398 -
## Other    0.01417    0.80224  1.00000 0.55751
##
## P value adjustment method: bonferroni
```

Result table (`pairwise.t.test`) above shows that Catholic-Protestant ($p < 0.01$), Jewish-Protestant ($p < 0.05$) and Other-Protestant ($p < 0.05$) had significantly different age at wedding.

D. Evaluate Hypothesis

We need to know the p-value and confidence interval to evaluate our hypothesis.

```r
amodel <- aov(agewed ~ relig, ar)
TukeyHSD(amodel)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = agewed ~ relig, data = ar)
##
## $relig
##                            diff        lwr       upr     p adj
## Catholic-Protestant  1.3811033  0.4163497 2.3458569 0.0009208
## Jewish-Protestant    3.3993150  0.5258226 6.2728073 0.0110537
## None-Protestant      1.0629305 -0.4124497 2.5383107 0.2822690
## Other-Protestant     3.1702180  0.4625833 5.8778526 0.0123250
## Jewish-Catholic      2.0182116 -0.9345488 4.9709721 0.3356751
## None-Catholic       -0.3181728 -1.9425442 1.3061986 0.9837028
## Other-Catholic       1.7891147 -1.0025015 4.5807309 0.4031473
## None-Jewish         -2.3363844 -5.4930902 0.8203214 0.2559719
## Other-Jewish        -0.2290970 -4.1174553 3.6592613 0.9998495
## Other-None           2.1072874 -0.8992252 5.1138001 0.3098737
```

We can see from the above that Protestant-Catholic ($p < 0.01$), Protestant-Jewish ($p < 0.05$), Protestant-Other ($p < 0.05$) have 95% confidence intervals that do not pass through zero.

We can also calculate the desired effect sizes (Cohen's D).

```r
# Effect Size Calculation
require(lsr)
```

```
## Loading required package: lsr
```

```
# Protestant-Catholic
d <- ar[ar$relig %in% c("Protestant", "Catholic"),]
d$relig <- factor(d$relig, c("Protestant", "Catholic"))
cohensD(agewed ~ relig, data = d)
```

```
## [1] 0.275601
```

```
# Protestant-Jewish
d <- ar[ar$relig %in% c("Protestant", "Jewish"),]
d$relig <- factor(d$relig, c("Protestant", "Jewish"))
cohensD(agewed ~ relig, data = d)
```

```
## [1] 0.6809443
```

```
# Protestant-Other
d <- ar[ar$relig %in% c("Protestant", "Other"),]
d$relig <- factor(d$relig, c("Protestant", "Other"))
cohensD(agewed ~ relig, data = d)
```

```
## [1] 0.6369082
```

As calculated above, Cohen's D indicates a significant effect size for mean age (higher) of Catholic, Jewish and Others vs. that of Protestants.