PSY308d.DA1

Daniel Pinedo

April 2, 2019

**Main Analyses:**

A set of researchers from a marketing company conducted a survey to investigate reading habits of Americans (this is a real dataset). Although a study has been published on this already, they have tasked you with investigating it further in order to understand the relationships between certain demographic variables of their sampled readers, inculding sex, employment, martial status, race, and age.

While your advisor thinks there might be some interesting results that could come from this data for an upcoming conference presentation, she only wants a sample write-up of a Results and Discussion section to start based on the following proposed RQs.

Research Questions to Investigate:  
 \*1:\* Is there a relationship between sex and employment?  
 \*2:\* Is there a relationship between sex and education?   
 \*3:\* Is there a relationship between marital status and employment?

**Conceptual Addition:** Following your analyses - please incorporate into the Discussion section of your write-up the following conceptual addition:

1.) Your advisor would also like your notes on a potentially interesting caveat for using this information to make generalizations to the general public. Specifically, your advisor wants you to test that the proportions of the sample match those which may be expected in the U.S. population for race, per the proportions below:

White: 80%  
Black or African American: 12%  
Asian or Pacific Islander: 3%  
 Mixed Race: 2%  
 Native American/American Indian: 1%  
 Other: 2%

2.) Following testing, interpret these results and discuss how this may affect your ability to interpret the data and make generalizations. What suggestions could you make moving forward?

*Please report all relevant statistics per APA format and write for a professional audience.*

library(pacman) #Package used to load all packages using p\_load(); will install missing packages

## Warning: package 'pacman' was built under R version 3.5.3

p\_load(vcd, MASS, jmv, gmodels, VIM)

dat <- read.csv("https://www.dropbox.com/s/zhhyiegj8gyakuu/Reading.csv?dl=1")  
head(dat) # check to see if labels are needed. In this case they are not. See Chi2Demo.Rmd for how-to.

## Age Sex Race Married Married.status  
## 1 66 Male <NA> No Divorced  
## 2 46 Male Native American/American Indian Yes Married  
## 3 32 Male Mixed race No Never been married  
## 4 27 Male Mixed race Yes Married  
## 5 16 Female Mixed race No Never been married  
## 6 55 Female Asian or Pacific Islander No Divorced  
## Education Employement  
## 1 College graduate Retired  
## 2 High school graduate Employed full-time  
## 3 High school graduate Employed full-time  
## 4 High school graduate Employed full-time  
## 5 High school incomplete Employed part-time  
## 6 Some college, no 4-year degree Have own business/self-employed  
## Incomes  
## 1 $20,000 to under $30,000  
## 2 Less than $10,000  
## 3 Less than $10,000  
## 4 $40,000 to under $50,000  
## 5 $10,000 to under $20,000  
## 6 $40,000 to under $50,000  
## How.many.books.did.you.read.during.last.12months.  
## 1 97  
## 2 97  
## 3 97  
## 4 97  
## 5 97  
## 6 97  
## Read.any.printed.books.during.last.12months.  
## 1 Yes  
## 2 Yes  
## 3 No  
## 4 Yes  
## 5 Yes  
## 6 Yes  
## Read.any.audiobooks.during.last.12months.  
## 1 No  
## 2 Yes  
## 3 Yes  
## 4 No  
## 5 Yes  
## 6 Yes  
## Read.any.e.books.during.last.12months.  
## 1 Yes  
## 2 Yes  
## 3 Yes  
## 4 Yes  
## 5 No  
## 6 Yes  
## Last.book.you.read..youâ..  
## 1 Purchased the book  
## 2 Purchased the book  
## 3 Borrowed the book from a friend or family member  
## 4 Borrowed the book from a library  
## 5 Purchased the book  
## 6 Purchased the book  
## Do.you.happen.to.read.any.daily.news.or.newspapers.  
## 1 No  
## 2 Yes  
## 3 Yes  
## 4 Yes  
## 5 Yes  
## 6 No  
## Do.you.happen.to.read.any.magazines.or.journals.  
## 1 Yes  
## 2 Yes  
## 3 Yes  
## 4 No  
## 5 No  
## 6 No

Take a look at the data set,

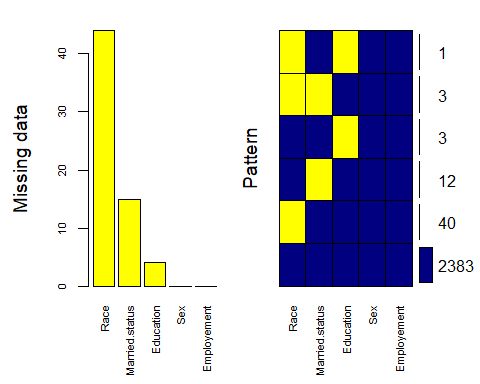
#Subset to use only necessary variables  
myvars <- c("Sex", "Race", "Married.status", "Education", "Employement")  
dat.subset <- dat[myvars]  
  
# Run descriptives for categorical variables  
  
glimpse <- descriptives(data = dat.subset,   
 vars = c('Sex', 'Employement', 'Education', 'Married.status', 'Race'))  
glimpse

##   
## DESCRIPTIVES  
##   
## Descriptives   
## -------------------------------------------------------------------------   
## Sex Employement Education Married.status Race   
## -------------------------------------------------------------------------   
## N 2442 2442 2438 2427 2398   
## Missing 0 0 4 15 44   
## Mean   
## Median   
## Minimum   
## Maximum   
## -------------------------------------------------------------------------

# First thing to note is the Row for "Missing" DATA --> Different N's and the line that indicates how many are missing shows there are missing cases in our sample of items to be tested  
# There are 44 missing total all from the "Race" category {no further analysis of missing values needed - see /Regression/Regression\_DIagnostics for how-to}  
#Option: (1) delete list-wise

Check missing data for patterns

#check the pattern of missing data  
VIM\_plot <-aggr(dat.subset,   
 col=c('navyblue', 'yellow'),   
 numbers = TRUE,   
 prop = FALSE,  
 sortVars = TRUE,   
 labels = names(dat.subset),   
 cex.axis = .7,   
 gap = 3,   
 ylab = c("Missing data", "Pattern"))



##   
## Variables sorted by number of missings:   
## Variable Count  
## Race 44  
## Married.status 15  
## Education 4  
## Sex 0  
## Employement 0

#yellow bar chart is percentage missing from each variable  
#blue and yellow chart shows pattern of missing data

Remove missing cases and view data again with **observed frequencies**

# Option: Listwise deletion of missing data. New dataset is named "dat.no.NA"  
dat.no.NA <- na.omit(dat.subset)  
  
# check descriptives again  
# no missing cases  
glimpse.no.NA <- descriptives(data = dat.no.NA,   
 vars = c('Sex', 'Employement', 'Education', 'Married.status', 'Race'),   
 freq = TRUE)  
glimpse.no.NA

##   
## DESCRIPTIVES  
##   
## Descriptives   
## -------------------------------------------------------------------------   
## Sex Employement Education Married.status Race   
## -------------------------------------------------------------------------   
## N 2383 2383 2383 2383 2383   
## Missing 0 0 0 0 0   
## Mean   
## Median   
## Minimum   
## Maximum   
## -------------------------------------------------------------------------   
##   
##   
## FREQUENCIES  
##   
## Frequencies of Sex   
## --------------------------------------------------   
## Levels Counts % of Total Cumulative %   
## --------------------------------------------------   
## Female 1302 54.6 54.6   
## Male 1081 45.4 100.0   
## --------------------------------------------------   
##   
##   
## Frequencies of Employement   
## ---------------------------------------------------------------------------   
## Levels Counts % of Total Cumulative %   
## ---------------------------------------------------------------------------   
## Disabled 45 1.9 1.9   
## Employed full-time 1066 44.7 46.6   
## Employed part-time 315 13.2 59.8   
## Have own business/self-employed 47 2.0 61.8   
## Not employed for pay 396 16.6 78.4   
## Other 10 0.4 78.9   
## Retired 482 20.2 99.1   
## Student 22 0.9 100.0   
## ---------------------------------------------------------------------------   
##   
##   
## Frequencies of Education   
## ----------------------------------------------------------------------------------------------------   
## Levels Counts % of Total Cumulative %   
## ----------------------------------------------------------------------------------------------------   
## College graduate 557 23.4 23.4   
## High school graduate 515 21.6 45.0   
## High school incomplete 202 8.5 53.5   
## None 28 1.2 54.6   
## Post-graduate training/professional school after college 468 19.6 74.3   
## Some college, no 4-year degree 565 23.7 98.0   
## Technical, trade or vocational school AFTER high school 48 2.0 100.0   
## ----------------------------------------------------------------------------------------------------   
##   
##   
## Frequencies of Married.status   
## -----------------------------------------------------------------   
## Levels Counts % of Total Cumulative %   
## -----------------------------------------------------------------   
## Divorced 192 8.1 8.1   
## Living with a partner 119 5.0 13.1   
## Married 1304 54.7 67.8   
## Never been married 535 22.5 90.2   
## Separated 35 1.5 91.7   
## Single 48 2.0 93.7   
## Widowed 150 6.3 100.0   
## -----------------------------------------------------------------   
##   
##   
## Frequencies of Race   
## ---------------------------------------------------------------------------   
## Levels Counts % of Total Cumulative %   
## ---------------------------------------------------------------------------   
## Asian or Pacific Islander 62 2.6 2.6   
## Black or African-American 277 11.6 14.2   
## Mixed race 54 2.3 16.5   
## Native American/American Indian 24 1.0 17.5   
## Other 48 2.0 19.5   
## White 1918 80.5 100.0   
## ---------------------------------------------------------------------------

Assumptions - 1. Adequate expected cell counts - 5 or more in 2 x 2 or 5 or more in 80% of cells for larger table - Otherwise, Fisher’s test - 2. Independence of Observations - otherwise McNemar’s test of dependent proportions

## Chi-square Test of Independence

H1: Is Sex dependent upon Employement? Is there a relationship between Sex and Employement? H2: Is Sex dependent upon Education? Is there a relationship between Sex and Education? H3: Is Married dependent upon Employement? Is there a relationship between Married and Employement?

Cramer’s V - small = .1; medium = .3, large = .5; indicates effect size of discrepancy between observed and expected scores

# Chi-square = Sum[(Observed - Expected)^2/Expected]  
# Expected = [(# of row entries for cel)/(# total entries)] \* (# of column entries for cel)  
# Expected indicates expected values for each category if there is no relationship between two categorical variables  
# df = (# rows - 1) \* (# columns - 1)  
# report APA, magnitude of effect (Cramer's V), direction of effect example (more or less than expected in each category - include Contingency Table)  
  
H1 <- jmv::contTables(dat = dat.no.NA,  
 rows = 'Employement',  
 cols = 'Sex',  
 exp = TRUE,  
 phiCra = TRUE)  
  
H1

##   
## CONTINGENCY TABLES  
##   
## Contingency Tables   
## ------------------------------------------------------------------------------   
## Employement Female Male Total   
## ------------------------------------------------------------------------------   
## Disabled Observed 25 20 45   
## Expected 24.59 20.41   
##   
## Employed full-time Observed 493 573 1066   
## Expected 582.43 483.57   
##   
## Employed part-time Observed 188 127 315   
## Expected 172.11 142.89   
##   
## Have own business/self-employed Observed 21 26 47   
## Expected 25.68 21.32   
##   
## Not employed for pay Observed 266 130 396   
## Expected 216.36 179.64   
##   
## Other Observed 9 1 10   
## Expected 5.46 4.54   
##   
## Retired Observed 283 199 482   
## Expected 263.35 218.65   
##   
## Student Observed 17 5 22   
## Expected 12.02 9.98   
##   
## Total Observed 1302 1081 2383   
## Expected 1302.00 1081.00   
## ------------------------------------------------------------------------------   
##   
##   
## <U+03C7>² Tests   
## -------------------------------   
## Value df p   
## -------------------------------   
## <U+03C7>² 73.3 7 < .001   
## N 2383   
## -------------------------------   
##   
##   
## Nominal   
## ----------------------------   
## Value   
## ----------------------------   
## Phi-coefficient NaN   
## Cramer's V 0.175   
## ----------------------------

H2 <- jmv::contTables(dat = dat.no.NA,  
 rows = 'Education',  
 cols = 'Sex',  
 exp = TRUE,  
 phiCra = TRUE)  
  
H2

##   
## CONTINGENCY TABLES  
##   
## Contingency Tables   
## -----------------------------------------------------------------------------------------------------   
## Education Female Male Total   
## -----------------------------------------------------------------------------------------------------   
## College graduate Observed 314 243 557   
## Expected 304.3 252.7   
##   
## High school graduate Observed 276 239 515   
## Expected 281.4 233.6   
##   
## High school incomplete Observed 108 94 202   
## Expected 110.4 91.6   
##   
## None Observed 13 15 28   
## Expected 15.3 12.7   
##   
## Post-graduate training/professional school after college Observed 245 223 468   
## Expected 255.7 212.3   
##   
## Some college, no 4-year degree Observed 322 243 565   
## Expected 308.7 256.3   
##   
## Technical, trade or vocational school AFTER high school Observed 24 24 48   
## Expected 26.2 21.8   
##   
## Total Observed 1302 1081 2383   
## Expected 1302.0 1081.0   
## -----------------------------------------------------------------------------------------------------   
##   
##   
## <U+03C7>² Tests   
## ------------------------------   
## Value df p   
## ------------------------------   
## <U+03C7>² 4.44 6 0.617   
## N 2383   
## ------------------------------   
##   
##   
## Nominal   
## -----------------------------   
## Value   
## -----------------------------   
## Phi-coefficient NaN   
## Cramer's V 0.0432   
## -----------------------------

H3 <- jmv::contTables(dat = dat.no.NA,  
 rows = 'Employement',  
 cols = 'Married.status',  
 exp = TRUE,  
 phiCra = TRUE)  
  
H3

##   
## CONTINGENCY TABLES  
##   
## Contingency Tables   
## ----------------------------------------------------------------------------------------------------------------------------------------------------------------   
## Employement Divorced Living with a partner Married Never been married Separated Single Widowed Total   
## ----------------------------------------------------------------------------------------------------------------------------------------------------------------   
## Disabled Observed 13 2 18 5 3 2 2 45   
## Expected 3.626 2.247 24.62 10.10 0.661 0.906 2.833   
##   
## Employed full-time Observed 92 65 672 184 13 18 22 1066   
## Expected 85.888 53.233 583.33 239.32 15.657 21.472 67.100   
##   
## Employed part-time Observed 14 15 134 134 3 7 8 315   
## Expected 25.380 15.730 172.37 70.72 4.627 6.345 19.828   
##   
## Have own business/self-employed Observed 3 3 35 4 0 0 2 47   
## Expected 3.787 2.347 25.72 10.55 0.690 0.947 2.958   
##   
## Not employed for pay Observed 14 28 138 176 12 13 15 396   
## Expected 31.906 19.775 216.69 88.90 5.816 7.977 24.927   
##   
## Other Observed 0 1 6 0 0 1 2 10   
## Expected 0.806 0.499 5.47 2.25 0.147 0.201 0.629   
##   
## Retired Observed 56 5 297 15 4 6 99 482   
## Expected 38.835 24.070 263.75 108.21 7.079 9.709 30.340   
##   
## Student Observed 0 0 4 17 0 1 0 22   
## Expected 1.773 1.099 12.04 4.94 0.323 0.443 1.385   
##   
## Total Observed 192 119 1304 535 35 48 150 2383   
## Expected 192.000 119.000 1304.00 535.00 35.000 48.000 150.000   
## ----------------------------------------------------------------------------------------------------------------------------------------------------------------   
##   
##   
## <U+03C7>² Tests   
## -------------------------------   
## Value df p   
## -------------------------------   
## <U+03C7>² 643 42 < .001   
## N 2383   
## -------------------------------   
##   
##   
## Nominal   
## ----------------------------   
## Value   
## ----------------------------   
## Phi-coefficient NaN   
## Cramer's V 0.212   
## ----------------------------

## Conceptual Question: Goodness-of-Fit

H0: Proportions fit expected values for Race Ha: Proportions do not fit expected values for Race

*Order Matters* Asian or Pacific Islander: 3% Black or African American: 12% Mixed Race: 2% Native American/American Indian: 1% Other: 2% White: 80%

# H0 = equal proportions in each category; Ha = unequal proportions in each category  
# Chi-square = Sum[(Observed - Expected)^2/Expected]  
# df = # of categories - 1  
  
# use print(levels(dat.no.NA$Race)) to check for order of list in order to apply ratios accurately  
# in this case it is in alphabetical order  
  
# First checking unweighted expected values  
# H0: equal proportions in each category  
# Ha: unequal proportions in each category  
goodness <- jmv::propTestN(data = dat.no.NA,  
 var = 'Race',  
 expected = TRUE,   
 ratio = c(1, 1, 1, 1, 1, 1))  
  
goodness

##   
## PROPORTION TEST (N OUTCOMES)  
##   
## Proportions   
## ----------------------------------------------------------------------   
## Level Count Proportion   
## ----------------------------------------------------------------------   
## Asian or Pacific Islander Observed 62 0.0260   
## Expected 397 0.167   
##   
## Black or African-American Observed 277 0.1162   
## Expected 397 0.167   
##   
## Mixed race Observed 54 0.0227   
## Expected 397 0.167   
##   
## Native American/American Indian Observed 24 0.0101   
## Expected 397 0.167   
##   
## Other Observed 48 0.0201   
## Expected 397 0.167   
##   
## White Observed 1918 0.8049   
## Expected 397 0.167   
## ----------------------------------------------------------------------   
##   
##   
## <U+03C7>² Goodness of Fit   
## ------------------------   
## <U+03C7>² df p   
## ------------------------   
## 7097 5 < .001   
## ------------------------

# Ha holds  
  
# check with weights added from conceptual question  
goodness.weighted <- jmv::propTestN(data = dat.no.NA,  
 var = 'Race',  
 expected = TRUE,   
 ratio = c(.03, .12, .02, .01, .02, .80))  
  
goodness.weighted

##   
## PROPORTION TEST (N OUTCOMES)  
##   
## Proportions   
## ----------------------------------------------------------------------   
## Level Count Proportion   
## ----------------------------------------------------------------------   
## Asian or Pacific Islander Observed 62 0.0260   
## Expected 71 0.0300   
##   
## Black or African-American Observed 277 0.1162   
## Expected 286 0.1200   
##   
## Mixed race Observed 54 0.0227   
## Expected 48 0.0200   
##   
## Native American/American Indian Observed 24 0.0101   
## Expected 24 0.0100   
##   
## Other Observed 48 0.0201   
## Expected 48 0.0200   
##   
## White Observed 1918 0.8049   
## Expected 1906 0.8000   
## ----------------------------------------------------------------------   
##   
##   
## <U+03C7>² Goodness of Fit   
## -----------------------   
## <U+03C7>² df p   
## -----------------------   
## 2.46 5 0.783   
## -----------------------

# H0 holds