PSYCH308A - Data Analysis 5 (DA5)

Brady C. Jackson

2024/10/15

Table of Contents

# Load packages. Set messages and warnings to FALSE so I don't have to see the  
# masking messages in the output.  
library(psych)  
library(jmv) # for descriptive  
library(ggplot2)  
library(dplyr)  
library(magrittr)  
library(stringr) # for sub\_str  
library(AER)

# Data Prep

First we need to load the data

# Load the data as a dataframe  
reading\_dat <- read.csv("./DA5a.csv")  
  
# Assert column names to be lowercase  
colnames(reading\_dat) <- tolower(colnames(reading\_dat))  
  
# Check the structure of the data:  
print(head(reading\_dat))

## age sex race married married.status  
## 1 41 Female Asian or Pacific Islander No Divorced  
## 2 26 Male Asian or Pacific Islander No Living with a partner  
## 3 21 Female Asian or Pacific Islander No Living with a partner  
## 4 25 Female Asian or Pacific Islander No Living with a partner  
## 5 35 Male Asian or Pacific Islander Yes Married  
## 6 37 Female Asian or Pacific Islander Yes Married  
## education employment  
## 1 Post-graduate training/professional school after college Employed full-time  
## 2 Some college, no 4-year degree Employed full-time  
## 3 Some college, no 4-year degree Employed full-time  
## 4 College graduate Employed full-time  
## 5 Post-graduate training/professional school after college Employed full-time  
## 6 Post-graduate training/professional school after college Employed full-time  
## incomes how.many.books.did.you.read.during.last.12months.  
## 1 $50,000 to under $75,000 2  
## 2 $20,000 to under $30,000 50  
## 3 $40,000 to under $50,000 4  
## 4 $50,000 to under $75,041 20  
## 5 $100,000 to under $150,000 16  
## 6 $100,000 to under $150,000 10  
## read.any.printed.books.during.last.12months.  
## 1 Yes  
## 2 Yes  
## 3 Yes  
## 4 Yes  
## 5 Yes  
## 6 Yes  
## read.any.audiobooks.during.last.12months.  
## 1 No  
## 2 No  
## 3 No  
## 4 No  
## 5 No  
## 6 No  
## read.any.e.books.during.last.12months.  
## 1 No  
## 2 Yes  
## 3 No  
## 4 No  
## 5 Yes  
## 6 Yes  
## last.book.you.read..you  
## 1 Borrowed the book from a library  
## 2 Purchased the book  
## 3 Purchased the book  
## 4 Borrowed the book from a friend or family member  
## 5 Purchased the book  
## 6 Borrowed the book from a library  
## do.you.happen.to.read.any.daily.news.or.newspapers.  
## 1 No  
## 2 Yes  
## 3 No  
## 4 Yes  
## 5 No  
## 6 Yes  
## do.you.happen.to.read.any.magazines.or.journals.  
## 1 No  
## 2 No  
## 3 Yes  
## 4 Yes  
## 5 No  
## 6 No

cat("\n\n")

str(reading\_dat)

## 'data.frame': 2442 obs. of 15 variables:  
## $ age : int 41 26 21 25 35 37 40 30 55 39 ...  
## $ sex : chr "Female" "Male" "Female" "Female" ...  
## $ race : chr "Asian or Pacific Islander" "Asian or Pacific Islander" "Asian or Pacific Islander" "Asian or Pacific Islander" ...  
## $ married : chr "No" "No" "No" "No" ...  
## $ married.status : chr "Divorced" "Living with a partner" "Living with a partner" "Living with a partner" ...  
## $ education : chr "Post-graduate training/professional school after college" "Some college, no 4-year degree" "Some college, no 4-year degree" "College graduate" ...  
## $ employment : chr "Employed full-time" "Employed full-time" "Employed full-time" "Employed full-time" ...  
## $ incomes : chr "$50,000 to under $75,000" "$20,000 to under $30,000" "$40,000 to under $50,000" "$50,000 to under $75,041" ...  
## $ how.many.books.did.you.read.during.last.12months. : int 2 50 4 20 16 10 3 3 3 2 ...  
## $ read.any.printed.books.during.last.12months. : chr "Yes" "Yes" "Yes" "Yes" ...  
## $ read.any.audiobooks.during.last.12months. : chr "No" "No" "No" "No" ...  
## $ read.any.e.books.during.last.12months. : chr "No" "Yes" "No" "No" ...  
## $ last.book.you.read..you : chr "Borrowed the book from a library" "Purchased the book" "Purchased the book" "Borrowed the book from a friend or family member" ...  
## $ do.you.happen.to.read.any.daily.news.or.newspapers.: chr "No" "Yes" "No" "Yes" ...  
## $ do.you.happen.to.read.any.magazines.or.journals. : chr "No" "No" "Yes" "Yes" ...

# Check uniqueness of fields of interrest for upcoming questions   
# (sex, education, and employment). We want to make sure that there are not   
# equal value inputs that would look different due to things like case or type:  
# (e.g. High school graduate and HIGH School Graduatre and High School Complete)  
uni\_edu = unique(reading\_dat$education)   
uni\_sex = unique(reading\_dat$sex)  
uni\_emp = unique(reading\_dat$employment)  
  
# Print the values for inspection  
cat("\nUnique values in Education:\n")

##   
## Unique values in Education:

cat(uni\_edu, sep="\n")

## Post-graduate training/professional school after college  
## Some college, no 4-year degree  
## College graduate  
## High school graduate  
## High school incomplete  
## Technical, trade or vocational school AFTER high school  
## None

cat("---\n\n")

## ---

cat("\nUnique values in Sex:\n")

##   
## Unique values in Sex:

cat(uni\_sex, sep="\n")

## Female  
## Male

cat("---\n\n")

## ---

cat("\nUnique values in Employment:\n")

##   
## Unique values in Employment:

cat(uni\_emp, sep="\n")

## Employed full-time  
## Employed part-time  
## Have own business/self-employed  
## Not employed for pay  
## Retired  
## Student  
## Disabled

cat("---\n\n")

## ---

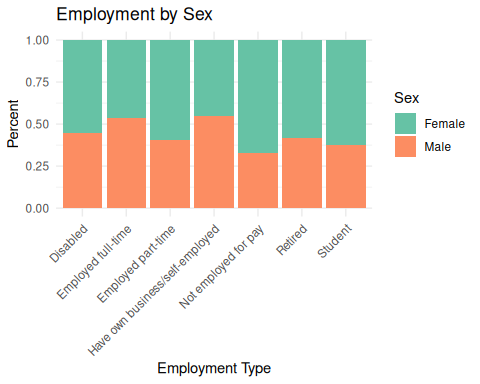
# Question 06

Data for Employment and Sex look appropriately unique. So we’ll create a  
visualization (bar chart) for employment sorted by sex. Then we will print  
contingency tables and run the test

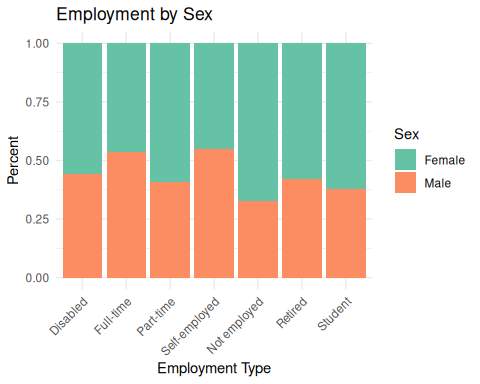
## Q06 Visualization

Code below will create two bar-charts, one with the breakdown named by the  
original employment categories in the data, and one with the same data but  
labeled by more readable display names. Both are printed so that they can  
be inspected to ensure names were mapped correctly.

# Create a mapping vector setting each long-name to a shortened display name  
emp\_disp\_name\_map = c( "Employed full-time" = "Full-time",  
 "Employed part-time" = "Part-time",  
 "Have own business/self-employed" = "Self-employed",  
 "Not employed for pay" = "Not employed",  
 "Retired" = "Retired",  
 "Student" = "Student",  
 "Disabled" = "Disabled"  
 )  
  
# Bar chart visualization. This code constructs a bar chart that shows each   
# employment type broken down by percentage of females vs. percentage of  
# males at that level. Note: Two versions of the same chart are produced, one   
# uses default education level names and one with shortened display names. This   
# is just to ensure I mapped the display names correctly.  
ggplot(reading\_dat, aes( x=employment, fill=as.factor(sex) ) ) +   
 geom\_bar(position = "fill") +   
 scale\_fill\_brewer(palette = "Set2") +   
 labs(y = "Percent", x = "Employment Type", fill= "Sex", title = "Employment by Sex") +  
 theme\_minimal() +   
 theme( axis.text.x = element\_text(angle = 45, hjust = 1) )



ggplot(reading\_dat, aes( x=employment, fill=as.factor(sex) ) ) +   
 geom\_bar(position = "fill") +   
 scale\_fill\_brewer(palette = "Set2") +   
 scale\_x\_discrete(labels = emp\_disp\_name\_map) +  
 labs(y = "Percent", x = "Employment Type", fill= "Sex", title = "Employment by Sex") +  
 theme\_minimal() +   
 theme( axis.text.x = element\_text(angle = 45, hjust = 1) )



## Q06 Contingency Table

# We create a table for consumption  
emp\_v\_sex\_tab <- prop.table( xtabs(~ sex + employment, data = reading\_dat), 1 )  
round(emp\_v\_sex\_tab, 2)

## employment  
## sex Disabled Employed full-time Employed part-time  
## Female 0.02 0.37 0.14  
## Male 0.02 0.50 0.12  
## employment  
## sex Have own business/self-employed Not employed for pay Retired Student  
## Female 0.02 0.20 0.22 0.02  
## Male 0.03 0.12 0.19 0.02

## Q06 NHST

jmv::contTables(data = reading\_dat,   
 rows="sex", cols="employment",   
 exp=TRUE,   
 phiCra=TRUE  
 )

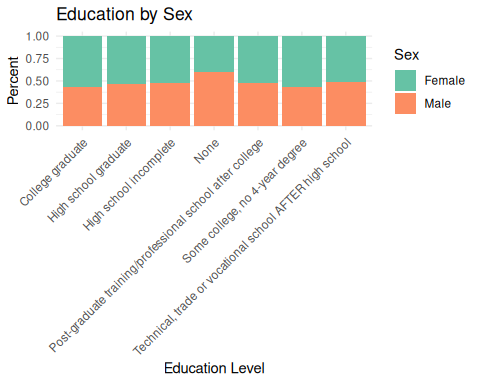
##   
## CONTINGENCY TABLES  
##   
## Contingency Tables   
## ─────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────   
## sex Disabled Employed full-time Employed part-time Have own business/self-employed Not employed for pay Retired Student Total   
## ─────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────   
## Female Observed 30 486 191 28 270 291 33 1329   
## Expected 29.38821 570.3489 175.2408 33.74201 218.7789 272.6572 28.84398 1329.000   
##   
## Male Observed 24 562 131 34 132 210 20 1113   
## Expected 24.61179 477.6511 146.7592 28.25799 183.2211 228.3428 24.15602 1113.000   
##   
## Total Observed 54 1048 322 62 402 501 53 2442   
## Expected 54.00000 1048.0000 322.0000 62.00000 402.0000 501.0000 53.00000 2442.000   
## ─────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────   
##   
##   
## χ² Tests   
## ──────────────────────────────────────   
## Value df p   
## ──────────────────────────────────────   
## χ² 62.98363 6 < .0000001   
## N 2442   
## ──────────────────────────────────────   
##   
##   
## Nominal   
## ────────────────────────────────   
## Value   
## ────────────────────────────────   
## Phi-coefficient NaN   
## Cramer's V 0.1605983   
## ────────────────────────────────

# Question 07

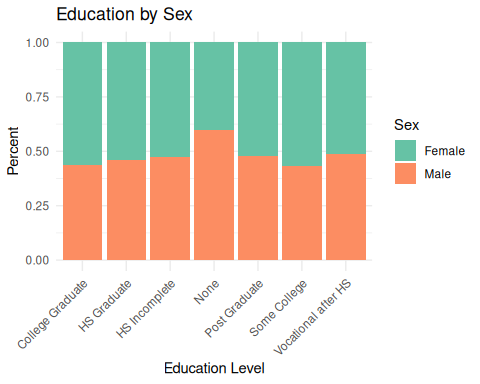
Data for Education and Sex look appropriately unique. So we’ll create a  
visualization (bar chart) for education sorted by sex. Then we will print  
contingency tables and run the test

## Q07 Visualization

# Create a mapping vector setting each long-name to a shortened display name  
edu\_disp\_name\_map = c( "Post-graduate training/professional school after college" = "Post Graduate",  
 "Some college, no 4-year degree" = "Some College",  
 "College graduate" = "College Graduate",  
 "High school graduate" = "HS Graduate",  
 "High school incomplete" = "HS Incomplete",  
 "Technical, trade or vocational school AFTER high school" = "Vocational after HS",  
 "None" = "None"  
 )  
  
# Bar chart visualization. This code constructs a bar chart that shows each   
# level of education broken down by percentage of females vs. percentage of  
# males at that level. Note: Two versions of the same chart are produced, one   
# default education level names and one with shortened display names. This is   
# just to ensure I mapped the display names correctly.  
ggplot(reading\_dat, aes( x=education, fill=as.factor(sex) ) ) +   
 geom\_bar(position = "fill") +   
 scale\_fill\_brewer(palette = "Set2") +   
 labs(y = "Percent", x = "Education Level", fill= "Sex", title = "Education by Sex") +  
 theme\_minimal() +   
 theme( axis.text.x = element\_text(angle = 45, hjust = 1) )



ggplot(reading\_dat, aes( x=education, fill=as.factor(sex) ) ) +   
 geom\_bar(position = "fill") +   
 scale\_fill\_brewer(palette = "Set2") +   
 scale\_x\_discrete(labels = edu\_disp\_name\_map) +  
 labs(y = "Percent", x = "Education Level", fill= "Sex", title = "Education by Sex") +  
 theme\_minimal() +   
 theme( axis.text.x = element\_text(angle = 45, hjust = 1) )



## Q07 Contingency Table

# We create a table for consumption  
edu\_v\_sex\_tab <- prop.table( xtabs(~ sex + education, data = reading\_dat), 1 )  
round(edu\_v\_sex\_tab, 2)

## education  
## sex College graduate High school graduate High school incomplete None  
## Female 0.24 0.21 0.08 0.01  
## Male 0.22 0.22 0.09 0.02  
## education  
## sex Post-graduate training/professional school after college  
## Female 0.19  
## Male 0.21  
## education  
## sex Some college, no 4-year degree  
## Female 0.25  
## Male 0.23  
## education  
## sex Technical, trade or vocational school AFTER high school  
## Female 0.02  
## Male 0.02

## Q07 NHST

jmv::contTables(data = reading\_dat,   
 rows="sex", cols="education",   
 exp=TRUE,   
 phiCra=TRUE  
 )

##   
## CONTINGENCY TABLES  
##   
## Contingency Tables   
## ─────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────   
## sex College graduate High school graduate High school incomplete None Post-graduate training/professional school after college Some college, no 4-year degree Technical, trade or vocational school AFTER high school Total   
## ─────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────   
## Female Observed 317 281 108 16 254 328 25 1329   
## Expected 305.8550 284.0860 111.56634 21.76904 263.9496 315.1069 26.66708 1329.000   
##   
## Male Observed 245 241 97 24 231 251 24 1113   
## Expected 256.1450 237.9140 93.43366 18.23096 221.0504 263.8931 22.33292 1113.000   
##   
## Total Observed 562 522 205 40 485 579 49 2442   
## Expected 562.0000 522.0000 205.00000 40.00000 485.0000 579.0000 49.00000 2442.000   
## ─────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────   
##   
##   
## χ² Tests   
## ─────────────────────────────────────   
## Value df p   
## ─────────────────────────────────────   
## χ² 6.778156 6 0.3418515   
## N 2442   
## ─────────────────────────────────────   
##   
##   
## Nominal   
## ─────────────────────────────────   
## Value   
## ─────────────────────────────────   
## Phi-coefficient NaN   
## Cramer's V 0.05268451   
## ─────────────────────────────────

# Question 08

R was not used in any capacity for this question.