Project Assignment 3

Elikem Asudo Tsatsu Gale-Zoyiku

2023-04-01

## **Research Question**

**Is there an association between the sexual orientation of cis-males, depression levels, and their income? Can education explain any association if there is?**

Explanatory Variables: Gender (Categorical), Sexual Orientation (Categorical), and Educational attainment (categorical)

Response Variables: Depression levels (Categorical), Anxiety levels (Categorical), Ratio of income to poverty level (Categorical)

### 1. Load data set(s) and libraries

load("C:/Users/egale/OneDrive - Ashesi University/Desktop/Statistics with Probability/IPUMS\_NHIS.RData")  
library(descr)

## Warning: package 'descr' was built under R version 4.2.3

library(stats)  
library(dplyr)

## Warning: package 'dplyr' was built under R version 4.2.3

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

### 2. Create variable subset

vars=c("AGE","SEX","SEXORIEN","EDUC","DEPFEELEVL","DEPFREQ","POVERTY","WORFEELEVL","WORFREQ")  
myHealthData = IPUMS\_NHIS[vars]  
str(myHealthData)

## 'data.frame': 116291 obs. of 9 variables:  
## $ AGE : int 47 61 77 37 8 34 68 46 36 5 ...  
## $ SEX : int 2 1 2 2 1 2 2 2 2 2 ...  
## $ SEXORIEN : int 2 2 2 2 0 2 2 2 2 0 ...  
## $ EDUC : int 400 201 301 303 0 400 201 400 301 0 ...  
## $ DEPFEELEVL: int 3 0 3 0 0 1 1 1 0 0 ...  
## $ DEPFREQ : int 4 5 4 5 5 3 4 4 5 5 ...  
## $ POVERTY : int 37 32 33 34 34 37 11 37 33 33 ...  
## $ WORFEELEVL: int 3 0 1 2 0 2 1 2 0 0 ...  
## $ WORFREQ : int 2 5 4 2 5 4 1 4 5 5 ...

myHealthData<-as.data.frame(myHealthData)  
str(myHealthData)

## 'data.frame': 116291 obs. of 9 variables:  
## $ AGE : int 47 61 77 37 8 34 68 46 36 5 ...  
## $ SEX : int 2 1 2 2 1 2 2 2 2 2 ...  
## $ SEXORIEN : int 2 2 2 2 0 2 2 2 2 0 ...  
## $ EDUC : int 400 201 301 303 0 400 201 400 301 0 ...  
## $ DEPFEELEVL: int 3 0 3 0 0 1 1 1 0 0 ...  
## $ DEPFREQ : int 4 5 4 5 5 3 4 4 5 5 ...  
## $ POVERTY : int 37 32 33 34 34 37 11 37 33 33 ...  
## $ WORFEELEVL: int 3 0 1 2 0 2 1 2 0 0 ...  
## $ WORFREQ : int 2 5 4 2 5 4 1 4 5 5 ...

### 3. Data management I: check for and recode errors and NAs

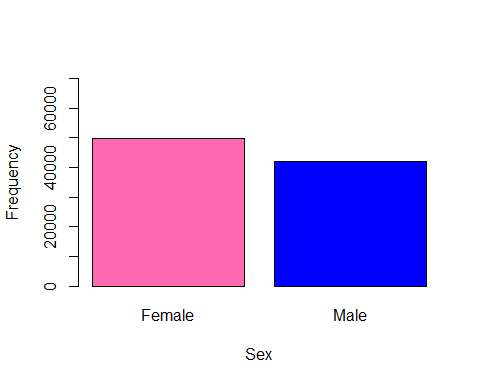
myHealthData$Gender <- rep(NA,nrow(myHealthData))  
myHealthData$Gender[myHealthData$SEX==1] <- 'Male'  
myHealthData$Gender[myHealthData$SEX==2] <- 'Female'  
  
  
myHealthData$WorkingAge <- rep(NA,nrow(myHealthData))  
myHealthData$WorkingAge[myHealthData$AGE<=18] <- 'Not working age'  
myHealthData$WorkingAge[myHealthData$AGE>18 & myHealthData$AGE<997] <- 'Working age'  
  
  
myHealthData$Orientation <- rep(NA,nrow(myHealthData))  
myHealthData$Orientation[myHealthData$SEXORIEN==1] <- 'Not Straight'  
myHealthData$Orientation[myHealthData$SEXORIEN==3] <- 'Not Straight'  
myHealthData$Orientation[myHealthData$SEXORIEN==4] <- 'Not Straight'  
myHealthData$Orientation[myHealthData$SEXORIEN==2] <- 'Straight'  
  
myHealthData$EDUC<-as.numeric(myHealthData$EDUC)  
myHealthData$Education <- rep(NA, nrow(myHealthData))  
myHealthData$Education[myHealthData$EDUC < 112] <- "No high school"  
myHealthData$Education[myHealthData$EDUC >= 112 & myHealthData$EDUC<116] <- "Some high school"  
myHealthData$Education[myHealthData$EDUC >= 200 & myHealthData$EDUC<=202] <- "High school diploma or equivalent"  
myHealthData$Education[myHealthData$EDUC >= 300 & myHealthData$EDUC<=301] <- "Some college"  
myHealthData$Education[myHealthData$EDUC >= 302 & myHealthData$EDUC<=303] <- "Associate's Degree"  
myHealthData$Education[myHealthData$EDUC == 400] <- "Bachelor's Degree"  
myHealthData$Education[myHealthData$EDUC == 504] <- "Other Degree"  
myHealthData$Education[myHealthData$EDUC >= 500 & myHealthData$EDUC<=503] <- "Postgraduate"  
  
myHealthData$DepressionFrequency <- rep(NA, nrow(myHealthData))  
myHealthData$DepressionFrequency[myHealthData$DEPFREQ == 1] <- "Daily"  
myHealthData$DepressionFrequency[myHealthData$DEPFREQ == 2] <- "Weekly"  
myHealthData$DepressionFrequency[myHealthData$DEPFREQ == 3] <- "Monthly"  
myHealthData$DepressionFrequency[myHealthData$DEPFREQ == 4] <- "Rarely"  
myHealthData$DepressionFrequency[myHealthData$DEPFREQ == 5] <- "Never"  
  
  
myHealthData$DepressionLevel <- rep(NA, nrow(myHealthData))  
myHealthData$DepressionLevel[myHealthData$DEPFEELEVL == 1] <- "A lot"  
myHealthData$DepressionLevel[myHealthData$DEPFEELEVL== 2] <- "A little"  
myHealthData$DepressionLevel[myHealthData$DEPFEELEVL == 3] <- "Somewhere in between"  
myHealthData$DepressionLevel[myHealthData$DEPFEELEVL == 0] <- NA  
  
  
myHealthData$WorryFrequency <- rep(NA, nrow(myHealthData))  
myHealthData$WorryFrequency[myHealthData$WORFREQ == 1] <- "Daily"  
myHealthData$WorryFrequency[myHealthData$WORFREQ == 2] <- "Weekly"  
myHealthData$WorryFrequency[myHealthData$WORFREQ == 3] <- "Monthly"  
myHealthData$WorryFrequency[myHealthData$WORFREQ == 4] <- "Rarely"  
myHealthData$WorryFrequency[myHealthData$WORFREQ == 5] <- "Never"  
  
  
myHealthData$WorryLevel <- rep(NA, nrow(myHealthData))  
myHealthData$WorryLevel[myHealthData$WORFEELEVL == 1] <- "A lot"  
myHealthData$WorryLevel[myHealthData$WORFEELEVL== 2] <- "A little"  
myHealthData$WorryLevel[myHealthData$WORFEELEVL == 3] <- "Somewhere in between"  
myHealthData$WorryLevel[myHealthData$WORFEELEVL == 0] <- NA  
  
  
myHealthData$PovertyRatio <- rep(NA, nrow(myHealthData))  
myHealthData$PovertyRatio[myHealthData$POVERTY<=14] <- "Below the poverty line"  
myHealthData$PovertyRatio[myHealthData$POVERTY<=25 & myHealthData$POVERTY>=20] <- "Between 1 to 1.99 units above the poverty line"  
myHealthData$PovertyRatio[myHealthData$POVERTY<=32 & myHealthData$POVERTY>=31] <- "Between 2 to 2.99 units above the poverty line"  
myHealthData$PovertyRatio[myHealthData$POVERTY<=34 & myHealthData$POVERTY>=33] <- "Between 3 to 3.99 units above the poverty line"  
myHealthData$PovertyRatio[myHealthData$POVERTY<=36 & myHealthData$POVERTY>=35] <- "Between 4 to 4.99 units above the poverty line"  
myHealthData$PovertyRatio[myHealthData$POVERTY==37] <- "5 units and above the poverty line"

### 4. Data management II: further subset and create secondary variable

# Subset the data frame to take out observations below the working age  
myHealthData <- myHealthData[myHealthData$WorkingAge == "Working age", ]

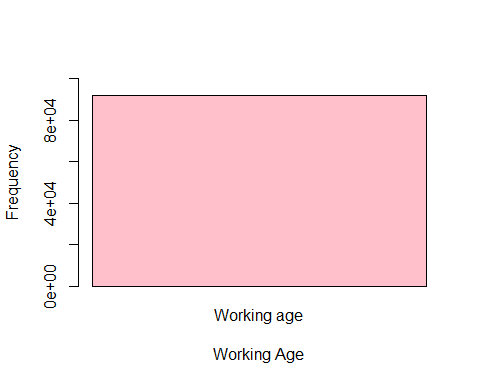
### 5. Descriptive statistics (sample means, standard deviations, proportions) and univariate displays

freq(myHealthData$Gender, ylab="Frequency",xlab="Sex",ylim=c(0,70000),col=c("hotpink","blue"))



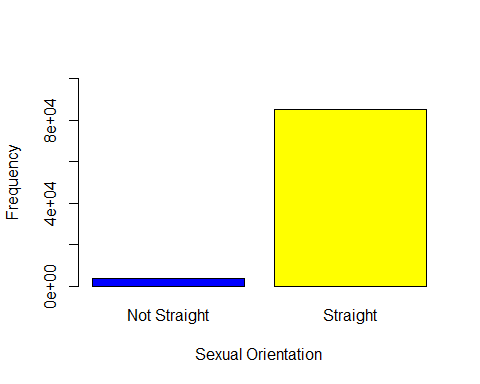
## myHealthData$Gender   
## Frequency Percent Valid Percent  
## Female 49932 54.0553 54.2  
## Male 42197 45.6816 45.8  
## NA's 243 0.2631   
## Total 92372 100.0000 100.0

freq(myHealthData$WorkingAge, ylab="Frequency",xlab="Working Age",ylim=c(0,100000),col=c("pink","blue"))



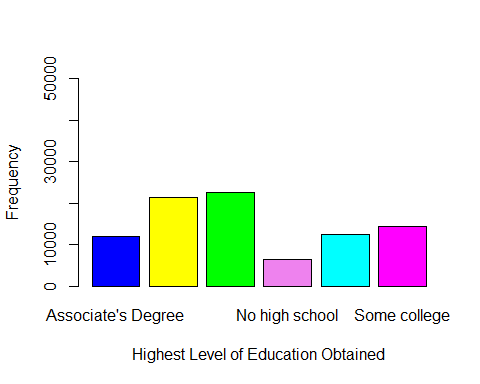
## myHealthData$WorkingAge   
## Frequency Percent Valid Percent  
## Working age 92135 99.7434 100  
## NA's 237 0.2566   
## Total 92372 100.0000 100

freq(myHealthData$Orientation,ylab="Frequency",xlab="Sexual Orientation",ylim=c(0,100000), col = c("blue","yellow"))



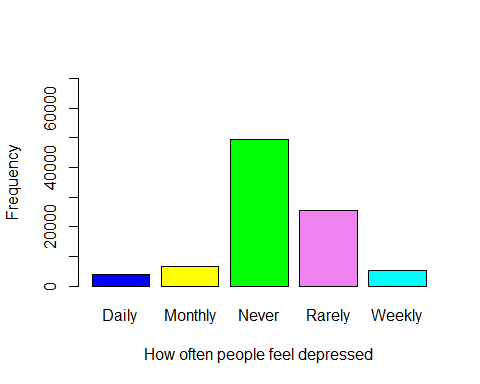
## myHealthData$Orientation   
## Frequency Percent Valid Percent  
## Not Straight 3418 3.700 3.86  
## Straight 85132 92.162 96.14  
## NA's 3822 4.138   
## Total 92372 100.000 100.00

freq(myHealthData$Education,ylab="Frequency",xlab="Highest Level of Education Obtained",ylim=c(0,50000),col=c("blue","yellow","green","violet","cyan","magenta"))



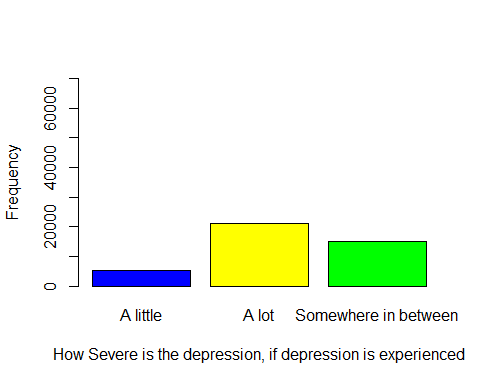
## myHealthData$Education   
## Frequency Percent Valid Percent  
## Associate's Degree 11966 12.954 13.403  
## Bachelor's Degree 21274 23.031 23.829  
## High school diploma or equivalent 22620 24.488 25.336  
## No high school 6430 6.961 7.202  
## Postgraduate 12525 13.559 14.029  
## Some college 14464 15.658 16.201  
## NA's 3093 3.348   
## Total 92372 100.000 100.000

freq(myHealthData$DepressionFrequency,ylab="Frequency",xlab="How often people feel depressed",ylim=c(0,70000),col=c("blue","yellow","green","violet","cyan","magenta"))



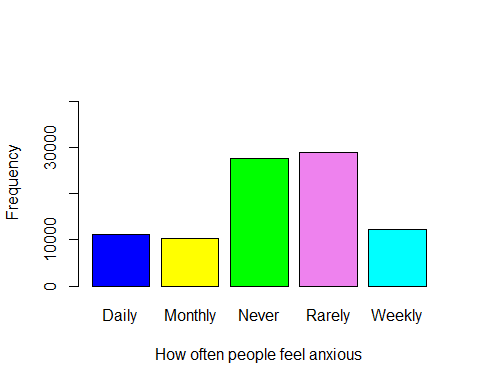
## myHealthData$DepressionFrequency   
## Frequency Percent Valid Percent  
## Daily 3750 4.060 4.151  
## Monthly 6494 7.030 7.188  
## Never 49367 53.444 54.642  
## Rarely 25423 27.522 28.140  
## Weekly 5312 5.751 5.880  
## NA's 2026 2.193   
## Total 92372 100.000 100.000

freq(myHealthData$DepressionLevel,ylab="Frequency",xlab="How Severe is the depression, if depression is experienced",ylim=c(0,70000),col=c("blue","yellow","green","violet","cyan","magenta"))



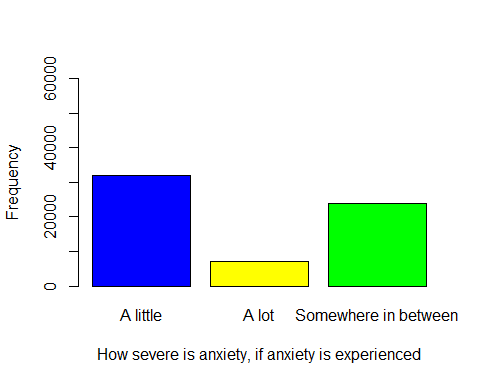
## myHealthData$DepressionLevel   
## Frequency Percent Valid Percent  
## A little 5338 5.779 12.80  
## A lot 21247 23.002 50.94  
## Somewhere in between 15126 16.375 36.26  
## NA's 50661 54.845   
## Total 92372 100.000 100.00

freq(myHealthData$WorryFrequency,ylab="Frequency",xlab="How often people feel anxious",ylim=c(0,45000),col=c("blue","yellow","green","violet","cyan","magenta"))



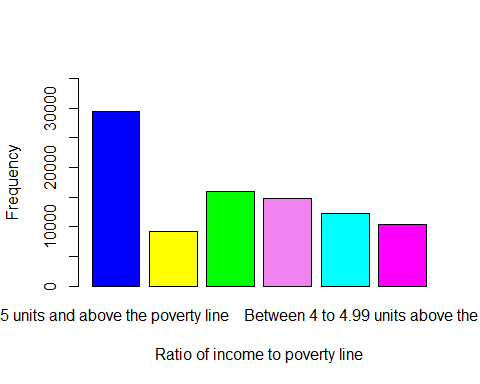
## myHealthData$WorryFrequency   
## Frequency Percent Valid Percent  
## Daily 11175 12.098 12.36  
## Monthly 10235 11.080 11.32  
## Never 27767 30.060 30.71  
## Rarely 29025 31.422 32.10  
## Weekly 12226 13.236 13.52  
## NA's 1944 2.105   
## Total 92372 100.000 100.00

freq(myHealthData$WorryLevel,ylab="Frequency",xlab="How severe is anxiety, if anxiety is experienced",ylim=c(0,60000),col=c("blue","yellow","green","violet","cyan","magenta"))



## myHealthData$WorryLevel   
## Frequency Percent Valid Percent  
## A little 31953 34.59 50.66  
## A lot 7187 7.78 11.39  
## Somewhere in between 23934 25.91 37.95  
## NA's 29298 31.72   
## Total 92372 100.00 100.00

freq(myHealthData$PovertyRatio,ylab="Frequency",xlab="Ratio of income to poverty line",ylim=c(0,35000),col=c("blue","yellow","green","violet","cyan","magenta"))



## myHealthData$PovertyRatio   
## Frequency Percent Valid Percent  
## 5 units and above the poverty line 29500 31.9361 32.018  
## Below the poverty line 9137 9.8915 9.917  
## Between 1 to 1.99 units above the poverty line 15952 17.2693 17.314  
## Between 2 to 2.99 units above the poverty line 14831 16.0557 16.097  
## Between 3 to 3.99 units above the poverty line 12312 13.3287 13.363  
## Between 4 to 4.99 units above the poverty line 10403 11.2621 11.291  
## NA's 237 0.2566   
## Total 92372 100.0000 100.000

t<-table(myHealthData$Gender)  
p<-prop.table(t)\*100;p

##   
## Female Male   
## 54.19792 45.80208

t<-table(myHealthData$WorkingAge)  
p<-prop.table(t)\*100;p

##   
## Working age   
## 100

t<-table(myHealthData$Orientation)  
p<-prop.table(t)\*100;p

##   
## Not Straight Straight   
## 3.859966 96.140034

t<-table(myHealthData$Education)  
p<-prop.table(t)\*100;p

##   
## Associate's Degree Bachelor's Degree   
## 13.402928 23.828672   
## High school diploma or equivalent No high school   
## 25.336305 7.202142   
## Postgraduate Some college   
## 14.029055 16.200898

t<-table(myHealthData$DepressionFrequency)  
p<-prop.table(t)\*100;p

##   
## Daily Monthly Never Rarely Weekly   
## 4.150709 7.187922 54.642153 28.139597 5.879618

t<-table(myHealthData$DepressionLevel)  
p<-prop.table(t)\*100;p

##   
## A little A lot Somewhere in between   
## 12.79758 50.93860 36.26382

t<-table(myHealthData$WorryFrequency)  
p<-prop.table(t)\*100;p

##   
## Daily Monthly Never Rarely Weekly   
## 12.35790 11.31840 30.70620 32.09736 13.52015

t<-table(myHealthData$WorryLevel)  
p<-prop.table(t)\*100;p

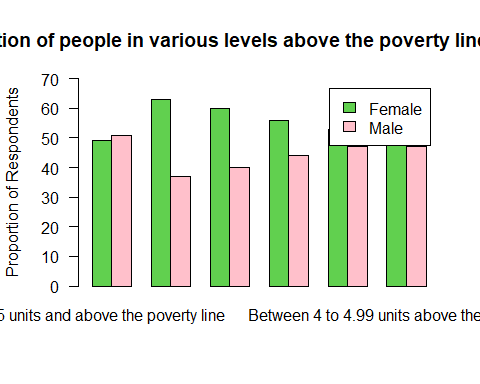
##   
## A little A lot Somewhere in between   
## 50.65954 11.39455 37.94590

t<-table(myHealthData$PovertyRatio)  
p<-prop.table(t)\*100;p

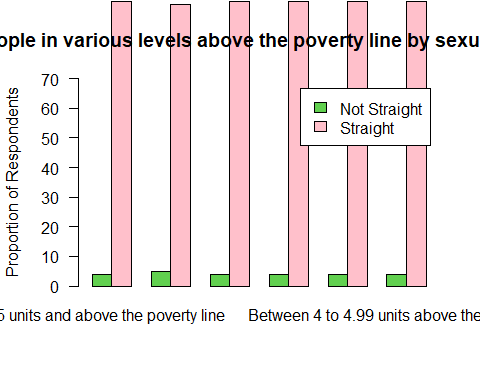
##   
## 5 units and above the poverty line   
## 32.01823   
## Below the poverty line   
## 9.91697   
## Between 1 to 1.99 units above the poverty line   
## 17.31372   
## Between 2 to 2.99 units above the poverty line   
## 16.09703   
## Between 3 to 3.99 units above the poverty line   
## 13.36300   
## Between 4 to 4.99 units above the poverty line   
## 11.29104

### 6. Bivariate tables and graphs

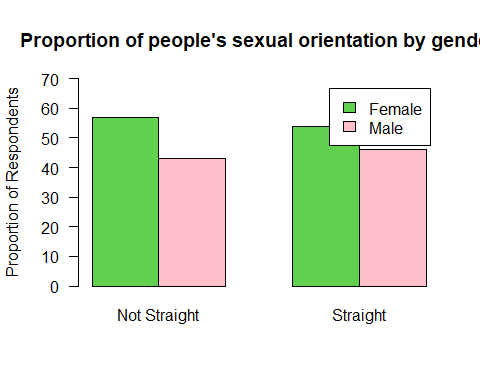
tab1<-table(myHealthData$Gender,myHealthData$PovertyRatio)  
prob\_tab1<-round(prop.table(tab1,2),2)\*100  
barplot(prob\_tab1,main="Proportion of people in various levels above the poverty line by gender",ylab="Proportion of Respondents", las=1, col=c(123,"pink"), beside = TRUE,legend=row.names(prob\_tab1),ylim=c(0,70))



tab2<-table(myHealthData$Orientation,myHealthData$PovertyRatio)  
prob\_tab2<-round(prop.table(tab2,2),2)\*100  
barplot(prob\_tab2,main="Proportion of people in various levels above the poverty line by sexual orientation gender",ylab="Proportion of Respondents", las=1, col=c(123,"pink"), beside = TRUE,legend=row.names(prob\_tab2),ylim=c(0,70))



tab3<-table(myHealthData$Gender,myHealthData$Orientation)  
prob\_tab3<-round(prop.table(tab3,2),2)\*100  
barplot(prob\_tab3,main="Proportion of people's sexual orientation by gender",ylab="Proportion of Respondents", las=1, col=c(123,"pink"), beside = TRUE,legend=row.names(prob\_tab3),ylim=c(0,70))



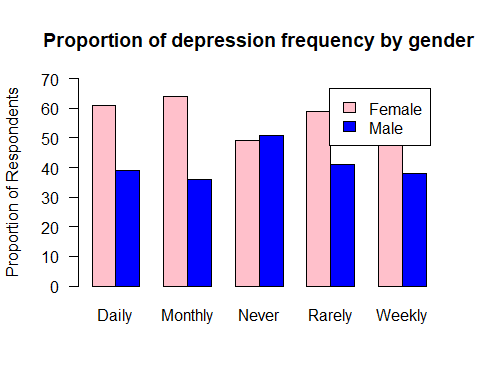
tab4 <- table(myHealthData$Gender,myHealthData$DepressionFrequency); tab4

##   
## Daily Monthly Never Rarely Weekly  
## Female 2287 4179 24215 14967 3274  
## Male 1462 2313 25150 10456 2037

prop\_tab4 <- round(prop.table(tab4,2),2)\*100; prop\_tab4

##   
## Daily Monthly Never Rarely Weekly  
## Female 61 64 49 59 62  
## Male 39 36 51 41 38

barplot(prop\_tab4,main="Proportion of depression frequency by gender",ylab="Proportion of Respondents", las=1, col=c("pink","blue"), beside = TRUE,legend=row.names(prop\_tab4),ylim=c(0,70))



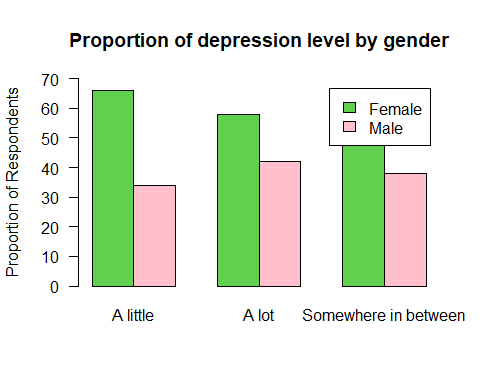
tab4 <- table(myHealthData$Gender,myHealthData$DepressionLevel); tab4

##   
## A little A lot Somewhere in between  
## Female 3498 12381 9359  
## Male 1839 8865 5765

prop\_tab4 <- round(prop.table(tab4,2),2)\*100; prop\_tab4

##   
## A little A lot Somewhere in between  
## Female 66 58 62  
## Male 34 42 38

barplot(prop\_tab4,main="Proportion of depression level by gender",ylab="Proportion of Respondents", las=1, col=c(123,"pink"), beside = TRUE,legend=row.names(prop\_tab4),ylim=c(0,70))



I have learnt that the percentage of the population above working age that identifies as gay, bisexual. or lesbian is under 4%. This figure was unexpected. I also observed that more women than men identify as not straight. I also found that the proportion of people who identify as gay, lesbian or bisexual is fairly constant across all poverty levels. These observations have interesting repercussions on my research questions.

### 7. Bivariate analysis (hypothesis tests and post-hoc tests)

chisq\_results <- chisq.test(myHealthData$DepressionFrequency, myHealthData$Gender)  
chisq\_results

##   
## Pearson's Chi-squared test  
##   
## data: myHealthData$DepressionFrequency and myHealthData$Gender  
## X-squared = 1209.2, df = 4, p-value < 2.2e-16

chisq\_results$observed

## myHealthData$Gender  
## Female Male  
## Daily 2287 1462  
## Monthly 4179 2313  
## Never 24215 25150  
## Rarely 14967 10456  
## Weekly 3274 2037

chisq\_results$expected

## myHealthData$Gender  
## Female Male  
## Daily 2030.203 1718.797  
## Monthly 3515.626 2976.374  
## Never 26732.727 22632.273  
## Rarely 13767.368 11655.632  
## Weekly 2876.076 2434.924

chisq\_results$residuals

## myHealthData$Gender  
## Female Male  
## Daily 5.699273 -6.194082  
## Monthly 11.188126 -12.159477  
## Never -15.398803 16.735724  
## Rarely 10.224043 -11.111692  
## Weekly 7.419925 -8.064121

chisq\_results <- chisq.test(myHealthData$DepressionLevel, myHealthData$Gender)  
chisq\_results

##   
## Pearson's Chi-squared test  
##   
## data: myHealthData$DepressionLevel and myHealthData$Gender  
## X-squared = 112.91, df = 2, p-value < 2.2e-16

chisq\_results$observed

## myHealthData$Gender  
## myHealthData$DepressionLevel Female Male  
## A little 3498 1839  
## A lot 12381 8865  
## Somewhere in between 9359 5765

chisq\_results$expected

## myHealthData$Gender  
## myHealthData$DepressionLevel Female Male  
## A little 3229.559 2107.441  
## A lot 12856.512 8389.488  
## Somewhere in between 9151.929 5972.071

chisq\_results$residuals

## myHealthData$Gender  
## myHealthData$DepressionLevel Female Male  
## A little 4.723650 -5.847518  
## A lot -4.193725 5.191511  
## Somewhere in between 2.164524 -2.679516

To test the relationship between gender and depression levels which are both categorical variables, I conducted a Chi square test of independence. To conduct a Chi square test of independence, I followed these steps:

1. I stated the null and alternative hypotheses:

* Ho: There is no association between gender and depression levels.  
  H1: There is an association between gender and depression levels.

Summary of my samples:

Depression Frequency by Gender

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Daily (%) | Monthly (%) | Never (%) | Rarely (%) | Weekly (%) |
| Female | 61 | 64 | 49 | 59 | 62 |
| Male | 39 | 37 | 51 | 41 | 38 |

Depression Frequency by Gender (Actual Counts)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Daily** | **Monthly** | **Never** | **Rarely** | **Weekly** | **Total** |
| **Female** | 2287 | 4179 | 24215 | 14967 | 3274 | 48922 |
| **Male** | 1462 | 2313 | 25150 | 10456 | 2037 | 41418 |
| **Total** | 3749 | 6492 | 49365 | 25423 | 5311 | **90340** |

Depression Level of respondents by Gender

|  |  |  |  |
| --- | --- | --- | --- |
|  | A little (%) | A lot (%) | Somewhere in between (%) |
| Female | 66 | 58 | 62 |
| Male | 34 | 42 | 38 |

Depression Level of respondents by Gender (Actual Counts)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **A little** | **A lot** | **Somewhere in between** | **Total** |
| **Female** | 3498 | 12381 | 9359 | 25238 |
| **Male** | 1839 | 8865 | 5765 | 16469 |
| **Total** | 5337 | 21246 | 15124 | **41707** |

//the counts in this table are less because I took out the respondents who said they had never felt depressed

The conditions for a Chi-Squared test are met because:

* Both gender and depression frequency / depression level are categorical variables.
* The cells in the contingency table are mutually exclusive, i.e. nobody reports that they experience two or more levels of depression, and nobody reports that they are both male and female.
* The expected value of all the cells are greater than 5, and no cell has an expected value less than 1.

Test Statistic (Chi-Square):

* Pearson's Chi-squared test  
    
  data: myHealthData$DepressionFrequency and myHealthData$Gender  
  X-squared = 1209.2, df = 4, p-value < 2.2e-16
* Pearson's Chi-squared test  
    
  data: myHealthData$DepressionLevel and myHealthData$Gender  
  X-squared = 112.91, df = 2, p-value < 2.2e-16
* The p-value of the test for the association between gender and depression frequency is significantly less than 0.05, so I have sufficient evidence to reject my null hypothesis. This implies that there is an association between a person’s gender and how frequently they experience depression.
* The p-value of the test for the association between gender and depression level is significantly less than 0.05, so I have sufficient evidence to reject my null hypothesis. This implies that there is an association between a person’s gender and how severe the person’s depression is.

1. I calculated the expected frequencies for each cell of the contingency table using the formula: expected frequency = (row total \* column total) / grand total.
2. I calculated the Chi square statistic using the formula: Chi square = sum of (observed frequency - expected frequency)^2 / expected frequency for each cell.
3. I determined the degrees of freedom using the formula: degrees of freedom = (number of rows - 1) \* (number of columns - 1).
4. I chose a significance level (alpha) and compared the Chi square statistic to the critical value from the Chi square distribution table with the corresponding degrees of freedom and alpha level.
5. I made a decision to reject or fail to reject the null hypothesis based on the comparison. If the Chi square statistic was greater than or equal to the critical value, I rejected the null hypothesis and concluded that there is an association between the two variables. If the Chi square statistic was less than the critical value, I failed to reject the null hypothesis and concluded that there is no evidence of an association between the two variables.

The type of error that could have been made in this test is a type I error, which occurs when the null hypothesis is rejected when it is actually true. This means that I could have falsely concluded that there is an association between gender and depression when there is none. The probability of making a type I error is equal to the significance level (alpha) that I chose for the test.

I repeated this test four (4) more times for different sets of hypothesis. These are:

**Set 2:**  
Ho: There is no association between sexual orientation and depression levels.  
H1: There is an association between sexual orientation and depression levels.

**Set 3:**  
Ho: There is no association between gender and poverty level.  
H1: There is an association between gender and poverty level.

**Set 4:**  
Ho: There is no association between sexual orientation and poverty level.  
H1: There is no association between sexual orientation and poverty level.

### 8. Moderation

### 9. Save

save(myHealthData, file = "myHealthData.RDATA")