Final Report - Greg Finkelberg

## Background/Motivation for the Study

As someone seeking a graduate degree who deals with depression and anxiety, I’m interested to see how education level impacts the prevalence of depression among those in the NHANES dataset. I tend to think that the more educated a person is the more likely they are to deal with depression, however, my preliminary findings show a drastic trend in the opposite direction. The relevant literature I’ve found seems to agree with this trend.

Sources:

Ross, C. E., & Van Willigen, M. (1997). Education and the subjective quality of life. *Journal of health and social behavior*, 275-297.

Lorant, V., Deliège, D., Eaton, W., Robert, A., Philippot, P., & Ansseau, M. (2003). Socioeconomic inequalities in depression: a meta-analysis. *American journal of epidemiology*, *157*(2), 98-112.

Miech, R. A., Eaton, W. W., & Brennan, K. (2005). Mental health disparities across education and sex: A prospective analysis examining how they persist over the life course. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *60*(Special\_Issue\_2), S93-S98.

Quesnel-Vallée, A., & Taylor, M. (2012). Socioeconomic pathways to depressive symptoms in adulthood: evidence from the National Longitudinal Survey of Youth 1979. *Social science & medicine*, *74*(5), 734-743.

Lundborg, P. (2013). The health returns to schooling—what can we learn from twins?. *Journal of population economics*, *26*, 673-701.

Ross, C. E. (2017). *Social causes of psychological distress*. Routledge.

## Research Question and Hypothesis

Is there a link between education level and the prevalence of depression?

Null hypothesis - There is no difference in the prevalence of depression between individuals belonging to different education groups.

Alternative hypothesis - There is a difference in the prevalence of depression between individuals belonging to different education groups.

## 

## Data Description and Exploratory Data Analysis

My target variables in the NHANES dataset are Education and Depressed. The sample size is 6430 adults. I loaded the following libraries and modified the Education variable to fit the labels on my bar chart later in my analysis. This variable will be named “Ed” and the data frame will be “df”. Education level will now range from 1-5, with 1 being the least educated (8th grade) and 5 being the most educated (College grad). You can see what the other education levels equate to in the following code.

library(NHANES)  
library(dplyr)  
library(ggplot2)  
  
df <- NHANES %>% mutate(Ed = case\_when(Education == '8th Grade' ~ '1',  
Education == '9 - 11th Grade' ~ '2',  
Education == 'High School' ~ '3',  
Education == 'Some College' ~ '4',  
Education == 'College Grad' ~ '5'))

I first wanted to visualize the distribution of subjects by education level and how often they felt depressed. I started with a contingency table of the raw distribution and then converted it to a proportional table.

my.table <- table(df$Ed, df$Depressed, dnn=c("Education Level","Days Spent Depressed"))  
addmargins(my.table)

Days Spent Depressed  
Education Level None Several Most Sum  
 1 250 48 54 352  
 2 551 128 96 775  
 3 1039 246 79 1364  
 4 1603 321 120 2044  
 5 1608 233 54 1895  
 Sum 5051 976 403 6430

prop\_table <- prop.table(my.table, margin = 1)  
prop\_table

Days Spent Depressed  
Education Level None Several Most  
 1 0.71022727 0.13636364 0.15340909  
 2 0.71096774 0.16516129 0.12387097  
 3 0.76173021 0.18035191 0.05791789  
 4 0.78424658 0.15704501 0.05870841  
 5 0.84854881 0.12295515 0.02849604

Proportions = as.data.frame(prop\_table)

Both variables are categorical. Depression is separated into three categories expressing how many days subjects experience depression (i.e. None, Several, and Most). Education is separated into five different levels that indicate how far into their education each subject got (i.e. 1-5). As you can see in the proportional table below, the proportion of subjects experiencing no depression increases with education level. Furthermore, the frequency of subjects being in the most severe depression category goes down with increased education level.

Proportional Table

|  |  |  |  |
| --- | --- | --- | --- |
| **Education Level** | **Days Spent Depressed** | | |
| None | Several | Most |
| 1 | 0.710 | 0.136 | 0.153 |
| 2 | 0.711 | 0.165 | 0.124 |
| 3 | 0.762 | 0.180 | 0.058 |
| 4 | 0.784 | 0.157 | 0.059 |
| 5 | 0.849 | 0.123 | 0.028 |

Table 1. Proportional distribution of depression for each education group.

Next, I wanted to visualize this trend in a bar chart.

Proportions %>%  
 ggplot(aes(x = Education.Level, y = Freq,   
 fill = Days.Spent.Depressed)) +   
 geom\_bar(stat = "identity", position = "dodge") +  
 labs(title = "Proportional Depression by Education Level",   
 x = "Education Level", y = "Proportional Frequency",   
 fill = "Days Spent Depressed") +   
 theme\_bw()

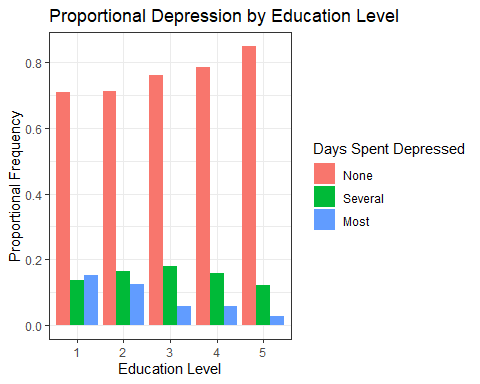


Figure 1. Bar chart of proportional depression for each education group.

This graph shows the same trend as the contingency tables. The proportional frequency of subjects experiencing no depression increases with education level, and severe depression is reduced with increased education level. Interestingly, the proportion of those in the “Several” category slightly increases through education level 3 (High School), then it declines with increased education level. Overall, the graph does seem to support the alternative hypothesis that there is a difference in the prevalence of depression between education levels. A chi-squared test should determine if this trend is statistically significant.

## Analysis

Because both variables are categorical and more than two levels, a chi-squared test is the most appropriate test to analyze these data. First, the conditions for test validity must be met. Because the dataset was collected randomly and the sample size is less than 10% of the population, the first two conditions are met. Finally, we need to make sure the sample is large enough for all the expected counts to be 5 or greater. I calculated these values using Excel and they can be seen in the following table.

Expected Values

|  |  |  |  |
| --- | --- | --- | --- |
| **Education Level** | **Days Spent Depressed** | | |
| None | Several | Most |
| 1 | 276.51 | 53.43 | 22.06 |
| 2 | 608.79 | 117.64 | 48.57 |
| 3 | 1071.47 | 207.04 | 85.49 |
| 4 | 1605.64 | 310.26 | 128.11 |
| 5 | 1488.59 | 287.64 | 118.77 |

Table 2. Expected sample distributions for each group.

Now that the conditions have been met, we can run the chi-squared test and view the residuals.

CStest <- chisq.test(df$Ed, df$Depressed, correct=FALSE)  
CStest

Pearson's Chi-squared test  
  
data: df$Ed and df$Depressed  
X-squared = 167.01, df = 8, p-value < 2.2e-16

CStest$stdres

df$Depressed  
df$Ed None Several Most  
 1 -3.5406671 -0.8295567 7.2239545  
 2 -5.3930864 1.1063457 7.4949885  
 3 -2.4133089 3.3122164 -0.8166345  
 4 -0.1720411 0.8019334 -0.8958432  
 5 7.9576841 -4.1652876 -7.3094992

Residuals

|  |  |  |  |
| --- | --- | --- | --- |
| Education Level | Days Spent Depressed | | |
| None | Several | Most |
| 1 | -3.54 | -0.83 | 7.22 |
| 2 | -5.39 | 1.11 | 7.49 |
| 3 | -2.41 | 3.31 | -0.82 |
| 4 | -0.17 | 0.80 | -0.90 |
| 5 | 7.96 | -4.17 | -7.31 |

Table 3. Chi-squared test residuals.

The test resulted in a p-value much smaller than 0.05, suggesting that there is a significant association between education and depression. I created a table of the residuals to see how the sample values differed from the expected values. As you can see from the residuals, depression prevalence appears to be higher than expected among those in lower education levels and less than expected in more educated individuals.

## Conclusions

The results of the chi-squared test seem to be unambiguous. The extremely small p-value combined with the residual table confirms the same trend these data have shown throughout the preliminary analysis. The p-value indicates that there is a statistically significant relationship between education and depression. The residuals combined with the bar graph in the earlier section show the negative direction of this relationship, meaning that the more educated a person is, the less likely they are to experience depression. Because of this, we can reject the null hypothesis as there is clearly a difference in the prevalence of depression between education groups.

I think the key take away is that while there is a clear link between education and depression, this analysis would benefit from the inclusion of other variables. Specifically, I think it would be interesting to include an economic variable representing financial well-being. It stands to reason that the more educated a person is the more likely they are to earn more money. Earning more money could lead to less stress from factors relating to poverty. Less stress could lead to a reduced prevalence of mental illness. Mental health is such a complex research area that a multi-variable analysis could be very illuminating.