Mini-Project 3

Stat 212: Interim 2021

Chamee Vang, Marco Ruiz, Ivana Kramarevsky

Introduction

NHANES is survey data collected by the US National Center for Health Statistics (NCHS) on about 5,000 individuals every year since 1999. These individuals have been identified to be from mostly racial minorities. The intended population is "the non-institutionalized civilian resident population of the United States". From this data set, we plan to compare education status and if the individual is physically active. Our research questions are as follows:

1) Does the college attainment level of the participant affect how active they are through physical activities?
2) Does the trend found in the question above hold when broken down by general educational level?

Methods

The "Education" variable we used was from the NHANES dataset which focused on participants aged 20 years and older. They were categorized into: 8th grade, 9-11th grade, High School, Some College, and College graduates. The "PhysActive" variable determined if the participant was physically active in the form of 'Yes' and 'No' collected from participants 12 years and older. We created a `College` variable by combining 8th grade, 9-11th grade, and high school responses in the `Education` variable to create a "no college" response. Similarly, we combined "some college" and "college graduate" responses to create a "college" response. We did this to have a binary response that both relates to and generalizes from the `Education` responses. Lastly, we filtered out participants that had no responses for their education level or physical activity.

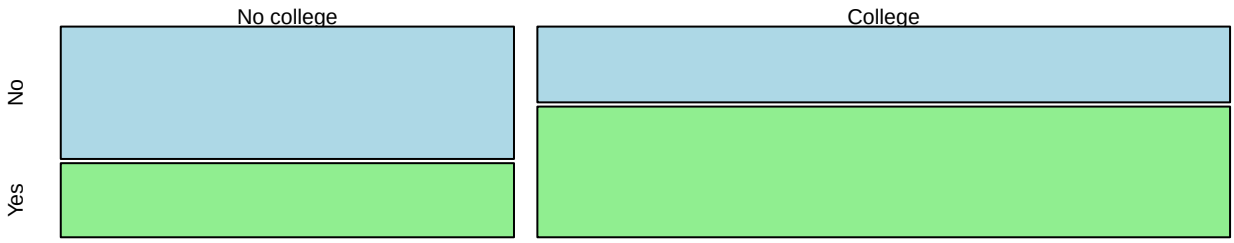
For our first question about college attainment and physical active participation, our null hypothesis is: there is no difference in proportion between participants with a college attainment and are physically active versus the participants without a college attainment and are physically active. Our alternative hypothesis is that there is a difference in the proportions. Our response variable is whether the participant is physically active, and our explanatory variable is college attainment. We will perform a hypothesis test for the difference between the two proportions. For the second question, our null hypothesis is that there is no trend between the education level acquired and the status of physical activity of each participant. Our alternative hypothesis is that there is a trend between the two variables. Our explanatory variable is the participant's education level and the response is their physical activity. We will perform a Chi-squared test of independence.

Results

Two tables and mosaic plots are produced to show a summary of our data. Both mosaic plots, named "Education and Physical Activity" and "College and Physical Activity," show a trend of increasing physical activity as participants further their education.

Table 1. Total counts broken by college attainment and physical activity

```
##
##                 No  Yes  Sum
##   No college  1831 1025 2856
##   College     1602 2763 4365
##   Sum         3433 3788 7221
```




```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  tally(PhysActive ~ College)
## X-squared = 520.07, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.2514261 0.2967662
## sample estimates:
##    prop 1    prop 2
## 0.6411064 0.3670103
```

For the first research question, the mosaic plot shows college grads have roughly double the exercise rate of the non-college group. The lower proportion has, very roughly, a 30% physical activity rate. The higher proportion has, roughly, a 60% physical activity rate. There is also a larger population of college students within the sample, as seen in Table 1.

The test statistic is an chi-squared value of 520.07. The p-value is 10^{-16}. The high x-squared value and unusual p-value indicate that the difference in proportion is very unusual, so we reject the null hypothesis. Since the proportion of college vs. non-college participants who exercise is higher as seen on the mosaic plot, this means that there is a significant boost in doing physical activity and attending college. The 95% confidence interval is (0.25, 0.30).

Table 2. Total counts broken by education level and physical activity

```
##
##                     No  Yes  Sum
##   8th Grade        339  112  451
##   9 - 11th Grade   610  278  888
##   High School      882  635 1517
##   Some College    1056 1211 2267
##   College Grad     546 1552 2098
##   Sum             3433 3788 7221
```

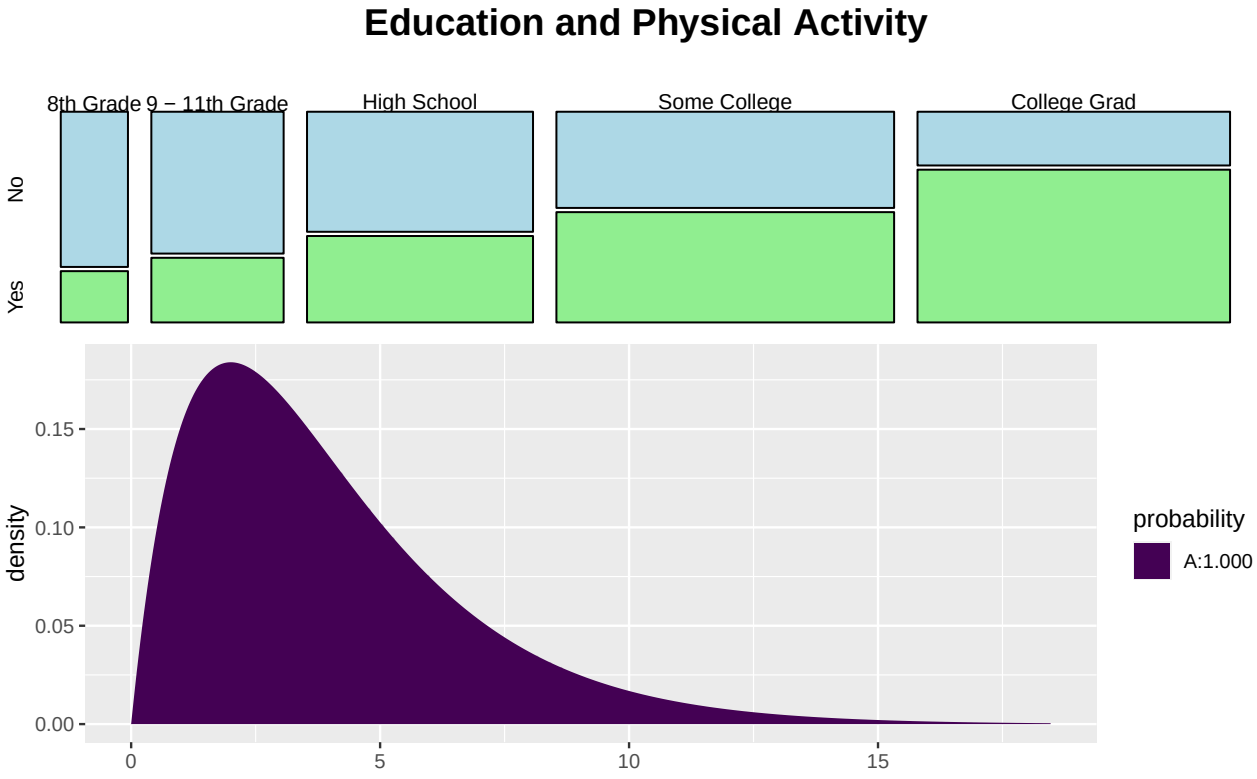
```
##
##  Pearson's Chi-squared test
##
## data:  NHANES_edu_table
## X-squared = 755.95, df = 4, p-value < 2.2e-16
```

Table 3. Chi-squared expected table

```
##
##                          No       Yes
##   8th Grade       214.4139  236.5861
##   9 - 11th Grade  422.1720  465.8280
##   High School     721.2105  795.7895
##   Some College   1077.7747 1189.2253
##   College Grad    997.4289 1100.5711
```

Table 4. Chi-squared residual table

```
##
##                           No         Yes
##   8th Grade        8.5083036  -8.0998112
##   9 - 11th Grade   9.1414613  -8.7025703
##   High School      5.9872401  -5.6997865
##   Some College    -0.6632662   0.6314221
##   College Grad   -14.2938222  13.6075611
```




```
##     X-squared
## 2.665786e-162
```

For the second question, the largest amount of data points are in the categories "some college" and "college grad," as seen in Table 2. In addition to this trend, our mosaic plot shows that the proportion of exercise increases for each increase in education level acquired by roughly 5/4 of the previous rate. We can also see that Table 2 and Table 3 differ quiet a bit, which is quantified in Table 4; lower levels of education should have higher values counts under the "No" response, assuming that there is no trend between the two variable.

The test statistic is an x-squared value of 755.95. The p-value is 10^{-16}. The high x-squared value with its unusual p-value means we reject the null hypothesis again: there is a significant relationship between education level and physical activity status. Because the proportion of participants who exercise visually corresponds to the education level on the mosaic plot, this means that there is a significant positive correlation between the level of education acquired and physical activity status.

Discussion

To recap, the first research question is "does attending and completing college affect how physically active participants are?" Since the null hypothesis is rejected, then the visually summarized relationship with the increased rate of exercise of college attendees is significant in the sample population of the NHANES data set. We are 95% confident that 25% to 30% of U.S. civilian residents with college attainment are more physically active than those without college attainment.

The second research question is "does the trend found in the question above hold when broken down by educational level?" Since the null hypothesis is rejected, then the visually summarized relationship between increased rates of exercise with higher educational status is significant in the sample population of the NHANES data set. So, the same trend holds when comparing rates of exercise to education levels below college. Overall, rates of greater educational achievement in the population produce a greater proportion of the population that participates in physical activity.

A limit on the sample population used in the NHANES set is that it is more diverse than the actual population. More work could be done by studying the two research questions by race or ethnicity. Another possible study is another observational one that looks into the factors that may contribute to increased rates of exercise with educational level.