Statistics 2024 Homework

Kun Dong – kun.dong@ucdconnect.ie

October 10, 2024

Problem 1 Gender

The variable “newgender” is a nominal variable(categorical variable) representing the gender of the respondents. It has two categories: “Female” with 1,094 respondents and “Male” with 929 respondents. The mode of the “newgender” variable is “Female”, with a frequency of 1094. As shown in Figure 1.

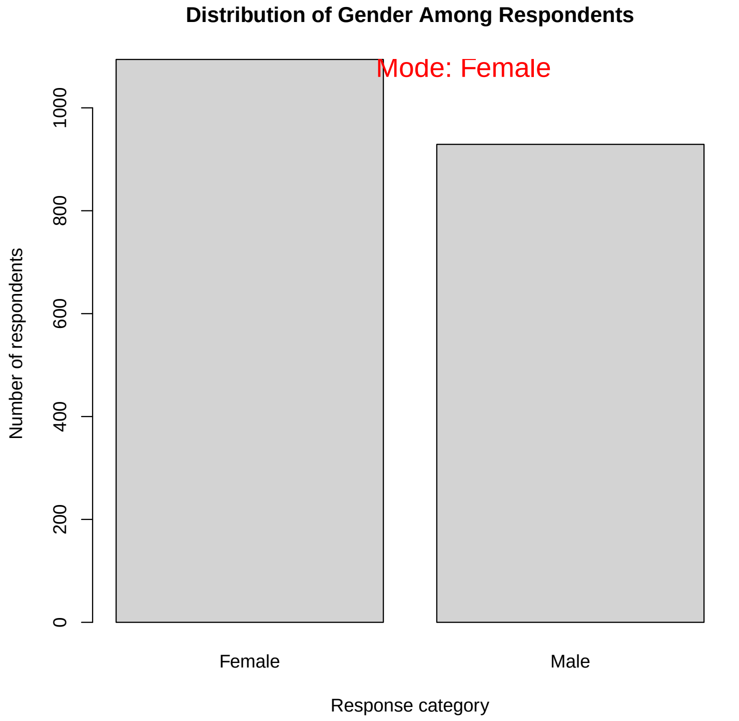

Figure 1: Distribution of Gender Among Respondents

As I mentioned, the variable ”newgender” is a nominal variable, because its categories represent distinct groups with no order. The appropriate measure of central tendency for nominal variables is the mode. Specifically, the mode is the gender category “Female” with 1094 respondents. As for dispersion, nominal variables do not have a meaningful measure of dispersion like variance, standard deviation even range.

Problem 2 Respondents' Educational Attainment

The variable "newdegree" represents the educational level of respondents and is a nominal variable categorized into five distinct educational levels: "LT HS" (Less than High School) with 297 respondents, "HS" (High School) with 1003 respondents, "Jun Coll" (Junior College) with 173 respondents, "Bachelor" with 355 respondents, and "Grad deg" (Graduate Degree) with 194 respondents.
The bar plot illustrates the distribution of these educational levels among respondents, with the height of each bar representing the number of respondents in each category. As shown in Figure 2
The plot is titled and includes labeled axes for clarity. The mode, which is the educational level with the highest frequency, the median, which illustrates the educational level at the midpoint, are both indicated in red, and the range of the variable is indicated in blue.

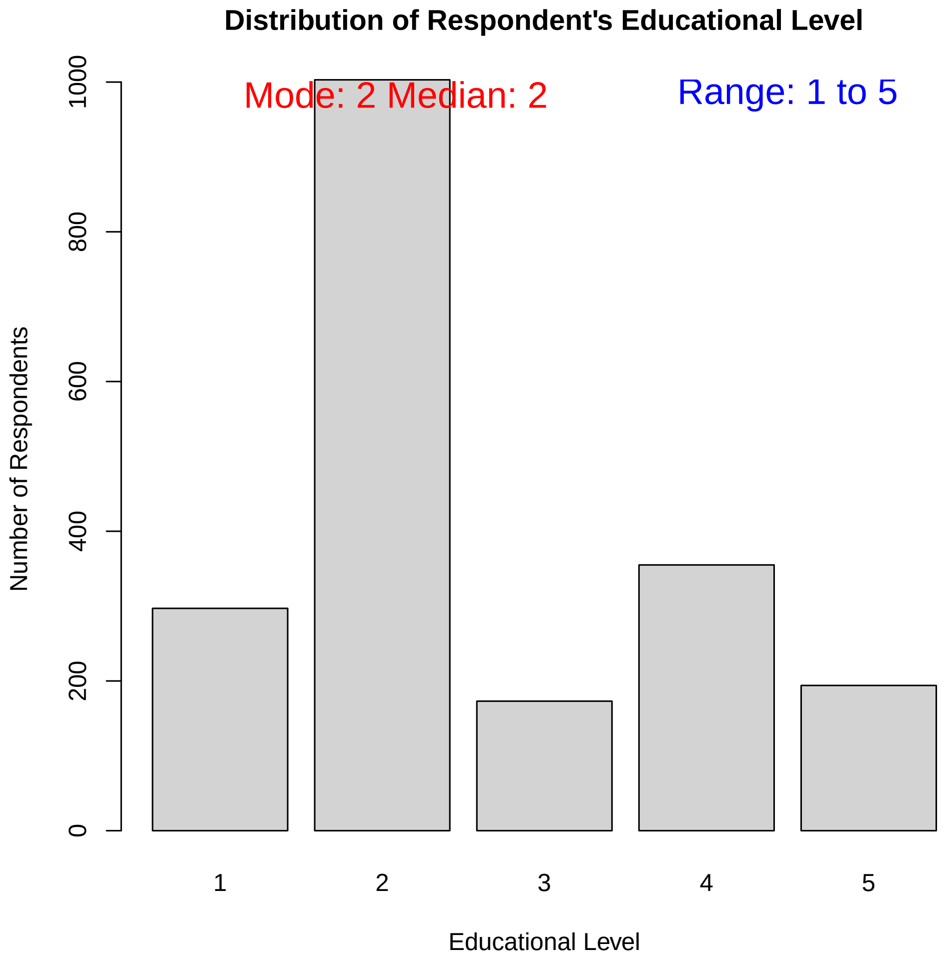

Figure 2: Distribution of Educational Attainment Among Respondents

The variable "newdegree" is an ordinal variable, because it categorizes respondents' educational levels in a meaningful order. The mode of "newdegree" is the educational level with the highest number of respondents, which is 2("HS"). In terms of the dispersion, the range of the variable "newdegree" is from 1("LT HS") to 5("Grad deg").

Problem 3 Hours per day watching TV

The variable “tvhours” is a continuous variable, and the central tendency and dispersion can be analyzed using the following measures. The mode of the variable “tvhours” is 2, which represents the number of hours most frequently watched by respondents. The median value of “tvhours” is 2, indicating that half of the respondents watch 2 hours or less, while the other half watch more than 2 hours. The mean of “tvhours” is approximately 2.98, with a range from 0 to 24 hours. Additionally, the variance of “tvhours” is about 7.07, and the standard deviation is about 2.66, providing insight into the degree of dispersion.
The chart illustrates the distribution of respondents’ TV watching hours. The mode is marked with a red line, the median with an orange line, and the mean with a blue line, with their corresponding values displayed on the plot. Additionally, the title and x-axis are clearly labeled for clarity. It is purple line that represents the probability density. As shown in Figure 3

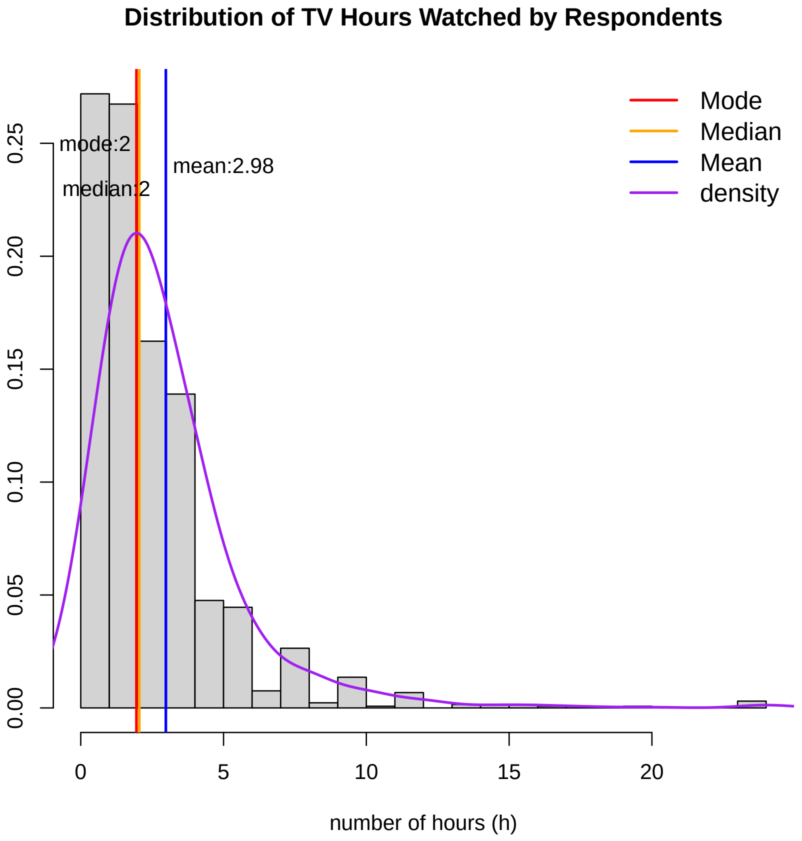

Figure 3: Distribution of TV Hours Watched by Respondents

Since the mean is greater than the median, which equals the mode, it is clear that this is a positively (right) skewed distribution. The data is concentrated on the left with a long right tail, making this characteristic quite apparent. This distribution pattern indicates that most people watch a small amount of TV, while only a few watch a large amount.

Problem 4 Education and Health Care System

The dataset used “newdegree” to describe educational levels and the variable “newhealthcare” to represent opinions about the healthcare system. According to the principle that the independent variable (cause) is educational level and the dependent variable (outcome) is the opinion on the healthcare system, we can formulate the null and alternative hypotheses. Specifically, the Null Hypothesis (H_0) states that there is no relationship between educational level (“newdegree”) and opinions on the healthcare system (“newhealthcare”). The Alternative Hypothesis (H_1) posits that there is a relationship between educational level and opinions on the healthcare system. To further investigate the relationship between educational level and health opinions, a contingency table is generated to display the frequency, as shown in Figure 4.

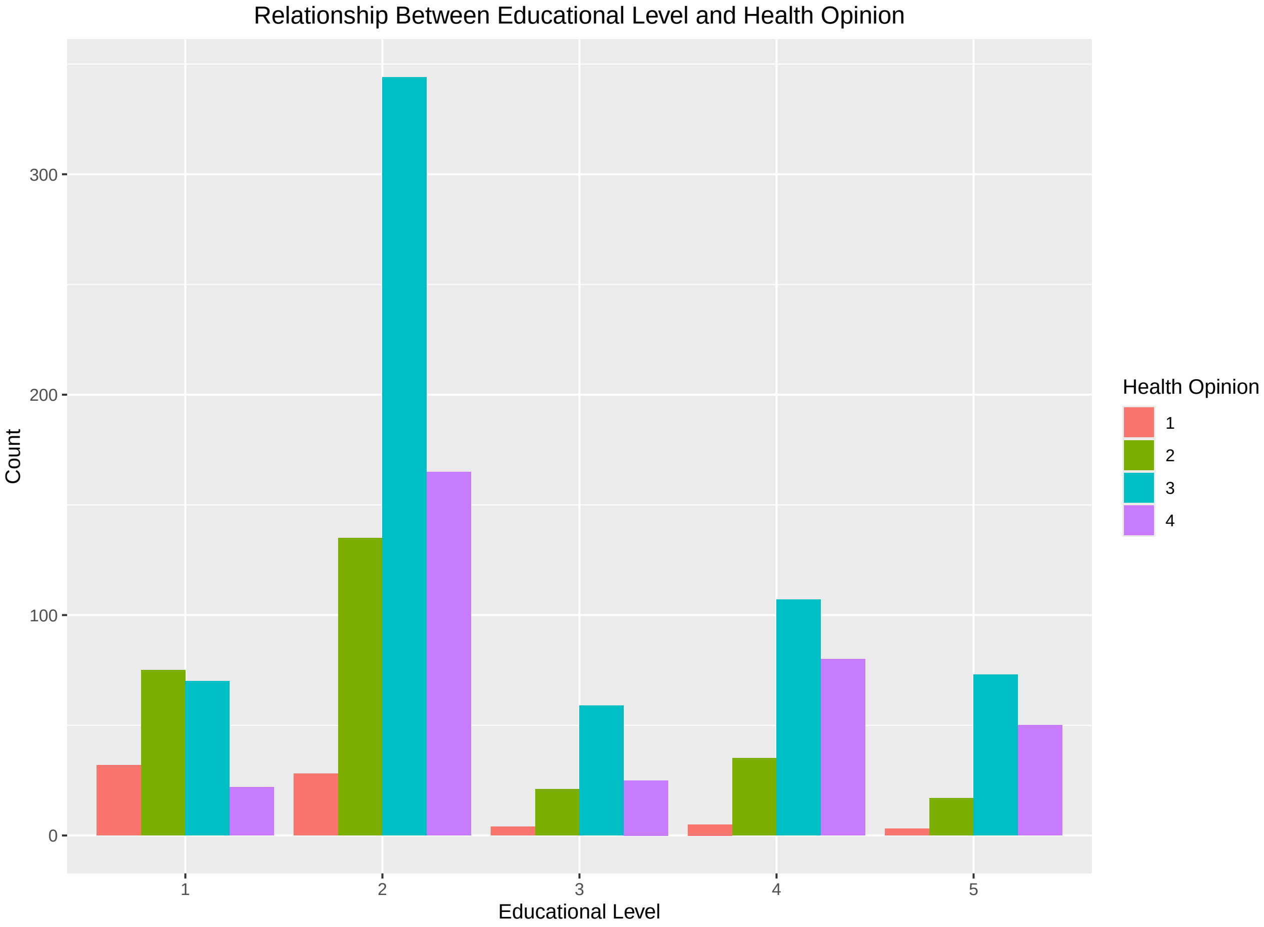

Figure 4: Relationship Between Educational Level and Health Opinion

Using the `chisq.test()` function, the test yielded $\chi^2 = 128.01$, degrees of freedom df $= 12$, and a p-value $< 2.2 \times 10^{-16}$. This result indicates that we can reject the null hypothesis at a significance level of 0.05. In other words, there is a strong relationship between educational level (“newdegree”) and opinions on the healthcare system (“newhealthcare”); they are dependent. Given that we reject the null hypothesis in favor of the alternative hypothesis, we conclude that individuals’ levels of education significantly influence their views on the healthcare system. However, further analysis is needed to understand the specific nature of this relationship. Finally, we save the modified data object as ourdata_lab_hw1 in the working directory.