Stat250 Project

Victor Huang, Nick Baeck

3/4/2022

Introduction/Methods

There are many ways for a state and country to measure the wellness of their citizens, two of the most prevalent ones being average physical health of the people and the average wealth of the country as a whole. As such, it makes sense to assume that the two might be correlated. For this project we will be investigating this theoretical correlation, more specifically, we will be analyzing the effects GDP, average income, and poverty rate might have on its citizens health. [1] More specifically, how low GDP, low average income, and high poverty rates might correlate with a higher percent of the population being obese (an indicator of poor health). The data we will be using in this lab will be based off of 200 data points of 50 states collected over 4 years (2014, 2015, 2016, and 2017). At the start of the year, data that measured state wealth (Average.Income, Poverty.Rate, Real.GDP, etc.) and unhealthiness (Adult.Obesity) was collected from each state. After the course of a year, the data for each state was collected again. From the data, we hypothesize that the states with a higher GDP, higher average income, and lower poverty rates will have a noticeable higher proportion of citizens that are obese, we seek to legitimize this hypothesis by using wealth indicators such as Average.Income, Poverty.Rate, Real.GDP, and Adult.Obesity to check if their is a correlation between wealth and health among U.S. states.

As stated above, our observation will focus on three separate wealth indicators and one health indicator:

1. Adult.Obesity: The proportion of adult citizens that are obese
2. Average.Income: The average income of citizens
3. Poverty.Rate: The proportion of citizens that are considered impoverished
4. Real.GDP: An index/measurement for a state's general wealth

Results

We start our analysis by finding the average income, real GDP, and Poverty Rate for all the data points of the country over time. By the definition of average, the data points above the average income and average real GDP are considered "rich" while those below the average are considered "poor" (the opposite holds true for poverty rates). After separating our data points into "rich" and "poor", we will compare the results for the two groups side-by-side for all three variables to see the effects of wealth on health. We conclude with the following null and alternative hypothesisg

$H_0 : p_{highGDP} = p_{lowGDP}$ verses $H_a : p_{highGDP} < p_{lowGDP}$

The null hypothesis states that adult obesity rates won't be affected by high and low average GDP while the alternative states that adult obesity rate will be higher for low GDP states when compared to high GDP states

[1] Data taken from Kaggle dataset provided by Anne Dune linked here: https://www.kaggle.com/annedunn/obesity-and-gdp-rates-from-50-states-in-20142017

$H_0 : p_{lowAvgInc} = p_{highAvgInc}$ verses $H_a : p_{lowAvgInc} < p_{highAvgInc}$

The null hypothesis states that adult obesity rates won't be affected by high and low average income while the alternative states that adult obesity rate will be higher for low average income states when compared to high average income states

$H_0 : p_{lowPovRate} = p_{highPovRate}$ verses $H_a : p_{lowPovRate} < p_{highPovRate}$

The null hypothesis states that adult obesity rates won't be affected by high and low poverty rates while the alternative states that adult obesity rate will be higher for high poverty rate states when compared to low poverty rate states

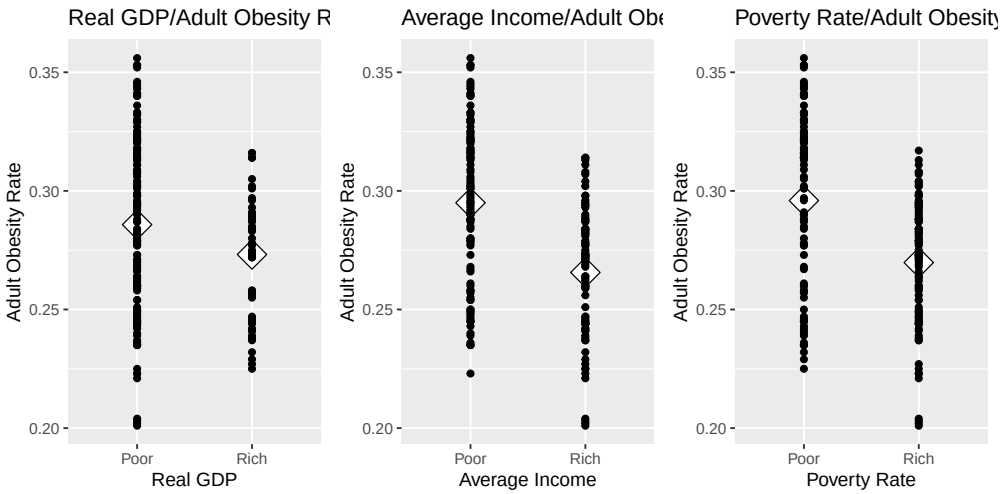

Figure 1: GDP, Average Income, Poverty Rate and Obesity Rate

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c out of crichGDPCount_obese out of richGDPCountpoorGDPCount_obese out of poorGDPCount
## X-squared = 0.50656, df = 1, p-value = 0.7617
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.0000000  0.2063661
## sample estimates:
##    prop 1    prop 2
## 0.5423729 0.4751773

##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c out of crichIncomeCount_obese out of richIncomeCountpoorIncomeCount_obese out of poorIncome
## X-squared = 26.126, df = 1, p-value = 1
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.0000000  0.4927034
## sample estimates:
##    prop 1    prop 2
## 0.7045455 0.3303571

##
```

```
##  2-sample test for equality of proportions with continuity correction
##
## data:  c out of crichPovertyCount_obese out of richPovertyCountpoorPovertyCount_obese out of poorPov
## X-squared = 24.976, df = 1, p-value = 1
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.0000000  0.4597811
## sample estimates:
##    prop 1    prop 2
## 0.6320755 0.2857143
```

For each of the wealth variables we split the dataset into two categories ("rich" and "poor") based on the aforementioned definition. After doing so we conducted a side-by-side plot comparing the obesity rate for each wealth variable mentioned. With this done, we proceeded to conduct proportion tests to see if there is indeed a statically discernible difference to prove our hypothesize. Setting the significance level to be $\alpha = 0.05$, we found that the χ-squared values are 0.50656, 26.126, and 24.976 and the p-values are 0.7617, 1, and 1 for the comparisons between Real GDP, Average Income, and Poverty Rate respectively. This completely goes against out inital side-by-side plots as none of the indicators show evidence that support our initial hypothesis. As such, we are unable to reject the null hypothesis for all three of our initial hypothesis. Thus, we do not have significant evidence that high real GDP, average income, and low poverty rates correlate to high obesity rates

Out of curiosity, we decided to dig a little deeper and see if the inverse applies. That is, if more wealth correlates to higher obesity rates. We proceeded to conduct proportion tests to see if there is indeed a statically discernible difference to prove our new hypothesis. Setting the significance level to be $\alpha = 0.05$, we found that the χ-squared values are 0.50656, 26.126, and 24.976 and the p-values are 0.2383, 1.599e-07, and 2.903e-07 for the comparisons between Real GDP, Average Income, and Poverty Rate respectively.

As such, we accepted the new null hypothesis for our second and third hypothesis and rejected our first new null hypothesis. Thus, we have significant evidence that high average income and poverty rates correlate with a higher obesity rate while not having any evidence that real GDP is related to obesity rates.

Discussion

To conclude everything above, using a significance level of $\alpha = 0.05$ we were unable to reject any of our initial hypothesis. However, looking at our "inverse" hypothesis, we concluded that a higher average income and lower poverty rates correlate to a higher obesity rate.

A problem our group wants to point out is that the dataset we are working with was very small and only consists of 27 subjects. Perhaps with a larger sample size we would have been able to get more accurate results. We also would like to point out that even though these mice were randomly picked, we are not sure they serve as an accurate representation of the population as a whole. Another issue could exists in the fact that each group did not have the same number of mice as the group of 27 mice were split into 8, 9, and 10 mice groups. If we were able to do this study again, we would prefer to have a larger sample size that represents the population at large.

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c out of crichGDPCount_obese out of richGDPCountpoorGDPCount_obese out of poorGDPCount
## X-squared = 0.50656, df = 1, p-value = 0.2383
## alternative hypothesis: greater
## 95 percent confidence interval:
##  -0.0719749  1.0000000
```

```
## sample estimates:
##    prop 1    prop 2
## 0.5423729 0.4751773


##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c out of crichIncomeCount_obese out of richIncomeCountpoorIncomeCount_obese out of poorIncome
## X-squared = 26.126, df = 1, p-value = 1.599e-07
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.2556732 1.0000000
## sample estimates:
##    prop 1    prop 2
## 0.7045455 0.3303571


##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c out of crichPovertyCount_obese out of richPovertyCountpoorPovertyCount_obese out of poorPov
## X-squared = 24.976, df = 1, p-value = 2.903e-07
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.2329413 1.0000000
## sample estimates:
##    prop 1    prop 2
## 0.6320755 0.2857143
```