SEIS631 Final Project

Nikolai

5/4/2022

Introduction

Personality has always been a fascination of mine. While many personality tests have little validity or predictive value, The Big Five personality tests have a reputation of being a good measure with predictive power. I have found a publicly available dataset of demographic information and personality data that I wish to analyze using techniques taught in class.

What

As mentioned earlier, the Big Five Personality test is thought to be one of the more accurate and predictive personality tests for human behavior. For example, having a combination of high Agreeableness and Conscientiousness usually means the person is going to be a good employee while having high neuroticism is related to a variety of bad outcomes. I would like to see if there are relationships between the expression of different personality traits across countries and gender.

Why

While we know that the expression of certain combinations of personality traits predict certain behaviors, these traits are often seen as being separate and distinct measures of personality. I argue that there are likely 'personality types' - combinations of these traits that are more likely to occur together - that define subsets of the population.

How

My plan is to subset the dataset while controlling for the expression of one or more personality traits. Using these traits as a control, I will look to see if the other traits expression in this subset is significantly different from the normal population.

Body

Predicting human behavior has always been the goal of personality research. Most personality tests started as observational interpretation of behavior types and tried to create a logical framework for them, such as the Meyers-Briggs test. Most of these tests turned out to have low predictive power or real application, rendering them important in the public eye and little else. The Big Five took a different approach. It started with analyzing language into groups of related adjectives. When a group of adjectives displayed distinct characteristics and a larger population, they were deemed a 'trait'. This eventually lead to five personality traits- Openness to experience, Conscientiousness, Extroversion, Agreeableness, and Neuroticism. These traits turned out to have higher predictive power than other personality tests.

Cleaning Data

The dataset was quite large and had demographic information and the actual responses to the individual questions. While this is ideal, it means the dataset must be cleaned to be useful. For example, to account for

the tendency to give high ratings over low ratings regardless of the question's content, several of the questions are stated negatively. For example, on a rating of 1-5, answering the question 'I like to talk' with a 5 would be high extroversion, but answering the question 'I don't like loud parties' with a 5 would actually be the opposite. I will write a function to inverse a question's rating and then manually code which questions are positive or negative and apply the function only to negative questions. Then, because there would be too many rows with this dataset if looking only at the individual questions, I will average the answers per trait to get a single score for the five traits.

Controlling Trait Expression

Each observation (respondent) is represented in one row in the dataset. This means that so long as the dataset is not pivoted or 'melted', it is possible to subset the dataset looking for observations that have certain traits or combinations of traits, and then see the effect on the remaining traits. In order to visualize the data nicely in one graph, the data will have to be pivoted after the subset operation.

Grouping Data on Gender, Country, and Age

After trait expression has been modified, it will become possible to group the data by a variety of other demographic information. The demographic information I'm most interested in is Gender, Country, and Age, as these three stand out to me as having significant impact on people's lives. Because this test has respondents from over 100 different countries, most with a low response rate, I plan on using data from only countries with a high response rate. The others will be aggregated into an 'Other' country group.

Comparing Subset Population to General Population

After, I plan to use statistical analysis

Considerations

There are many considerations to be aware of with this dataset. The dataset represents an online self-reported personality test, which means it does not represent a true random sampling of the population and the responses are not an objective determination of how the participants might really act. It's possible that certain personality traits will be more probable given that these personality traits might make responding to a survey more likely. Also, while age is included in the data, it is not a longitudinal study. This means that we cannot determine how personality might change over time with this dataset. Also, the vast majority of respondents are young (age 25 or younger), which could have an influence on the results. This might be due to how old the test is, or the fact that younger populations might be forced to take the test in an introduction to psychology class. Lastly, when subsetting the population will become much smaller than the general population, which means the results could be skewed and the sample error could be quite high.

Topics From Class

(a) Git
(b) RMarkdown
(c) Statistical concepts such as normal distributions, mean, standard deviations, percentiles, and areas under the curve
(d) Geometric distributions
(e) Statistics to determine if two populations are significantly different.

Conclusion

In conclusion, I expect this to be an interesting analysis of the data in a way that could find macro personality types using the method outlined above.