SDGB-7844 Homework #3

Prof. Matthew Murphy

March 4, 2020

The USA Today Diversity Index

The USA TODAY Diversity Index is a number – on a scale from 0 to 100 – that represents the chance that two people chosen randomly from an area will be different by race and ethnicity. In more personal terms: "What is the chance that the next person I meet will be different from me?" A higher number means more diversity, a lower number, less. The index was invented in 1991 by Phil Meyer of the University of North Carolina and Shawn McIntosh of USA TODAY.

The index is calculated using current federal standards for race and ethnicity, which are treated as separate concepts. That leads to a multi-step calculation:

Create each racial group's share of the population. We will treat it as the probability that ONE person chosen at random will be of that race.

$$RaceProportion_i = \frac{Race_i}{Total_i}$$

The probability that TWO people chosen at random will be of that particular race is the single probability multiplied by itself (squared). Take each racial group's share of the population and square it. Sum the squared probabilities for the separate races. This is the probability that two people are of the same race.

$$P(Racial_i) = \sum_{i=1}^{n} RaceProportion_i^2$$

In the current federal scheme, there are five named races – white, black/African-American, American Indian/Alaska Native, Asian and Native Hawaiian/Other Pacific Islander. The Census also includes a category called "Some other race." Because studies show that people who check it are overwhelmingly Hispanic, that category is not used. Hispanics' effect on diversity is calculated in the next step.

Because Hispanic origin is a separate Census question, the probability that someone is Hispanic or not must be figured separately. Take the Hispanic and non-Hispanic percentages of the population, square each and add them to get the probability that any two people will be Hispanic or not.

$$P(Ethnic_i) = Hispanic_i^2 + NonHispanic_i^2$$

To calculate whether two people are the same on both measures, multiply the results of the first two steps. This is the probability that any two people are the same by race and ethnicity.

$$P(Same_i) = P(Racial_i) \times P(Ethnic_i)$$

Subtract the result from 1 to get the chance that two people are different – diverse. For ease of use, multiply the result by 100 to place it on a scale from 0 to 100.

$$DiversityIndex_i = \Big(1 - P(Same_i)\Big) \times 100$$

Exploratory Analysis

1. At what level (State, County, City) is the American Community Survey data available? How is this different than the deccenial census?
2. What variable and variable codes are available to describe race and ethnicity in the US Census? Describe how these variables are represented in the data (Variables: B2001_001-B2001_006 & B03002_001-B03002_002 & B03002_012).
3. How does the American Community Survey define race and ethnicity? Is this important information to report under assumptions for your analysis?
4. Does the American Community Survey provide the margin of error for their estimates of the proportion of the prevalence of each race and ethnicity? How might this impact the validity of our results?
5. Use the *tidycensus* API to assign the race and ethnicity data for New York, New Jersey and Connecticut (at the County level) to a data frame.

Computing The USA Today Diversity Index

Each of the calculations below will be done **by county** and not in aggregate.

Step 1:

In the current federal scheme, there are five named races – white, black/African-American, American Indian/Alaska Native, Asian and Native Hawaiian/Other Pacific Islander and an estimate for total population (B2001_001-B2001_006). Ensure that you have collected the proper data from the *tidycensus* API for these values, as well as the values for the Hispanic population (B03002_001-B03002_002 & B03002_012).

Use the *spread* function to create columns for each racial group (and the total population). Rename these columns to better reflect the data if you have not already done so.

Calculate each group's share of the population. This is done by dividing the value for each racial column by the total population column. Create new variables for your data frame for each calculation.

$$RaceProportion_i = \frac{Race_i}{Total_i}$$

Step 2:

Take each racial group's share of the population, square it and sum the results.

$$P(Racial_i) = \sum_{i=1}^{n} RaceProportion_i^2$$

The Census also includes a category called "Some other race." Because studies show that people who check it are overwhelmingly Hispanic, that category is not used. Hispanics' effect on diversity is calculated in Step 3.

Step 3:

Because Hispanic origin is a separate Census question, the probability that someone is Hispanic or not must be figured separately. Take the Hispanic and non-Hispanic percentages of the population, square each and add them to get the chance that any two people will be Hispanic or not. Use this calculation to create a new variable in your data frame.

$$P(Ethnic_i) = Hispanic_i^2 + NonHispanic_i^2$$

Step 4:

To calculate whether two people are the same on both measures, multiply the results of the first two steps. Use this calculation to create a new column in your data frame. This is the probability that any two people are the SAME by race and ethnicity.

$$P(Same_i) = P(Racial_i) \times P(Ethnic_i)$$

Step 5:

Subtract the result from 1 to get the chance that two people are different – diverse. For ease of use, multiply the result by 100 to place it on a scale from 0 to 100. Create a new column with your USA Today Diversity Index value.

$$DiversityIndex_i = \Big(1 - P(Same_i)\Big) \times 100$$

Geo-spatial Analysis and Visualization

Be sure to properly label your plots and axes. Points will be deducted for incorrect plot titles or axes.

6. Create a histogram of USA Today Diversity Index values. Describe the shape of the histogram in statistical terms (Hint: skewness).
7. Create a visualization which compares the top 10 counties and their diversity index value using ggplot2.
8. Using the *leaflet* mapping library for R (or another mapping library of your choice), visualize the USA Today Diversity Index by county for New York, New Jersey and Connecticut.
9. Display the following data in the "tooltip" when mousing over your plot: USA Today Diversity Index Value and County Name.
10. Does there appear to be any relationship between geography and diversity? Which state appears to be the most diverse?

Extra Credit

11. Create a new data frame using the *tidycensus* API with data on median household income by county for New York, New Jersey and Connecticut. Join this data together with the data from New York County. Use ggplot2 (or another visualization library) to visualize the USA Today Diversity Index value and median household incomeon the same plot (Hint: try facet wrap!).
12. Does there appear to be any relationship between median household income and diversity? How do counties differ on these two measures?