PhenX Toolkit

Data Collection Worksheet

Please Note: The Data Collection Worksheet (DCW) is a tool to aid integration of a PhenX protocol into a study. The PhenX DCW is not designed to be a data collection instrument. Investigators will need to decide the best way to collect data for the PhenX protocol in their study. Variables captured in the DCW, along with variable names and unique PhenX variable identifiers, are included in the PhenX Data Dictionary (DD) files.

The ACS data used in this protocol can be accessed by using Excel to read the Summary Files or using the "Download Center" at the U.S. Census Bureau's American FactFinder portal at http://factfinder.census.gov. Users can find additional information on these tools at the following locations:

Using Excel to Access Summary Files: http://www2.census.gov/programs-surveys/acs/summary_file/2014/documentation/tech_docs/ACS_SF_Excel_Import_Tool.pdf

Using the Download Center: http://www2.census.gov/programs-surveys/acs/summary_file/2014/documentation/tech_docs/How_to_Access_ACS_Estimates_AFF.pdf

The technical documentation for the ACS summary files is available online at http://www.census.gov/programs-surveys/acs/technical-documentation.html. Select the "Summary File Documentation" link, and then select the data set of interest. Users not familiar with Census data should consult the technical materials.

The key race/ethnicity data in the ACS are found in "Table B03002: Hispanic or Latino by Race." This table is preferred over other possible race and race/ethnic tables available, as it provides data on the main race/ethnic groups in the United States and explicitly incorporates data on Hispanic or Latino populations, otherwise not available in the race-only tables.

Variable Code	Variable Name
B03002001	Total:
B03002002	Not Hispanic or Latino:

B03002003	White alone
B03002004	Black or African American alone
B03002005	American Indian and Alaska Native alone
B03002006	Asian alone
B03002007	Native Hawaiian and Other Pacific Islander alone
B03002008	Some other race alone
B03002009	Two or more races:
B03002010	Two races including Some other race
B03002011	Two races excluding Some other race, and three or more races
B03002012	Hispanic or Latino:
B03002013	White alone
B03002014	Black or African American alone
B03002015	American Indian and Alaska Native alone
B03002016	Asian alone
B03002017	Native Hawaiian and Other Pacific Islander alone

B03002018	Some other race alone
B03002019	Two or more races:
B03002020	Two races including Some other race
B03002021	Two races excluding Some other race, and three or more races

The race/ethnic data are available for all small census geographies—such as census block, census block group, and census tract—and can be easily extracted for almost any geographic level. Note: Although block group data have long been available from the Census File Transfer Protocol site, the Census Bureau did not make block groups available for download at American FactFinder until the release of the 2009-2013 ACS. Information about accessing block group data for earlier years is available at http://www.census.gov/library/video/acs_block_group.html.

Researchers can use the data in this table to easily calculate basic variables (e.g., the percentage of any race and/or ethnicity group) or to combine groups (e.g., all minorities).

Unbiased Versions of S via Difference of Means Calculations

Index Score = $\bar{Y}_1 - \bar{Y}_2 = (1/N_1)\Sigma n_{1i} y_i - (1/N_2)\Sigma n_{2i} y_i$

where

n_1 and n_2 are the counts for the reference and comparison groups, respectively, in spatial unit i,

N_1 and N_2 are the counts for the reference and comparison groups, respectively, for the larger area as a whole,

y_i is a score for “scaled contact with the reference group” assigned on the basis of an index-specific function of the reference group proportion in the population of spatial unit i given by $p_i = n_{1i}/(n_{1i}+n_{2i})$, and

$\bar{Y}_1 and\ \bar{Y}_2$ are group means for “scaled contact with reference group”.

In the case of S, the functions for assigning scores on scaled contact with the reference group (y_i) based on the reference group proportion in the population of spatial unit i (p_i) is simple and easy to implement.

For S, $y_i = p_i$. Accordingly, S registers the simple group difference in average contact with the reference group.

S takes value of 0 when the two groups have identical levels of contact with the reference group. This occurs when the two groups live together in smaller areas in the same proportions seen for the larger area as a whole. S takes value of 1 when the comparison group has no contact with the reference group and the reference group has only contact with itself. This occurs when the two groups live apart in areas that are homogeneous.

These formulations of S are mathematically equivalent to the "standard" formulas for S given earlier (derivations are provided in Fossett 2017). They thus yield scores that are identical to the scores obtained using the standard formulas and thus will have the same bias components.

Obtaining Unbiased Index Scores for S

Bias is eliminated from S by calculating the value of p_i as follows:

for members of the reference group, $p_i = (n_{1i}-1)/(n_{1i}+n_{2i}-1)$, and

for members of the comparison group, $p_i = (n_{1i}-0)/(n_{1i}+n_{2i}-1)$.

The resulting adjusted values of p_i are applied as before. The values of S obtained using the adjusted values of p_i in the difference of means formula will be free of bias (Fossett 2017).[1]

The adjustment to p_i shown above removes the impact of self-contact on the value p_i. In so doing, it completely eliminates index bias at the point of initial measurement. The basis for this welcome result is simple. The expected value of contact with the reference group among neighbors (excluding the individual under consideration) is unbiased; it is the same for both groups. But the expected value of contact with the reference group based from self-contact is biased; it is always positive for members of the reference group (larger in value when counts involved are small) and always zero for members of the comparison group. Extending the Dissimilarity Index and the Separation Index: The Multigroup Analog

While much early research on segregation looked at two groups (e.g., black and white, or majority and minority), today's society is multiethnic. Two-group measures are useful but limited for describing complex patterns of segregation. The choice to use a two-group or multigroup D or S depends on the specific question of interest. In a region where the population is composed of three groups (e.g., white non-Hispanic, black non-Hispanic, and Hispanic), we may be interested in

a) segregation between two specific groups (e.g., How segregated are white from

black residents?); or

b) segregation among all three groups (e.g., How segregated are white non-Hispanic, black non-Hispanic, and Hispanic residents from each other?).

The two-group measure can still be used by comparing all possible pairs of population groups (Morrill, 1995), but these are not comprehensive, and multiple groups are not treated simultaneously. To address segregation among multiple groups requires a multigroup analog to D (Morgan et al., 1975; Sakoda, 1981). The multigroup analog describes the extent to which two or more population groups are similarly distributed among subareas. The formulas for multigroup dissimilarity (D) and multigroup separation (S) (from Reardon & Firebaugh, 2002) are:

$$D = \sum_{m=1}^{M}\sum_{j=1}^{J}\frac{t_j}{2TI}|\pi_{jm} - \pi_m| \quad S = \sum_{m=1}^{M}\sum_{j=1}^{J}\frac{t_j}{TI}(\pi_{jm} - \pi_m)^2$$

where

T is total population,

M is the number of groups m,

J is the number of subareas or units j,

t_j is number of individuals in subarea j,

π_m is the proportion in group m,

π_{jm} is the proportion in group m, of those in unit j, and

I is the Simpson's Interaction Index, given by

$$I = \sum_{m=1}^{M}\pi_m(1 - \pi_m)$$

In the Stata statistical software package, the command seg (installed by typing "ssc install seg" from within Stata) will compute both two-group and multigroup versions of S (Reardon & Firebaugh, 2002).[2]

Researchers have extended segregation measures by incorporating the spatial dimension (White, 1983; Wong, 1993; Reardon & O'Sullivan, 2004). Fossett (2017) introduces spatial formulations of S and other popular measures of uneven distribution.

Unbiased versions of multigroup indices have not been developed.

[1]There is one further adjustment. Singleton individuals—individuals who happen to be the only member of either group residing in the spatial unit, are excluded from the calculations as the adjusted calculation of p_i will be undefined for them. In practice, this is a rare occurrence.

[2] The seg program calculates S under multiple mathematically equivalent formulations including the “normalized exposure index” and the “squared coefficient of variation index”.

Protocol source: https://www.phenxtoolkit.org/protocols/view/211404