Measuring Diversity in the United States

The nation's diversity increased dramatically over the past decade ... because of a huge increase in immigrants, particularly Hispanics, in more regions of the country. There is nearly a 1 in 2 chance that two people selected at random are racially or ethnically different, according to the index. (USA Today, March 14, 2001)

MARK SCHILLING
California State University, Northridge

Diversity is a word we hear frequently these days. But what does it actually mean? Sociologically, the word is used in reference to the number and degree of representation of racial and ethnic groups in a university, a city neighborhood, and so forth. Still, the notion is somewhat vague. What do people really have in mind when they say, for example, that the United States is more diverse than it has been in the past?

In order to come up with some sort of mathematical definition of diversity, consider a population of individuals comprised of k groups that are represented in the population in proportions p_i, $i = 1, 2, \ldots, k$. A reasonable objective is to come up with some function of the p_i's that measures the extent to which the population is spread across these groups.

If the groups had an *ordinal* relationship where we could assign values to them corresponding to a numerical scale, then one possible measure would be the variance of the distribution, or equivalently its square root, the standard deviation. But data on race and ethnicity are not ordinal, so these measures make no sense here.

The definitions of race and ethnicity used by the Census Bureau are influenced greatly by self-identification and do not represent any clear-cut scientific definition of biological stock. In fact, the categories used for the decennial census have changed from census to census. This poses a challenge in comparing diversity from one census to another.

The 2000 Census used the following racial categories: White, Black or African American, American Indian and Alaska Native, Asian, Native Hawaiian and Other Pacific Islander, and "Some Other Race." In the 1990 Census the Asian and Pacific Islander groups formed one category, while the 2000 Census category American Indian and Alaska Native was comprised of three separate categories (American Indian, Eskimo and Aleut) in 1990. More significantly, the 2000 Census was the first to allow respondents to indicate that they were members of more than one race. With six individual racial categories (including "Some Other Race"), this means the census needed to allow for fifty-seven possible mixed-race categories (see if you can verify this).

The Census Bureau treats ethnicity as a separate factor from race, with only two categories—Hispanic and non-Hispanic. Thus for the 2000 Census there were $(6 + 57) \times 2 = 126$ distinct racial/ethnic combinations that an individual could conceivably be classified into.

How then can one quantify the racial and ethnic diversity of the United States population? One idea is to simply focus on the largest group size and define the measure $D_1 = 1 - \max_i p_i$ (subtracting from one makes sense because one would think of diversity as increasing when $\max_i p_i$ decreases). For the U.S. population, $\max_i p_i$ is the proportion of non-Hispanic whites in the population. Its value for the 2000 Census data is .691, so $D_1 = .309$. In a sense this indicates that 30.9% of Americans are members of a minority. The obvious weakness of this measure is that it ignores the racial and ethnic structure of this minority population.

An inverse measure, $D_2 = \min_i p_i$, could also be considered. This measure judges diversity by the *rarest* group in the population. Besides having the same sort of drawback as D_1, D_2 is fatally flawed by its dependence on how finely the population is classified into distinct racial/ethnic groups.

More elaborate methods for measuring diversity are found in ecology, where the diversity of ecosystems and of individual species is often of interest. If there are many species in an ecosystem, with no small number of species being much more abundant than the rest, then the ecosystem is highly diverse as the typical species is at least somewhat rare. Let R_i represent the rareness of species i. We will assume that R_i is defined in some way so that the more rare a species is, the larger its R value is. Then the *average* rareness of all species—or, in our application, of racial/ethnic groups

in the population—is $\sum_{i=1}^{k} p_i R_i$. This class of functions, for different definitions of rareness R_i, measures diversity.

One simple way to define the rareness of a group is as the complement of the frequency with which the group appears in the population, that is, $R_i = 1 - p_i$. This produces *Simpson's diversity index*

$$D_3 = \sum_{i=1}^{k} p_i(1-p_i).$$

This measure has the following appealing interpretation: pick two members of the population at random, then D_3 represents the probability that the two individuals are from different groups.

Another possible choice for the definition of rareness is $R_i = \ln(1/p_i)$. A group that constitutes only 1% of the population is thus rated twice as rare as one which constitutes 10%, while one that comprises just 0.1% of the population is counted as three times as rare. This definition of rareness leads to *Shannon's diversity index*

$$D_4 = -\sum_{i=1}^{k} p_i \ln p_i.$$

This quantity plays a central role in information theory, a subject whose theoretical foundations were laid by the American mathematician and electrical engineer Claude E. Shannon. Physicists and probabilists know it as *entropy*.

Both D_3 and D_4 have certain desirable properties. Each attains its minimum possible value when there is only one group and attains its maximum possible value for a given k in the case when all of the p_i are equal, i.e., when all groups occur with equal frequency. In addition, D_3 and D_4 each become larger if any group is divided into two new groups. (You may want to try to verify these assertions. Lagrange multipliers are useful for showing one part.)

Often in applications odds are used in place of probabilities. Suppose we define the rareness of a group as the *odds* that a randomly selected member of the population is from a different group, as the probability that a randomly selected member of the population is from a different group as in D_3. That is, we let $R_i = (1 - p_i)/p_i$. This yields the measure

$$D_5 = \sum_{i=1}^{k} (1-p_i) = k-1,$$

which is simply the number of groups comprising the population less one.

The table contains the raw numbers for the 2000 Census, from which the diversity measures above can be computed.

	Ethnicity	
	Not Hispanic	Hispanic
Race	% of population	% of population
W	69.1	6.0
B	12.1	0.3
AI	0.7	0.1
AS	3.6	—
NH	0.1	—
O	0.2	5.3
T	1.6	0.8

W = White, B = Black or African American, AI = American Indian and Alaska Native, AS = Asian, NH = Native Hawaiian and Other Pacific Islander, O = Some Other Race, T = Two or More Races.

There is of course no one "right" mathematical definition of diversity. One of the measures described above, however, has achieved prominence in media reports on the U.S. Census. The national newspaper *USA Today* has chosen to quantify diversity based on census data with $100D_3$, which it touts as the *USA Today* Diversity Index.

The value of this index stood at 49 in 2000, up substantially from the 1990 value of 40. You can check the 2000 calculation using the data above if you wish. Note that "Some Other Race" and "Two or More Races" are each treated as single racial categories, even though there is obviously great variation in the racial composition of these two groups. It is not hard to see (try it!) that the effect on the value of Simpson's diversity index D_3 of consolidating people of different infrequently occurring races and race combinations into these two groups is no more than $.002^2 + .016^2 + .053^2 + .008^2 = .003$. Thus the *USA Today* index would likely round to 49 with or without this grouping of uncommon races.

For the same reason, changing category definitions between 1990 and 2000 does not greatly interfere with comparing the diversity of the United States population in those two years. In the words of *USA Today*, "The nation's diversity increased dramatically over the past decade...because of a huge increase in immigrants, particularly Hispanics, in more regions of the country. There is nearly a 1 in 2 chance that two people selected at random are racially or ethnically different, according to the index." (*USA Today*, March 14, 2001) Of course, everyday encounters between individuals are not random, and the proportion of such interactions that involve people from different racial or ethnic groups is undoubtedly much less than 49%.

We have not directly addressed the question of why *USA Today* chose to use Simpson's diversity index rather than, say, Shannon's. I will leave it as a challenge for you to compare the stability of these two measures as a small group is divided into smaller groups. For instance, suppose that at least one of the Census Bureau's "Some Other Race" and "Two or More Races" categories is split into some number of subgroups. We noted above that the effect on Simpson's index would not be great. Can the same be said for Shannon's index?

Endnote

There was one other difference between the censuses of 1990 and 2000 that is not mentioned above. The order of the questions on race and Hispanic origin was different, with the one on Hispanic origin placed first in 2000. It is conceivable that many more respondents may have identified themselves as Hispanics in 2000 than if the ethnicity question had remained *after* the question on race as in 1990. Hence conclusions about the large increase in diversity from 1990 to 2000 should be drawn with some measure of caution. ■