10

STATISTICAL ANALYSIS OF QUALITATIVE VARIATION

Alan Agresti
UNIVERSITY OF FLORIDA

Barbara F. Agresti
UNIVERSITY OF FLORIDA

Many variables of interest in the social sciences are measurable only at the nominal level. That is, they represent types of phenomena such as race, ethnicity, religious affiliation, or political party preference. It is sometimes of interest to measure the amount of variation, or heterogeneity, within a population with respect to one or more of these variables. By a measure of variation for a qualitative variable, we mean a description of the dispersion of the population over a number of nominal categories. We shall refer to such measures as indices of qualitative variation or diversity.

The ecological sciences have made considerable use of

such measures to determine the species diversity in a geographical or spatial area. Rex (1973) studied the diversity of gastropod species at deep-sea levels, for example, and Sargent and Owen (1975) studied similar diversity among moth wing patterns. (See also Pielou, 1969, Chap. 18, for a discussion of this type of application.) Identical measures have been developed in linguistics for estimating the heterogeneity of populations with respect to the native languages of their members (Greenberg, 1956; Lieberson, 1964).

In addition, sociological uses of measures of diversity have appeared—two primary references are Lieberson (1969) and Mueller, Schuessler, and Costner (1977). These measures have been applied mainly to the degree of division of labor of a society. (Gibbs and Martin, 1962; Grandjean, 1974; Frisbie, 1975; and Labovitz and Gibbs, 1964, represent a few examples of this application.) Amemiya (1963) has employed the same type of measure to describe "economic differentiation." These measures could also be applied to racial or ethnic diversity, religious diversity, diversity in political party preference, or diversity with respect to other qualitative variables. As far as we can tell, they have not been widely used in sociology, except in measuring the division of labor based on occupational diversity. Since much of sociology deals with qualitative variables, the measures to be discussed in this chapter might have a broader range of application.

We shall describe two of the most important measures of qualitative variation: (1) Simpson's index of diversity and (2) Mueller and Schuessler's index of qualitative variation. In the next section we define these indices, discuss some of their basic properties, and illustrate their use. The sampling distributions are given in the following section along with examples of large-sample confidence intervals and tests of hypotheses. Then statistical methods are presented for comparing two measures of qualitative variation and for measuring diversity "between groups." Finally, we offer some generalizations concerning the measurement and statistical analysis of qualitative variation for a cross-classification of two or more nominal variables. Derivations of the sampling distributions used in the inferential analyses are given in three appendices at the end of the chapter.

MEASURING QUALITATIVE VARIATION

A commonly used measure of qualitative variation was developed by Simpson (1949). This "index of diversity" has been widely used in the ecological sciences in studies of species variation. (See, for example, Sargent and Owen, 1975, and Rex, 1973.) Simpson's index is apparently identical to one first proposed by Corrado Gini in 1912 (Lieberson, 1969, p. 815). The same measure of diversity has also been "discovered" by others, including Greenberg (1956), who calls it the index of linguistic diversity A. The same index is referred to as A_w (Lieberson, 1969) and $M1$ (Gibbs and Martin, 1962) in some sociological applications.

In order to measure diversity according to this method, one must first classify the individual observations into a number of categories, say k. One then computes the value

$$D = 1 - \sum_{i=1}^{k} p_i^2 \tag{1}$$

where p_i is the proportion of observations in the ith category ($i = 1, \ldots, k$). This index gives us the probability that two individuals selected at random from the population would be in different categories. This interpretation strictly applies if the population size is infinite or if the sampling is done with replacement. In that case, an index of $D = 0.6$ for a classification of ethnicity means that the probability is 0.6 that two randomly selected individuals have different ethnicity. Conversely, the probability of selecting two individuals from the same ethnic category would be 0.4. The larger the number of categories and the more uniformly dispersed the observations over the categories, the higher this index of diversity tends to be.

To be precise about what is meant by "diversity" as measured by this index, we now consider a few of the basic properties of D. These can be easily noted from its definition in Formula (1). First, the minimum possible value of D is zero, which occurs if and only if some $p_i = 1$, and hence $p_j = 0$ for $j \neq i$. That is, the least diverse population is one in which all members of the population are classified in just one category. Second, for a fixed number of categories k, the maximum possible value of the diversity

index is $D = 1 - 1/k = (k - 1)/k$, which occurs if and only if $p_1 = p_2 = \cdots = p_k = 1/k$. That is, for a nominal classification with k levels, the most diverse population is one that is evenly spread over the k categories.

A third property is concerned with how the value of D relates to the relative heterogeneity of the proportions. Let D be the value of the index of diversity corresponding to the proportions $(p_1, \ldots, p_k)$ for some population, and let D' be the value of the index of diversity corresponding to the proportions $(p'_1, \ldots, p'_l, p_{l+1}, \ldots, p_k)$ for some other population. That is, the two distributions are identical except for l of the categories. Now let

$$\bar{p} = \sum_{i=1}^{l} p_i/l = \sum_{i=1}^{l} p'_i/l$$

and suppose that

$$|p_i - \bar{p}| \leq |p'_i - \bar{p}| \qquad i = 1, 2, \ldots, l$$

That is, the proportions $(p_1, \ldots, p_l)$ are each closer to their average than the corresponding proportions $(p'_1, \ldots, p'_l)$. Then it is easily seen that this implies that $D \geq D'$, with $D > D'$ if some of the inequalities are strict. For example, $D > D'$ when $(p_1, p_2, p_3, p_4) = (0.2, 0.2, 0.3, 0.3)$ and $(p'_1, p'_2, p_3, p_4) = (0.1, 0.3, 0.3, 0.3)$. As a special case, for $l = k$, it follows that $D \geq D'$ if

$$|p_i - 1/k| \leq |p'_i - 1/k| \qquad i = 1, \ldots, k$$

That is, one population is more diverse than another, according to this index of diversity, if each proportion for that population is at least as close to the overall average $1/k$ as the corresponding proportion for the other population.

Notice that the upper limit for D of $(k - 1)/k$ approaches 1 asymptotically as the number of categories k increases. In other words, the potential diversity increases as the number of levels in the classification increases. A community with residents falling in six ethnic categories has a greater potential for diversity than a community with residents falling into four such categories. For additional properties of D, see Bhargava and Doyle (1974) and Bhargava and Uppuluri (1975).

In some situations, a researcher might prefer to use a mea-

sure that can attain an upper limit of 1 no matter what the number of categories. A standardized version of the diversity index D is given by

$$I = \left(1 - \sum_{i=1}^{k} p_i^2\right) / (1 - 1/k) \qquad (2)$$

$$= [k/(k-1)]\, D$$

That is, we can control for the number of categories in the classification by dividing D by its maximum possible value: $(k-1)/k$. This measure has the same properties as those just discussed for D, except that the maximum possible value is 1, which occurs if and only if $p_i = 1/k$, $i = 1, \ldots, k$. The measure I was apparently first introduced by Mueller and Schuessler (1961, pp. 171–179) and is usually referred to as the index of qualitative variation (IQV). It was also used as an index of economic diversity by Amemiya (1963) and by Labovitz and Gibbs (1964) in measuring division of labor.

Advantages and disadvantages of standardizing for the number of categories are discussed in Lieberson (1969, pp. 860–861) and in Mueller, Schuessler, and Costner (1977, p. 177). If comparisons are made between groups with the same number of categories, then either measure is appropriate, since in each case I is the same constant multiple of D. If the groups have different numbers of categories, however, and we believe that a larger number of categories contributes to greater diversity, then we would wish to use the unstandardized index D. Otherwise we would conclude, for example, that a population with proportions (0.5,0.5), for which $I = 1$, is more diverse than one with proportions (0.2,0.2,0.2,0.15,0.25), for which $I = 0.99$. Basically D is a function of both the number of categories *and* the dispersion of the population among the categories, whereas I is just a measure of the dispersion of the population among the categories, whatever they are.

An alternative approach to measuring population diversity is based on the so-called information index, $\sum_{i=1}^{k} p_i \log p_i$, originally developed for use in communications theory. Although its properties are considered especially useful by ecologists, it is not quite so easy to interpret as D or I. The interested reader is

referred to Good (1953), Renyi (1961), and the works of Pielou (1967; 1969, pp. 224–233) for details of this index.

Given a sample of n observations from some population classification, one could compute the sample analogs of the diversity measures,

$$\hat{D} = 1 - \sum_{i=1}^{k} \hat{p}_i^2 \tag{3}$$

and

$$\hat{I} = [k/(k-1)]\left(1 - \sum_{i=1}^{k} \hat{p}_i^2\right) \tag{4}$$

where the $\hat{p}_i$ are the sample proportions of observations in these categories. Through simple algebra, it can be seen that $\hat{D}$ and $\hat{I}$ are related to the statistic used in the well-known chi-square goodness-of-fit test when the null hypothesis is

$$H_0: \quad p_1 = p_2 = \cdots = p_k = 1/k$$

For example, if χ^2 denotes the value of that statistic in the test for this particular null hypothesis, then

$$\hat{I} = 1 - \chi^2/n(k-1) \tag{5}$$

and

$$\hat{D} = 1 - (1/k)(\chi^2/n + 1) \tag{6}$$

Thus $\hat{I}$ ($\hat{D}$) is a function of the χ^2 statistic that falls between 0 and $1[(k-1)/k]$, equaling the upper limit in the extreme case when $\chi^2 = 0$ (that is, when each sample proportion $\hat{p}_i$ equals the hypothesized value $p_i = 1/k$). In fact, Weiler (1966) proposed an "index of discrepancy" designed to measure the deviation of an actual population from the population as stated in the null hypothesis of the goodness-of-fit test. In the special case in which the null hypothesis is that of equal proportions, Weiler's measure ϕ^2 reduces to

$$\phi^2 = \chi^2/n(k-1) = 1 - \hat{I} \tag{7}$$

That is, a high index of discrepancy corresponds to a low amount of diversity in the population. This complement of the index of

qualitative variation is also shown by Mueller, Schuessler, and Costner (1977, p. 180), to equal the ratio of the variance of the k frequencies about their mean to the maximum possible such variance.

To illustrate the calculation of these measures, we have taken some data from Agresti (1976) that describe the major occupational-class breakdown of Walton County, a farming county in Florida's panhandle, in the years 1870 and 1885. The data came from samples of census manuscripts for those years.[1] Table 1 shows the sample distributions in the classifications for

TABLE 1
Occupational Status by Race and Year in Walton County, Florida

Occupational Status	White 1870	White 1885	Black 1870	Black 1885
Professional	0.019	0.099	0.007	0.000
Manager, clerical, proprietor	0.029	0.093	0.000	0.000
Skilled	0.086	0.040	0.007	0.000
Unskilled	0.053	0.073	0.046	0.049
Laborer	0.455	0.517	0.776	0.896
Farmer	0.359	0.179	0.164	0.056
Sample size	209	151	152	144

the four samples: white 1870, white 1885, black 1870, black 1885. Records for the entire black population were obtained in each year, whereas a systematic random sample of records for the white households yielded the sample of white individuals with occupations described in Table 1.

To compute the index of diversity, we find for the white population in 1870 that

$$\sum_{i=1}^{6} \hat{p}_i^2 = 0.019^2 + 0.029^2 + 0.086^2 + 0.053^2 + 0.455^2 + 0.359^2 = 0.347$$

Then the index of diversity is $\hat{D} = 1 - 0.347$, or 0.653. This means that the probability of a randomly selected pair of individuals coming from different occupational levels, for the white sample in 1870, is 0.653. Similarly, one finds that the indices of

[1] Florida held a special census in 1885.

diversity for the white sample in 1885, the black population in 1870, and the black population in 1885 are 0.675, 0.368, and 0.192, respectively.

The maximum possible value for $\hat{D}$ when there are $k = 6$ categories is $1 - 1/k = 1 - 1/6 = 0.833$. Thus the standardized value of $\hat{D}$ for the 1870 white sample, which is the index of qualitative variation, is

$$\hat{I} = 0.653/0.833 = 0.784$$

Similarly, the values of the index of qualitative variation for the white sample of 1885, the black population of 1870, and the black population of 1885, respectively, are 0.811, 0.442, and 0.230. Notice that the black population became much less diverse in occupational distribution while the white population stayed at about the same level of diversity over this 15-year period. In addition, in each year, the black population appears to have been much less heterogeneous than the white with respect to occupations.

If one knows, theoretically, the total set of categories for a variable, then that total set should be used as the basis for standardizing the D index, even though there may be no observations in some of the categories. For example, even though the entire black population of 1885 was concentrated in just three categories, the maximum possible diversity for this classification corresponds to a uniform distribution over the six categories of occupational status, so $k = 6$. In the sampling problems we consider next, we shall assume that k is known; that is, before obtaining the sample, the nominal classification is well defined. If k is unknown and the number of categories sampled is naturally treated as a random variable (as is often the case in ecological problems, where the total number of species may be unknown), the following sampling theory is not appropriate.

STATISTICAL INFERENCE

In this section we explain how to make the standard statistical inferences about the values of the diversity index D and its standardized value, the index of qualitative variation I. We shall suppose that a random sample of n measurements is selected from

the population of interest. It can be shown that $\hat{D}$ and $\hat{I}$, as defined in Formulas (3) and (4), are just slightly biased estimates of D and I and that unbiased estimates (Good, 1953) are $[n/(n-1)]\hat{D}$ and $[n/(n-1)]\hat{I}$. The fact that $\hat{D}$ and $\hat{I}$ are slightly biased need not concern us, since we are dealing with large-sample approximate sampling distributions for these indices.

The exact variance of $\hat{D}$ was given by Simpson (1949). For large samples, this variance is approximately $\sigma_D^2 = \sigma^2/n$, where σ^2 can be estimated from the sample observations by

$$\hat{\sigma}^2 = 4\left[\sum_{i=1}^{k} \hat{p}_i^3 - \left(\sum_{i=1}^{k} \hat{p}_i^2\right)^2\right] \tag{8}$$

Suppose, moreover, that $0 < D < (k-1)/k$, corresponding to $0 < I < 1$. Then, from standard distribution-theory arguments, it follows (see Appendix A) that for large sample sizes $\sqrt{n}(\hat{D} - D)/\hat{\sigma}$ has approximately the standard normal distribution. This result is also reported in a paper by Van Belle and Ahmad (1974).

Therefore a large-sample $100(1-\alpha)$ percent confidence interval for D, when $0 < \hat{D} < k/(k-1)$, is given by the interval

$$\hat{D} \pm z_{\alpha/2}\hat{\sigma}/\sqrt{n} \tag{9}$$

where $z_{\alpha/2}$ is the z value corresponding to the $\alpha/2$ and $1-\alpha/2$ quantiles of the standard normal distribution. In calling this a large-sample $100(1-\alpha)$ percent confidence interval, we mean that the probability that this interval includes D converges to $1-\alpha$ as the sample size increases indefinitely. Alternatively, if one wished to test the null hypothesis H_0: $D = D_0$, where $0 < D_0 < (k-1)/k$, one could use the test statistic

$$z = \sqrt{n}(\hat{D} - D_0)/\hat{\sigma} \tag{10}$$

which has the standard normal distribution under H_0, for large sample sizes. Conducting this test with alternative hypothesis H_a: $D \neq D_0$ at the α significance level is equivalent to noting whether the number D_0 falls within the $100(1-\alpha)$ percent confidence interval for D.

The sample size n that is needed for the standard normal

approximation for $\sqrt{n}(\hat{D} - D)/\hat{\sigma}$ to be adequate has not been investigated. We would conjecture that the convergence of the sampling distribution to the standard normal would be relatively rapid, unless one of the p_i is very close to 1 (in which case $\hat{D}$ could equal zero with moderate probability even for large n). The small-sample distributions of some indices of diversity have been investigated by Bowman and others (1971); though primarily for the case $p_1 = \cdots = p_k$.

To illustrate the foregoing discussion, we return to the data presented in Table 1. For the white sample in 1870, we noted that the index of diversity was $\hat{D} = 0.653$. For the sample[2] proportions given in the table,

$$\sum_{i=1}^{k} \hat{p}_i^3 = 0.141 \qquad \sum_{i=1}^{k} \hat{p}_i^2 = 0.347$$

and thus

$$\hat{\sigma}^2 = 4[0.141 - (0.347)^2] = 0.083$$

$$\hat{\sigma} = 0.288$$

A 95 percent confidence interval for D is

$$\hat{D} \pm 1.96\hat{\sigma}/\sqrt{n} = 0.653 \pm 1.96(0.288)/\sqrt{209}$$

$$= 0.653 \pm 0.039 = (0.614,0.692)$$

Similarly, using the sample proportions for the white population in 1885, one attains $\hat{\sigma}^2 = 0.163$, so that a 95 percent confidence interval for the index of diversity for that group is $0.675 \pm 0.064 = (0.611,0.739)$.

Occupational data were obtained for the entire black population in 1870. For that group, the index of diversity was 0.368. To test whether the 1870 white population had the same diversity as the black population, we could test the null hypothesis H_0: $D = 0.368$ against H_a: $D \neq 0.368$, where D represents the index of diversity for the entire white population of

[2]For simplicity of exposition, we are omitting the use of a finite population correction, although 20 percent of the white population was sampled in this example.

Walton County in 1870. The test statistic is

$$z = \sqrt{209}(0.653 - 0.368)/(0.288) = 14.31$$

Thus we would feel very confident in concluding that the white population enjoyed a greater degree of occupational diversity than the black population in 1870. A similar conclusion applies to 1885.

Since the index of qualitative variation I is related to D by $I = Dk/(k - 1)$, inference procedures for I follow directly from those for D. For example, a $100(1 - \alpha)$ percent confidence interval for I is

$$\hat{I} \pm z_{\alpha/2}[k/(k - 1)]\hat{\sigma}/\sqrt{n} \tag{11}$$

The hypothesis H_0: $I = I_0$ can be tested by using the test statistic

$$z = \frac{\sqrt{n}(\hat{I} - I_0)}{[k/(k - 1)]\hat{\sigma}} \tag{12}$$

For the 1870 white population $\hat{I} = 0.784$, and a 95 percent confidence interval for I is (0.737,0.831), which is simply the interval for D multiplied by $k/(k - 1) = 1.2$. The test of H_0: $I = I_0$ yields the same z value as that of H_0: $D = D_0$, where $I_0 = D_0k/(k - 1)$ is the hypothesized value of the index of qualitative variation corresponding to the hypothesized value D_0 of the diversity index.

Notice that the normal sampling distributions of $\hat{D}$ and $\hat{I}$ discussed in this section apply when $0 < D < (k - 1)/k$, or $0 < I < 1$. If $D = 0$ $(I = 0)$, then $p_i = 1$ for some i, so $\hat{p}_i = 1$ with probability 1 for all $n \geq 1$ for that category i, and hence $\hat{D} = \hat{I} = 0$ with probability 1. If one observes a $\hat{D} > 0$ in some sample, then of course one can reject H_0: $D = 0$, with a zero probability of a type I error. If $D > 0$, the probability that $\hat{D} = 0$ converges to zero as the sample size increases.

Now if $D = (k - 1)/k (I = 1)$, $\hat{D}$ could not be normally distributed about D, since $\hat{D} \leq D$ $(\hat{I} \leq I)$ with probability 1. In this extreme case, $p_1 = p_2 = \cdots = p_k = 1/k$. And since $\chi^2 = n[k(1 - \hat{D}) - 1]$ is the statistic used in the chi-square goodness-of-fit test of H_0: $p_1 = p_2 = \cdots = p_k = 1/k$, it follows that for a large sample size

$$n[k(1 - \hat{D}) - 1] = n(k - 1)(1 - \hat{I}) \tag{13}$$

has approximately the chi-square distribution with $k - 1$ degrees of freedom. Thus one can test the hypothesis H_0: $D = (k - 1)/k (I = 1)$ of greatest possible diversity using this test statistic. Again, using the data from the white population in 1870 to illustrate, the test statistic for testing H_0: $\hat{D} = 0.833$ $(I = 1)$ is

$$n[k(1 - \hat{D}) - 1] = 209[6(1 - 0.653) - 1] = 225.85$$

which is based on 5 degrees of freedom. This statistic indicates that the white population of 1870 did not have maximum heterogeneity with respect to occupation. Of course, in nearly all practical applications we would not expect the population to exhibit the maximum or the minimum diversity. It would usually be much more informative to compute a confidence interval for the value of D or I.

QUALITATIVE VARIATION FOR TWO GROUPS

In many situations, it is of interest to compare D or I values for two or more populations. We might wish to compare the diversity on occupational status of two groups in some community, or the diversity of the occupational structures of two communities, or the diversity of the same community at two points in time. If independent random samples are chosen from the populations of interest, then confidence intervals and tests of hypotheses for the difference between D values or I values for two populations can be obtained using the sampling variances given in the previous section.

Suppose that a random sample of size n_1 is selected from the first population and that these observations are classified according to a scheme with k_1 categories. Denote the population proportions for the k_1 categories by $\{p_i,\ 1 \leq i \leq k_1\}$ and the index of diversity by D_1. Now suppose that another random sample of size n_2 is selected from the second population and classified into a set of k_2 categories. The corresponding proportions and diversity index for the second population are denoted by $\{q_i,\ 1 \leq i \leq k_2\}$ and D_2. Now if n_1 and n_2 are relatively large, it follows from the previous section that a $100(1 - \alpha)$ percent con-

fidence interval for the difference $D_2 - D_1$ between the two population diversity indices is

$$(\hat{D}_2 - \hat{D}_1) \pm z_{\alpha/2}\sqrt{(\hat{\sigma}_1^2/n_1) + (\hat{\sigma}_2^2/n_2)} \tag{14}$$

where

$$\hat{\sigma}_1^2 = 4\left[\sum_{i=1}^{k_1} \hat{p}_i^3 - \left(\sum_{i=1}^{k_1} \hat{p}_i^2\right)^2\right] \qquad \hat{\sigma}_2^2 = 4\left[\sum_{i=1}^{k_2} \hat{q}_i^3 - \left(\sum_{i=1}^{k_2} \hat{q}_i^2\right)^2\right] \tag{15}$$

The corresponding $100(1 - \alpha)$ percent confidence interval for the difference $I_2 - I_1$ between the two indices of qualitative variation is

$$(\hat{I}_2 - \hat{I}_1) \pm z_{\alpha/2}\sqrt{\left(\frac{k_1}{k_1 - 1}\right)^2 \frac{\hat{\sigma}_1^2}{n_1} + \left(\frac{k_2}{k_2 - 1}\right)^2 \frac{\hat{\sigma}_2^2}{n_2}} \tag{16}$$

If interest focuses primarily on testing the hypothesis H_0: $D_1 = D_2$ of identical diversity for the two groups, one could use the test statistic

$$z = \frac{\hat{D}_1 - \hat{D}_2}{\sqrt{(\hat{\sigma}_1^2/n_1) + (\hat{\sigma}_2^2/n_2)}} \tag{17}$$

which has approximately the standard normal distribution under H_0. To test the hypothesis H_0: $D_2 - D_1 = c$, the numerator of the z test statistic would be replaced by $(\hat{D}_2 - \hat{D}_1) - c$. Notice that H_0 will be rejected in favor of H_a: $D_2 - D_1 \neq c$ at the α significance level if c does not fall in the $100(1 - \alpha)$ percent confidence interval for $D_2 - D_1$. If there are s samples ($s \geq 2$), one could test H_0: $D_1 = D_2 = \cdots = D_s$ using the test statistic

$$\chi^2 = \sum_{i=1}^{s} [n_i(\hat{D}_i - \bar{D})^2/\hat{\sigma}_i^2] \tag{18}$$

which has approximately the chi-square distribution with $s - 1$ degrees of freedom under H_0 if all s sample sizes are large, where

$$\bar{D} = \left[\sum_{i=1}^{s} n_i\hat{D}_i/\hat{\sigma}_i^2\right] \Big/ \left[\sum_{i=1}^{s} n_i/\hat{\sigma}_i^2\right] \tag{19}$$

The reasoning for this test (see Appendix A) is similar to that given by Goodman and Kruskal (1963, p. 318) for a several-

sample test for measures of association. The analogous tests about I values yield the exact same values of the test statistics if the number of categories in each classification is the same. Again, these inference procedures are appropriate as long as the indices are not equal to their boundary values.

To illustrate these procedures, let us return to the data in Table 1 and compare the diversity indices D_1 and D_2 of the white population of Walton County in 1870 and 1885. The samples were independently selected from the two populations. We have seen that $\hat{D}_1 = 0.653$ and $\hat{D}_2 = 0.676$ and that $\hat{\sigma}_1^2 = 0.083$ and $\hat{\sigma}_2^2 = 0.163$. Thus a 95 percent confidence interval for the change in diversity of the white occupational structure is

$$(0.676 - 0.653) \pm 1.96\sqrt{(0.083/209) + (0.163/151)}$$

$$= 0.023 \pm 0.075 = (-0.052, 0.098)$$

The corresponding 95 percent confidence interval for the change in the index of qualitative variation $I_2 - I_1$ is

$$(0.811 - 0.784) \pm 1.96\sqrt{\left(\frac{6}{5}\right)^2\left(\frac{0.083}{209}\right) + \left(\frac{6}{5}\right)^2\left(\frac{0.163}{151}\right)}$$

$$= 0.027 \pm 0.090 = (-0.063, 0.117)$$

Thus, for these sample sizes, there is insufficient evidence to conclude that there was a change in occupational diversity from 1870 to 1885 for the white population in Walton County.

In examples such as this one, a hypothesis test might be employed for testing for a change in D or for testing whether the change is different from the observed change for the black population ($D_2 - D_1 = 0.192 - 0.368 = -0.176$). For these data, we would not reject the null hypothesis of no change in diversity for the white population using the usual significance levels, but we would reject the null hypothesis that the change in diversity for the white population is the same as the corresponding change for the black population. Even if the white population became less diverse in occupational status from 1870 to 1885, it seems unlikely that the decrease would be as great as that for the black population.

This simple approach to making inferences about differences in qualitative variation is completely unlike the one

recently proposed by Swanson (1976). He appears to have been unaware of the work that has been done on the sampling variances of these measures (Simpson, 1949; Good, 1953; Van Belle and Ahmad, 1974), and the test statistic he proposes is invalid.[3]

A somewhat different way to measure diversity for two groups involves measuring diversity "between groups." Lieberson (1969) defined the diversity between two populations as being the probability, when one member is randomly selected from each population, that the two members are classified differently on the variable of interest. Now if the same classification scheme of k categories is used for each of the two populations, then the probability that the two individuals are in the same category is $\sum_{i=1}^{k} p_i q_i$, where $\{p_i\}$ and $\{q_i\}$ are the sets of population proportions for the two groups. Thus the index of between-groups diversity suggested by Lieberson is just

$$D_b = 1 - \sum_{i=1}^{k} p_i q_i \tag{20}$$

A related index, $1 - D_b$, was used by Holgate (1971) to study the drift over time in the proportion of isonymous marriages—that is, marriages between individuals having the same surname. In that case, the two populations are the collections of men and women in a given generation.

The following properties of the measure D_b are easily observed. First, the two groups are the most diverse with respect to each other if $p_i = 0$ whenever $q_i > 0$ and $q_i = 0$ whenever $p_i > 0$—that is, if the categories into which the observations in the first population fall are mutually exclusive of the categories into which the observations from the second population fall. For this extreme case, $D_b = 1$. The two groups are least diverse relative to each other when $p_i = q_i = 1$ for some category i—that is, when all the members in both the populations are classified in the same category. In that extreme case, $D_b = 0$. Notice that when $p_i = q_i$, $i = 1, \ldots, k$, then

$$D_b = 1 - \sum_{i=1}^{k} p_i q_i = 1 - \sum_{i=1}^{k} p_i^2 = 1 - \sum_{i=1}^{k} q_i^2$$

[3]Among other difficulties, Swanson's assumption that the distribution of the expanded binomial random variable is normal is not even approximately true when the number of categories is small.

so that $D_b = D_1 = D_2$. In other words, when each population has exactly the same distribution on the nominal classification of interest, then the "within-group" indices of diversity D_1 and D_2 are identical to the "between-groups" diversity D_b. In particular, $p_i = q_i = 1$ for some category i leads to $D_1 = D_2 = D_b = 0$, and $p_i = q_i = 1/k$ for $1 \leq i \leq k$ leads to $D_1 = D_2 = D_b = 1 - 1/k$. Notice, however, that the fact that $D_1 = D_2$ is not enough by itself to imply that D_b has the same value (as when $p_i = 1$, $q_j = 1$ $(i \neq j)$, so that $D_1 = D_2 = 0$, but $D_b = 1$). Also, unlike the within-group index of diversity D, the measure D_b can attain an upper value of 1, so there is no need to define a standardized version.

Using similar arguments (see Appendix B) to those that led to the sampling distribution of $\hat{D}$ discussed in the previous section, we can show that $\sqrt{n_1 + n_2}(\hat{D}_b - D_b)/\hat{\sigma}_b$ has approximately the standard normal distribution for large samples n_1 and n_2, when $0 < D_b < 1$, where

$$\hat{D}_b = 1 - \sum_{i=1}^{k} \hat{p}_i \hat{q}_i \tag{21}$$

and

$$\hat{\sigma}_b^2 = \left(\frac{n_1 + n_2}{n_1}\right) \sum_{i=1}^{k} \hat{p}_i \hat{q}_i^2 + \left(\frac{n_1 + n_2}{n_2}\right) \sum_{i=1}^{k} \hat{q}_i \hat{p}_i^2 - \frac{(n_1 + n_2)^2}{n_1 n_2} \left(\sum_{i=1}^{k} \hat{p}_i \hat{q}_i\right)^2 \tag{22}$$

It follows that a large-sample $100(1 - \alpha)$ percent confidence interval for D_b is

$$D_b \pm Z_{\alpha/2} \sqrt{\frac{\sum_{i=1}^{k} \hat{p}_i \hat{q}_i^2}{n_1} + \frac{\sum_{i=1}^{k} \hat{q}_i \hat{p}_i^2}{n_2} - \frac{(n_1 + n_2)(1 - \hat{D}_b)^2}{n_1 n_2}} \tag{23}$$

Furthermore, the null hypothesis H_0: $D_b = D_b^{(0)} (0 < D_b^{(0)} < 1)$ can be tested using the test statistic

$$Z = \sqrt{n_1 + n_2}(\hat{D}_b - D_b^{(0)})/\hat{\sigma}_b \tag{24}$$

for large sample sizes. Clearly, if $D_b = 1$, then $\hat{D}_b = 1$ with probability 1; and if $D_b = 0$, then $\hat{D}_b = 0$ with probability 1. Thus one would reject H_0: $D_b = 1$ whenever $\hat{D}_b < 1$ and reject H_0: $D_b = 0$

whenever $\hat{D}_b > 0$. The probability of a type I error would be zero in each case.

To illustrate these formulas, let us consider the between-groups diversity for the white population in 1870 relative to that in 1885. Using the values of $\{\hat{p}_i\}$ and $\{\hat{q}_i\}$ calculated when $\hat{D}_1$ and $\hat{D}_2$ were obtained, we have

$$\sum_{i=1}^{k} \hat{p}_i\hat{q}_i = (0.019)(0.099) + \cdots + (0.359)(0.179) = 0.311$$

so that $\hat{D}_b = 0.689$. That is, the probability is 0.689 that two individuals selected at random, one from the white sample of 1870 and one from the white sample of 1885, will be in different occupational classes. Also $n_1 = 209$, $n_2 = 151$, $\sum_{i=1}^{k} \hat{p}_i\hat{q}_i^2 = 0.134$, and $\sum_{i=1}^{k} \hat{q}_i\hat{p}_i^2 = 0.130$, so that a 95 percent confidence interval for D_b is

$$0.689 \pm 1.96\sqrt{\frac{0.134}{209} + \frac{0.130}{151} - \frac{(360)(0.311)^2}{(209)(151)}}$$

$$= 0.689 \pm 0.039 = (0.650, 0.728)$$

One might note that the true value of the between-groups diversity index as calculated for the black population of Walton County in 1870 and 1885 is

$$1 - [(0.007)(0.000) + \cdots + (0.164)(0.056)] = 0.293$$

which reflects the very high concentration of blacks in the laborer category at both times. It is again clear that the measure for the white populations is different from the known value of 0.293 for the black population, although in some examples one might wish to compute the Z test statistic to report the attained significance level at which such a conclusion holds.

In some situations, it may be of interest to compare the diversity index D for a group to the between-groups diversity index D_b for that group and some other group. For a particular classification, for example, one might wish to compare the probability that two members selected at random from the group are in different categories to the corresponding probability that a pair of individuals selected at random from the two groups (one from each) are in different categories. Lieberson (1969, p. 852)

mentions this type of comparison as a means of analyzing political cleavage. He notes that one could compare the political cohesion within each socioeconomic group with the political bonds existing between socioeconomic groups. Another natural application arises in comparing language diversity within a country to language diversity between countries.

As a statistical description of such a difference, one could calculate a confidence interval for $D_1 - D_b$, or $D_2 - D_b$, where D_1 and D_2 are the within-group indices of diversity for the two groups. Using the same notation as applied previously, a large-sample $100(1 - \alpha)$ percent confidence interval for

$$D_1 - D_b = \left(1 - \sum_{i=1}^{k} p_i^2\right) - \left(1 - \sum_{i=1}^{k} p_i q_i\right) = \sum_{i=1}^{k} p_i(q_i - p_i) \tag{25}$$

is (see Appendix C)

$$(\hat{D}_1 - \hat{D}_b) \pm \mathcal{Z}_{\alpha/2} \cdot \sqrt{\frac{\sum_{i=1}^{k} \hat{p}_i(\hat{q}_i - 2\hat{p}_i)^2 - [2\hat{D}_1 - (1 + \hat{D}_b)]^2}{n_1} + \frac{\sum_{i=1}^{k} \hat{q}_i\hat{p}_i^2 - [1 - \hat{D}_b]^2}{n_2}} \tag{26}$$

The $\mathcal{Z}$ test statistic for testing H_0: $D_1 = D_b$ is just the sample difference $(\hat{D}_1 - \hat{D}_b)$ divided by its estimated standard error, which is the square root term in Formula (26). These results are appropriate if $0 < D_1 < (k - 1)/k$ or $0 < D_b < 1$ and the two samples are random and independently obtained.[4] The analogous confidence interval and test for $D_2 - D_b$ has the same form with the $\{\hat{p}_i\}$ and $\{\hat{q}_i\}$, $\hat{D}_1$ and $\hat{D}_2$, and n_1 and n_2 interchanged.

We shall again use the white populations in 1870 and 1885 to illustrate the computations. Now $\hat{D}_1 = 0.653$ and $\hat{D}_b = 0.689$, and $\sum_{i=1}^{k} \hat{p}_i(\hat{q}_i - 2\hat{p}_i)^2 = 0.176$, $\sum_{i=1}^{k} \hat{q}_i\hat{p}_i^2 = 0.130$, and $\sum_{i=1}^{k} \hat{p}_i\hat{q}_i = 0.311$. Thus a 95 percent confidence interval for $D_b - D_1$ is

[4] Notice, though, that $\hat{D}_1$ and $\hat{D}_b$ are not statistically independent.

$$(0.689 - 0.653) \pm 1.96$$

$$\sqrt{\frac{0.176 - [2(0.653) - (1.689)]^2}{209} + \frac{0.130 - (0.311)^2}{151}}$$

$$= 0.036 \pm 0.037 = (-0.001, 0.073)$$

Since $\hat{D}_2 = 0.676$, $\sum_{i=1}^{k} \hat{q}_i(\hat{p}_i - 2\hat{q}_i)^2 = 0.179$, and $\sum_{i=1}^{k} \hat{p}_i\hat{q}_i^2 = 0.134$, one can verify that a 95 percent confidence interval for $D_b - D_2$ is $0.013 \pm 0.049 = (-0.036, 0.062)$. That is, two white individuals chosen at random from Walton County, one each in 1870 and 1885, are not necessarily more likely to have different occupational levels than two white individuals both chosen at random in 1870 or both chosen at random in 1885 (because the intervals contain zero).

MULTIVARIATE GENERALIZATIONS

Lieberson (1969) has introduced generalizations of the index of diversity D and the between-groups index of diversity D_b for the situation in which the populations are cross-classified according to two or more nominal variables. In this section we introduce these indices, describe their properties, give an example of their computation, and illustrate large-sample confidence intervals for their values.

Lieberson (1969, p. 853) defines the multivariate extension of the index of diversity D to be the "average proportion of disagreement between pairs on the characteristics under study" when two observations are randomly chosen. To be more precise about the meaning of this index, we must introduce additional notation. Suppose we are considering the cross-classification of m variables and the lth of these variables has k_l levels, $l = 1, 2, \ldots, m$. Then there are $k_1 \times k_2 \times \cdots \times k_m$ cells in the cross-classification. An m-tuple $\mathbf{i} = (i_1, i_2, \ldots, i_m)$ is used to represent the cell in the cross-classification corresponding to level i_1 of the first variable, level i_2 of the second variable, ..., and level i_m of the mth variable. The symbol $p_\mathbf{i}$ denotes the proportion of the population classified in cell $\mathbf{i}$. If two individuals are selected at random from the population, then the probability that the first is in cell $\mathbf{i}$ and the second is in cell $\mathbf{j}$ is $p_\mathbf{i}p_\mathbf{j}$.

To form the multivariate index of diversity, the probability of each possible type of pairing is multiplied by the proportion of the m variables for which the pairs have a nonidentical categorization, and then this is summed over all possible pairs. To elaborate: For a member in cell **i** and a member in cell **j**, the proportion of nonidentical attributes is the proportion of positions in **i** and **j** for which $i_l \neq j_l$. We consider all such pairs of cells for the two members, and we weight the proportion of nonidentical attributes for a particular pair by the probability $p_{\mathbf{i}}p_{\mathbf{j}}$ of classification in those cells. When this weighted proportion of nonidentical attributes is summed over all possible pairs of cells, we get the average proportion of disagreement between pairs selected randomly (with replacement) on the m characteristics.

Lieberson (1969, p. 855) gives a formula for the multivariate index of diversity D that is somewhat simpler to use for computation than the definition just given. His formula is

$$D = 1 - \left(\sum_{i=1}^{k_1} p_{i\cdots}^2 + \sum_{i=1}^{k_2} p_{\cdot i\cdots}^2 + \cdots + \sum_{i=1}^{k_m} p_{\cdots i}^2 \right) \Big/ m \qquad (27)$$

Notice that $\sum_{i=1}^{k_1} p_{i\cdots}^2$ is simply the sum of squares of the marginal proportions in the k_1 categories of the first variable; the other sums refer to the corresponding sum of squared marginal proportions over the categories of each of the other variables.

The basic properties of D are readily seen from the computing formula (27). Notice that the value of D just depends on the marginal proportions of each of the m variables, not on the distribution among the cells of the cross-classification. Let $D^{(1)}$, $D^{(2)}, \ldots, D^{(m)}$ denote the univariate indices of diversity corresponding to the m marginal distributions—that is, $D^{(1)} = 1 - \sum_{i=1}^{k_1} p_{i\cdots}^2$ and so forth. Then

$$\begin{aligned} D &= 1 - [(1 - D^{(1)}) + \cdots + (1 - D^{(m)})]/m \\ &= (D^{(1)} + \cdots + D^{(m)})/m \end{aligned} \qquad (28)$$

In other words, Lieberson's multivariate index of diversity is just the simple average of the marginal univariate indices. Thus, for the univariate case ($m = 1$), D is exactly the Simpson index of diversity discussed in previous sections of this chapter. It fol-

lows that the multivariate index $D = 0$ if and only if $p_{\mathbf{i}} = 1$ for some $\mathbf{i}$, in which case $D^{(i)} = 0$ for all $i = 1, \ldots, m$. Also

$$D \leq 1 - (1/k_1 + 1/k_2 + \cdots + 1/k_m)/m \tag{29}$$

which is the average of the upper bounds $(1 - 1/k_l)$ for the univariate indices. This upper bound for D is achieved when the marginal distribution for each variable is evenly dispersed over the categories for that variable. The index D could be divided by its maximum possible value to get an adjusted index that takes on values between 0 and 1. Such a measure would be a multivariate extension of the index of qualitative variation.[5] Of course, as the numbers of categories $(k_1, \ldots, k_m)$ get larger, the upper bound for D approaches 1.

To illustrate the computation of the multivariate version of D, we consider the data in Table 2 representing the relation of occupational status to place of birth for the white sample in 1870.[6]

TABLE 2
Occupational Status by Place of Birth and Year for White Adults in Walton County, Florida

	1870			1885		
Occupational Status	Florida natives	Non-natives	Total	Florida natives	Non-natives	Total
Professional	0.000	0.015	0.015	0.042	0.042	0.085
Manager, clerical, proprietor	0.000	0.021	0.021	0.025	0.051	0.076
Skilled	0.010	0.077	0.087	0.008	0.034	0.042
Unskilled	0.031	0.026	0.056	0.000	0.042	0.042
Laborer	0.215	0.210	0.426	0.195	0.381	0.576
Farmer	0.082	0.313	0.395	0.076	0.102	0.178
Total	0.338	0.662		0.347	0.653	
Sample size		195			118	

Let variable 1 represent occupational status and variable 2 represent place of birth. Denote the multivariate index of diversity for this group by D_1. Then we have $m = 2$ variables, with $k_1 = 6$ levels for the first and $k_2 = 2$ for the second. Also

[5]The index defined in this manner would not, in general, be the same as the average of the univariate indexes of qualitative variation.

[6]This is a smaller sample than those in previous examples, since information on place of birth was not available for all individuals in the original sample.

$$\sum_{i=1}^{6} \hat{p}_{i\cdot}^2 = (0.015)^2 + \cdots + (0.395)^2 = 0.349$$

$$\sum_{i=1}^{2} \hat{p}_{\cdot i}^2 = (0.338)^2 + (0.662)^2 = 0.552$$

so that

$$\hat{D}_1 = 1 - (0.349 + 0.552)/2 = 0.549$$

That is, for two individuals selected at random with replacement from the white sample in 1870, the average proportion of the two variables on which the two individuals differ in classification is 0.549. Similarly, using the data for the white sample in 1885, one can verify that $\sum_{i=1}^{6} \hat{q}_{i\cdot}^2 = 0.380$, $\sum_{i=1}^{2} \hat{q}_{\cdot i}^2 = 0.547$, and $\hat{D}_2 = 0.537$.

The inference procedures for the multivariate index are direct extensions of the corresponding procedures for the univariate diversity index. A large-sample $100(1 - \alpha)$ percent confidence interval for D, when a simple random sample is taken from the population cross-classification, is given by

$$\hat{D} \pm \zeta_{\alpha/2}\hat{\sigma}/\sqrt{n} \tag{30}$$

where the expression (see Appendix A) for $\hat{\sigma}^2$ is

$$\hat{\sigma}^2 = (4/m^2)\left[\sum_{\mathbf{i}} \hat{p}_{\mathbf{i}}(\hat{p}_{i_1\cdots} + \cdots + \hat{p}_{\cdots i_m})^2\right] - 4(1 - \hat{D})^2 \tag{31}$$

The term in the brackets is a sum taken over all $k_1k_2 \ldots k_m$ cells in the cross-classification. For each cell, one multiplies the proportion in that cell by the square of the sum of the proportion of observations having the same category on variable 1, the proportion of observations having the same category on variable 2, and so forth.

Consider, for example, the data in Table 2 for the white population in 1870. For these sample proportions,

$$\begin{aligned}\sum_{\mathbf{i}} \hat{p}_{\mathbf{i}}(\hat{p}_{i_1\cdot} + \hat{p}_{\cdot i_2})^2 &= (0.010)(0.338 + 0.087)^2 \\ &\quad + \cdots + (0.313)(0.662 + 0.395)^2 \\ &= 0.848\end{aligned}$$

Also $m = 2$, $n_1 = 195$, and $\hat{D}_1 = 0.549$, so that $\hat{\sigma}^2 = 0.848 - 4(1 - 0.549)^2 = 0.034$ and $\hat{\sigma} = 0.185$. An approximate 95 per-

cent confidence interval for D_1 is given by

$$0.549 \pm 1.96(0.185)/\sqrt{195} = 0.549 \pm 0.026 = (0.523, 0.575)$$

To compare the multivariate diversity indices D_1 and D_2 for two populations based on independent random samples of size n_1 and n_2, one could compute the confidence interval for their difference, which is

$$(\hat{D}_2 - \hat{D}_1) \pm z_{\alpha/2}\sqrt{(\hat{\sigma}_1^2/n_1) + (\hat{\sigma}_2^2/n_2)} \tag{32}$$

where $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are computed according to Formula (31) for the two samples. To test for the equality of s multivariate diversity indices, one could use the same chi-square test as described by Formula (18). This test and the intervals (Expressions 30 and 32) are valid as long as the values of the indices are strictly between their boundary points.

Lieberson (1969) also introduced a multivariate analog of the measure of diversity D_b between two populations classified according to the same scheme. This index represents the average proportion of nonidentical attributes between pairs for the characteristics under study when one observation is randomly drawn from each of the two populations. Lieberson presents the computing formula for this multivariate version of D_b of

$$D_b = 1 - \left(\sum_{i=1}^{k_1} p_{i\cdots}q_{i\cdots} + \sum_{i=1}^{k_2} p_{\cdot i\cdots}q_{\cdot i\cdots} + \cdots + \sum_{i=1}^{k_m} p_{\cdots i}q_{\cdots i}\right)\Big/m \tag{33}$$

The terms in parentheses represent the sum, over all the marginal categories of all the m variables, of the product of the proportions of the first and second population in each category.

It is easily seen that D_b is just the simple average of the between-groups diversity indices for the m pairs of marginal distributions of the m variables. In the univariate case, it reduces to the index D_b previously considered. The index can take values between 0 and 1. The value of 0 occurs if and only if $p_i = q_i = 1$ for some cell $\mathbf{i}$. The value of 1 occurs if and only if the observations for one population fall in completely different categories on *all* variables than the observations for the other population. The larger the value of D_b, the less is the tendency for pairs chosen at

random from the two populations to share the same attributes on the m variables—that is, the greater the diversity of the two groups relative to each other.

For the bivariate classification of occupational status and place of birth for the white samples of 1870 and 1885, $m = 2$,

$$\sum_{i=1}^{6} \hat{p}_{i\cdot}\hat{q}_{i\cdot} = (0.015)(0.085) + \cdots + (0.395)(0.178) = 0.324$$

and

$$\sum_{i=1}^{2} \hat{p}_{\cdot i}\hat{q}_{\cdot i} = (0.338)(0.347) + (0.662)(0.653) = 0.550$$

Thus

$$\hat{D}_b = 1 - (0.324 + 0.550)/2 = 0.563$$

In other words, for a randomly selected white individual from the sample of 1870 and a randomly selected one from the sample of 1885, the expected proportion of nonidentical attributes is 0.563 for this bivariate classification.

For large samples n_1 and n_2 from the two populations, a $100(1 - \alpha)$ percent confidence interval for $D_b (0 < D_b < 1)$ is given (see Appendix B) by

$$\hat{D}_b \pm \mathcal{Z}_{\alpha/2}\hat{\sigma}_b/\sqrt{n_1 + n_2} \tag{34}$$

where

$$\hat{\sigma}_b^2 = \left[\left(\frac{n_1 + n_2}{n_1}\right)\sum_{\mathbf{i}} \hat{p}_{\mathbf{i}}(\hat{q}_{i_1\cdots} + \cdots + \hat{q}_{\cdots i_m})^2 + \left(\frac{n_1 + n_2}{n_2}\right)\sum_{\mathbf{j}} \hat{q}_{\mathbf{j}}(\hat{p}_{j1\cdots} + \cdots + \cdots + \hat{p}_{\cdots j_m})^2\right] \Big/ m^2 - (1 - \hat{D}_b)^2(n_1 + n_2)^2/n_1 n_2 \tag{35}$$

In Table 2, for which $n_1 = 195$ and $n_2 = 118$, the first summation in the estimated variance expression is

$$(0.010)(0.042 + 0.347)^2 + \cdots + (0.313)(0.178 + 0.653)^2 = 0.814$$

The second summation is

$$0.042\,(0.015 + 0.338)^2 + \cdots + (0.102)(0.395 + 0.662)^2 = 0.813$$

Thus

$$\hat{\sigma}_b^2 = \left[\frac{313}{195}(0.814) + \frac{313}{118}(0.813)\right]^2 \Big/ 4 - (1 - 0.563)^2(195 + 118)^2/(195)(118) = 0.053$$

and $\hat{\sigma}_b = 0.230$. A 95 percent confidence interval for D_b is given by

$$0.563 \pm 1.96(0.230)/\sqrt{313} = 0.563 \pm 0.026 = (0.537, 0.589)$$

In the multivariate case, one could also give a confidence interval for the difference in the indices $D_1 - D_b$ or $D_2 - D_b$. This interval would describe the difference between the average proportion of disagreement within a population and the average proportion of disagreement between that population and another one for the attributes in the cross-classification. The derivation of the sampling distribution in this case and the corresponding large-sample confidence interval are given in Appendix C.

CONCLUSION

We have now completed our presentation of methods of describing and making large-sample inferences about population heterogeneity with respect to qualitative variables. We have shown how to measure and estimate population diversity on a one-variable classification for one, two, and several groups. A measure of between-groups diversity was defined, and its sampling properties were considered. In addition, we have considered the statistical properties of analogous multivariate indices, where diversity is measured with respect to a cross-classification of nominal variables.

These measures of diversity, as noted, have been applied in several disciplines, particularly ecology. We believe they can also be of value to sociologists in a variety of applications. In the examples analyzed in this chapter, some interesting conclusions can be made about the postbellum status of blacks in a Southern county. Most notably, the occupational diversity for blacks declined sharply in the decades following the emancipation of slaves, while the corresponding white occupational diversity re-

mained about constant. It might also be useful in many applications to correlate indices such as these with indices representing other social phenomena, such as crime rates or measures of inequality. It is hoped that drawing together the various measures of diversity and presenting their sampling theory within one chapter will be helpful to researchers interested in describing or estimating qualitative diversity in a population.

APPENDICES

The derivations of the large-sample behavior of $\hat{D}$, $\hat{D}_b$, and $\hat{D} - \hat{D}_b$ follow the standard methods used by Goodman and Kruskal (1963, p. 359) for bivariate measures of association. The derivation in each of the three appendices is based on the fact that the diversity indices are functions of a collection of sample proportions that are jointly normally distributed for large samples. We just consider the case of "multinominal" random sampling from the cells of the classification. This case corresponds to random sampling from an infinite population or random sampling with replacement from a finite population. None of the marginal frequencies in the cross-classification (for the multivariate case) are treated as fixed.

Appendix A: Sampling Distribution of $\hat{D}$

We shall derive the sampling distribution of $\hat{D}$ for the general multivariate version, for which

$$\sqrt{n}(\hat{D} - D) = \sqrt{n} \left[\sum_{i=1}^{k_1} (p^2_{i\cdots} - \hat{p}^2_{i\cdots}) + \cdots + \sum_{i=1}^{k_m} (p^2_{\cdots i} - \hat{p}^2_{\cdots i}) \right] \Big/ m$$

from Equation (27). This quantity is a continuous function of the sample proportions $\{\hat{p}_i\}$, with continuous first partial derivatives. Since the sample proportions are asymptotically jointly normally distributed, $\sqrt{n}(\hat{D} - D)$ is itself asymptotically normally distributed with mean zero and variance $\sigma^2 = \mathbf{d}' \, \boldsymbol{\Sigma} \, \mathbf{d}$, where $\boldsymbol{\Sigma}$ is

the covariance matrix of the $\{\sqrt{n}\hat{p}_{\mathbf{i}}\}$ and $\mathbf{d}$ is the vector of first partial derivatives of $\hat{D}$ with respect to the $\{\hat{p}_{\mathbf{i}}\}$ evaluated at $\{p_{\mathbf{i}}\}$.

Now the covariance between $\sqrt{n}\hat{p}_{\mathbf{i}}$ and $\sqrt{n}\hat{p}_{\mathbf{j}}$ is $\delta_{\mathbf{ij}}p_{\mathbf{i}} - p_{\mathbf{i}}p_{\mathbf{j}}$, where $\delta_{\mathbf{ij}} = 1$ if $\mathbf{i} = \mathbf{j}$ (that is, if $i_l = j_l$, for $1 \leq l \leq m$) and $\delta_{\mathbf{ij}} = 0$ otherwise. Also, for arbitrary $\mathbf{i} = (i_1, \ldots, i_m)$,

$$\left.\frac{\partial \hat{D}}{\partial \hat{p}_{\mathbf{i}}}\right|_{\{p_{\mathbf{i}}\}} = -2(p_{i_1\ldots} + \cdots + p_{\ldots i_m})/m$$

Thus the asymptotic variance of $\sqrt{n}(\hat{D} - D)$ is

$$\mathbf{d}' \Sigma \mathbf{d} = \frac{4}{m^2}\sum_{\mathbf{i}}\sum_{\mathbf{j}}(p_{i_1\ldots} + \cdots + p_{\ldots i_m})(p_{j_1\ldots} + \cdots + p_{\ldots j_m})(\delta_{\mathbf{ij}}p_{\mathbf{i}} - p_{\mathbf{i}}p_{\mathbf{j}})$$

$$= \frac{4}{m^2}\left\{\sum_{\mathbf{i}} p_{\mathbf{i}}(p_{i_1\ldots} + \cdots + p_{\ldots i_m})^2 - \left[\sum_{\mathbf{i}} p_{\mathbf{i}}(p_{i_1\ldots} + \cdots + p_{\ldots i_m})\right]^2\right\}$$

Lastly,

$$\sum_{\mathbf{i}} p_{\mathbf{i}}(p_{i_1\ldots} + \cdots + p_{\ldots i_m}) = \sum_{i=1}^{k_1} p_{i\ldots}^2 + \cdots + \sum_{i=1}^{k_m} p_{\ldots i}^2 = m(1 - D)$$

so that

$$\sigma^2 = \frac{4}{m^2}\sum_{\mathbf{i}} p_{\mathbf{i}}(p_{i_1\ldots} + \cdots + p_{\ldots i_m})^2 - 4(1 - D)^2$$

For the univariate case ($m = 1$), this reduces to

$$\sigma^2 = 4\left[\sum_{i=1}^{k} p_i^3 - (1 - D)^2\right]$$

Thus, for large samples, $\hat{D}$ is approximately normally distributed about D with variance σ^2/n. This fact is of use as long as $\sigma^2 > 0$; for example, in the univariate case, whenever $0 < D <$

$(k - 1)/k$. For two independent random samples of sizes n_1 and n_2, it follows that $\hat{D}_1 - \hat{D}_2$ is approximately normally distributed about $D_1 - D_2$ with variance $\sigma_1^2/n_1 + \sigma_2^2/n_2$. For large samples, the estimated variances can be substituted in formulas for confidence intervals and tests of hypotheses by virtue of Slutzky's theorem.

If $\hat{D}_1, \hat{D}_2, \ldots, \hat{D}_s$ are the values of the index of diversity for s independent random samples, and if $\sigma_1^2, \ldots, \sigma_s^2$ are the values of the preceding expression for σ^2 for each of the corresponding populations, then the $\{\sqrt{n_i}[(\hat{D}_i - D_i)/\sigma_i]\}$ are independent and have approximately the standard normal distribution. If $D_1 = \cdots = D_s = D$, then

$$\sum_{i=1}^{s} n_i(\hat{D}_i - D)^2/\sigma_i^2$$

has approximately the chi-square distribution with s degrees of freedom. Now, letting $\bar{D} = (\sum_{i=1}^{s} D_i n_i/\sigma_i^2)/(\sum_{i=1}^{s} n_i/\sigma_i^2)$,

$$\sum_{i=1}^{s} n_i(\hat{D}_i - D)^2/\sigma_i^2 = \sum_{i=1}^{s} n_i(\hat{D}_i - \bar{D})^2/\sigma_i^2 + (\bar{D} - D)^2 \sum_{i=1}^{s} n_i/\sigma_i^2$$

The asymptotic variance of $\bar{D}$ is $(\sum_{i=1}^{s} n_i/\sigma_i^2)^{-1}$, so the term $(\bar{D} - D)(\sum_{i=1}^{s} n_i/\sigma_i^2)^{1/2}$ has approximately the standard normal distribution and its square has approximately the chi-square distribution with 1 degree of freedom. A standard application of Cochran's theorem can be used to show that $\sum_{i=1}^{s} n_i(\hat{D}_i - \bar{D})/\sigma_i^2$ has approximately the chi-square distribution with $s - 1$ degrees of freedom. Also, by Slutzky's theorem, this term has the same asymptotic distribution if the estimated variances $\{\hat{\sigma}_i^2\}$ are substituted for $\{\sigma_i^2\}$. Thus to test the null hypothesis H_0: $D_1 = D_2 = \cdots = D_s$, based on s independent random samples, one can use the test statistic

$$\chi^2 = \sum_{i=1}^{s} n_i(\hat{D}_i - \bar{D})^2/\hat{\sigma}_i^2$$

which has approximately (if $n_1, \ldots, n_s$ are large) the chi-square distribution with $s - 1$ degrees of freedom under the null hypothesis.

Appendix B: Sampling Distribution of $\hat{D}_b$

For the general multivariate case,

$$\sqrt{n_1 + n_2}(\hat{D}_b - D_b) = \sqrt{n_1 + n_2} \left[\sum_{i=1}^{k_1} (p_{i\cdots} q_{i\cdots} - \hat{p}_{i\cdots}\hat{q}_{i\cdots}) + \cdots + \cdots + \sum_{i=1}^{k_m} (p_{\cdots i} q_{\cdots i} - \hat{p}_{\cdots i}\hat{q}_{\cdots i}) \right] \Big/ m$$

from Equation (33). Now for any cell $\mathbf{i} = (i_1, \ldots, i_m)$,

$$\left.\frac{\partial \hat{D}_b}{\partial \hat{p}_{\mathbf{i}}}\right|_{\{p_{\mathbf{i}}, q_{\mathbf{i}}\}} = -(q_{i_1\cdots} + \cdots + q_{\cdots i_m})/m$$

$$\left.\frac{\partial \hat{D}_b}{\partial \hat{q}_{\mathbf{i}}}\right|_{\{p_{\mathbf{i}}, q_{\mathbf{i}}\}} = -(p_{i_1\cdots} + \cdots + p_{\cdots i_m})/m$$

Let $\hat{\mathbf{r}} = (\{\hat{p}_i\}, \{\hat{q}_i\})$ be a vector with $2k_1k_2\cdots k_m$ positions, the first $k_1k_2\cdots k_m$ consisting of the $\{\hat{p}_i\}$ and the last $k_1k_2\cdots k_m$ consisting of the $\{\hat{q}_i\}$. Also let $\mathbf{d}' = (\mathbf{d}_1', \mathbf{d}_2')$, where $\mathbf{d}_1$ is the vector of length $k_1\cdots k_m$ consisting of the

$$\left.\frac{\partial \hat{D}_b}{\partial \hat{p}_{\mathbf{i}}}\right|_{\{p_{\mathbf{i}}, q_{\mathbf{i}}\}}$$

and $\mathbf{d}_2$ is the vector of length $k_1\cdots k_m$ consisting of the

$$\left.\frac{\partial \hat{D}_b}{\partial \hat{q}_{\mathbf{i}}}\right|_{\{p_{\mathbf{i}}, q_{\mathbf{i}}\}}$$

Then $\sqrt{n_1 + n_2}(\hat{D}_b - D_b)$ is asymptotically normally distributed about zero with variance $\sigma_b^2 = \mathbf{d}' \Sigma \mathbf{d}$, where Σ is the covariance matrix of $\sqrt{n_1 + n_2}\hat{\mathbf{r}}$. Now

$$\operatorname{cov}(\sqrt{n_1 + n_2}\hat{p}_{\mathbf{i}}, \sqrt{n_1 + n_2}\hat{p}_{\mathbf{j}}) = (n_1 + n_2)\operatorname{cov}(\hat{p}_{\mathbf{i}}, \hat{p}_{\mathbf{j}}) = [(n_1 + n_2)/n_1](\delta_{\mathbf{ij}} p_{\mathbf{i}} - p_{\mathbf{i}} p_{\mathbf{j}})$$

and similarly

$$\operatorname{cov}(\sqrt{n_1 + n_2}\hat{q}_{\mathbf{i}}, \sqrt{n_1 + n_2}\hat{q}_{\mathbf{j}}) = [(n_1 + n_2)/n_2](\delta_{\mathbf{ij}} q_{\mathbf{i}} - q_{\mathbf{i}} q_{\mathbf{j}})$$

Since the two random samples are independent, $\operatorname{cov}(\hat{p}_i, \hat{q}_j) = 0$ and the covariance matrix $\hat{\Sigma}$ may be partitioned as

$$\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_1 & 0 \\ 0 & \hat{\Sigma}_2 \end{pmatrix}$$

where the **i-j**th element of $\hat{\Sigma}_1$ is $[(n_1 + n_2)/n_1](\delta_{ij}p_i - p_ip_j)$ and the **i-j**th element of $\hat{\Sigma}_2$ is $[(n_1 + n_2)/n_2](\delta_{ij}q_i - q_iq_j)$. Thus the asymptotic variance of $\sqrt{n_1 + n_2}(\hat{D}_b - D_b)$ is

$$\sigma_b^2 = \mathbf{d}'\hat{\Sigma}\mathbf{d} = \mathbf{d}_1'\hat{\Sigma}_1\mathbf{d}_1 + \mathbf{d}_2'\hat{\Sigma}_2\mathbf{d}_2$$

$$= \frac{n_1 + n_2}{n_1 m^2} \sum_{\mathbf{i}} \sum_{\mathbf{j}} (q_{i_1\cdots} + \cdots + q_{\cdots i_m})(q_{j_1\cdots} + \cdots + q_{\cdots j_m})$$

$$(\delta_{ij}p_i - p_ip_j)$$

$$+ \frac{n_1 + n_2}{n_2 m^2} \sum_{\mathbf{i}} \sum_{\mathbf{j}} (p_{i_1\cdots} + \cdots + p_{\cdots i_m})(p_{j_1\cdots} + \cdots + p_{\cdots j_m})$$

$$(\delta_{ij}q_i - q_iq_j)$$

$$= \frac{n_1 + n_2}{n_1 m^2} \left\{ \sum_{\mathbf{i}} p_i(q_{i_1\cdots} + \cdots + q_{\cdots i_m})^2 \right.$$

$$\left. - \left[\sum_{\mathbf{i}} p_i(q_{i_1\cdots} + \cdots + q_{\cdots i_m}) \right]^2 \right\}$$

$$+ \frac{n_1 + n_2}{n_2 m^2} \left\{ \sum_{\mathbf{i}} q_i(p_{i_1\cdots} + \cdots + p_{\cdots i_m})^2 \right.$$

$$\left. - \left[\sum_{\mathbf{i}} q_i(p_{i_1\cdots} + \cdots + p_{\cdots i_m}) \right]^2 \right\}$$

Finally, since

$$\sum_{\mathbf{i}} p_i(q_{i_1\cdots} + \cdots + q_{\cdots i_m})$$

$$= \sum_{\mathbf{i}} q_i(p_{i_1\cdots} + \cdots + p_{\cdots i_m}) = m(1 - D_b)$$

the asymptotic variance reduces to

$$\sigma_b^2 = \left\{ [(n_1 + n_2)/n_1] \sum_i p_i(q_{i_1\ldots} + \cdots + q_{\ldots i_m})^2 \right.$$

$$\left. + [(n_1 + n_2)/n_2] \sum_i q_i(p_{i_1\ldots} + \cdots + p_{\ldots i_m})^2 \right\}/m^2$$

$$- (1 - D_b)^2(n_1 + n_2)^2/n_1n_2$$

For the univariate case, the expression reduces to

$$\sigma_b^2 = [(n_1 + n_2)/n_1] \sum_{i=1}^{k} p_iq_i^2 + [(n_1 + n_2)/n_2]$$

$$\sum_{i=1}^{k} q_ip_i^2 - (1 - D_b)^2(n_1 + n_2)^2/n_1n_2$$

Thus, as $n_1 \to \infty$ and $n_2 \to \infty$ so that $n_1/(n_1 + n_2) \to \lambda (0 < \lambda < 1)$, the sampling distribution of $\sqrt{n_1 + n_2}(\hat{D}_b - D_b)$ converges to the normal distribution with mean zero and variance σ_b^2 given by this formula with $n_1/(n_1 + n_2)$ replaced by λ. From the properties of D_b, it is seen that $\sigma_b^2 = 0$ if and only if $D_b = 0$ or $D_b = 1$. Thus this formula leads to confidence intervals $(\hat{D}_b \pm z_{\alpha/2}\hat{\sigma}_b/\sqrt{n_1 + n_2})$ and tests of hypotheses about D_b as long as $0 < D_b < 1$.

Appendix C: Sampling Distribution of $\hat{D}_1 - \hat{D}_b$

We shall again treat the general multivariate case. When the random samples from the two populations are independent, the derivation is similar to the one in Appendix B. The asymptotic distribution of $\sqrt{n_1 + n_2}[(\hat{D}_1 - \hat{D}_b) - (D_1 - D_b)]$ is normal about zero with variance $\mathbf{d}' \Sigma \mathbf{d}$. Here Σ is the same covariance matrix as in Appendix B and $\mathbf{d}' = (\mathbf{d}_1', \mathbf{d}_2')$, with $\mathbf{d}_1$ the vector of the

$$\left.\frac{\partial(\hat{D}_1 - \hat{D}_b)}{\partial \hat{p}_i}\right|_{\{p_i, q_i\}} = [(q_{i_1\ldots} + \cdots + q_{\ldots i_m})$$

$$- 2(p_{i_1\ldots} + \cdots + p_{\ldots i_m})]/m$$

and $\mathbf{d}_2$ the vector of the

$$\left.\frac{\partial(\hat{D}_1 - \hat{D}_b)}{\partial \hat{q}_i}\right|_{\{p_i, q_i\}} = (p_{i_1\ldots} + \cdots + p_{\ldots i_m})/m$$

Following the same steps as in Appendix B, we obtain the asymptotic variance of

$$\begin{aligned}\sigma^2_{1,b} = \frac{n_1 + n_2}{n_1 m^2}\Bigg(&\sum_i p_i[q_{i_1\ldots} + \cdots + q_{\ldots i_m} \\ &- 2(p_{i_1\ldots} + \cdots + p_{\ldots i_m})]^2 \\ &- \Bigg\{\sum_i p_i[q_{i_1\ldots} + \cdots + q_{\ldots i_m} \\ &- 2(p_{i_1\ldots} + \cdots + p_{\ldots i_m})]\Bigg\}^2\Bigg) \\ &+ \frac{n_1 + n_2}{n_2 m^2}\Bigg\{\sum_i q_i(p_{i_1\ldots} + \cdots + p_{\ldots i_m})^2 \\ &- \Bigg[\sum_i q_i(p_{i_1\ldots} + \cdots + p_{\ldots i_m})\Bigg]^2\Bigg\} \\ = \frac{n_1 + n_2}{n_1 m^2}\Bigg\{&\sum_i p_i[q_{i_1\ldots} + \cdots + q_{\ldots i_m} - 2(p_{i_1\ldots} + \cdots + p_{\ldots i_m})]^2 \\ &- [m(2D_1 - D_b - 1)]^2\Bigg\} \\ &+ \frac{n_1 + n_2}{n_2 m^2}\Bigg\{\sum_i q_i(p_{i_1\ldots} + \cdots + p_{\ldots i_m})^2 - [m(1 - D_b)]^2\Bigg\}\end{aligned}$$

Thus, for large samples, an approximate $100(1 - \alpha)$ percent confidence interval for $D_1 - D_b$ is $(\hat{D}_1 - \hat{D}_b) \pm z_{\alpha/2}\hat{\sigma}_{1,b}/\sqrt{n_1 + n_2}$. For the univariate case, this reduces to Formula (26). The variance $\sigma^2_{1,b}$ is strictly positive as long as not both D_1 and D_b are at their boundary values.

REFERENCES

AGRESTI, B. F.

1976 "Household and family in the postbellum South: Walton County, Florida, 1870 and 1885." Ph.D. dissertation. University of Florida.

AMEMIYA, E. C.
1963 "Measurement of economic differentiation." *Journal of Regional Science* 5:85–87.

BHARGAVA, T. N., AND DOYLE, P. H.
1974 "A geometric study of diversity." *Journal of Theoretical Biology* 43:241–251.

BHARGAVA, T. N., AND UPPULURI, V. R. R.
1975 "On diversity in human ecology." *Metron* 34:1–13.

BOWMAN, K. O., AND OTHERS.
1971 "Comments on the distribution of indices of diversity." Pp. 315–366 in G. P. Patil, E. C. Pielou, and W. E. Waters (Eds.), *Statistical Ecology*. Vol. 3: *Many Species Populations, Ecosystems, and Systems Analysis*. University Park: Pennsylvania State University Press.

FRISBIE, P.
1975 "Measuring the degree of bureaucratization at the societal level." *Social Forces* 53:563–573.

GIBBS, J. P., AND MARTIN, W. T.
1962 "Urbanization, technology, and the division of labor: International patterns." *American Sociological Review* 27:667–677.

GOOD, I. J.
1953 "The population frequencies of species and the estimation of population parameters." *Biometrika* 40:237–264.

GOODMAN, L. A., AND KRUSKAL, W. H.
1963 "Measures of association for cross-classifications. III: Approximate sampling theory." *Journal of the American Statistical Association* 58:310–364.

GRANDJEAN, B. D.
1974 "The division of labor, technology, and education: Cross-national evidence." *Social Science Quarterly* 55:297–309.

GREENBERG, J. H.
1956 "The measurement of linguistic diversity." *Language* 32:109–115.

HOLGATE, P.
1971 "Drift in the random component of isonymy." *Biometrics* 27:448–451.

LABOVITZ, S., AND GIBBS, J. P.
1964 "Urbanization, technology, and the division of labor: Further evidence." *Pacific Sociological Review* 7:3–9.

LIEBERSON, S.
1964 "An extension of Greenberg's linguistic diversity measure." *Language* 40:526–531.

1969 "Measuring population diversity." *American Sociological Review* 34:850–862.

MUELLER, J. H., AND SCHUESSLER, K. F.

1961 *Statistical Reasoning in Sociology.* Boston: Houghton Mifflin.

MUELLER, J. H., AND SCHUESSLER, K. F., AND COSTNER, H. L.

1977 *Statistical Reasoning in Sociology.* (3rd ed.) Boston: Houghton Mifflin.

PIELOU, E. C.

1967 "The use of information theory in the study of the diversity of biological populations." *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 4:163–177.

1969 *An Introduction to Mathematical Ecology.* New York: Wiley-Interscience.

RENYI, A.

1961 "On measures of entropy and information." *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* 1:547–561.

REX, M. A.

1973 "Deep-sea species: Decreased gastropod diversity at abyssal depths." *Science* 181:1051–1053.

SARGENT, T. D., AND OWEN, D. F.

1975 "Apparent stability in hindwing diversity in samples of moths of varying species composition." *OIKOS* 26:205–210.

SIMPSON, E. H.

1949 "Measurement of diversity." *Nature* 163:688.

SWANSON, D. A.

1976 "A sampling distribution and significance test for differences in qualitative variation." *Social Forces* 55:182–184.

VAN BELLE, G., AND AHMAD, I.

1974 "Measuring affinity of distributions." Pp. 651–668 in F. Proschan and R. J. Serfling (Eds.), *Reliability and Biometry: Statistical Analysis of Lifelength.* Philadelphia: Society for Industrial and Applied Mathematics.

WEILER, H.

1966 "A coefficient measuring the goodness of fit." *Technometrics* 8:327–334.