This article examines the sampling distributions of popular indexes of segregation: the dissimilarity index and the Gini index. Although applications of segregation indexes are common in the social sciences, researchers have usually failed to recognize their stochastic nature. This study addresses that failure by deriving the exact sampling distribution of two popular indexes of segregation and developing a convenient asymptotic test for hypotheses about changes in segregation. Monte Carlo simulations show that the proposed test has appropriate significance levels. An example applies the methods developed in this article to test whether segregation of men and women across college majors decreased during the 1980s. The analysis finds that gender segregation has decreased, and that the decrease is statistically significant.

Sampling Distributions of Segregation Indexes

MICHAEL R. RANSOM
Brigham Young University

Segregation indexes measure the segregation by race, ethnicity, or sex across occupations, residential areas, and other dimensions. For example, a segregation index summarizes in a single number the differences between the distribution of men and women across occupations. Many indexes have been suggested. Duncan and Duncan (1955), James and Taeuber (1985), and Hutchens (1991) compare and criticize several alternative measures. These indexes have been used widely in the social sciences. For example, they have been used to measure school segregation by race (Zoloth 1976), residential segregation by race (Taeuber and Taeuber 1972), segregation of college professors (Ransom 1990) or students (Jacobs 1995) across academic fields by sex, and occupational segregation by sex and race (Albeda 1986).

AUTHOR'S NOTE: *I acknowledge helpful comments by James B. McDonald, Whitney Newey, and Eric Eide. Eric Eide graciously provided the data on college majors that is used in the example. I am grateful for financial support by the College of Family, Home and Social Sciences at Brigham Young University.*

SOCIOLOGICAL METHODS & RESEARCH, Vol. 28 No. 4, May 2000 454-475

Segregation indexes based on sample data are merely estimators of the level of segregation in a population. However, researchers have generally ignored the stochastic nature of segregation indexes. The few exceptions have used rather different approaches to modeling stochastic variability in the indexes. Massey (1978) used the jackknife technique to compute variances of the dissimilarity index. Boisso et al. (1994) used another computer-intensive method, the bootstrap, to obtain confidence intervals for estimated indexes. Farley and Johnson (1985) derived a sampling distribution for the dissimilarity under the hypothesis of no segregation. Cortese, Falk, and Cohen (1976) also derived a distribution for the no-segregation case using a somewhat different statistical model. However, the hypothesis of no segregation is seldom an interesting one, as Massey (1978) notes in his comment on the Cortese et al. article.[1]

Previous attempts to model the sampling variability in segregation indexes have failed to become part of the researcher's toolbox because of limited applicability or computational burden, or both. This article provides a useful, computationally convenient framework for statistical inference in the analysis of segregation.

SEGREGATION INDEXES

Most measures of segregation are mathematically related to the concept of the segregation curve, introduced in an influential article by Duncan and Duncan (1955). The segregation curve plots the cumulative proportion of females against the cumulative proportion of males, and the occupations are ranked in order of the male/female ratio. Figure 1 shows an example of a segregation curve. The curve is piecewise linear, with a line segment for each occupation. The slope of each segment is the ratio of the fraction of all males in the occupation to the fraction of all females in the occupation. The curve has a convex shape due to the ranking of the occupations. Segments near the origin are female dominated, whereas segments in the upper right corner are male dominated. For example, the lower left line segment represents an occupation that contains 25 percent of the women but only 5 percent of the men.

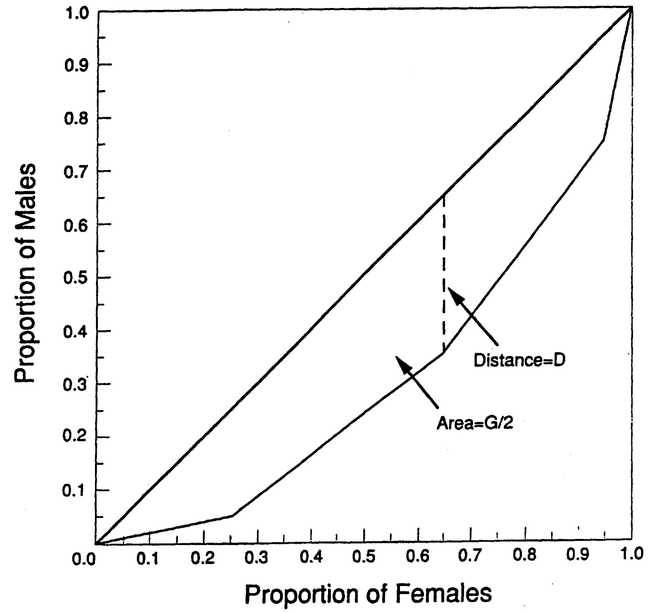


Figure 1: Segregation Curve

Perfect integration of occupations would occur if the same proportion of men and women were found in each occupation. This would be represented by the diagonal line running from the origin to the upper right corner of the box. Complete segregation would occur if occupations contained only females or only males. The corresponding segregation curve would be the lines that form the lower right side of the box. The curve is independent of the total number of men or women being analyzed, since each segment of the curve depends only on the proportions of men and women in each field. Thus, it makes sense to compare segregation curves across different years, as long as the definitions of the occupation groups have not changed over that time.[2]

Segregation indexes summarize the level of segregation represented by the segregation curve in a single numerical value. Hutchens (1991) points out that each index implicitly incorporates value judgments about the nature of occupational inequality. Such value judgments may be controversial. Nevertheless, most attempts to measure segregation have used one of two segregation indexes.

By far the most common measure of segregation in the applied literature is the dissimilarity index:

$$D = \left(\frac{1}{2}\right)\sum_{i=1}^{K} \left| p_i^F - p_i^M \right|, \qquad (1)$$

where p_i^F is the proportion of the sample of females that is in occupation i, p_i^M is the proportion of the sample of males that is in occupation i, and K is the number of occupations.

The dissimilarity index is easy to compute and interpret. The index varies between 0 (no segregation) and 1 (complete segregation). D can be interpreted as the fraction of women (or men) that must change occupations to achieve proportional representation in each occupation. In terms of the segregation curve, D is the maximum vertical distance between the diagonal line and the segregation curve (see Figure 1). An equivalent interpretation is that D is twice the area of the largest triangle that can be inscribed between the diagonal line and the segregation curve.

The Gini index of segregation, G, has been used less widely than D. The Gini index is defined as

$$G = \left(\frac{1}{2}\right)\sum_{i=1}^{K}\sum_{j=1}^{K} \left| p_i^F p_j^M - p_j^F p_i^M \right| = \sum_{i=1}^{K-1}\sum_{j=i+1}^{K} \left| p_i^F p_j^M - p_j^F p_i^M \right|. \qquad (2)$$

These formulas can be derived directly from the geometrical definition of G, or from alternative formulas such as that in James and Taeuber (1985:5). Equation (2) shows that the Gini index depends only on the proportion of all men and the proportion of all women in each occupation. Geometrically, the Gini index is twice the area between the segregation curve and the diagonal line, or the area between the diagonal and the segregation curve as a fraction of the total area beneath the diagonal (which is one half). Hutchens (1991) argues convincingly that G is a better measure of segregation because D is not sensitive to some changes in the occupational distribution.

A STATISTICAL MODEL FOR SEGREGATION INDEXES

To compute a segregation index, we observe from a sample of size N the number of men and women in each of K occupational categories. Denote these numbers as n_i^F and n_i^M. The multinomial distribution provides a convenient model for this sampling process:

$$P(n_1^F, n_2^F, \mathrm{K}, n_K^F, n_1^M, n_2^M, \mathrm{K}\ n_K^M) = N! \prod_{i=1}^{K} \prod_{J=F,M} \frac{(\pi_i^J)^{n_i^J}}{n_i^J!}, \qquad (3)$$

where the parameter π_i^J is the probability of observing an individual of sex J and occupation i in the sample (for a discussion of the multinomial distribution, see Johnson and Kotz 1969, chap. 11). This model provides a completely general description of any segregation process.

The indexes, D or G, in equations (1) and (2) are estimators because they are based on the sample proportions,

$$p_i^J = n_i^J / \sum_{t=1}^{K} n_t^J .$$

The population values of the indexes, which measure the "true" level of segregation in the population, can be expressed in terms of the parameters of the multinomial model by replacing p_i^J with

$$\pi_i^J / \sum_{t=1}^{K} \pi_t^J ,$$

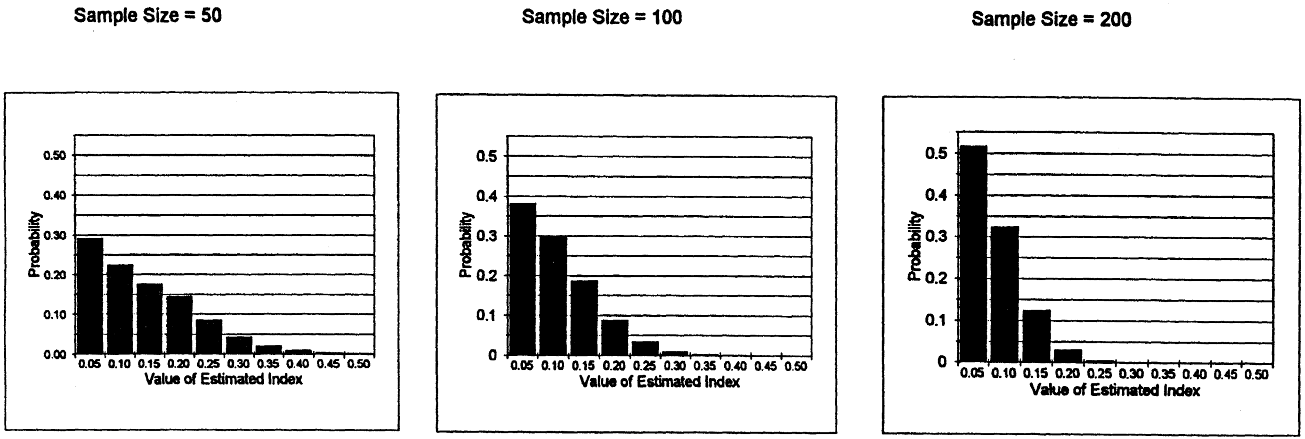
which is the corresponding concept from the population.[3] The statistical model derives from the sampling uncertainty—one cannot know the actual value of the underlying parameters of the multinomial distribution. It does not appear that this model would be appropriate for analysis of segregation based on full-count census enumeration data, such as is common in analyses of residential segregation by race, for example. In such cases there is no question of "statistical significance," since the population is observed.

EXACT SAMPLING DISTRIBUTIONS OF SEGREGATION INDEXES

Conceptually, it is straightforward to calculate the exact sampling distribution of D and G by enumerating all of the possible combinations of n_i^F and n_i^M, evaluating the index values that correspond, and calculating the probability of this occurring using the probability model in equation (3). Unfortunately, even with few occupational groups and moderate sample sizes, the computations are intractable. However, a simple case of only two occupations illustrates interesting sampling properties of the estimators. With only two occupations, the Gini index and the dissimilarity index must be identical—equations (1) and (2) must simplify to $G = D = |p_1^F - p_1^M|$.

Figure 2 shows histograms and reports moments of the exact sampling distribution for a case in which there is no segregation in the underlying population. In this example, $\pi_1^F = \pi_1^M = .25$, so on average there will be half of the men and half of the women in each occupation and the true values for D and G are both 0. As Figure 2 shows, for small samples, the estimates of D and G are not likely to be close to 0. With a sample size of 50, the estimates of G and D will be greater than .1 with probability of almost .5, and the expected value of the estimator is .114. However, with increases in the sample sizes, the distribution of the estimators collapses toward the population value of 0. For example, with a sample size of 200, more than half of the samples will yield estimates that are less than .05 and about 85 percent of the samples will yield estimates that are less than .1. The expected value of the estimator is about .06. Samples of such size (100 per occupation) are small relative to samples used in most social science applications.

The bias results from the fact that the index is an absolute value. Any sampling variation that causes p_1^F to deviate from p_1^M results in a nonzero value for the index. Thus, virtually *any* sampling variability in p_1^F and p_1^M will result in an upward bias to the estimator of D or G. The size of the bias depends on the variability of p_1^F and p_1^M, which declines with increasing sample size. This bias is not limited to the case in which the index value is 0. For example, if expected values for p_1^F and p_1^M are .6 and .4, respectively, the population segregation

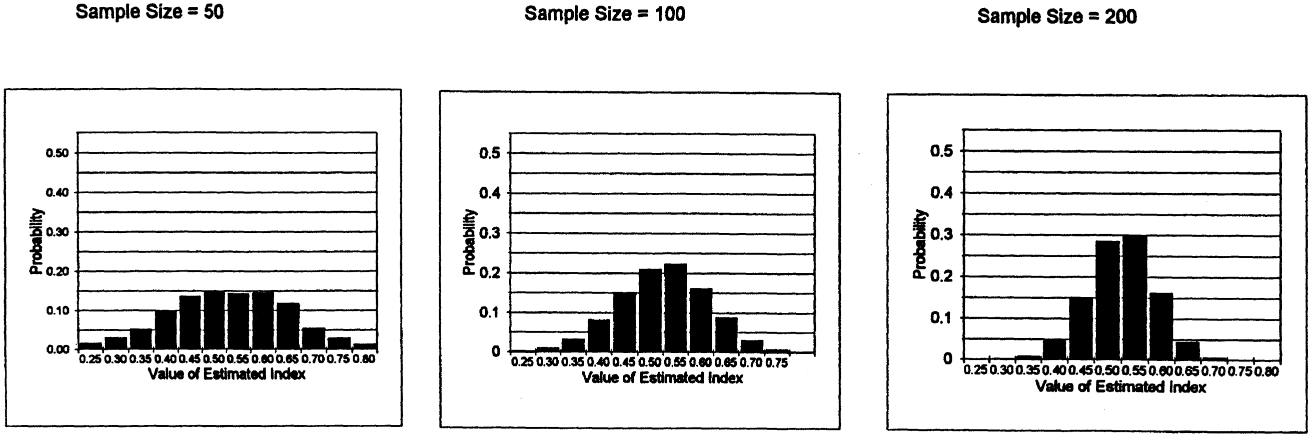


Expected Value	0.1140
Variance	0.0074
Coef of Skewness	0.9583
Coef of Kurtosis	3.7571

Expected Value	0.0802
Variance	0.0037
Coef of Skewness	0.9774
Coef of Kurtosis	3.8143

Expected Value	0.0566
Variance	0.0018
Coef of Skewness	0.9865
Coef of Kurtosis	3.8419

Figure 2: Exact Sampling Distributions for Dissimilarity or Gini Indexes of Segregation: Population Value of Index = 0



Expected Value	0.5000
Variance	0.0153
Coef of Skewness	-0.1667
Coef of Kurtosis	2.9779

Expected Value	0.5000
Variance	0.0076
Coef of Skewness	-0.1173
Coef of Kurtosis	2.9937

Expected Value	0.5000
Variance	0.0038
Coef of Skewness	-0.0823
Coef of Kurtosis	2.9968

Figure 3: Exact Sampling Distributions for Dissimilarity or Gini Indexes of Segregation: Population Value of Index = .5

index value is .2. But if the sample size is small, it will not be unusual to observe p_1^F smaller than p_1^M in some samples. Whereas $p_1^F - p_1^M$ will be centered on .2, $D = |p_1^F - p_1^M|$ will be centered on a larger number. Of course, in large samples, or in populations in which the proportions differ greatly across occupations, this bias will be small.

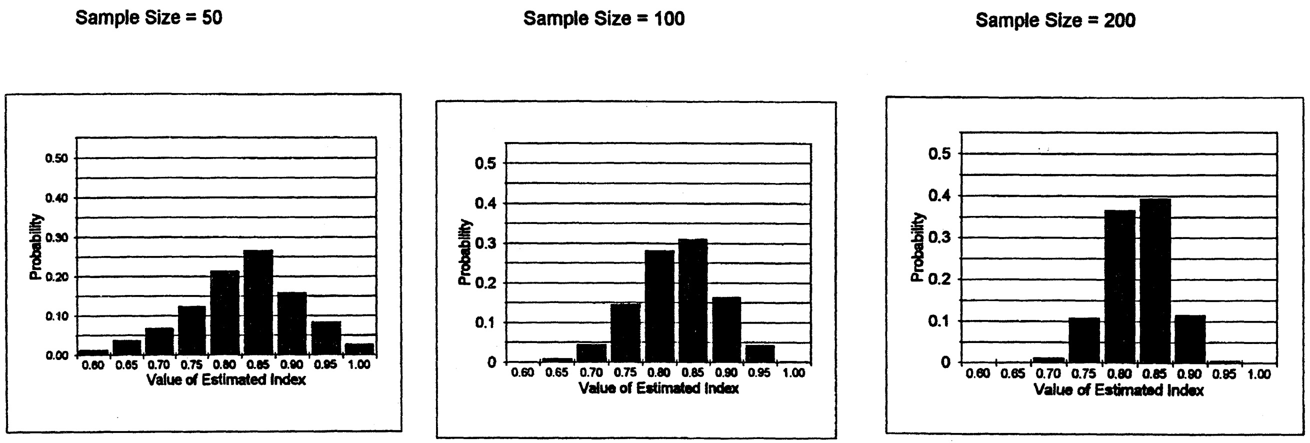
This upward bias is discussed in Farley and Johnson (1985) and Cortese et al. (1976). However, the practical importance of the bias may have been overstated by these authors, since they studied the case of no segregation, an extreme case of little practical interest.

Figure 3 presents histograms and moments of the exact sampling distribution for the case in which D and G are .5 ($\pi_1^F = \pi_2^M = .125, \pi_2^F = \pi_1^M = .375$). Figure 4 presents the same information for D and G equal to .8 ($\pi_1^F = \pi_2^M = .05, \pi_2^F = \pi_1^M = .45$). These distributions are very symmetric and unbiased to the limits of the precision reported, even for very small sample sizes. As the samples increase in size, skewness approaches 0 and kurtosis approaches 3. This suggests that the normal distribution is an appropriate model for the sampling distributions in large samples.

Evaluating the distribution of the estimated index for the case of multiple occupations requires much more computation than for two occupations. Even for just three occupations, it takes several hours of computing time to evaluate the probability function of D and G for a sample size of 100. Nevertheless, it is important to examine this case, since with three or more groups the Gini index and the dissimilarity index will usually differ. Here, I examine a case of three occupations.

Figure 5 shows histograms and moments for the exact sampling distribution for D and G for a case of no segregation in the population: $\pi_i^F = \pi_i^M = 1/6$ for all occupations in this example. This demonstrates again the substantial bias of the estimators when segregation is absent from the underlying population. In this case, the index is the sum of absolute values—any element of this sum can be a source of upward bias to the total.

Figure 6 contains histograms and moments of the exact sampling distribution for D and G for the following case: $\pi_i^F = (.175, .225, .1)$, $\pi_i^M = (.025, .175, .3)$. The population includes equal numbers of men and women, but women are more concentrated in the first occupation and men in the third occupation, with women just slightly more likely to be in the second occupation. The corresponding population values



Sample Size = 50

Sample Size = 100

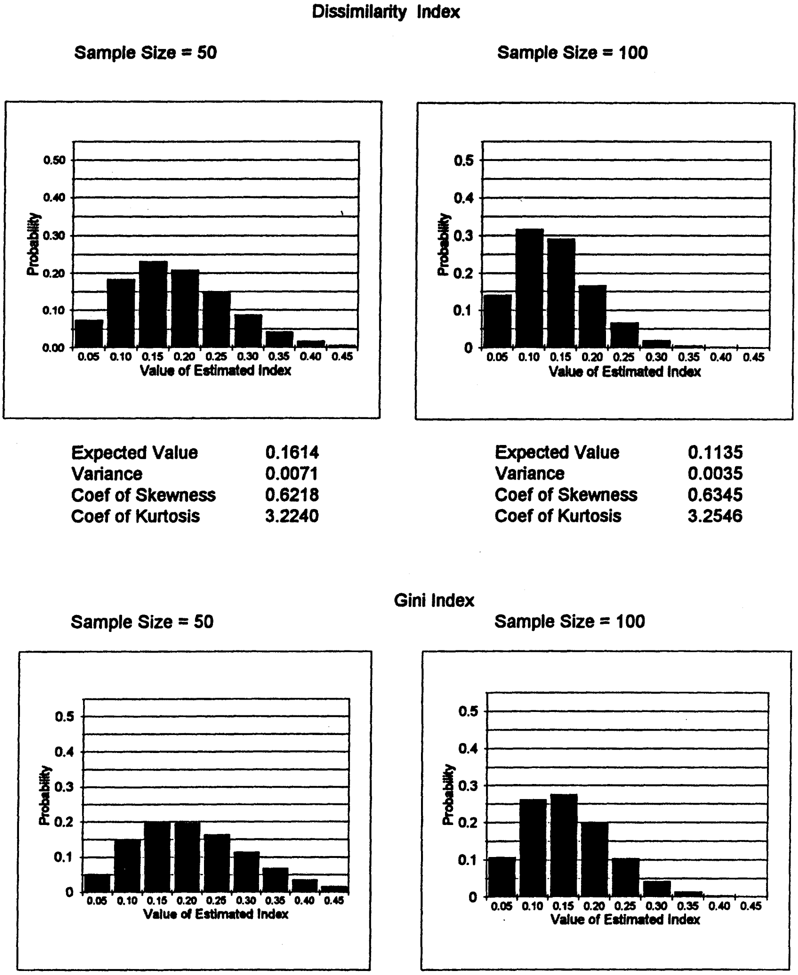
Sample Size = 200

Sample Size = 50	
Expected Value	0.8000
Variance	0.0074
Coef of Skewness	-0.3898
Coef of Kurtosis	3.1148

Sample Size = 100	
Expected Value	0.8000
Variance	0.0036
Coef of Skewness	-0.2709
Coef of Kurtosis	3.0540

Sample Size = 200	
Expected Value	0.8000
Variance	0.0018
Coef of Skewness	-0.1900
Coef of Kurtosis	3.0262

Figure 4: Exact Sampling Distributions for Dissimilarity or Gini Indexes of Segregation: Population Value of Index = .8



Dissimilarity Index, Sample Size = 50

Expected Value	0.1614
Variance	0.0071
Coef of Skewness	0.6218
Coef of Kurtosis	3.2240

Dissimilarity Index, Sample Size = 100

Expected Value	0.1135
Variance	0.0035
Coef of Skewness	0.6345
Coef of Kurtosis	3.2546

Gini Index, Sample Size = 50

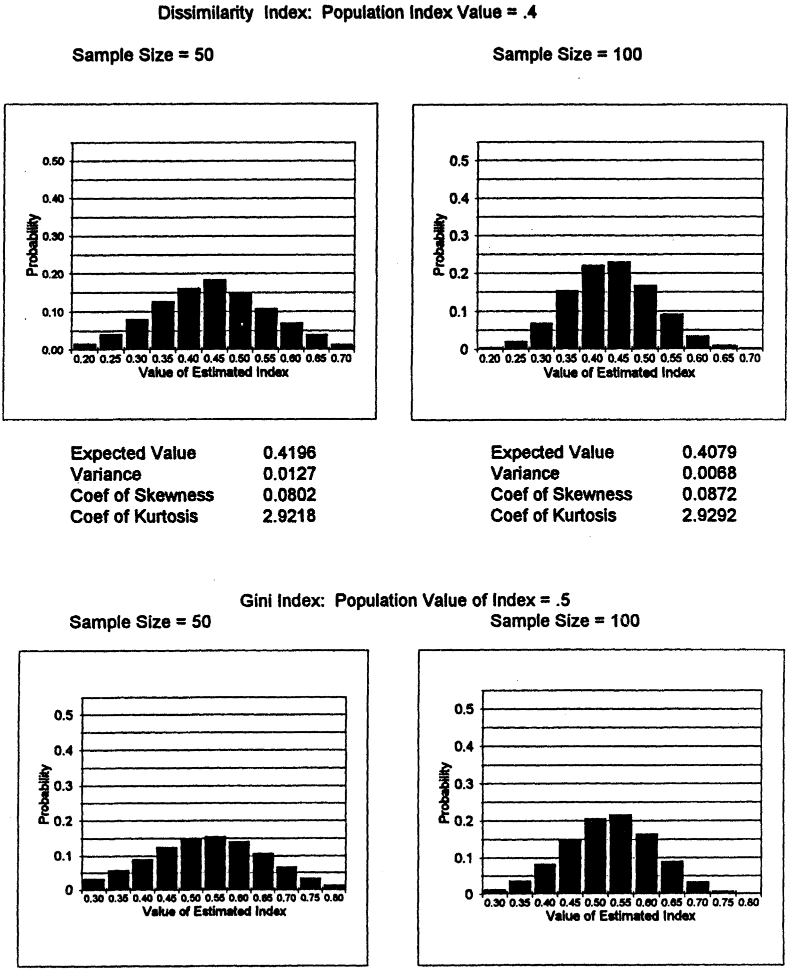
Expected Value	0.1864
Variance	0.0094
Coef of Skewness	0.6020
Coef of Kurtosis	3.1651

Gini Index, Sample Size = 100

Expected Value	0.1310
Variance	0.0047
Coef of Skewness	0.6248
Coef of Kurtosis	3.2266

Figure 5: Exact Sampling Distributions for Dissimilarity or Gini Indexes of Segregation (three occupation groups): Population Value of Index = 0

are .4 for *D* and .5 for *G*. Both *D* and *G* are biased upward, but the bias is slight, even in such small samples. The bias depends not only on the sample size but also on the parameters of the underlying occupational



Dissimilarity Index, Sample Size = 50:

Expected Value	0.4196
Variance	0.0127
Coef of Skewness	0.0802
Coef of Kurtosis	2.9218

Dissimilarity Index, Sample Size = 100:

Expected Value	0.4079
Variance	0.0068
Coef of Skewness	0.0872
Coef of Kurtosis	2.9292

Gini Index, Sample Size = 50:

Expected Value	0.5027
Variance	0.0156
Coef of Skewness	-0.1638
Coef of Kurtosis	2.8706

Gini Index, Sample Size = 100:

Expected Value	0.5002
Variance	0.0080
Coef of Skewness	-0.1574
Coef of Kurtosis	2.9799

Figure 6: Exact Sampling Distributions for Dissimilarity or Gini Indexes of Segregation (three occupation groups)

distribution. It is worth noting that there are other parameter values that would yield the same index values for the population but might result in somewhat different sampling distributions for D and G. The moments reported in Figure 6 also suggest approximately normal sampling distributions for large sample sizes.

ASYMPTOTIC DISTRIBUTION OF ESTIMATORS

Define $s_i^J = n_i^J / N$, the share of the sample (of both genders) having gender J and occupation i. The asymptotic sampling distributions of D and G follow from the asymptotic distribution of s_i^J. Because s_i^J is the maximum likelihood estimator for π_i^J, s_i^J is consistent and asymptotically normal.

Proposition 1. $S = (s_1^F, s_2^F, \ldots, s_K^F, s_1^M, s_2^M, \ldots, s_K^M)$ is a $2K$-dimensional statistic. The asymptotic distribution of $u = (n^{1/2}(s_1^F - \pi_1^F), \ldots, n^{1/2}(s_K^M - \pi_K^M))$ is a $2K$ variate normal with mean zero and covariance matrix Ω:

$$\Omega = \begin{bmatrix} \Omega_{FF} & \Omega_{FM} \\ \Omega_{FM}^T & \Omega_{MM} \end{bmatrix},$$

where

$$\Omega_{FF} = \begin{bmatrix} \pi_1^F(1-\pi_1^F) & -\pi_1^F\pi_2^F & \mathrm{K} & -\pi_1^F\pi_K^F \\ -\pi_2^F\pi_1^F & \pi_2^F(1-\pi_2^F) & \mathrm{K} & -\pi_2^F\pi_K^F \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ -\pi_K^F\pi_1^F & -\pi_K^F\pi_2^F & \mathrm{K} & \pi_K^F(1-\pi_K^F) \end{bmatrix}$$

$$\Omega_{MM} = \begin{bmatrix} \pi_1^M(1-\pi_1^M) & -\pi_1^M\pi_2^M & \mathrm{K} & -\pi_1^M\pi_K^M \\ -\pi_2^M\pi_1^M & \pi_2^M(1-\pi_2^M) & \mathrm{K} & -\pi_2^M\pi_K^M \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ -\pi_K^M\pi_1^M & -\pi_K^M\pi_2^M & \mathrm{K} & \pi_K^M(1-\pi_K^M) \end{bmatrix}$$

$$\Omega_{FM} = \begin{bmatrix} \pi_1^F \pi_1^M & -\pi_1^F \pi_2^M & \ldots & -\pi_1^F \pi_K^M \\ -\pi_2^F \pi_1^M & \pi_2^F \pi_2^M & \ldots & -\pi_2^F \pi_K^M \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ -\pi_K^F \pi_1^M & -\pi_K^F \pi_2^M & \ldots & -\pi_K^F \pi_K^M \end{bmatrix}.$$

This result follows from the asymptotic distribution of n_i^J (see Johnson and Kotz 1969:288).

The derivation of the asymptotic distributions of D and G from the distribution of s_i^J is straightforward using the δ-method of Rao (1973:388-89).

Proposition 2. Denote $D = h(S) =$

$$\left(\frac{1}{2}\right)\sum_{i=1}^{K}\left|\frac{s_i^F}{\sum\limits_{t=1}^{K} s_t^F} - \frac{s_i^M}{\sum\limits_{t=1}^{K} s_t^M}\right| = \left(\frac{1}{2}\right)\sum_{i=1}^{K}\left|p_i^F - p_i^M\right|.$$

If $p_i^F \neq p_i^M$ for all i, then $N^{1/2}(D - h(\Pi))$ is asymptotically normal distributed with a mean of 0 and covariance $\lambda\Omega\lambda'$, where λ is the vector of the partial derivatives of h with respect to Π: $\lambda = (\partial h(\Pi)/\partial\pi_1^F\ \partial h(\Pi)/\partial\pi_2^F \ldots)$. Thus, the asymptotic distribution of D can be approximated by a normal with mean $h(S)$, the usual "estimated" value for D, and variance $(N^{-1})HVH'$, where V is constructed by replacing the π_i^J of Ω with the corresponding sample statistics, s_i^J, and H is the corresponding "sample" version of λ. Typical elements of H will be of the form

$$\frac{\partial D}{\partial s_j^F} = \frac{1}{2}\left[\left(\frac{1}{S_F}\right)\left[\sum_{i=1}^{K} sign(p_i^F - p_i^M)p_i^F - sign(p_j^F - p_j^M)\right]\right] \quad (4a)$$

$$\frac{\partial D}{\partial s_j^M} = \frac{1}{2}\left[\left(\frac{1}{S_M}\right)\left[\sum_{i=1}^{K} sign(p_i^F - p_i^M)p_i^M - sign(p_j^F - p_j^M)\right]\right], \quad (4b)$$

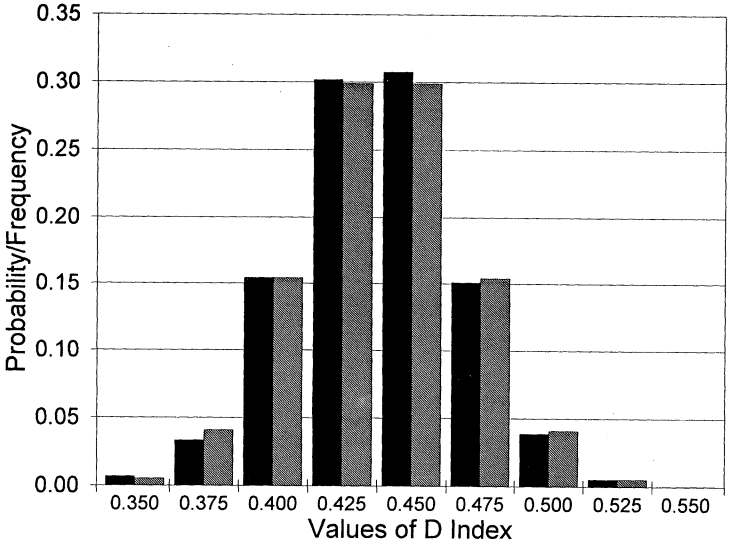


Figure 7: Asymptotic Versus Simulated Distribution

where $S_J = \sum_{t=1}^{K} S_t^J$ and $sign(a) = 1$ if $a > 0$, $sign(a) = -1$ if $a < 0$.

This result follows from Proposition 1 by applying theorems (iii) and (iv) of Rao (1973:388-89). A crucial condition for this result is that $p_i^F \neq p_i^M$, which ensures the differentiability of $h(S)$. The condition need not be satisfied for any particular problem or sample. However, with a sufficiently large sample, the requirement is equivalent to $\pi_i^F \neq \pi_i^M$, which will almost surely be satisfied, since the converse is very strict.

Proposition 3. Denote $G = f(S) =$

$$\left(\frac{1}{2}\right)\sum_{i=1}^{K}\sum_{j=1}^{K}\left|p_i^F p_j^M - p_j^F p_i^M\right|,$$

where $p_i^J = s_i^J / \Sigma s_i^J$. If $p_i^F p_j^M \neq p_i^M p_j^F$ for all $i \neq j$, then $N^{1/2}(G - f(\Pi))$ is asymptotically normal distributed with a mean of 0 and covariance

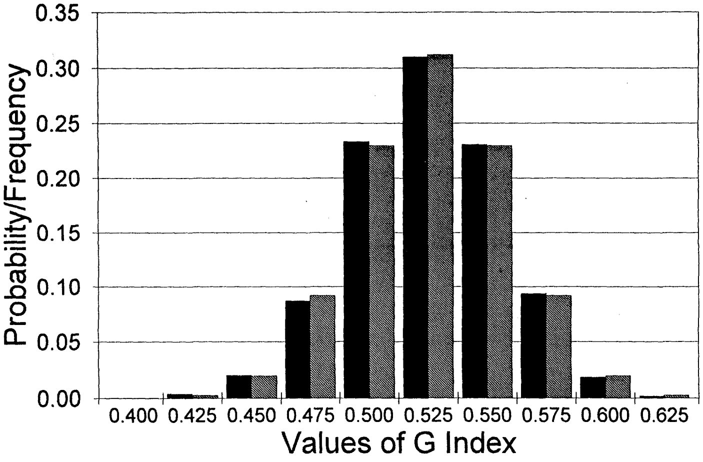


Figure 8: Asymptotic Versus Simulated Distribution

$\gamma\Omega\gamma'$, where γ is the vector of the partial derivatives of f with respect to Π: $\gamma = (\partial f(\Pi)/\partial\pi_1^F \; \partial f(\Pi)/\partial\pi_2^F \; \ldots)$. Practically, the asymptotic distribution of G can be approximated by a normal distribution with mean $f(S)$, the usual estimator of G, and variance $(N^{-1})FVF'$, where V is constructed as in Proposition 2 and F is the sample counterpart of γ. Typical elements of F are

$$\frac{\partial G}{\partial s_j^J} = \frac{\partial G}{\partial p_j^J}\frac{1}{S_J} - \sum_{t=1}^{K}\frac{\partial G}{\partial p_t^J}\frac{p_t^J}{S_J}, J = F, M, \qquad (5a)$$

where

$$\frac{\partial G}{\partial p_t^F} = \left(\frac{1}{2}\right)\left(\sum_{i\neq t}^{K} sign(p_t^F p_i^M - p_i^F p_t^M)p_i^M - \sum_{i\neq t}^{K} sign(p_i^F p_t^M - p_t^F p_i^M)p_i^M\right) \qquad (5b)$$

and

$$\frac{\partial G}{\partial p_t^M} = \left(\frac{1}{2}\right)\left(-\sum_{i\neq t}^{K} sign(p_t^F p_i^M - p_i^F p_t^M)p_i^M + \sum_{i\neq t}^{K} sign(p_i^F p_t^M - p_t^F p_i^M)p_i^F\right). \qquad (5c)$$

The reasoning for this result is identical to that of Proposition 2. Differentiability for the case of the Gini index requires that $p_i^F\, p_j^M$ not equal $p_i^M\, p_j^F$. Another way to express this restriction is that $p_i^F / p_i^M \neq p_j^F / p_j^M$—different occupations cannot have the same female/male ratios. Geometrically, this means that different segments of the segregation curve cannot have the same slope.

SIMULATIONS

Here, I compare the asymptotic distribution of *D* and *G* with the results of a simulation experiment. The case examined is based on the parameters $\pi^F = .175, .15, .075, .1$ and $\pi^M = .025, .075, .1, .3$, with a sample size of 1,500. The corresponding population values for *D* and *G* are .45 and .5375, respectively. For the simulation, 5,000 samples were generated. *D* and *G* were calculated for each sample. In Figures 7 and 8, dark bars show the frequency with which these simulated values fall within the indicated ranges. Corresponding probabilities from the asymptotic distributions (hatched bars) match these simulated sampling distributions closely.

In practice, the most common use of these sampling distributions will be to test whether segregation levels differ across different samples, for example, to test whether the segregation of men and women across occupations changed between the 1980 and the 1990 census. The test statistic for independent samples is

$$t = [D_1 - D_0] / [V(D_1) + V(D_0)]^{1/2}. \tag{6}$$

Under the null hypothesis of no change ($D_0 = D_1$), this will have a standard normal distribution. $V(D_i)$ is the asymptotic variance of D_i, calculated from the sample statistics. D_1 and D_0 are similarly calculated from sample information.

Table 1 reports results of a Monte Carlo experiment. The experiment examines two cases. In each case, the segregation index values of the corresponding populations are the same. Case 1 compares samples from two populations that have a fairly low level of segregation—a dissimilarity index of .25 and a Gini index of .2825. One of the

TABLE 1: Simulation Results: Rejection Frequencies for Tests of Equal Segregation (based on 5,000 replications)

	Case 1		*Case 2*	
Multinomial parameters				
Population 1	$\Pi^F = .07, .06, .04, .03$ $\Pi^M = .48, .16, .12, .04$		$\Pi^F = .07, .06, .04, .03$ $\Pi^M = .04, .16, .12, .48$	
Population 2	$\Pi^F = .175, .15, .10, .075$ $\Pi^M = .30, .10, .075, .025$		$\Pi^F = .175, .15, .10, .075$ $\Pi^M = .025, .10, .075, .30$	
Segregation index values	$D_1 = D_2 = .2500$ $G_1 = G_2 = .2825$		$D_1 = D_2 = .4500$ $G_1 = G_2 = .5525$	
	D	G	D	G
Rejection frequencies (percentage) at the 5-Percent Significance Level				
Sample size				
$N_1 = N_2 = 250$	4.90	4.88	4.46	5.16
$N_1 = N_2 = 1{,}000$	4.30	3.98	4.96	4.84
$N_1 = N_2 = 2{,}000$	5.14	4.70	4.88	4.70

populations consists of 20 percent women and 80 percent men, the other of 50 percent women. Case 2 is similar, but the underlying level of segregation is higher—$D = .45$ and $G = .5375$.

The test statistic described in (6) was calculated for 5,000 samples of pairs of D and pairs of G. Table 1 presents the frequency with which the test statistics lead to rejection of the hypothesis of equal segregation indexes. The rejection region of the test is defined as $|t| > 1.96$, corresponding to a theoretical significance level of 5 percent. As Table 1 shows, the empirical significance level of these tests is very close to the theoretical level of 5 percent when sample sizes are moderately large.[4]

EXAMPLE

This example examines changes in the segregation of men and women across college major. The data used are from two large longitudinal surveys conducted by the U.S. Department of Education. The National Longitudinal Survey of the High School Class of 1972 (NLS72) is a sample of students who graduated from high school in 1972. Data on college major comes from a follow-up survey conducted in 1979. The High School and Beyond (HSB) survey samples students who graduated from high school in 1980. Information on college major comes from a 1986 follow-up survey. The analysis here is based on respondents who graduated from college and held full-time, nonmilitary jobs. Further details are available in Eide (1995), on which this example is based.

Table 2 presents the distribution of college majors of men and women from the two surveys. Columns (1) and (2) report the fraction of men and women in each college major from the NLS72, and columns (3) and (4) report comparable statistics from the HSB survey. Note that these are p_i^J—the corresponding s_i^J can be calculated from p_i^J and the sample sizes.[5] For example, s_1^M for the NLS72 sample is .0849, which is computed from the data for the NLS72 sample reported in columns (1) and (2): $0.164 \times 1252/(1252 + 1166)$. The variance computation for D requires calculation of the vector H and matrix V from Proposition 2. In this example, the vector H has 38 elements (19 fields by 2 sexes) and V is a 38×38 matrix. H is constructed by applying the appropriate data from Table 2 to equation (4a) and (4b). V is likewise fashioned by inserting the appropriate s_i^J into the corresponding positions of the matrix defined in Proposition 1. The variance of G is calculated similarly from the formulas in Proposition 3.

The bottom panel of the table reports the Gini index and the dissimilarity index and their variances. t_D and t_G are test statistics for the hypothesis that segregation by college major among college graduates over this time period has not decreased. The critical value for the (one-tailed) test is 1.645 for a significance of 5 percent. Thus, the Gini index indicates that segregation has declined. The dissimilarity index has also fallen, but the null hypothesis of no decline cannot be rejected at this significance level, although the value of the test is very close to the critical value.

TABLE 2: Tests of Changes in Segregation by College Major Distributions

	National Longitudinal Survey 1972		*High School and Beyond 1984*	
Field of Study	*Men (1)*	*Women (2)*	*Men (3)*	*Women (4)*
Business management	.164	.059	.137	.156
Accounting	.059	.021	.101	.075
Marketing	.021	.007	.054	.046
Engineering	.091	.004	.155	.046
Architecture	.012	.005	.012	.014
Computer science	.010	.003	.062	.027
Mathematics	.009	.020	.034	.009
Physical science	.066	.029	.028	.010
Biological science	.077	.047	.062	.034
Health science	.026	.125	.010	.069
Social science	.164	.158	.097	.125
Psychology	.028	.039	.021	.043
Communications	.029	.024	.062	.055
Public affairs	.014	.022	.017	.013
Education	.093	.235	.041	.180
Foreign language	.002	.024	.006	.002
Letters	.037	.049	.028	.037
Fine arts	.022	.042	.035	.030
Other major	.077	.086	.039	.028
Observations	1,252	1,166	503	670
Gini index	.46857		.39816	
Variance (G)	.000396		.000910	
Dissimilarity index	.33495		.27897	
Variance (D)	.000367		.000799	
$t_G = 1.948$				
$t_D = 1.639$				

CONCLUSION

In this article, I have examined the sampling distributions for two indexes of segregation: the dissimilarity index and the Gini index. I have also derived the asymptotic sampling distribution for these indexes based on a general statistical model. The asymptotic distributions for both the dissimilarity index and the Gini index are normal under conditions that are likely to be satisfied in typical applications.

Simulations confirm that tests based on these asymptotic distributions have the appropriate significance level. This method provides a computationally simple technique to test whether changes in segregation across time or location are statistically significant.

NOTES

1. Massey (1978) also points out that the computational burden of the model renders the method infeasible.

2. This may be a particular problem when segregation indexes are used in the analysis of residential segregation, since the units upon which census data are collected often change.

3. Note the difference in interpretation between p_i^J and π_i^J. For example, p_2^F is the fraction of all females in occupation 2, but π_2^F is the probability of observing a female in occupation 2 from a population of both males and females in all occupations.

4. In some cases, samples were generated that did not satisfy one of the differentiability conditions. Such cases were discarded. For case 1, the total number of discarded replications were 82 for a sample size of 250, 5 for a sample size of 1,000, and 2 for a sample size of 2,000. For case 2, the total number discarded were 52 for a sample size of 250, 4 for a sample size of 1,000, and 1 for a sample size of 2,000.

5. In this example, the sample proportions actually are based on weighted data because of the stratified nature of the sample.

REFERENCES

Albeda, Randy P. 1986. "Occupational Segregation by Race and Gender, 1958-1981." *Industrial and Labor Relations Review* 39:404-11.

Boisso, Dale, Kathy Hayes, Joseph Hirschberg, and Jacques Silber. 1994. "Occupational Segregation in the Multidimensional Case: Decomposition and Tests of Significance." *Journal of Econometrics* 61:164-71.

Cortese, Charles F., R. Frank Falk, and Jack Cohen. 1976. "Further Considerations on the Methodological Analysis of Segregation Indices." *American Sociological Review* 41:630-37.

Duncan, Otis D. and Beverly Duncan. 1955. "A Methodological Analysis of Segregation Indices." *American Sociological Review* 20:210-17.

Eide, Eric. 1995. "Changes in Gender Segregation by College Major and Occupation Among College Graduates." Unpublished manuscript, Department of Economics, Brigham Young University.

Farley, Reynolds and Robert Johnson. 1985. "On the Statistical Significance of the Index of Dissimilarity." Pp. 415-20 in *1985 Proceedings of the Social Statistics Section.* Washington, DC: American Statistical Association.

Hutchens, Robert M. 1991. "Segregation Curves, Lorenz Curves, and Inequality in the Distribution of People Across Occupations." *Mathematical Social Sciences* 21:31-51.

Jacobs, Jerry A. 1995. "Gender and Academic Specialties: Trends Among Recipients of College Degrees in the 1980s." *Sociology of Education* 66:81-98.

James, David R. and Karl E. Taeuber. 1985. "Measures of Segregation." Pp. 1-32 in *Sociological Methodology 1985*, edited by Nancy B. Tuma. San Francisco: Jossey-Bass.

Johnson, Norman L. and Samuel Kotz. 1969. *Discrete Distributions*. New York: Wiley.

Massey, Douglas S. 1978. "On the Measurement of Segregation as a Random Variable." *American Sociological Review* 43:587-90.

Ransom, Michael R. 1990. "Gender Segregation by Field in Higher Education." *Research in Higher Education* 31:477-94.

Rao, C. R. 1973. *Linear Statistical Inference and Its Applications*. 2d ed. New York: Wiley.

Taeuber, Karl E. and Alma F. Taeuber. 1972. *Negroes in Cities: Residential Segregation and Neighborhood Change*. New York: Atheneum.

Zoloth, Barbara. 1976. "Alternative Measures of School Segregation." *Land Economics* 52:278-98.

Michael R. Ransom is a professor of economics at Brigham Young University, where he teaches econometrics and microeconomic theory. His research focuses on the labor market, including issues related to male/female differences in pay and the labor market for highly educated workers.