Appendices

Appendix A: Summary of Notation and Conventions

This appendix reviews the notation and conventions for terms used in this monograph. Where appropriate it provides commentary to clarify usage by context.

Pairwise Calculations

In standard applications, indices of uneven distribution are based on pairwise population counts and group proportions. The adjective “pairwise” indicates that calculations use only population counts for the two groups in the segregation comparison. If other groups are present in the population, their counts are excluded and have no impact on index scores. Accordingly, unless indicated by direct statement or by obvious context, references here to total counts and terms based on total counts (e.g., group proportions) should be taken as being based on pairwise comparisons; that is, based on the sum of the population counts for just the two groups in the comparison.

Reference and Comparison Groups (Groups 1 and 2)

When index scores are calculated using the difference of means formulation introduced in this monograph it is necessary to designate one of the two groups in the segregation comparison as the “reference” or “focal” group. The second group is then designated the “comparison” group. The choice of which group is designated as the “reference” is arbitrary and it has no impact on the resulting index scores. The choice is necessary to organize calculations and facilitate presentation. For subscripting purposes it is convenient to designate the reference group as “Group 1” and the comparison group as “Group 2”.

M. Fossett, *New Methods for Measuring and Analyzing Segregation*,
The Springer Series on Demographic Methods and Population Analysis 42,
DOI 10.1007/978-3-319-41304-4

The empirical literature on residential segregation in US urban areas overwhelmingly focuses on majority-minority segregation comparisons such as White-Black, White-Latino, and White-Asian comparisons. Based on substantive concerns regarding majority-minority inequality and assimilation, it is customary to assess residential distributions for different minority groups – e.g., Blacks, Latinos, and Asians – in relation to the residential distribution of the majority group – Whites. I follow this custom and thus designate the majority group – Whites – as the reference group.

This has no consequence for index scores or for their substantive implications. But it does structure discussion and interpretation of results to focus on implications for majority-minority inequality and residential assimilation.

City-Wide Terms for Pairwise Calculations

$N_1 =$ the city-wide population count for Group 1, the "reference" or "focal" group.
$N_2 =$ the city-wide population count for Group 2, the "comparison" group.
$T =$ the combined city-wide pairwise population count ($T = N_1 + N_2$).
$P =$ the city-wide proportion for Group 1 ($P = N_1/[N_1 + N_2]$).
$Q =$ the city-wide proportion for Group 2 ($Q = N_2/[N_1 + N_2]$; $Q = 1 - P$).

Area-Specific Terms for Pairwise Calculations

$i =$ index for the areas of the city; applied where appropriate, omitted to reduce clutter when unnecessary (e.g., when clear based on context).
$j =$ a second index for the areas of the city used in formulas where one area (denoted by i) is compared to other areas (denoted by j).
$n_1 =$ the area population count for Group 1, the reference group.
$n_2 =$ the area population count for Group 2, the comparison group.
$t =$ the combined area pairwise count ($t = n_1 + n_2$).
$p =$ the area proportion for Group 1 ($p = n_1/[n_1 + n_2]$).
$q =$ the area proportion for Group 2 ($q = n_2/[n_1 + n_2]$; $q = 1 - p$).
$s_1 =$ the area share of the city-wide Group 1 population ($s_1 = n_1/N_1$).
$s_2 =$ the area share of the city-wide Group 2 population ($s_2 = n_2/N_2$).

Terms for Individuals or Households

$k =$ an index for individuals in a group or, depending on context, in the city-wide population.

m = an index similar to k for individuals in a group or in the city-wide population. This is relevant for some formulas for the Gini Index (G) where individuals indexed by k are compared to all other individuals in the population indexed by m.

Selected Terms and Conventions Relevant for the Gini Index (G)

X_i = cumulative proportion of Group 1 based on ordering areas from low to high on p_i and then summing area group share terms ($X_i = \Sigma s_1$ over relevant areas).
Y_i = cumulative proportion of Group 2 based on ordering areas from low to high on p_i and then summing area group share terms ($Y_i = \Sigma s_2$ over relevant areas).

Selected Terms and Conventions Relevant for the Theil Entropy Index (H)

The original derivation of the Theil index is grounded in an information theory framework (Shannon 1948; Theil and Finizza 1971; Theil 1972) drawing on a notion of entropy (E) quantified as given below.

E = entropy for the city overall given by $E = P \cdot \mathrm{Log}_2\left(1/P\right) + Q \cdot \mathrm{Log}_2\left(1/Q\right)$.

E_i = entropy for area i given by $E_i = p_i \cdot \mathrm{Log}_2\left(1/p_i\right) + q_i \cdot \mathrm{Log}_2\left(1/q_i\right)$.

Note that Log_2 denotes the base 2 logarithm. Many applications use natural logarithms in place of base 2 logarithms.

Selected Terms and Conventions Relevant for the Atkinson Index (A)

Formulas for the Atkinson index (A) include two constants – α and β. Values for α are restricted to fall between 0 and 1 exclusive of end points (i.e., $0 < \alpha < 1$). β is obtained by $1-\alpha$. The Atkinson index is symmetric when α is 0.5 and is asymmetric otherwise. When A is asymmetric it yields different index values depending on which of the two groups in the comparison is adopted as the reference group in the comparison. This leads some to view asymmetric versions of A as unacceptable for use as a general measure of segregation (White 1986). I agree with this view. Accordingly, discussion of the Atkinson index in this monograph is limited to the symmetric version where $\alpha = \beta = 0.5$. This version of the Atkinson index has close relations with the Hutchens square root index (R) which is more tractable mathematically.

Appendix B: Formulating Indices of Uneven Distribution as Overall Averages of Individual-Level Residential Outcomes

This appendix chapter reviews alternative formulations of indices of uneven distribution to clarify how aggregate segregation is related to individual residential outcomes. This is useful for at least two inter-related reasons; one substantive and one methodological. The substantive reason is that sociological interest in segregation usually rests on the assumption that it has important implications for individual life chances associated with area of residence. Based on this concern, it would be useful to better understand how indices of uneven distribution register individual residential outcomes. The methodological reason is that formulating indices of uneven distribution in terms of individual residential outcomes is a necessary step for clarifying how segregation emerges from individual-level residential attainment processes.

The view that segregation emerges from micro-level attainment processes and carries important implications for group differences in residential outcomes is hardly new or controversial. In light of this it is surprising that methodological discussions of indices of uneven distribution give little attention to this issue. For example, consider two familiar formulas for the widely used Gini Index (G) and the Delta or Dissimilarity Index (D) shown in Fig. B.1.[1] These formulas were featured five decades ago in Duncan and Duncan's (1955) landmark methodological study. These formulas and close variations on them are widely used in empirical studies in part because they are computationally efficient and are easy to implement. However, Duncan and Duncan raised the concern that "[i]n none of the literature on segregation indices is there a suggestion of how to use them to study the *process* of segregation" (1955:216, emphasis in original). The reason for this is that the formulas given in Fig. B.1 provide little basis for understanding how segregation is connected to the residential outcomes of individuals. Indeed, individual-level residential outcomes are "invisible" in these formulas.

Advances in computing technology have rendered the issue of computing efficiency mostly irrelevant. Yet it is still typical for the measurement of segregation using G, D, and other indices to be discussed in relation to convenient computing formulas. It is fine to use efficient computing formulas for the narrow purpose of obtaining index values. But researchers and broad audiences who gain their understanding of segregation based solely on these formulas will have, at best, only vague notions regarding how segregation arises from micro-level attainment processes. This problem can be addressed by considering alternative formulations of popular segregation indices that clarify how index scores are connected to individual residential outcomes.

[1] Figure B.1 also includes a similar style formula for the more recently introduced Hutchens square root index (R) (2001).

$G = 100 \cdot (\Sigma X_{i-1}Y_i - \Sigma X_iY_{i-1})$ where X and Y are group proportions cumulated over areas ranked low to high on p_i (Duncan and Duncan 1955)

$D = 100 \cdot ½\, \Sigma \,|\, (n_{1i}/N_1) - (n_{2i}/N_2) \,|$ (Duncan and Duncan 1955)

$R = 100 \cdot \left(1.0 - \Sigma \sqrt{(n_{1i}/N_1) \cdot (n_{2i}/N_2)}\right)$ (Hutchens 2001:23)

Fig. B.1 Area-based computing formulas for indices of uneven distribution that do not draw on individual-level residential outcomes (Note: N denotes city-wide population count, n denotes area population count, subscripts 1 and 2 denote the two groups in the segregation comparison, subscript i denotes area, X_i and Y_i denote the cumulative proportions of groups 1 and 2 over areas ranked from low to high on p_i – the group 1 (reference group) proportion in the combined group population in area i ($p_i = n_{1i}/[n_{1i} + n_{2i}]$))

Focusing Attention on Individual-Level Residential Outcomes

All widely used indices of uneven distribution can be formulated in terms of individual-level residential outcomes (y) that are scored from area group (e.g., racial) proportions (p). This can be done in two distinct ways. One is to formulate index scores as simple *overall averages* of individual-level residential outcomes (y). The other is to formulate index scores as a *difference of group means* on individual-level residential outcomes (y). Both approaches can be used to obtain "correct" index values. But that is a minor benefit as convenient formulas for obtaining correct index values are readily available. The main benefit of these formulations is that they can be used to gain insight into how different indices register and summarize individual residential outcomes. In addition, formulating indices in terms of individual attainments brings certain practical advantages which I note below.

Figure B.2 presents computing formulas that highlight how individual-level residential outcomes are registered by six popular measures of uneven distribution – the Gini Index (G), the Delta or Dissimilarity Index (D), the Atkinson Index (A), the Hutchens Square Root Index (R), the Theil Entropy index (H), and the Separation Index (S) (also known as the variance ratio [V], and eta squared [η^2]). The calculations indicated in these formulas involve first computing area-specific scores (i.e., neighborhoods) based on pairwise group proportions and then averaging these scores over individuals. More specifically, the formulas have the following features:

- the core terms in the calculations are scores computed for areas (indexed here by "i") based on calculations involving area group proportions; that is involving the values of p_i and q_i as given in Appendix A,
- the area-specific scores are summed over all *individuals* based on weighting the score for each area by the area-specific combined population count (t) for the two groups in the segregation comparison,
- the population-weighted sum of area-specific scores is then divided by the combined population of the two groups for the city (T) to obtain an overall average, and

$G = 100 \cdot (1/2T^2PQ) \cdot \Sigma\Sigma\, t_i \cdot t_j \, |p_i - p_j|$

$100 \cdot (1/2TPQ) \cdot \Sigma\, t_i \cdot [(1/T) \cdot \Sigma\, t_j \, |p_i - p_j|]$ (noted to clarify area-specific term)

$D = 100 \cdot (1/2TPQ) \cdot \Sigma t_i \, (|p_i - P|)$ (noted to highlight similarities with S)

$A = 100 \cdot [1 - (Q/P) \{\Sigma\, t_i \cdot (p_i^{\alpha} q_i^{\beta}/QT)\}^{1/\alpha}]$ where $0 < \alpha < 1$ and $\beta = 1 - \alpha$

Setting $\alpha=\beta=0.5$ yields the "symmetric" version of A, the version most relevant for use in segregation analysis. Using this setting for α, the formula of A can be expressed in the two formulations shown below to highlight similarities with formulas for the Hutchens square root index (R) and the separation index (S).

$100 \cdot \left[1 - \left\{\Sigma\, (t_i/T) \sqrt{p_i q_i}\right\}^2 / PQ\right]$ (noted to highlight similarities with R & S)

$100 \cdot \left[1 - \left\{(1/T) \cdot \Sigma\, t_i \cdot \sqrt{p_i q_i}\right\}^2 / PQ\right]$ (noted to highlight similarities with R & S)

$R = 100 \cdot \left[1.0 - \Sigma (t_i/T) \sqrt{p_i q_i / PQ}\,\right]$ (noted to highlight similarities with A & S)

$100 \cdot \left[1.0 - (1/T)\, \Sigma\, t_i \sqrt{p_i q_i / PQ}\,\right]$ (noted to highlight similarities with A & S)

$H = 100 \cdot \Sigma\, [(E - E_i)/E] \cdot (t_i/T)$

$100 \cdot (1/T) \cdot \Sigma\, t_i \cdot [(E - E_i)/E]$

where E is entropy for the city overall given by $E = P \cdot \mathrm{Log}_2(1/P) + Q \cdot \mathrm{Log}_2(1/Q)$ per information theory (Shannon 1948; Theil 1972) and E_i is entropy for area i and is given by $E_i = p_i \cdot \mathrm{Log}_2(1/p_i) + q_i \cdot \mathrm{Log}_2(1/q_i)$. If desired, one can use natural logarithms as well as base 2 logarithms.

$S = 100 \cdot (1/TPQ) \cdot \Sigma t_i \, (p_i - P)^2$ (noted to highlight similarities with D)

$100 \cdot [1 - \Sigma\, (t_i/T)(p_i q_i/PQ)]$ (noted to highlight similarities with A & R)

$100 \cdot [1 - (1/T) \cdot \Sigma\, t_i \cdot (p_i q_i/PQ)]$ (noted to highlight similarities with A & R)

Fig. B.2 Area-based computing formulas for indices of uneven distribution that implicitly feature averages for individual-level residential outcomes

- any other terms present in the formula serve only to rescale the resulting overall average to the range 0–1.

Based on these features, it is appropriate to describe the resulting index value as an overall individual-level average on area-specific residential outcomes scored from area group composition (p).

Figure B.3 reorganizes the expressions in Fig. B.2 to present them in a form that explicitly casts each index in terms of an index-specific, individual-level residential outcome (y) that is averaged over all individuals in the two groups in the comparison. The formulas in this figure are not necessarily the most convenient for computing index scores. But they make it clear that aggregate segregation index scores can be understood as simple summary measures (i.e., means) for individual residential outcomes.

The individual level residential outcomes (y) identified in Fig. B.3 can be characterized as follows: the outcomes register the degree to which the group proportion for the area (p_i) departs from the group proportion for the city as a whole. The specific way in which this departure is quantified varies from one index to another and that becomes the basis for each one's unique way of registering uneven distribution.

	Averaging Scores for y Over Individuals	Scores Assigned to Individuals
G =	$100 \cdot (1/T) \cdot \Sigma y_k$	$y_k = \Sigma \mid p_k - p_m \mid / 2TPQ$ where k and m index individuals, p_k denotes the pairwise area proportion for the reference group (p_i) for the k'th individual, p_m denotes area proportion for the reference group (p_i) for the m'th individual (note, this reorganizes the terms in the second formula for G in Figure B.2)
D =	$100 \cdot (1/T) \cdot \Sigma y_k$	$y_k = \mid p_i - P \mid / 2PQ$
A =	No comparable solution is available but the value of the "symmetric" version of A (given by setting $\alpha=\beta=0.5$) can be obtained from $2R\text{-}R^2$	
R =	$100 \cdot [1 - (1/T) \cdot \Sigma y_k]$	$y_k = \sqrt{p_i q_i / PQ}$
H =	$100 \cdot (1/T) \cdot \Sigma y_k$	$y_k = (E - E_i)/E$ with E_i and E as given in Figure B.2
S =	$100 \cdot (1/T) \cdot \Sigma y_k$ or $100 \cdot [1 - (1/T) \cdot \Sigma y_k]$	$y_k = (p_i - P)^2/PQ$ $y_k = p_i q_i / PQ$

Fig. B.3 Alternative formulas for uneven distribution that explicitly cast indices as overall averages of residential outcomes (y) for individuals (Note: k and m index individuals, p_k denotes the pairwise area proportion for the reference group (p_i) for the k'th individual, p_m denotes area proportion for the reference group (p_i) for the m'th individual)

But all of the indices can be understood as registering average exposure to departures from the group mix that would obtain under even distribution. If all neighborhoods have the group mix of the city as a whole, all of the values of y will be 0 and the final index value also will be 0. If members of the two groups never reside in the same areas, the values of y move to the extreme values that can apply to individuals residing in neighborhoods where p_i is 1 or 0 and the sum of y goes to the maximum value possible for the city given its group composition. The resulting sum is then rescaled to yield an index value of 1 by incorporating index-specific constant terms (e.g., 2PQ for D).

Options for Spatial Versions of Indices of Uneven Distribution

These index formulations carry at least one practical benefit; they can be used to calculate spatial segregation scores as well as aspatial segregation scores for any of the indices. That is,

> Formulas that cast segregation index values as overall averages on individual-level residential outcomes can readily be adapted for computing *spatial* as well as *aspatial* versions of the segregation indices.

Aspatial versions of segregation indices are familiar and widely used in empirical studies. They are obtained by applying the computing formulas introduced here, or any of the formulas introduced earlier, using data for non-overlapping "bounded" areas such as school districts, census tracts, block groups, or blocks. In the aspatial formulation, each bounded area represents a particular neighborhood and every individual or household in the area is treated as having the residential outcome calculated for this area.

When index values are cast as overall averages of individual-level residential outcomes as in Fig. B.3, the indices also can be implemented in spatial measures. This is accomplished by computing averages for individual residential outcomes (y) that scored for "overlapping" spatially-defined neighborhoods that are specified uniquely for each individual based on the population residing within a spatially defined neighborhood. For example, the spatial formulation could be implemented using census data by taking small bounded areas such as census blocks and defining the spatial neighborhood as the population residing in the "focal" block plus the surrounding adjacent blocks. In this approach the population in any particular block will be part of uniquely-defined, spatially-delimited neighborhood.

When using these formulas, the question of whether the index is viewed as aspatial or spatial depends only on how "neighborhoods" are conceived. This can be stated in general terms as follows. Whether or not the index values obtained using these formulas are properly described as spatial or aspatial is determined by the definitions of the neighborhoods used to calculate the individual-level residential outcomes used in the relevant index calculations. If the residential outcomes are for non-overlapping bounded areas, the index values are aspatial. If the residential outcomes are for individual-specific, overlapping neighborhoods, then the index values are spatial.

Summary of Difference of Means Formulations

I now review a second way in which indices of uneven distribution can be formulated in terms of individual-level residential outcomes. This is to cast each index as a difference of group means on individual-level residential outcomes. Groups are designated as groups 1 and 2 with group 1 being taken as the reference group.[2] Each segregation index value (S) is then given as the difference of group means ($\overline{Y}_1 - \overline{Y}_2$) on individual residential outcomes (y) that are scored as a function of the pairwise proportion for group 1 in the area in which the individual resides (i.e., $\mathrm{y} = \mathrm{f}\left(\mathrm{p}\right)$).

Figure B.4 gives formulas for calculating values of popular segregation indices in this manner. My intent here is only to introduce formulas that place popular indices of uneven distribution in the general "difference of group means" framework. Appendices C-F provide detailed discussions of the mathematical basis for the formulas given here. The body of the monograph provides a more general discussion of this new measurement approach and the benefits associated with adopting it.

[2] The choice of which group serves as the reference is arbitrary in the sense that the index score obtained is the same either way.

Index Formulated as a Difference of Means	Residential Outcome Scores (y) Assigned to Individuals Based on y = f(p)
$G = 100 \cdot 2(\overline{Y}_1 - \overline{Y}_2)$	$y_i = f(p_i) =$ relative rank (quantile scoring) on p_i
$D = 100 \cdot (\overline{Y}_1 - \overline{Y}_2)$	$y_i = f(p_i) = 0$ if $p_i < P$, 1 if $p_i \geq P$
Alternatively, compute D as a simplified version of G based on collapsing area values for p_i into a two-category rank scheme consisting of areas where $p_i < P$ and areas where $p_i \geq P$.	
A = No direct solution is yet found but $A = 2R{-}R^2$ for the "symmetric" version of A given by on setting $\alpha{=}\beta{=}0.5$.	
$R = 100 \cdot (\overline{Y}_1 - \overline{Y}_2)$	$y_i = Q + \left(1 - \sqrt{p_i q_i / PQ}\right) / (p_i/P - q_i/Q)$
$H = 100 \cdot (\overline{Y}_1 - \overline{Y}_2)$	$y_i = Q + [(E{-}e_i)/E] / (p_i/P - q_i/Q).$
$S = 100 \cdot (\overline{Y}_1 - \overline{Y}_2)$	$y_i = p_i$

Fig. B.4 Formulas casting indices of uneven distribution (S) as group differences of means ($\overline{Y}_1 - \overline{Y}_2$) on individual residential outcomes (y) (Note: p_i denotes the pairwise area proportion for the reference group (p_i) in the area where individual i resides and y_i is the residential outcome score generated by the index-specific scoring function $f(p_i)$)

For the moment I note that the approach is attractive on conceptual grounds because these formulas clarify that segregation indices measure whether groups to experience similar or different averages on specific residential outcomes. Additionally, the formulas reveal that differences between indices arise from a single source; the specific nature of the scaling function $y = f(p)$ that scores residential outcomes (y) from values of area group proportion (p). Area group proportion (p) reflects simple group contact or exposure in its original or "natural" metric. The scoring function $y = f(p)$ rescales group contact and maps it onto an alternative scaling metric for residential outcomes (y) specific to the index in question. From this perspective all popular indices of uneven distribution register group differences of means on "scaled" pairwise group contact.

Appendix C: Establishing the Scaling Functions $y = f(p)$ Needed to Cast the Gini Index (G) and the Dissimilarity Index (D) as Differences of Group Means on Scaled Pairwise Contact

This is the first of several appendix chapters which establish how popular indices of uneven distribution can be placed in the "difference of group means" framework. The feature of this framework is that the values of each index are obtained as a simple difference of group means on individual residential outcomes (y) that are scored from to 0 to 1 based on area group proportion (p) computed from pairwise population counts. Taking the familiar example of White-Black segregation, area group proportion (p) can be set to proportion White of the combined White and Black population in the area; that is, $p = w/(w + b)$ where w and b are the counts of Whites and Blacks, respectively, in the area.[3] Residential outcome scores (y) are then obtained from an index-specific scaling function $y = f(p)$ that takes values of p that range from 0 to 1 and rescales them to new values that also range from 0 to 1. The segregation index score is then obtained from the difference ($Y_W - Y_B$) where Y_W and Y_B are the group means for Whites and Blacks, respectively, on residential outcomes (y).

For individuals, p registers simple pairwise "contact" or "exposure" to the reference group based on residing in a given area. In the example under consideration the reference group is Whites and p thus registers "contact with" or "exposure to" Whites. The residential outcome score (y) can be described as "scaled pairwise contact" or "scaled pairwise exposure". Accordingly, the segregation index score can be described as a difference of group means on scaled contact; in the example under consideration, it is the White-Black difference in average scaled contact with Whites.

The General Task

The key to placing a particular index of uneven distribution in the difference of means framework is to identify a scaling function $y = f(p)$ that accomplishes the goal of scoring residential outcomes (y) from area group proportions (p) such that the scores for y fall over the range 0–1 and yield the value of the index of interest as a difference of means on y for the two groups in the segregation comparison. I have identified scaling functions meeting these criteria for all popular indices of uneven distribution including: the gini index (G), the delta or dissimilarity index (D), the Hutchens square root index (R), the Theil entropy index (H) and the separation index (S). Placing these various indices in the difference of means framework gives them a common basis for interpretation and a specific basis for comparison. The common basis for interpretation is that all indices measure White-Black differences in average scaled contact with Whites. The specific basis for comparison is that the differences between index scores arise ***solely*** from differences in how index-specific scaling functions $y = f(p)$ map values of pairwise contact from its original or "natural" metric based on area group proportion (p) onto values of residential outcomes (y).

The main task of this appendix chapter and the ones that follow it is to establish the particular scaling function $y = f(p)$ that will yield the value of the index in question. The general way task is to start with a generic expression of the difference of means formulation.

[3] Alternatively, p can be set to area proportion Black. The choice is arbitrary as the index score is the same either way.

$$\text{Difference of Means Formula} = \left(\mathrm{Y_W} - \mathrm{Y_B}\right) = \left(1/\mathrm{W}\right)\cdot\Sigma\, \mathrm{w_i y_i} - \left(1/\mathrm{B}\right)\cdot\Sigma\, \mathrm{b_i y_i}$$

Then equate this formula to a standard formula for the index of interest and then manipulate the full expression to obtain a solution for y. In this appendix chapter and the ones that follow it I review steps that accomplish this task and establish a basis for an index specific scaling function $y = f(p)$ relevant for G, D, R, H, and S.

I expect that many readers will not be especially interested in the derivations of the relevant scaling functions. With this in mind, I presented only the final formulas in the main body of this monograph and in the overview discussion just provided in Appendix B. Readers who are not interested in the details of these derivations can rely on these earlier presentations and skip the remainder of this chapter and the additional appendix chapters that follow. For those who elect to slog through the technical details, I thank you in advance for your patience and forbearance. I claim only that the derivations accomplish what is needed and apologize for the fact that they are tedious and inelegant.

Introducing the Function $y = f(p)$ for the Gini Index (G)

For the Gini Index (G) the relevant scaling function $y = f(p)$ is relatively simple; it is the quantile (percentile) or *relative rank* transformation.

$$\mathrm{y} = \text{quantile(p), or, more exactly}$$
$$\mathrm{y} = 2\cdot\text{quantile}(\mathrm{p}).$$

Under this scaling approach, households are assigned values on residential outcomes (y) based on the population-weighted relative rank position of their area of residence on area group proportion (p); more specifically, the quantile score on p for individuals.

I review the quantile scaling function in more detail below. For the moment I note briefly that the scaling function $y = f(p)$ for G is a continuous, monotonic, nonlinear transformation of p that changes p from its original or "natural" metric to a new scaling metric. The nonlinear transformation produces a curve that tends to rise faster when p is low and when p is high and tends to rise more slowly when p is in the middle ranges. As a result, the scaling transformation serves to exaggerate group differences on p over portions of the lower and upper ranges of the scale of p (i.e., $p < 0.25$ and $p > 0.75$) while compressing group differences on p over middle portions of the range of p (i.e., $0.30 < p < 0.70$). Thus, the quantile transformation can and often does change small quantitative differences between Whites and Blacks on p into large differences on rank-order quantile scores. This in turn makes average White-Black differences on y larger than average White-Black differences on p. The tendency is moderate when groups are approximately equal in size. It becomes more and more pronounced when groups become increasingly unequal in size.

As formulated for the difference of group means framework, the Gini Index (G) for White-Black segregation can be given by

$$\begin{aligned} &\mathrm{Y_W} - \mathrm{Y_W} = \mathrm{G}/2,\ \text{or} \\ &\left(\mathrm{Y_W} - \mathrm{Y_B}\right)/0.5 = 2\left(\mathrm{Y_W} - \mathrm{Y_B}\right) = \mathrm{G} \end{aligned} \tag{C.1}$$

for $\mathrm{y} = \mathrm{quantile}(\mathrm{p})$, or, alternatively, for $\mathrm{y} = 2 \cdot \mathrm{quantile}(\mathrm{p})$,

$$\left(\mathrm{Y_W} - \mathrm{Y_B}\right) = \mathrm{G}. \tag{C.1a}$$

In this formulation residential outcomes (y) register each household's *relative rank* position on area proportion White (p), $\mathrm{Y_W}$ is the mean on y for White households, and $\mathrm{Y_B}$ is the mean on y for Black households. One way to describe the formulation is that the value of G is the observed difference of group means on quantile scores for p divided by 0.5, the maximum value possible when scoring y as quantile scores. Alternatively, if y is scored as twice the quantile score (i.e., $2 \cdot \mathrm{quantile(p)}$), G is the simple difference of means.[4]

G Is a Measure of Rank Order Inequality on Contact

Surprisingly, methodological reviews of segregation indices rarely make, much less emphasize, the point that the Gini Index (G) assesses uneven distribution in terms of group differences in rank order standing on area group proportion scores (p). This quality of G has been noted in methodological studies that review the application of G as a measure of inter-group inequality on ordinal variables. Lieberson (1976) introduced a measure of inter-group inequality on ordinal outcomes which he termed the index of net difference (ND). He characterized ND as being "analogous" to G (1976:281). Fossett and South (1983) noted that ND and G are more than analogous; they are mathematically equivalent (this is established in expressions (C.2a) and (C.2b) below). Accordingly, ND can be characterized as an alternative computing formula for G that supports an explicit and potentially attractive substantive interpretation in terms of group difference in rank advantage.

This provides an initial basis for interpreting G for White-Black segregation as an index of relative rank difference between Whites and Blacks in their distribution on residential contact with Whites (p). Specifically, in the ND formulation, the value of G is the difference of two probabilities; (a) the probability that a randomly chosen White will have greater residential contact with Whites than will a randomly

[4] Under maximum uneven distribution all Whites live in neighborhoods that are 100% White and all Blacks live in neighborhoods that are 100% Black. Their respective average quantile scores on area proportion White will be $1 - \mathrm{P}/2$ for Whites and $\mathrm{Q}/2$ for Blacks. The group (White-Black) difference of means will be $(1 - \mathrm{P}/2) - \mathrm{Q}/2$ which resolves to $1 - \left(\mathrm{P}/2 + \mathrm{Q}/2\right) = 1 - \left(\mathrm{P} + \mathrm{Q}\right)/2 = 1 - 1/2 = 0.5$.

chosen Black, and (b) the probability that a randomly chosen Black will have greater residential contact with Whites than will a randomly chosen White.

Fossett and South (1983:861) note that the value of ND, and therefore G, can be obtained from the following computing formula

$$\mathrm{ND} = \mathrm{G} = \Sigma_i \Sigma_j x \cdot (w_i/W)(b_j/B)$$

where i and j index areas ranked on area proportion White (p), and x is scored: 1 if $(i>j)$, 0 if $(i=j)$, and −1 if $(i<j)$. This formula highlights that G responds solely to White-Black comparisons on rank order standing on area proportion White (p). Thus, it gives insight into why G is insensitive to the quantitative magnitude of group differences on p; G treats all White-Black differences on p as either 1 or −1, regardless of the difference involved is large or small.

Fossett and Siebert (1997, Appendix A) also explore the formulation of G as a measure of inter-group inequality on ranked outcomes. They showed that G is a special case of Somers' d_{yx}, a measure of ordinal (rank-order) association. Consequently, G can be interpreted as an ordinal slope coefficient that indicates the impact of race (i.e., group membership) on the rank order standing of individuals on residential contact with Whites (p). Of more direct relevance for the present discussion, Fossett and Siebert also noted that the value of G can be given as twice the difference of group means on percentile (or quantile) scores for ranked outcomes. In application to White-Black segregation this means that G registers the White-Black difference of means on quantile scores for contact with Whites (p).

Calculating G as a Difference of Means

The procedure for obtaining the value of G for White-Black segregation as a difference of means on residential outcomes (y) can be given as follows. First implement the relative rank scoring function $y = f(p)$ by ordering areas from low to high based on values of area proportion White (p_i).[5] Note that p_i is calculated using only counts for Whites and Blacks (i.e., $p_i = w_i/(w_i + b_i)$). Designate the number of households in the area ranked lowest on area proportion White (p_1) by t_1 based on $t_1 = w_1 + b_1$ where w_1 and b_1 are the counts for Whites and Blacks, respectively, in the area. Then calculate the average relative rank position (y_1) on area proportion White (p_1) for households in this area as $y_1 = (t_1/2)/T$ where T is the combined population of Whites and Blacks in the city based on $T = W + B$. The calculation reflects the fact that households in this area occupy ranks 1 through t_1 on area proportion White (p) and so they all are assigned the average for this range of relative rank positions. The number of households in the area ranked next lowest on area

[5] Areas that are identical on area proportion White (p) can be combined and treated as single areas, or they can be handled separately. There is no practical difference as the average score for y will be the same either way.

proportion White (p_2) is designated by t_2. The average relative rank position (y_2) for these households on area proportion White (p) is $\left[t_1 + \left(t_2 / 2\right)\right] / T$, the average for the relative rank position for households in the area. Continue with this procedure until all areas are scored on y.

The resulting White-Black difference of means on y is then given by

$$Y_W - Y_B = \Sigma w_i y_i / W - \Sigma b_i y_i / B.$$

This result takes a value equal to G/2.

Deriving G as a Difference of Means

The next several sections establish that the difference of means formulation of the Gini Index (G) maps exactly onto the usual computing formulas for G. Unfortunately, the discussion is long and tedious. Readers who are not interested in these details should skip forward to the section that discusses the differences of means formulation of the Dissimilarity Index (D).

Specifying Some Useful Terms and Relationships

To begin, it is helpful to introduce several terms and establish certain relationships among them. I start by introducing the following three terms:

$pt_i = t_i / T$, this term registers the i'th area's proportion (share) of the city's combined population of Whites and Blacks,
$pw_i = w_i / W$, this term registers the i'th area's proportion (share) of the city's White population, and
$pb_i = b_i / B$, this term registers the i'th area's proportion (share) of the city's Black population.

When calculating G the areas of the city are ordered from lowest to highest value on area proportion White (p). This leads to the following terms

$cpt_i = \Sigma pt_i = \Sigma t_i / T$, cumulative proportion (share) of the city's combined population of Whites and Blacks residing in areas ranked 1 through i on area proportion White (p),
$cpw_i = \Sigma pw_i = \Sigma w_i / W$, cumulative proportion (share) of the city's White population residing in areas ranked 1 through i on area proportion White (p), and
$cpb_i = \Sigma pb_i = \Sigma b_i / B$, cumulative proportion (share) of the city's Black population residing in areas ranked 1 through i on area proportion White (p).

These terms can be used to give the familiar computing formula for G introduced by Duncan and Duncan (1955: 211) as

$$G = \Sigma pw_i \cdot \Sigma pb_{i-1} - \Sigma pb_i \cdot \Sigma pw_{i-1}. \tag{C.2}$$

This can be restated with alternative notation as

$$G = \Sigma\left(cpw_i \cdot cpb_{i-1}\right) - \Sigma\left(cpb_i \cdot cpw_{i-1}\right). \tag{C.2a}$$

Recognizing that $\left(cpw_i \cdot cpb_{i-1}\right) = \left(pw_i \cdot cpb_{i-1}\right) + \left(cpw_{i-1} \cdot cpb_{i-1}\right)$, and that $\left(cpb_i \cdot cpw_{i-1}\right) = \left(pb_i \cdot cpw_{i-1}\right) + \left(cpb_{i-1} \cdot cpw_{i-1}\right)$, (C.2a) can be restated as

$$G = \Sigma\left(pw_i \cdot cpb_{i-1}\right) - \Sigma\left(pb_i \cdot cpw_{i-1}\right) \tag{C.2b}$$

Expressions (C.2), (C.2a), and (C.2b) are mathematically equivalent variations of the standard computing formula for G. Expression (C.2a) corresponds to the traditional computing formulas for G given in Duncan and Duncan (1955). Expression (C.2b) is an alternative computing formula for G which Lieberson (1976) termed ND.

A Brief Demonstration

I begin with an example that applies the terms introduce above to obtain G by the conventional formula and also demonstrates how the value of G can be obtained by the simpler approach of computing the difference of group means from percentile scores. The example case has just five areas, each one with 100 people. These are listed from high to low based on proportion White in the area. Appendix Fig. C.1 lists for basic terms for each area. These include the group count terms (t_i, w_i, b_i), proportion White for the area (p_i), the proportion of the group population residing in the area (pt_i, pw_i, pb_i), and the cumulative proportion of the group population residing in areas with area proportion White at or below p_i (cpt_i, $cptw_i$, $cptb_i$).

Appendix Fig. C.2 presents terms that are used directly to calculate the value of G. The second and third columns in the figure present the terms used to calculate the value of G via the Lieberson (1976) “net difference” variation of the formula given in Duncan and Duncan (1955) (expression (C.2b) above). The difference between the sums for the two columns (i.e., 0.903–0.027) yields the value of G as 0.876. The fourth column gives the percentile score for each area as ranked on area proportion

Area	t_i	w_i	b_i	p_i	pt_i	pw_i	pb_i	cpt_i	cpw_i	cpb_i
5	100	100	0	1.000	0.200	0.286	0.000	1.000	1.000	1.000
4	100	95	5	0.950	0.200	0.271	0.033	0.800	0.714	1.000
3	100	90	10	0.900	0.200	0.257	0.067	0.600	0.443	0.967
2	100	65	35	0.650	0.200	0.186	0.233	0.400	0.186	0.900
1	100	0	100	0.000	0.200	0.000	0.667	0.200	0.000	0.667
	500	350	150			1.000	1.000			

Fig. C.1 Example of calculating the Gini index – intermediate terms

Area	$pw_i \cdot cpb_{i-1}$	$pb_i \cdot cpw_{i-1}$	y_i	$pw_i \cdot y_i$	$pb_i \cdot y_i$
5	0.286	0.000	0.900	0.257	0.000
4	0.262	0.015	0.700	0.190	0.023
3	0.231	0.012	0.500	0.129	0.033
2	0.124	0.000	0.300	0.056	0.070
1	---	---	0.100	0.000	0.067
	0.903	0.027		0.631	0.193

Fig. C.2 Example of calculating the Gini index – final terms

White (p). This is the residential outcome (y) relevant for computing G in the difference of means framework. The fifth and sixth columns give weighted sum calculations for obtaining separate group means on y for Whites and Blacks. Twice the difference of the sums for the two columns (i.e., $2 \cdot \left(Y_W - Y_B\right) = 2 \cdot \left(0.631 - 0.193\right) = 2 \cdot 0.438$) also yields the value of G as 0.876.

Getting on with the Derivation

This example illustrates that the difference of means approach for obtaining G is simple and straight forward. The next task is to show how these formulas for G (C.2, C.2a, and C.2b) map onto the terms in the formulation of G as the White-Black difference of means $Y_W - Y_B$ on relative rank position on area proportion White (p). I apologize in advance for the fact that the derivation to follow is long and tedious. I suspect a simpler derivation can be given but I have not discovered it. What follows is one way to accomplish the task.

My first step is to introduce the term RRT_i as an alternative designation of y_i as “relative rank” standing on area proportion White (p). Thus,

$$RRT_i = y_i = \left(\Sigma pt_{i-1} + pt_i/2\right) = \left(\Sigma t_{i-1} + t_i/2\right)/T.$$

The “RR” in “RRT” refers to *relative rank* and the “T” indicates that it is calculated for the total of the combined population of White and Black households (ignoring other households). Multiplying relative rank by 100 gives a percentile score. Given these terms, the White-Black difference of means for y_i is given by

$$\begin{aligned} Y_W - Y_B &= \Sigma pw_i \cdot y_i - \Sigma pb_i \cdot y_i, \text{ or, alternatively,} \\ Y_W - Y_B &= \Sigma pw_i \cdot RRT_i - \Sigma pb_i \cdot RRT_i. \end{aligned} \tag{C.3}$$

Next I introduce two related terms – RRW_i and RRB_i. RRW_i registers average relative rank position on area proportion White (p) based on the distribution of *White households only* and is given by

$$\mathrm{RRW}_i = (\Sigma pw_{i-1} + pw_i/2) = (\Sigma w_{i-1} + w_i/2)/W.$$

RRB_i registers the relative rank position on area proportion White (p) based on the distribution of *Black households only* and is given by

$$\mathrm{RRB}_i = (\Sigma pb_{i-1} + pb_i/2) = (\Sigma b_{i-1} + b_i/2)/B.$$

The terms RRT_i, RRW_i, and RRB_i, are closely interrelated. Specifically, each one can be defined in terms of the other two according to the following expressions.

$$\mathrm{RRT}_i = P \cdot \mathrm{RRW}_i + Q \cdot \mathrm{RRB}_i \tag{C.4a}$$

$$\mathrm{RRW}_i = (\mathrm{RRT}_i - Q \cdot \mathrm{RRB}_i)/P \tag{C.4b}$$

$$\mathrm{RRB}_i = (\mathrm{RRT}_i - P \cdot \mathrm{RRW}_i)/Q \tag{C.4c}$$

The basis for expression (C.4a) can be clarified as follows

$$\begin{aligned}
\mathrm{RRT}_i &= (\Sigma pt_{i-1} + pt_i/2) \\
&= (\Sigma t_{i-1} + t_i/2)/T \\
&= (\Sigma w_{i-1} + \Sigma b_{i-1} + w_i/2 + b_i/2)/T \\
&= (\Sigma w_{i-1} + w_i/2)/T + (\Sigma b_{i-1} + b_i/2)/T \\
&= (\Sigma w_{i-1} + w_i/2)/[W \cdot (T/W)] + (\Sigma b_{i-1} + b_i/2)/[B \cdot (T/B)] \\
&= (W/T) \cdot (\Sigma w_{i-1} + w_i/2)/W + (B/T) \cdot (\Sigma b_{i-1} + b_i/2)/B \\
&= (W/T) \cdot (\Sigma w_{i-1} + w_i/2)/W + (B/T) \cdot (\Sigma b_{i-1} + b_i/2)/B \\
&= P \cdot (\Sigma pw_{i-1} + pw_i/2) + Q \cdot (\Sigma pb_{i-1} + pb_i/2) \\
&= P \cdot \mathrm{RRW}_i + Q \cdot \mathrm{RRB}_i .
\end{aligned}$$

Expressions (C.4b) and (C.4c) are simple rearrangements of (C.4a).

The relationships among RRT_i, RRW_i, and RRB_i help clarify how G relates to $Y_W - Y_B$. Expression (C.3) shows that the values of RRT_i are directly used in computing Y_W and Y_B. Expression (C.4a) establishes that RRT_i can be given in terms of RRW_i and RRB_i. These two terms can be incorporated into familiar computing expressions for G (yielding Eq. (C.5) below).

Before reviewing this in more detail I first digress to note that values of RRW_i and RRB_i define points on the segregation curve, the well-known graphical representation of uneven distribution that supports an appealing geometric interpretation of G. The segregation curve is constructed by taking areas in ascending order of area proportion White (p) and then plotting cumulative proportion White ($cpw_i = \Sigma w_i/W$) against cumulative proportion Black ($cpb_i = \Sigma b_i/B$). The curve is contrasted with the diagonal line that would result under conditions of exact even distribution and the value of G is given by ratio of the area between the diagonal and the curve to the total area under the diagonal. The values of RRW_i by RRB_i fall on the midpoints of the line segments that form the segregation curve.

The values of RRW_i and RRB_i can be used to directly calculate the value of G. To see this, start with the following familiar computing formula for G given by Duncan and Duncan (1955: 211)

$$G = \Sigma pw_i \cdot \Sigma pb_{i-1} - \Sigma pb_i \cdot \Sigma pw_{i-1}. \qquad \text{(C.2, restated)}$$

Then add 0 in the form of $\Sigma pw_i \cdot pb_i/2 - \Sigma pw_i \cdot pb_i/2$ to obtain

$$G = \left[\Sigma pw_i \cdot \Sigma pb_{i-1} - \Sigma pb_i \cdot \Sigma pw_{i-1}\right] + \left[\Sigma pw_i \cdot pb_i/2 - \Sigma pw_i \cdot pb_i/2\right].$$

Rearrange terms

$$G = \Sigma pw_i \cdot \left[\Sigma pb_{i-1} + pb_i/2\right] - \Sigma pb_i \cdot \left[\Sigma pw_{i-1} + pw_i/2\right].$$

Drawing on terms given earlier, substitute RRB_i for $\left[\Sigma pb_{i\ 1} + pb_i/2\right]$ and RRW_i for $\left[\Sigma pw_{i\ 1} + pw_i/2\right]$ to obtain

$$G = \Sigma\, pw_i \cdot RRB_i - \Sigma\, pb_i \cdot RRW_i. \qquad \text{(C.5)}$$

For later notational convenience, I designate $\Sigma\, pw_i \cdot RRB_i$ as G_W and $\Sigma\, pb_i \cdot RRW_i$ as G_B to get the compact expression

$$G = G_W - G_B. \qquad \text{(C.5a)}$$

Note that the terms G_W and G_B support straightforward substantive interpretations. Specifically, G_W indicates the proportion of total comparisons between White and Black households where the White household is higher on area proportion White (p) and G_B similarly indicates the proportion of comparisons where the Black household is higher.[6]

$$Y_W - Y_B = \Sigma\, pw_i \cdot RRT_i - \Sigma\, pb_i \cdot RRT_i. \qquad \text{(C.3, restated)}$$

Expression (C.5) is very similar in form to expression (C.3) (restated here for convenience). This suggests that the relationship of G to $Y_W - Y_B$ can be expressed in terms of specific relationships between the core terms in (C.3) and (C.5). This is indeed the case. The first relationship involves the terms $\Sigma\, pw_i \cdot RRB_i$ from (C.5) and $\Sigma\, pw_i \cdot RRT_i$ from (C.3). Their relationship can be given as

$$\Sigma\, pw_i \cdot RRB_i = \left(\Sigma\, pw_i \cdot RRT_i - P/2\right)/Q. \qquad \text{(C.6)}$$

[6] This corresponds closely to Lieberson's (1976) index of net difference (ND) interpretation of G. The only difference computationally is how ties are handled in the computations. In Lieberson's calculations, ties are dealt with separately. In this calculation, ties are apportioned in equal halves to each outcome. The resulting value of G (or ND) is identical.

The second relationship involves the terms $\Sigma\, pb_i \cdot RRW_i$ from (C.5) and $\Sigma\, pw_i \cdot RRT_i$. from (C.3). Their relationship can be given as

$$\Sigma\, pb_i \cdot RRW_i = \left(\Sigma\, pb_i \cdot RRT_i - Q/2\right)/P. \tag{C.7}$$

Similarly, the central terms in (C.2) for $Y_W - Y_B$ can be expressed in relation to the terms in (C.5) for G based on

$$Y_W = \Sigma\, pw_i \cdot RRT_i = Q \cdot \Sigma\, pw_i \cdot RRB_i + P/2 = Q \cdot G_W + P/2, \text{ and} \tag{C.8}$$

$$Y_B = \Sigma\, pb_i \cdot RRT_i = P \cdot \Sigma\, pb_i \cdot RRW_i + Q/2 = P \cdot G_B + Q/2. \tag{C.9}$$

Restating these using more compact notation yields

$$G_W = \left(Y_W - P/2\right)/Q. \tag{C.6a}$$

$$G_B = \left(Y_B - Q/2\right)/P. \tag{C.7a}$$

$$Y_W = Q \cdot G_W + P/2 \tag{C.8a}$$

$$Y_B = P \cdot G_B + Q/2 \tag{C.9a}$$

Establishing Expressions (C.6, C.6a) and (C.8, C.8a)

For the sake of completeness I show here how expressions (C.6, C.6a) and (C.8, C.8a) can be obtained. I begin by drawing on (C.4b) to restate the term $\Sigma\, pw_i \cdot RRB_i$ from (C.5) and then rearrange the result as follows.

$$\Sigma\, pw_i \cdot RRB_i = \Sigma\, pw_i \cdot \left[\left(RRT_i - P \cdot RRW_i\right)/Q\right]$$

$$\Sigma\, pw_i \cdot RRB_i = \Sigma\, pw_i \cdot \left[\left(RRT_i /Q\right) - \left(RRW_i \cdot P/Q\right)\right]$$

$$\Sigma\, pw_i \cdot RRB_i = \Sigma\, pw_i \cdot \left(RRT_i /Q\right) - \Sigma\, pw_i \cdot \left(RRW_i \cdot P/Q\right)$$

$$\Sigma\, pw_i \cdot RRB_i = \left(\Sigma\, pw_i \cdot RRT_i\right)/Q - \left(\Sigma\, pw_i \cdot RRW_i\right)\left(P/Q\right)$$

The value of the term $\Sigma\, pw_i \cdot RRW_i$ is 0.5 because the mean of relative rank position is necessarily $0.5 = ½$. Accordingly, the last expression can be simplified by substituting (½) for $\Sigma\, pw_i \cdot RRW_i$ to obtain (C.6) as follows

$$\Sigma\, pw_i \cdot RRB_i = \left(\Sigma\, pw_i \cdot RRT_i\right)/Q - \left(½\right)\left(P/Q\right)$$

$$\Sigma\,\mathrm{pw_i}\cdot\mathrm{RRB_i} = \left(\Sigma\,\mathrm{pw_i}\cdot\mathrm{RRT_i} - \mathrm{P}/2\right)/\mathrm{Q} \qquad \text{(C.6, restated)}$$

Or in more compact notation

$$\mathrm{G_W} = \left(\mathrm{Y_W} - \mathrm{P}/2\right)/\mathrm{Q} \qquad \text{(C.6a, restated)}$$

Reversing sides and rearranging terms to isolate $\mathrm{Y_W}$ yields

$$\left(\mathrm{Y_W} - \mathrm{P}/2\right)/\,\mathrm{Q} = \mathrm{G_W} = \mathrm{G_W}$$

$$\mathrm{Y_W}/\mathrm{Q} - \mathrm{P}/2\mathrm{Q} = \mathrm{G_W}$$

$$\mathrm{Y_W}/\mathrm{Q} = \mathrm{G_W} + \mathrm{P}/2\mathrm{Q}$$

$$\mathrm{Y_W} = \mathrm{Q}\cdot\left(\mathrm{G_W} + \mathrm{P}/2\mathrm{Q}\right)$$

$$\mathrm{Y_W} = \mathrm{Q}\cdot\mathrm{G_W} + \mathrm{P}/2 \qquad \text{(C.8a, restated)}$$

Expanding to less compact notation

$$\Sigma\,\mathrm{pw_i}\cdot\mathrm{RRT_i} = \mathrm{Q}\cdot\Sigma\,\mathrm{pw_i}\cdot\mathrm{RRB_i} + \mathrm{P}/2. \qquad \text{(C.8, restated)}$$

Establishing Expressions (C.7, C.7a) and (C.9, C.9a)

Next I show here how expressions (C.7, C.7a) and (C.9, C.9a) can be obtained. I begin by drawing on (C.4a) to restate the term $\Sigma\,\mathrm{pb_i}\cdot\mathrm{RRW_i}$ from (C.5) and then rearrange the result as follows.

$$\Sigma\,\mathrm{pb_i}\cdot\mathrm{RRW_i} = \Sigma\,\mathrm{pb_i}\cdot\left[\left(\mathrm{RRT_i} - \mathrm{Q}\cdot\mathrm{RRB_i}\right)/\mathrm{P}\right]$$

$$\Sigma\,\mathrm{pb_i}\cdot\mathrm{RRW_i} = \Sigma\,\mathrm{pb_i}\cdot\left[\left(\mathrm{RRT_i}\ /\mathrm{P}\right) - \left(\mathrm{RRB_i}\cdot\mathrm{Q}/\mathrm{P}\right)\right]$$

$$\Sigma\,\mathrm{pb_i}\cdot\mathrm{RRW_i} = \Sigma\,\mathrm{pb_i}\cdot\left(\mathrm{RRT_i}\ /\mathrm{P}\right) - \Sigma\,\mathrm{pb_i}\cdot\left(\mathrm{RRB_i}\cdot\mathrm{Q}/\mathrm{P}\right)$$

$$\Sigma\,\mathrm{pb_i}\cdot\mathrm{RRW_i} = \left(\Sigma\,\mathrm{pb_i}\cdot\mathrm{RRT_i}\right)/\mathrm{P} - \left(\Sigma\,\mathrm{pb_i}\cdot\mathrm{RRB_i}\right)\left(\mathrm{Q}/\mathrm{P}\right)$$

Since $\Sigma\,\mathrm{pb_i}\cdot\mathrm{RRB_i}$ is $0.5 = ½$, the last expression can be simplified by substituting (½) for $\Sigma\,\mathrm{pb_i}\cdot\mathrm{RRB_i}$ to obtain (C.8) as follows

$$\Sigma\,\mathrm{pb_i}\cdot\mathrm{RRW_i} = \left(\Sigma\,\mathrm{pb_i}\cdot\mathrm{RRT_i}\right)/\mathrm{P} - \left(½\right)\left(\mathrm{Q}/\mathrm{P}\right)$$

$$\Sigma\,\mathrm{pb_i}\cdot\mathrm{RRW_i} = \left(\Sigma\,\mathrm{pb_i}\cdot\mathrm{RRT_i} - \mathrm{Q}/2\right)/\mathrm{P}. \qquad \text{(C.8, restated)}$$

Or more compactly

$$G_B = \left(Y_B - Q/2\right)/P. \qquad \text{(C.8a, restated)}$$

Reversing sides and rearranging terms to isolate Y_B yields

$$Y_B/P\ - Q/2P = G_B$$

$$Y_B/P\ = G_B + Q/2P$$

$$Y_B = P \cdot \left(G_B + Q/2P\right)$$

$$Y_B = P \cdot G_B + Q/2 \qquad \text{(C.9a, restated)}$$

$$\Sigma\, pb_i \cdot RRT_i = P \cdot \Sigma\, pb_i \cdot RRW_i + Q/2. \qquad \text{(C.9, restated)}$$

Some Implications of Expressions (C.6) and (C.7)

Based on (C.6) and (C.7), G as given in (C.5) can be obtained from the core terms that define $Y_W - Y_B$ in (C.3) as follows

$$G = \left(\Sigma\, pw_i \cdot RRT_i - P/2\right)/Q - \left(\Sigma\, pb_i \cdot RRT_i - Q/2\right)/P \qquad \text{(C.10)}$$

or, in more compact notation,

$$G\ =\ \left(Y_W - P/2\right)/Q - \left(Y_B - Q/2\right)/P. \qquad \text{(C.10a)}$$

Similarly, based on (C.8) and (C.9), the term $Y_W - Y_B$ in (C.3) can be obtained from the terms that define G in (C.5) as follows

$$Y_W - Y_B\ = \left(Q \cdot \Sigma\, pw_i \cdot RRB_i + P/2\right) - \left(P \cdot \Sigma\, pb_i \cdot RRW_i + Q/2\right) \qquad \text{(C.11)}$$

or, in more compact notation,

$$Y_W - Y_B = \left(Q \cdot G_W + P/2\right) - \left(P \cdot G_B + Q/2\right). \qquad \text{(C.11a)}$$

These results establish that the value of the Gini Index (G) can be directly and exactly mapped onto the terms of the group difference of means ($Y_W - Y_B$) on residential outcomes (y) scored on the basis of relative rank position on area group proportion (p).

The Role of P and Q in Scaling Terms when Groups Differ in Relative Size

The results just reviewed show that, while the relationship between G and ($Y_W - Y_B$) is exact, it also is complex. Expressions (C.10, C.10a) and (C.11, C.11a) clarify how scores for G map onto scores for ($Y_W - Y_B$). In this, it is clear that the terms for relative group size – P and Q – play important roles. How can this be understood? One answer to that question is that the operations involving P and Q in these expressions rescale the core terms of G so they will map onto the core terms of $Y_W - Y_B$, and vice versa. This is necessary because the core terms in G and $Y_W - Y_B$ have different logical ranges. Accordingly, the operations involving P and Q in expression (C.10) rescale the core terms of $Y_W - Y_B$ so they will take the same value as their corresponding terms in G. Similarly, the operations involving P and Q in expression (C.11) rescale the core terms of G so they will take the same value as their corresponding terms in $Y_W - Y_B$.

The logical ranges for both G and its core terms are constant across all combinations of P and Q. The core term $\Sigma\, pw_i \cdot RRB_i$ (i.e., G_W) has a logical range of 0.5 based on having a minimum possible value of 0.5 under even distribution and a maximum value of 1.0 under complete segregation. The core term $\Sigma\, pb_i \cdot RRW_i$ (i.e., G_B) also has a logical range of 0.5 based on having a minimum possible value of 0.0 under complete segregation and a maximum value of 0.5 under even distribution. Thus, G ranges from a minimum of 0.0 under even distribution based on

$$G = \Sigma\, pw_i \cdot RRB_i - \Sigma\, pb_i \cdot RRW_i = 0.5 - 0.5 = 0.0$$

to a maximum of 1.0 under complete segregation based on

$$G = \Sigma\, pw_i \cdot RRB_i - \Sigma\, pb_i \cdot RRW_i = 1.0 - 0.0 = 1.0$$

The logical range for $Y_W - Y_B$ also is always constant but it is 0.5 not 1.0. This accounts for why G is divided by 2 in expression (C.1). Note, however, that the logical ranges for the two core terms Y_W and Y_B are not constants. In each case one boundary of their logical range is a constant but the other boundary varies with the values of P and Q. For the term $Y_W = \Sigma\, pw_i \cdot RRT_i$ the fixed boundary is its minimum possible value of 0.5, which occurs under even distribution. Its upper boundary (i.e., maximum possible value) is given by $Q + P/2$, which occurs under complete segregation and varies in exact value with city ethnic composition. For the term $Y_B = \Sigma pb_i \cdot RRT_i$, the fixed boundary of its logical range is 0.5, its maximum possible value which occurs under even distribution. Its lower boundary (i.e., the minimum possible value) is given by Q/2 which occurs under complete segregation and varies in exact value with city ethnic composition.

Thus, $Y_W - Y_B$ ranges from a minimum of 0.0 under even distribution based on

$$Y_W - Y_B = \Sigma\, pw_i \cdot RRT_i - \Sigma\, pb_i \cdot RRT_i = 0.5 - 0.5 = 0.0$$

to a maximum of 0.5 under complete segregation based on

$$\begin{aligned} Y_W - Y_B &= \Sigma\, pw_i \cdot RRT_i - \Sigma\, pb_i \cdot RRT_i = (Q + P/2) - (Q/2) \\ &= Q/2 + P/2 = (Q + P)/2 = 0.5. \end{aligned}$$

In light of these points, Expression (C.10) can now be understood as follows. The values of P/2 and Q in the term ($\Sigma\, pw_i \cdot RRT_i - P/2$) / Q rescale the value of the core term $\Sigma\, pw_i \cdot RRT_i$ used in computing Y_W in $Y_W - Y_B$ to map its position in the logical range of 0.5 to ($Q + P/2$) onto the correct position in the logical range of 0.5 to 1.0 for the parallel core term $\Sigma\, pw_i \cdot RRB_i$ used in computing G. Similarly, the values of Q/2 and P in the term ($\Sigma\, pb_i \cdot RRT_i - Q/2$) / P rescale the core term $\Sigma\, pb_i \cdot RRT_i$ used in computing Y_B in $Y_W - Y_B$ to map its position in the logical range of Q/2 to 0.5 onto the correct position in the logical range of 0.0 to 0.5 for the parallel core term $\Sigma\, pb_i \cdot RRW_i$ used in computing G.

Expression (C.11) can be interpreted in a similar way. P/2 and Q in the term ($Q \cdot \Sigma\, pw_i \cdot RRB_i - P/2$) / Q rescale the core term $\Sigma\, pw_i \cdot RRB_i$ used in computing G to map its position in the logical range of 0.5 to 1.0 on to the correct position in the logical range of 0.5 to $Q + P/2$ for the core term $\Sigma\, pw_i \cdot RRT_i$ used in computing $Y_W - Y_B$. Similarly, Q/2 and P in the term ($P \cdot \Sigma\, pb_i \cdot RRW_i - Q/2$) / P rescale the core term $\Sigma\, pb_i \cdot RRW_i$ used in computing G to map its position in the logical range of 0.0 to 0.5 onto the correct position in the logical range of Q/2 to 0.5 for the core term $\Sigma\, pb_i \cdot RRT_i$ used in computing $Y_W - Y_B$.

The Special Circumstance When $P = Q$

Things are relatively simple when $P = Q$. This can be seen by rearranging terms in (C.10) to obtain the alternative expression.

$$G = (1/Q) \cdot \Sigma\, pw_i \cdot RRT_i - (1/P) \cdot \Sigma\, pb_i \cdot RRT_i + Q/2P - P/2Q \qquad \text{(C.12)}$$

When $P = Q$, this resolves to

$$\begin{aligned} G = {} & 1/(1/2) \cdot \Sigma\, pw_i \cdot RRT_i - 1/(1/2) \cdot \Sigma\, pb_i \cdot RRT_i \\ & + (1/2)/[2 \cdot (1/2)] - (1/2)/[2 \cdot (1/2)] \end{aligned}$$

$$G = 2 \cdot \Sigma\, pw_i \cdot RRT_i - 2 \cdot \Sigma\, pb_i \cdot RRT_i + (1/2) - (1/2)$$

$$G = 2 \cdot (\Sigma\, pw_i \cdot RRT_i - \Sigma\, pb_i \cdot RRT_i)$$

$$G/2 = \Sigma\, pw_i \cdot RRT_i - \Sigma\, pb_i \cdot RRT_i$$

$$G/2 = Y_W - Y_B\,. \qquad \text{(C.1, restated)}$$

This corresponds to expression (C.1) presented at the beginning of this section.

Similarly, rearranging terms in (C.11) leads to the following alternative expression.

$$\mathrm{Y_W - Y_B \;=\; Q\cdot\Sigma\,pw_i\cdot RRB_i - P\cdot\Sigma\,pb_i\cdot RRW_i + P/2 - Q/2} \qquad \text{(C.13)}$$

When $\mathrm{P = Q}$, this resolves to

$$\mathrm{Y_W - Y_B \;=\; (1/2)\cdot\Sigma\,pw_i\cdot RRB_i - (1/2)\cdot\Sigma\,pb_i\cdot RRW_i + (1/2)/2 - (1/2)/2}$$

$$\mathrm{Y_W - Y_B = (1/2)\cdot(\Sigma\,pw_i\cdot RRB_i - \Sigma\,pb_i\cdot RRW_i)}$$

$$\mathrm{Y_w - Y_B = G/2} \qquad \text{(C.1, restated)}$$

And this also corresponds to expression (C.1).

Summary Comments on Formulating G as a Difference of Means ($Y_W - Y_B$) on Relative Rank

The relationship in expression (C.1) now can be placed in broader context as follows. The core terms that define G in expression (C.2) map directly and exactly onto the core terms that define $\mathrm{Y_W - Y_B}$ in expression (C.3). Consequently, G can be described as registering the White-Black difference in average relative rank on area proportion White (p). Examined in the "natural" metric of relative rank scores, the difference of means $\mathrm{Y_W - Y_B}$ has a logical range of 0.0–0.5 while the logical range of G is 0.0–1.0. Hence, expression (C.1) equates the two measures based on $\mathrm{Y_W - Y_B = G/2}$.

The Dissimilarity Index (D) – A Special Case of the Gini Index (G)

The dissimilarity or delta index (D) is closely related to the Gini Index (G). More specifically, D can be described as a special case of G where G is computed after areal units ordered on area group proportion scores (p) are collapsed into two categories: areas where the group proportion score exceeds the city-wide group proportion (i.e., $\mathrm{p > P}$) and areas where it does not (i.e., $\mathrm{p \le P}$). Based on this, D can be expressed as a difference of group means on residential outcomes (y) scored from area group proportions (p) in a manner comparable to that just outlined for G.

D and G both are intimately related to the segregation curve, a graphical device for depicting uneven distribution popularized by Duncan and Duncan (1955). An example of a standard segregation curve is shown in Fig. C.3. The curve is based on block group data for Whites and Blacks in the Houston, Texas metropolitan area in 2000 and is constructed as follows. First the areas (in this case block groups) are placed in ascending order based on proportion White (p) in the area. Then the curve

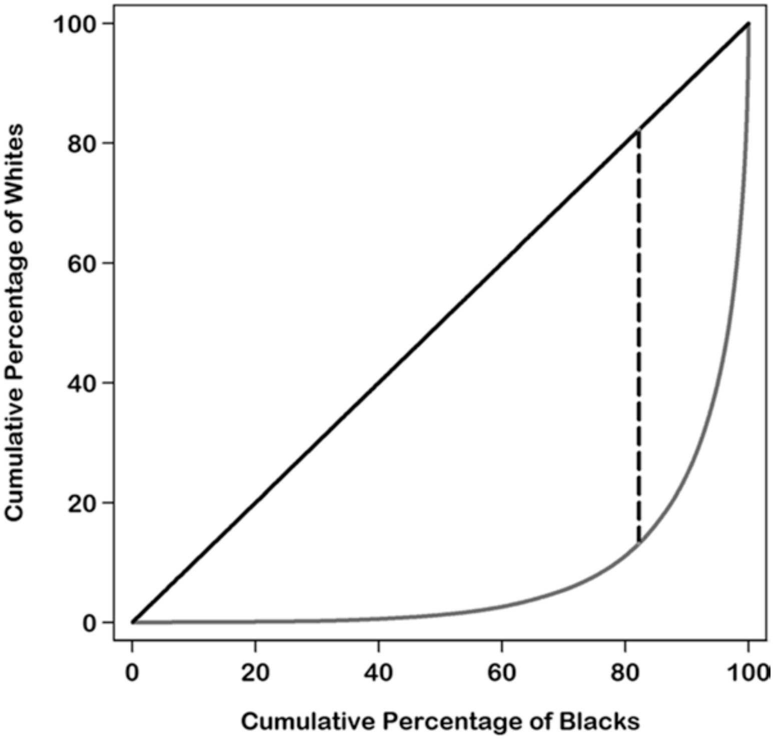


Fig. C.3 Example Segregation Curve for White-Black Comparison (Note: Units are ordered from low to high on area proportion White. Gini index is 84.7, Delta is 69.0)

is traced by drawing line segments connecting the sequence of (x,y) pairings for the cumulated proportion of the White population (on the y-axis) and the cumulated proportion of the Black population (on the x-axis) as areas are taken in ascending order on the value of p. The resulting curve is contrasted with the diagonal line between the starting point (0,1) and ending point (1,1) of the curve. The diagonal represents the segregation curve that would obtain under the condition of *exact* even distribution. The gap between the curve and the diagonal visually indicates the degree of departure from even distribution.

As is well known, G and D both have direct quantitative and geometric relations to the curve's departure from the diagonal. G registers the departure quantitatively based on the ratio of the area between the curve and the diagonal to the total area under the diagonal. In the example shown, the value of G is 84.7. D registers the degree of departure quantitatively based on the maximum vertical difference between the curve and the diagonal and in the example shown has a value of 69.0.

The geometric relationships to the segregation curve for G and D highlight an important difference between the two measures. The area interpretation of G makes it clear that its value is determined by the shape of the full curve. In contrast, the vertical line interpretation of D makes it clear that its value is determined by a single

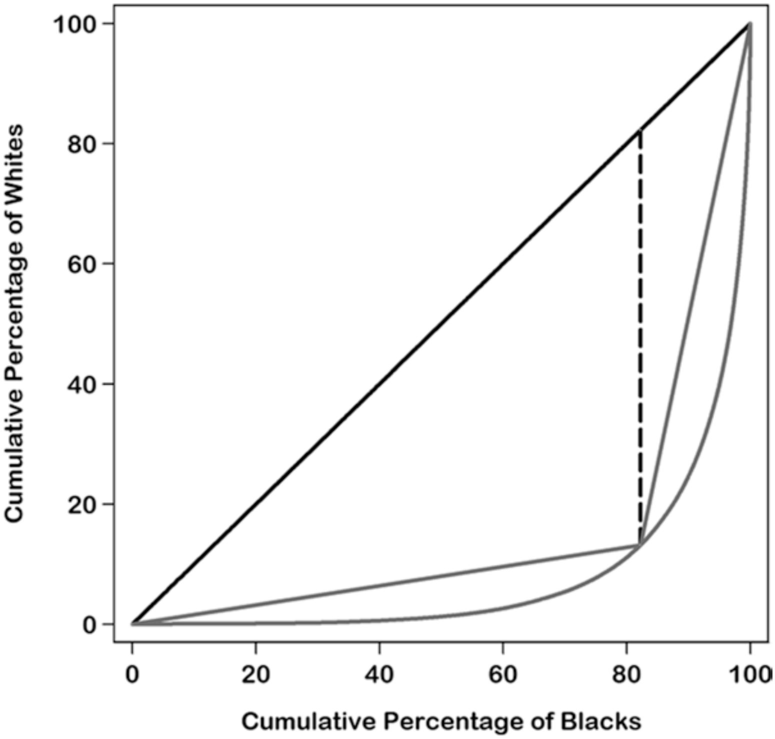


Fig. C.4 Example segregation curves for white-black comparison (Note: Units are ordered from low to high on area proportion White. Gini index is 84.7, Delta is 69.0)

point on the segregation curve. Accordingly, G responds to any residential shifts that promote more even distribution (i.e., that reduce the area between the diagonal and the curve) while D responds to such changes only if they affect the position of a particular point on the curve. The difference is highlighted in Fig. C.4. Here the segregation curve in the first graph is supplemented with a second segregation curve. This is a three point segregation curve defined by the triangle involving three points from the full segregation curve; the two end points (0,0) and (1,1) of the diagonal and the point on the full curve where the vertical distance between the curve and the diagonal is at its maximum. This last point determines the value of D so I designate it as (x_D, y_D). In the example shown it is (0.132, 0.822).[7]

[7] Becker et al. (1978) present a similar graphical analysis of D.

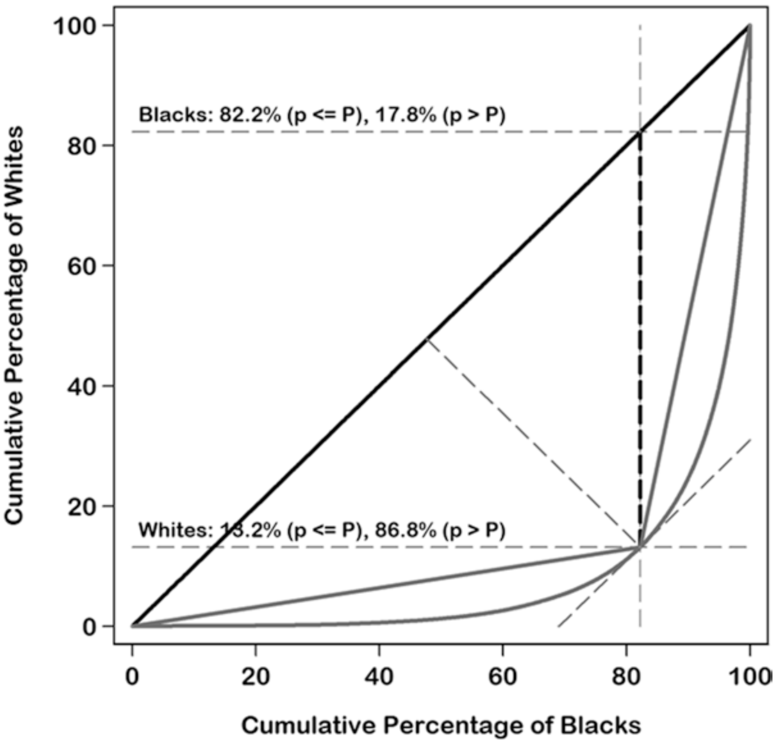


Fig. C.5 Example segregation curves for G and D with details (Note: Units are ordered from low to high on area proportion White. Gini index is 84.7, Delta is 69.0 = (86.8−17.8))

D Is G Calculated from a Special Three-Point Segregation Curve

D can be seen as a special case of G calculated for the three-point segregation curve defined by the points (0,0), (x_D, y_D), and (1,1). More specifically, D represents the minimum value of G that can obtain for a curve that has the point (x_D, y_D). This is depicted graphically in the detailed example in Appendix Fig. C.5. The relationships involved can be outlined in a general way as follows. Recall that the value of G is given by A/T where A is the area between the diagonal and the segregation curve and T is the total area under the diagonal which is ½. For the three point segregation curve associated with D, A is equal to the area of the triangle that forms the three-point segregation curve. Accordingly, $A = ½ \cdot b \cdot h$ where A is the area of the triangle, b is the length of the base of the triangle, and h is the height of the triangle. The base of the triangle is the diagonal and thus b is equal to the length of the diagonal which is $\sqrt{2}$. The height of the triangle (h) is equal to the length of the line that extends perpendicular from the diagonal and ends on the segregation curve at the point (x_D, y_D). This line is a side of a right isosceles triangle whose base has a length equal to the value of D – the maximum vertical distance from the segregation curve to the diagonal. Thus, $h = D/\sqrt{2}$.

It follows that the area (A) between the diagonal and the three point segregation curve for D is given by $A = ½ \cdot b \cdot h = ½ \cdot \surd 2 \cdot \left(D/\surd 2\right) = ½ \cdot D$. It also follows that the value of the Gini Index (G) for the three point segregation curve is given by $G = A/T = \left(½ \cdot D\right)/½$ which resolves to D. This establishes that the value of D is equivalent to the value of G for a simplified segregation curve analysis in which all areas of the city are grouped into just two categories; all areas where $p \leq P$, and all areas where $p > P$.

The comparison of the three-point segregation curve with the full curve highlights two characteristics of D. One is that $D \leq G$ because the full segregation curve for G can never be "inside" the three-point segregation curve for D. Another is that D is insensitive to variations in residential distribution other than the distinction between residing in areas where $p > P$ or not. Finally, D can be understood as the minimum possible value of G for a curve containing the point (x_D, y_D) because D treats Whites and Blacks as experiencing only two relative rank scores and this maximizes ties between Whites and Blacks on relative ranks. Expanding the curve to consider more points cannot reduce the value of G as the construction principles are such that the segregation curve can only stay the same or expand outward from the three-point curve if more points are added to the curve.

D Is a Simple Difference of Group Proportions Residing in Areas Where $p \geq P$

There is an alternative computing approach for D that is simple and carries an appealing substantive interpretation. It is based on understanding D as the difference in group proportions residing in areas where $p \geq P$. This interpretation traces to the fact that the maximum vertical difference between the curve and the diagonal occurs at a particular point on the segregation curve. Specifically, it is first encountered at the end of the line segment on the curve for the last areal unit where $p < P$. It then is maintained for all subsequent points on the curve for areas where $p = P$. It is last encountered at the beginning of the line segment on the curve for the first areal unit where $p \geq P$.

When there are no areas where $p = P$, the maximum vertical difference between the curve and the diagonal will be at a single point; the point where the line segment for the last area where $p < P$ connects with the line segment for the first area where $p > P$. When some areas have $p = P$, the maximum vertical difference will be found at the beginning and end of the line segment formed for these areas. So it is correct to say that the maximum vertical distance corresponding to the value of D can be found at the following locations on the line segments that create the segregation curve.

- the end point of the line segment for the first area where $p < P$
- any point on line segments for areas where $p = P$
- beginning of the line segment for the first area where $p > P$

Because the vertical distance is at its maximum at the beginning and end of line segments where $p = P$, one can say the maximum vertical distance is found

- the end point of the line segment for the last area where $p \leq P$
- the beginning point of the line segment for the first area where $p \geq P$

This can be seen by reviewing the construction of the segregation curve in more detail. Starting at (0,0) the curve is formed by plotting line segments connecting (x,y) points for group population shares that are being cumulated over areas taken in ascending order of p. Except in the unusual case of exact even distribution, $p < P$ for the initial areas and the line segments plotted for these areas will have a slope of less than 1. Accordingly, the curve initially falls away from the diagonal and the vertical distance between the curve and the diagonal increases with each successive area so long as $p < P$ with the vertical distance being greatest at the end point of the line segment for the area. The maximum vertical distance is first reached when the sequence arrives at the first area where $p \geq P$. If the next area plotted is one where $p = P$ (*exactly*), the line segment for that area will have a slope of 1 and will run parallel to the diagonal. The maximum vertical distance is maintained for all subsequent areas where $p = P$ (exactly).[8] This changes when the sequence reaches the first area where $p > P$. At this point, the slope of the line segment plotted for that area will be greater than 1 and the segregation curve begins rising faster than the diagonal. Accordingly, the vertical distance between the curve and the diagonal will start to decline. It will continue to decline with each successive area in the sequence and the curve ultimately rises back to the diagonal to connect with the end point (1,1).

This discussion makes it clear that the value of D can be understood as a simple difference of group proportions. Specifically, the value of D is equal to the difference between the proportions of Whites and Blacks, respectively, that reside in areas where Whites are represented at or above the level for the city overall (i.e., $p \geq P$). For convenience, I designate the (x,y) pair for the beginning point of the line segment for the first area where $p \geq P$ as (x_D, y_D). Applying the subscript "D" indicates that the values of x_D and y_D determine the value of D. The values of x_D and y_D register the proportions of Blacks and Whites, respectively, that reside in areas where Whites are under-represented (i.e., areas where $p < P$). Under even distribution the value of y_D would be equal to x_D. In light of this, the value of D is given by ($x_D - y_D$), the vertical distance between the diagonal and the curve at this point. The values $(1 - x_D)$ and $(1 - y_D)$ similarly indicate the proportions of Blacks and Whites, respectively, who reside in areas where Whites are represented at parity or higher (i.e., areas where $p \geq P$). D also can be obtained from ($\left[1 - y_D\right] - \left[1 - x_D\right]$). This expression supports an appealing substantive interpretation of D; it is the White-Black difference in the proportions that reside in areas where proportion White is at or above the level of the city overall.

The example presented in Fig. C.5 shows that 82.2 % of Blacks and 13.2 % of Whites reside in areal units where Whites are under-represented (i.e., $p < P$). It

[8] These points are noted in Becker et al. (1978) and Duncan and Duncan (1955).

likewise shows that 86.8 % of Whites and 17.8 % of Blacks reside in area units where the presence of Whites equals or exceed the citywide level (i.e., $p \geq P$). The value of D can be obtained in either of two ways. It can be obtained from the Black-White difference in percentages in residing in areas where Whites under-represented (i.e., $D = 82.2 - 13.2 = 69.0$). Alternatively and more appropriately for the purposes of the present task, it can be obtained from the White-Black difference in percentages in residing in areas where Whites are represented at or above the level for the city overall (i.e., $D = 86.8 - 17.8 = 69.0$).

The Dissimilarity or Delta Index (D) – Alternative Functions for Scaling Contact

The above discussion establishes at least two viable ways to score individual residential outcomes (y) based on area group proportion scores (p) such that delta (D) can be obtained as a simple difference of group means. The first option is based on viewing D as a special case of the Gini Index (G). In this approach, y is scored as the relative rank (percentile) transformation of p applied to the two-category residential scheme for the special case of the three-point segregation curve described above. In this case delta (D) can be given by an expression comparable to Expression (C.1) introduced earlier for G. Specifically,

$$Y_W - Y_B = D/2, \text{ or, alternatively, } 2\left(Y_W - Y_B\right) = D$$

where D can be understood as a special case of G.

The second alternative involves an even simpler scoring scheme for y. This scaling function draws on the mundane fact that a proportion is equivalent to the mean for a variable that is scored 0 or 1. The above discussion established that D is equal to the White-Black difference in proportions residing in areas where $p \geq P$. Accordingly, the group proportions involved can be restated as group means on a variable that is scored 1 for individuals who reside in an area that reaches or exceeds parity on contact with whites White (i.e., areas where $p \geq P$) and 0 otherwise (i.e., when $p < P$). This provides the basis for obtaining D by scoring residential outcomes for individuals (y) as 1 for areas where proportion White are at or above parity (i.e., $p \geq P$) and 0 otherwise. Then compute the means for Whites and Blacks separately to obtain the value of D according to

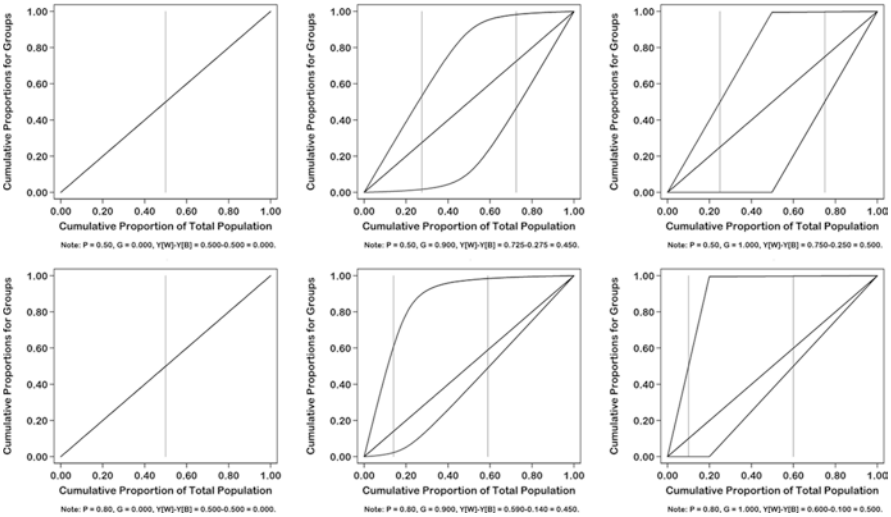
$$Y_W - Y_B = D.$$

One benefit of the resulting difference of means formulation of D is that it calls attention to how segregation as measured by D is linked with individual residential outcomes. Specifically, this formulation highlights the fact that D registers group differences in average contact with Whites when contact is rescaled from its original, “natural” metric of p – which can vary continuously over the range of 0–1 (inclusive) – to a binary scoring of either 0 or 1. Seeing D formulated in this way may raise questions concerning the methodological implications and desirability of collapsing p to a dichotomy when assessing group differences in exposure. I leave these issues for discussion elsewhere.

Alternative Graphical Explorations of Relative Rank Position

Before concluding this appendix chapter, I offer additional comments on the topic of relative rank position. The preceding discussion establishes that the values of G and D reflect group differences in relative rank position on area proportion White (p). It is surprising that this is not already more widely appreciated because G and D have close relationships with the segregation curve which is an appealing graphical device for comparing group differences in distribution over areas ranked on proportion White (p). With this in mind it is instructive to directly consider group distributions on relative rank position.

To that end, Fig. C.6 presents graphs that help provide additional insight into how relative rank position relates to group distributions. The figure presents 6 graphs. Each graph plots three curves that are constructed by first ordering areas from low to high on area proportion White (p) and then plotting the cumulated proportions of the White and Black population against the cumulated proportion of the total (combined White and Black) population and then also plotting the cumulated proportion of the total population against itself to form a diagonal line rising from (0,0) to (1,1). These plotted values are designated here designated as



Note: P = 0.50, G = 0.000, Y[W]-Y[B] = 0.500-0.500 = 0.000.

Note: P = 0.50, G = 0.900, Y[W]-Y[B] = 0.725-0.275 = 0.450.

Note: P = 0.50, G = 1.000, Y[W]-Y[B] = 0.750-0.250 = 0.500.

Note: P = 0.80, G = 0.000, Y[W]-Y[B] = 0.500-0.500 = 0.000.

Note: P = 0.80, G = 0.900, Y[W]-Y[B] = 0.590-0.140 = 0.450.

Note: P = 0.80, G = 1.000, Y[W]-Y[B] = 0.600-0.100 = 0.500.

Fig. C.6 Plots of cumulative proportions of whites, blacks, and combined total by cumulative proportion of combined total

$$cpw_i = \Sigma pw_i = \Sigma w_i / W,$$
$$cpb_i = \Sigma pb_i = \Sigma b_i / B, \text{ and}$$
$$cpt_i = \Sigma pt_i = \Sigma t_i / T.$$

The graph that results from plotting these values as described is similar to the segregation curve in one key respect; under conditions of exact even distribution, the curves for the White and Black population will coincide with the diagonal line for the total population. So the diagonal is a reference point for even distribution. A key difference from the segregation curve is that under conditions of uneven distribution, the curve for the cumulating proportion of the Black population will rise above the diagonal and the curve for the cumulating proportion of the White population will fall below the diagonal. Like the segregation curve, the areas between the curves and the diagonal in this graph have relationships to the values of G and D. This should not be surprising since the information plotted is very similar to the information plotted in the segregation curve. However, the visual representation here is distinct.

One feature of this graphical device is that the diagonal directly reflects relative rank position on area proportion White (p). Thus, the contrast between the diagonal and the curves for Whites and Blacks provides a basis for grasping their differences in relative rank position. A curve that rises above the diagonal is skewed toward below average rank positions. A curve that falls below the diagonal is skewed toward above average rank positions. The implications of the curves for group means on relative rank position are depicted graphically by plotting two vertical lines; one indicates the value of mean relative rank for Whites (Y_W) and the other indicates mean relative rank for Blacks (Y_B). Under conditions of exact even distribution, these will necessarily coincide at the value of 0.50, the overall mean on relative rank for area proportion White (p). Where these two values differ, the value for Y_W exceeds 0.50 and is necessarily higher than the value of Y_B which falls below 0.50. As noted earlier, the logical range for Y_W is from 0.5 to $Q+(P/2)$ and the logical range of Y_B is from Q/2 to 0.5, and the maximum value for ($Y_W - Y_B$) is 0.5 which occurs under complete segregation.

The graphs in the figure are organized by two rows and three columns. The three columns are for three conditions for segregation. The graphs in the first (leftmost) column are for the extreme condition of exact even distribution where the value of G is 0. The graphs in the third (rightmost) column are for the opposite extreme condition of complete segregation where the value of G is 100. The graphs in the middle column are for substantial, but not complete, segregation where the value of G is 0.900.[9] The two rows are for two conditions of city racial composition. The top row is for a city where P and Q are both 0.50. The bottom row is for a city where P is 0.80 and Q is 0.20.

[9] These segregation curves are based on simulated data generated using the hyperbola model for the segregation curve described in Duncan and Duncan (1955: 214).

The graphs on both rows of the first column look the same. This is because under conditions of even distribution $cpt_i = cpw_i = cpb_i$ and the graph will necessarily consist of three identical diagonal lines rising from the lower left to the top right and this pattern holds regardless of the values of P and Q. Similarly, the vertical lines depicting the values of Y_W and Y_B coincide and both are plotted at the value of 0.50.

When segregation exists, each of the three curves will be distinct. This is seen in the two graphs in the middle column of the figure which are for examples where the value of G is 0.900. The diagonal lines in the two graphs are produced by plotting cpt_i against itself. Because areas are ordered from low to high on area proportion White (p), the curves plotting cpb_i by cpt_i rise faster than the diagonals. In contrast, the curves plotting cpw_i by cpt_i rise slower than the diagonals. The vertical lines in these graphs indicate that, as noted above, the means on relative rank (y) for Blacks (Y_B) are below 0.50 and the means on relative rank (y) for Whites (Y_W) are above 0.50. The variation in location in the top and bottom rows documents how the particular values of the group means depend not only on the level of segregation involved but also on the values of P and Q. In both cases, however, the difference of means $Y_W - Y_B$ is 0.450 and is equal to G/2.

The graphs in the third (rightmost) column depict the extreme condition of complete segregation where G is 1.00. Again the diagonal lines in the graphs reflect the curves plotting cpt_i by cpt_i. The curves plotting cpb_i by cpt_i rise from 0.0 when cpt is 0.0 to 1.0 when cpt is Q (which is 0.5 in the top graph and 0.2 in the bottom graph) and then remain at 1.0 until cpt is 1.0. The curves plotting cpw_i by cpt_i stay at 0.0 until cpt_i reaches Q, then climbs to 1.0 when cpt reaches 1.0. Here the vertical lines depicting the means on relative rank (y) for Blacks (Y_B) are at the value Q/2 which is 0.25 in the top graph and 0.10 in the bottom graph. In contrast, the vertical lines depicting means on relative rank (y) for Whites (Y_W) are at the value $Q + P/2$ which is 0.75 in the top graph and 0.60 in the bottom graph. In both of these example cases, the difference between the two means is 0.5, the maximum possible value the difference can take. This is one half of G's maximum value of 1.0, consistent with relationship in Expression (C.1).

The graphs in Fig. C.6 illustrate an important implication of expressions (C.4b) and (C.4c); namely, that the height of the curves for cpb_i and cpw_i at a given value of cpt_i will depend on two factors. One, obviously, is the extent of segregation between Whites and Blacks. That is made clear by the progression across columns for either row of the figure. The other factor is the relative sizes of the groups in the comparison; that is, the ratio of P and Q. That is made clear by how the curves for cpb_i and cpw_i, and the group means associated with these curves (plotted as vertical lines), differ with the value of P.

I offer one last set of comments on the graphs in this figure. G and D have definite relationships to the graphs in Fig. C.6. The area between the curve plotting cpb_i by cpt_i and the diagonal equals the value of G for the comparison of Blacks against total (G_{TB}). The area between the curve plotting cpw_i by cpt_i and the diagonal equals the value of G for the comparison of Whites against total (G_{TW}). The sum of these two determines the value of G for the comparison of Whites to Blacks. Specifically, G is given by the ratio of the sum of these two areas to 0.5, the maximum possible value for the sum. D is equal to the maximum vertical distance between the curves for cpb_i and cpw_i and, exactly as is the case for the segregation curve, this is value is seen at the last area where $p_i \leq P$.

One implication I stress here is that the segregation curve, while familiar and appealing in many ways, is not the only graphical device for comparing group distributions over areas ranked on area proportion White (p). The graphs presented here contain the same information as the segregation curve and like the segregation curve they support a geometric interpretation of the values of G and D. In addition, they provide a more direct basis for assessing group differences on residential outcomes (y) that are scored to reflect relative rank position on area proportion White (p).

The Nature of the Y-P Relationship for G

The nature of the y-p relationship for the Gini Index (G) is complex and difficult to summarize. Since the relationship is based on a relative rank (percentile or quantile) transformation, the y-p relationship is monotonic and positive. But few general statements beyond that can be offered.

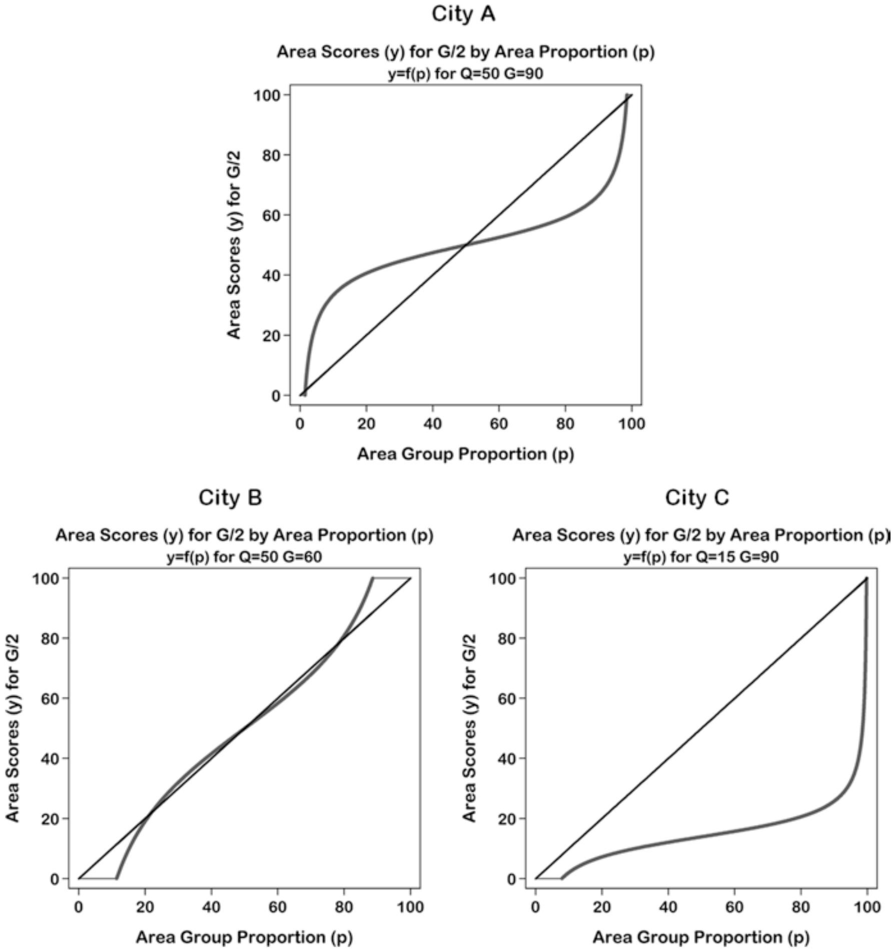
I have explored the relationship by performing simulation studies to gain insight into the nature of the y-p relationship. I cannot provide a full review of these explorations here. But I will provide a brief summary of key points. The simulations assumed a model city with the following characteristics. It has 1000 neighborhoods with 10,000 persons in each neighborhood and only two groups – Whites and Blacks. I populated individual neighborhoods based on a model segregation curve; specifically, a segregation curve defined by the “hyperbola model” described in Duncan and Duncan (1955: 213–215). By using the hyperbola model I was able to establish particular values of G in a given simulation and thus can vary city racial composition (P) and the value of G independently across simulation trials.

Each unique combination of values for P and G produces a unique distribution of Whites and Blacks across the neighborhoods of the city. Based on the resulting distributions, I calculated the scores of p and y for each neighborhood using procedures outlined earlier. I then performed graphical analyses to gain insight into how the y-p relationship varies across different combinations of values for P and G. I offer the following to summarize key findings from my explorations.

- The relationship between y – relative rank position on p – and p is always nonlinear.
- The value of y always increases as p increases but generally rises faster (has a steeper slope) at the beginning and at the end and rises slower (has a shallower slope) in between.
- The nonlinear y-p relationship is variable, not fixed. Its exact form varies with the values of city racial composition (P) and the value of G.
- City racial composition (P) determines whether the y-p relationship is symmetrical or asymmetrical. It is symmetrical when P is 50 and increasingly asymmetrical as P departs further from 50.

- The value of G determines whether the nonlinearity in the y-p relationship describe above is mild or pronounced. When G is high, the “steeper” portions of the y-p curve occur over short ranges on p and the “flatter” portion of the y-p curve occurs over an extended range of p. As the value of G declines, the “flatter” portion of the y-p curve becomes less distinct from the “steeper” portions of the curve.

I conclude this discussion by describing how the principles just listed play out in selected example cases. I start with an example for a hypothetical “City A” where the racial composition of the city is balanced (i.e., $P = 50$) and the level of segregation as measured by G is high (i.e., $G = 90$). As shown in the top panel of Appendix Fig. C.7, the y-p relationship is symmetrical (because P is 50) and strongly nonlin-


City A

City B

City C

Fig. C.7 Examples of y-p relationship under varying combinations of G and P

ear (because G is high). Specifically, y rises rapidly over a short portion of the lower range for p ($p = 0 - 15$); y then rises slowly over an extended portion of the intermediate range of p ($p = 15 - 85$); and y then rises rapidly again over a short portion of the upper range of p ($p = 85 - 100$). More specifically, y increases about 40 points over the range of 0–20 for p, then increases only 20 points over the range of 20–80 for p, and then increases another 40 points over the range of 80–100 for p.

The example labeled City B lowers G to 60 but leaves P unchanged at 50. The resulting y-p curve is shown in the lower left panel of the figure. The relationship remains symmetrical, as in City A, because P is 50. But the lower value for G produces a less strong nonlinear relationship evident in the fact that the differences between the steeper and flatter portions of the curve now are smaller. The example labeled City C leaves G unchanged for City A, but increases P to 85, a value more typical for US urban areas. The y-p curve continues to have distinct steep and flat portions as in City A. But now the curve is asymmetrical with most of the rise in y taking place over the last portion of the range of p ($p = 90 - 100$).

The pattern seen in City C becomes even more dramatic when relative minority group size is at low levels (i.e., below 5) and P is high. This provides a basis for understanding a finding that is discussed in Chaps. 6, 7, and 8 of the main text. The finding is that scores for G and D can be and often are much higher than scores for S when the two groups in the comparison are imbalanced in size. As the pattern for City C shows, this possibility arises because the two groups can differ by relative small amounts on p – the area outcome that determines S – and at the same time can differ by large amounts on y as scored for G and D. The pattern for City A, and especially the pattern for City B, yield insight into why discrepancies between G and D in comparison with S tend to be much smaller when city racial composition is balanced.

Appendix D: Establishing the Scaling Function $\mathbf{y = f(p)}$ Needed to Cast the Separation Index (S) as a Difference of Group Means on Scaled Pairwise Contact

In this appendix I establish the scaling function $y = f(p)$ that accomplishes the goal of scoring residential outcomes (y) from area group proportions (p) such that the scores for y fall over the range 0–1 and yield the value of the separation index (S) as a difference of means on y for the two groups in the segregation comparison. The end result is that, in the example of using S to assess White-Black segregation, $S = Y_W - Y_B$ where Y_W and Y_B are the group means for Whites and Blacks, respectively, on individual residential outcomes (y) scored from the value of the area group proportion (p) for the areas in which the individuals reside.

The value of p for an area reflects pairwise group contact or exposure. Accordingly, the value of y for an area can be described as reflecting scaled pairwise group contact or exposure and the expression ($Y_W - Y_B$) can described as the difference of group means on scaled pairwise group contact. The scaling function $y = f(p)$ that places S in the desired difference of group means framework is developed below. The scaling function is simple and substantively attractive. Specifically it is the exact one-to-one linear function $f(p_i) = p_i$ which means that S can be placed in the difference of means framework without rescaling p from its original or "natural" metric of pairwise group contact.

The separation index (S) has been known by many names including: the variance ratio index (V, James and Taeuber 1985), the correlation ratio (r, Stearns and Logan 1986; White 1986), eta squared (η^2 , Duncan and Duncan 1955; James and Taeuber 1985), the mean square deviation (MSD, White 1986; Zoloth 1976), r_{ij} (Coleman et al. 1975), and S (Zoloth 1976; Becker et al. 1978). The index is well established in the literature on segregation measurement and has been widely used in empirical segregation studies for many decades. S is particularly attractive when cast in the difference of means framework used here because S can be expressed as a difference of means on scaled pairwise group contact where group contact is based on area group proportion (p) in its "natural" metric – that is, without rescaling p as is required for the other indices considered here.

As best I have been able to determine, Becker et al. (1978: 353) were the first to show that in the two group case S can be given as the simple difference between the focal group's contact with itself (i.e., generically, P_{XX}, for White contact with Whites, P_{WW}) and the comparison group's contact with the focal group (i.e., generically, P_{YX}, for Black contact with Whites, P_{BW}) based on

$$S = P_{XX} - P_{YX} \text{ in generic form and}$$
$$S = P_{WW} - P_{BW} \text{ for White-Black segregation.}$$

Note that this relationship holds only when the population consists of only two groups and it does not generalize to situations where the population consists of three or more groups. The relationship can be adapted to all circumstances by restating contact as "pairwise" contact instead of "overall" contact as follows

$$S = P_{XX.XY} - P_{YX.XY}.$$

Here the suffix ".XY" in the subscripts contact indicates that the contact calculations are based only on the counts of the two groups in the segregation comparison. Thus, $P_{XX.XY}$ denotes the focal group's pairwise contact with itself and $P_{YX.XY}$ denotes the comparison group's pairwise contact with the focal or "reference" group.

For White-Black segregation, conventional or "overall" contact indices as introduced by Bell (1954) are given by

$$P_{WW} = 1/W \cdot \Sigma w_i p_i = 1/W \cdot \Sigma w_i (w_i / t_i)$$

for White contact with Whites and

$$P_{BW} = 1/B \cdot \Sigma\, b_i p_i = 1/B \cdot \Sigma\, b_i \left(w_i / t_i\right)$$

for Black contact with Whites. The corresponding pairwise contact indices are given as follows.

$$P_{WW.WB} = 1/W \cdot \Sigma\, w_i p_i = 1/W \cdot \Sigma\, w_i \left(w_i / \left(w_i + b_i\right)\right), \text{ and}$$

$$P_{BW.WB} = 1/B \cdot \Sigma\, b_i p_i = 1/B \cdot \Sigma\, b_i \left(w_i / \left(w_i + b_i\right)\right)$$

The difference key difference between overall and pairwise contact is that $t_i \neq \left(w_i + b_i\right)$ when the population includes groups other than Whites and Blacks.

All popular indices of uneven distribution [are usually applied as "pairwise"] measures. That is, their calculations draw only on counts for the two groups in the segregation comparison. So formulating contact indices in this way is not unusual. One simply must bear in mind that contact in this formulation is interpreted in terms of the pair of groups involved in the comparison. When the population also includes groups other than Whites and Blacks, the separation index is given by

$$S = P_{WW.WB} - P_{BW.WB}$$

where $P_{WW.WB}$ is White's average *pairwise* contact with Whites and $P_{BW.WB}$ is Black's average *pairwise* contact with Whites. When the population consists only of Whites and Blacks, the same expression obviously continues to hold but the "$_{.\,WB}$" subscript is not necessary.

The distinction between *overall* and *pairwise* contact is important but it is cumbersome. Since all indices of uneven distribution are based on pairwise comparisons, I drop the "$_{.\,XY}$" suffix notation from this point forward. Thus, for convenience, the expression

$$S = P_{WW} - P_{BW}$$

indicates a pairwise construction unless otherwise noted. Likewise, pairwise constructions are assumed for city and area proportion White (P and p_i, given respectively by $P = W/\left(W + B\right)$ and $p_i = w_i / \left[w_i + b_i\right]$) and city and area proportion Black (Q and q_i, given respectively as $Q = B/\left(W + B\right)$ and $q_i = b_i / \left[w_i + b_i\right]$). These conventions are in keeping with the literature on segregation measurement which lets context dictate when area proportion White (p_i) should computed using "overall" calculations (i.e., $p_i = w_i / t_i$) or "pairwise" calculations (i.e., $p_i = w_i / \left[w_i + b_i\right]$).

To conclude this discussion, the separation index (S) can be given as the group difference of means on average pairwise contact with the reference group. In the case of White-Black segregation, $S = P_{WW} – P_{BW}$. The terms P_{WW} and P_{BW} assess White and Black group averages on area proportion White (p_i). Setting residential outcomes (y_i) to the value of area proportion White (p_i) allows one to place S in the notation of the difference of means framework restating it as $S = Y_W – Y_B$. The next sections review terms from the "variance ratio" formulation of S and then demonstrates that the differences of means formulation of S and the variance ratio formulation of S are equivalent.

Variance Analysis

I now consider the relationship $S = \eta^2$ in more detail. I acknowledge that the expressions and relationships I introduce below are not particularly original. They have been noted elsewhere including, for example, in papers by Becker et al. (1978): 353) and White (1986:207) and also in statistical texts such as Blalock (1979: 81). The contribution of the discussion here is that it collects and calls attention to points not emphasized in most previous discussions.

Duncan and Duncan (1955) noted that the separation index (S) (which they termed the variance ratio) is equivalent to the eta squared (η^2) statistic from analysis of variance. More specifically, S is equal to η^2 for the analysis of how X, an individual-level binomial variable for race (coded 1 for Whites and 0 to Blacks), varies over areas. The value of S thus indicates the proportion of variation in race (X) that is "explained" by area of residence. Under even distribution S will be 0 because the representation of Whites and Blacks in each area will exactly reflect each group's representation in the city overall and knowledge of area will not improve the prediction of race above the baseline of assuming the overall city average. Under complete segregation S will be 1 because area of residence will be homogeneous – either all White or all Black – and thus area will perfectly predict race. Intermediate success in prediction is quantified as the ratio BSS/TSS from analysis of variance where BSS is the "between group sum of squares" for individual deviations from the overall mean and TSS is "total sum of squares" for individual deviations from the overall mean. The overall mean for X is the proportion White in the city population (P) so $TSS = \sum\left(X_k - P\right)^2$ with k used here to index individuals. Predictions for X are based on category means for X which in this case are equal to area proportion White (p_i) so $BSS = \sum\left(p_{ik} - P\right)^2$ with i here serving to index areas. Finally, for completeness, inability to explain X is quantified by WSS/TSS where WSS is the "within group sum of squares" given by $WSS = \sum\left(X_i - p_{ik}\right)^2$.

It is useful to note at this point that the value of η^2 also is equal to the square of the individual-level bivariate correlation of race (X) and area proportion White (p_i). Thus, one can interpret S as indicating the degree to which race determines area proportion White (p) for individuals as quantified by r^2 from the regression of p_i on X or of η^2 from the analysis of how p_i varies by race. Either way, it is clear that the value of S revolves around the impact of race on contact with Whites at the individual level as reflected in the White-Black difference of means in contact with Whites (p_i). Under even distribution explanation S will be 0 because all $p_i = P$ so the White and Black means for contact with Whites (p_i) are the same and knowledge of race will not improve the prediction of contact with Whites (p) above the baseline of assuming the overall city average (P). Under complete segregation S will be 1 because race will perfectly predict contact with Whites with all Whites living in areas where $p_i = 1$ and all Blacks living in areas where $p_i = 0$.

The more general relationship including intermediate outcomes is set forth in more detail below. Relevant relationships from analysis of variance can be summarized as follows.

$$TSS = BSS + WSS$$

$$\eta^2 = BSS/TSS$$

$$\eta^2 = 1 - WSS/TSS$$

$$TSS = \Sigma\Sigma\left(X_{ik} - P\right)^2$$

$$WSS = \Sigma\Sigma\left(X_{ik} - p_i\right)^2$$

$$BSS = \Sigma t_i\left(p_i - P\right)^2$$

with “i” serving as an index of areas and “k” serving as an index of individuals within areas.

The following expressions are adapted from discussions in White (1986: 207) and Becker et al. (1978: 353) and indicate how TSS, WSS, and BSS also can be obtained from terms that found in standard computing formulas for S.

$$TSS = TPQ$$

$$BSS = \Sigma t_i p_i^2 - TP^2$$

$$WSS = \Sigma t_i p_i q_i$$

$$BSS/TSS = 1/TPQ\left(\Sigma t_i p_i^2 - TP^2\right)$$

The basis for the three expressions is established as follows. First, the equivalence of TSS and TPQ can be established as follows based on Whites and Blacks being scored 0 and 1 on race (X).

$TSS = \Sigma\left(X_{ik} - P\right)^2$	(a standard formula for TSS)
$= W\left(1 - P\right)^2 + B\left(0 - P\right)^2$	(restate as separate operations for Whites and Blacks)
$= TP\left(1 - P\right)^2 + TQ\left(0 - P\right)^2$	(replace W with TP and B with TQ)
$= TPQ^2 + TQ\left(0 - P\right)^2$	(replace $(1 - P)^2$ with Q^2)
$= TPQ^2 + TQP^2$	(replace $(0 - P)^2$ with P^2)
$= TPQ\left(Q\right) + TPQ\left(P\right)$	(reorganize terms)
$= TPQ\left(Q + P\right)$	(reorganize terms)
$TSS = TPQ$	(based on $Q + P = 1$)

Next, the equivalence of WSS and Σ $t_i p_i q_i$ can be established as follows.

$$\begin{aligned}
\text{WSS} &= \Sigma\Sigma\left(X_{ik} - p_i\right)^2 && \text{(standard formula for WSS)} \\
&= \Sigma\, w_i\left(1-p_i\right)^2 + \Sigma\, b_i\left(0-p_i\right)^2 && \text{(restate as separate operations for Whites and Blacks)} \\
&= \Sigma\, t_i p_i\left(1-p_i\right)^2 + \Sigma\, t_i q_i\left(0-p_i\right)^2 && \text{(replace } w_i \text{ with } t_i p_i \text{ and } b_i \text{ with } t_i q_i\text{)} \\
&= \Sigma\, t_i p_i\left(q_i\right)^2 + \Sigma\, t_i q_i\left(0-p_i\right)^2 && \text{(replace } 1-p_i \text{ with } q_i\text{)} \\
&= \Sigma\, t_i p_i q_i^2 + \Sigma\, t_i q_i p_i^2 && \text{(replace } \left(0-p_i\right)^2 \text{ with } p_i^2\text{)} \\
&= \Sigma\, t_i p_i q_i\left(q_i\right) + \Sigma\, t_i p_i q_i\left(p_i\right) && \text{(reorganize terms)} \\
&= \Sigma\, t_i p_i q_i\left(q_i + p_i\right) && \text{(reorganize terms)} \\
\text{WSS} &= \Sigma\, t_i p_i q_i && \text{(based on } q_i + p_i = 1\text{)}
\end{aligned}$$

Then the equivalence of BSS and $\Sigma\, t_i p_i^2 - TP^2$ can be established as follows

$$\begin{aligned}
\text{BSS} &= \Sigma\, t_i\left(p_i - P\right)^2 && \text{(standard formula for BSS)} \\
&= \Sigma\, t_i\left(p_i^2 - 2p_i P + P^2\right) && \text{(multiply out } \left(p_i - P\right)^2\text{)} \\
&= \Sigma\, t_i p_i^2 - \Sigma\, t_i 2p_i P + \Sigma\, t_i P^2 && \text{(reorganize as multiple summations)} \\
&= \Sigma\, t_i p_i^2 - 2P \cdot \Sigma\, t_i p_i + P^2 \cdot \Sigma\, t_i && \text{(move constants outside of summations)} \\
&= \Sigma\, t_i p_i^2 - 2P\,\Sigma\, t_i p_i + TP^2 && \text{(substitute T for } \Sigma\, t_i\text{)} \\
&= \Sigma\, t_i p_i^2 - 2PTP + TP^2 && \text{(substitute TP for } \Sigma\, t_i p_i \text{ based on } P = \Sigma\, t_i p_i / T\text{)} \\
&= \Sigma\, t_i p_i^2 - 2TP^2 + TP^2 && \text{(reorganize terms)} \\
\text{BSS} &= \Sigma\, t_i p_i^2 - TP^2 && \text{(combine terms)}
\end{aligned}$$

From these expressions, η^2 and S can be obtained from the following computing formulas

$$S = \eta^2 = BSS/TSS$$

$$S = \eta^2 = \left(\Sigma\, t_i p_i^2 - P^2\right)/TPQ.$$

Formulation as a Difference of Means

S also can obtained from the simple difference between pairwise White contact with Whites (P_{WW}) and pairwise Black contact with Whites (P_{BW}); that is, $S = P_{WW} - P_{BW}$. Because y_i for S is scored directly from p_i, $Y_W = P_{WW}$ and $Y_B = P_{BW}$ and the following equalities hold.

$$Y_W - Y_B = BSS/TSS$$

$$P_{WW} - P_{BW} = BSS/TSS$$

I provide a derivation establishing these equivalences below. I initially developed the derivation independently. However, I later discovered that a similar derivation had been given in Becker et al. (1978: 353).

$S = Y_W - Y_B$ $= P_{WW} - P_{BW}$	(follows because $y_i = p_i$)
$= (\Sigma w_i p_i) / W - (\Sigma b_i p_i)/B$	(standard expressions for P_{WW} & P_{BW})
$= (\Sigma t_i p_i p_i)/ W - (\Sigma t_i p_i q_i) / B$	(replace w_i with $t_i p_i$ and b_i with $t_i q_i$)
$= (\Sigma t_i p_i^2) / TP - (\Sigma t_i p_i q_i) / TQ$	(replace W with TP and B with TQ)
$= (Q/Q)(\Sigma t_i p_i^2) / TP - (P/P) \cdot \Sigma t_i p_i q_i) / TQ$	(introduce 1 in the form of Q/Q and P/P)
$= (Q \cdot \Sigma t_i p_i^2) / TPQ - (P \cdot \Sigma t_i p_i q_i) / TPQ$	(reorganize terms)
$= (Q \cdot \Sigma t_i p_i^2 - P \cdot \Sigma t_i p_i q_i) / TPQ$	(reorganize terms)
$= [Q \cdot \Sigma t_i p_i^2 - P \cdot \Sigma t_i p_i (1 - p_i)]/ TPQ$	(reorganize terms)
$= [Q \cdot \Sigma t_i p_i^2 - (P \cdot \Sigma t_i p_i - P \cdot \Sigma t_i p_i^2)] / TPQ$	(restate $P \cdot \Sigma t_i p_i (1 - p_i)$ as $P \cdot \Sigma t_i p_i - P \cdot \Sigma t_i p_i^2$)
$= (Q \cdot \Sigma t_i p_i^2 + P \cdot \Sigma t_i p_i^2 - P \cdot \Sigma t_i p_i) / TPQ$	(reorganize terms)
$= [(P + Q) \cdot \Sigma t_i p_i^2 - P \cdot \Sigma t_i p_i]/ TPQ$	(reorganize terms)
$= (\Sigma t_i p_i^2 - P \cdot \Sigma t_i p_i) / TPQ$	((P + Q = 1 and drops out)
$= (\Sigma t_i p_i^2 - P \cdot TP)/ TPQ$	(substitute TP for $\Sigma\, t_i p_i$)
$= (\Sigma t_i p_i^2 - TP^2) / TPQ$	(reorganize terms)
$S = BSS / TSS$	(substitute BSS for $\Sigma t_i p_i^2 - TP^2$ and TSS for TPQ as established earlier)

As a last comment, I note that the discussion here shows that S simultaneously registers two separate and distinct aspects of the relationship between race and contact with Whites (p).

- Under the traditional eta squared or variance ratio interpretation, S indicates the strength of the association between race (i.e., group membership) and contact with Whites (p).
- Under the new interpretation of S as a difference of group means for contact with Whites, S indicates the "impact" or "effect" of race (i.e., group membership) on contact with Whites.

Thus, S equals both the regression coefficient (b) for race and the square of the correlation coefficient (r) from the bivariate regression analysis predicting contact with Whites (p_i) based on race (X). Interestingly, both options allow for applying significance tests for the value of S.

Appendix E: Establishing the Scaling Function $y = f(p)$ Needed to Cast the Theil Entropy Index (H) as a Difference of Group Means on Scaled Pairwise Contact

In this appendix I establish the scaling function $y = f(p)$ that accomplishes the goal of scoring residential outcomes (y) from area group proportions (p) such that the scores for y fall over the range 0–1 and yield the value of the Theil entropy index (H) as a difference of means on y for the two groups in the segregation comparison. The end result is that, in the example of using H to assess White-Black segregation, $H = Y_W - Y_B$ where Y_W and Y_B are the group means for Whites and Blacks, respectively, on individual residential outcomes (y) scored from the value of the area group proportion (p) for the areas in which the individuals reside.

The scaling function $y = f(p)$ that places H in the difference of group means framework is developed below. Discussion of this function in the main body of this monograph notes that y is a smooth continuous, nonlinear transformation of p that changes p from its original or "natural" metric to a new metric that exaggerates group differences on p over portions of the lower and upper ranges of p (i.e., roughly $p < 0.25$ and $p > 0.75$) and compresses group differences on p over middle portions of the range of p (i.e., roughly $0.30 < p < 0.70$).

Please note that the primary credit for discovering the scaling function for H should be given to Warner Henson, III. Warner derived the first version of the scaling function for H while working with me as an undergraduate research fellow completing his BS in sociology at Texas A&M University.[10] I have subsequently added refinements and extensions to his work to serve the needs of this monograph,

[10] That was in the 2007. Soon after, Mr. Henson graduated and enrolled in the Sociology doctoral program at Stanford University.

but these are minor changes. Mr. Henson established the essential features of the derivation.

Continuing with the familiar example of White-Black segregation, a basis for scoring residential outcomes (y) such that the scores of y fall over the same range as p (i.e., 0–1) and yield the Theil index (H) as the difference of means ($Y_W - Y_B$) can be established as follows. First, start with the desired equivalence

$$H = Y_W - Y_B = (1/T) \cdot \Sigma t_i (E - e_i)/E.$$

The expression on the far right side is an adaptation of the formula for H given in James and Taeuber (1985). Next replace the terms Y_W and Y_B with alternative computing expressions as follows

$$(1/W) \cdot \Sigma w_i y_i - (1/B) \cdot \Sigma b_i y_i = (1/T) \cdot \Sigma t_i (E - e_i)/E.$$

Then replace W and B with alternative expressions based on T, P, and Q. Specifically, replace W with PT and replace B with QT. Similarly, replace w_i and b_i with alternative expressions based on t_i, p_i, and q_i. Specifically, replace w_i with $p_i t_i$ and b_i with $q_i t_i$. This yields

$$(1/PT) \cdot \Sigma p_i t_i y_i - (1/QT) \cdot \Sigma q_i t_i y_i = (1/T) \cdot \Sigma t_i (E - e_i)/E.$$

Then rearrange terms as follows

$$(1/T) \cdot \Sigma (p_i/P) t_i y_i - (1/T) \cdot \Sigma (q_i/Q) t_i y_i = (1/T) \cdot \Sigma t_i (E - e_i)/E$$

$$\Sigma (p_i/P) t_i y_i - \Sigma (q_i/Q) t_i y_i = \Sigma t_i (E - e_i)/E$$

$$\Sigma t_i y_i \left[(p_i/P) - (q_i/Q) \right] = \Sigma t_i (E - e_i)/E$$

$$\Sigma t_i y_i = \Sigma t_i \left[(E - e_i)/E \right] / (p_i/P - q_i/Q) \cdot$$

From the above expression, it is evident that

$$y_i = \left[(E - e_i)/E \right] / (p_i/P - q_i/Q).$$

For actual calculations, E and e_i would be expanded to their full expressions using the following substitutions

$$E = P \cdot \ln(1/P) + Q \cdot \ln(1/Q), \text{ and}$$

$$e_i = p_i \cdot \ln(1/p_i) + q_i \cdot \ln(1/q_i).$$

Adjusting the Range to 0–1

At this point a small additional adjustment is needed. The scores for y_i will yield H as a difference of group means thus achieving one important goal of the exercise. However, the scores for y will not fall in the range 0–1. They instead will fall in the range –Q to P as p_i varies from its minimum value of 0 to its maximum value of 1. This is because, when p_i is either 0 or 1, e_i evaluates to 0 and the term $\left(E-e_i\right)/E$ evaluates to 1. This reduces the expression

$$y_i=\left[\left(E-e_i\right)/E\right]/\left(p_i/P-q_i/Q\right).$$

to

$$y_i=1/\left[\left(p_i/P\right)-\left(q_i/Q\right)\right].$$

When p_i is 0, this expression becomes

$$y_i=1/\left[\left(0/P\right)-\left(1/Q\right)\right]$$

which evaluates to -Q. Similarly, when p_i is 1, the resulting expression is

$$y_i=1/\left[\left(1/P\right)-\left(0/Q\right)\right]$$

which evaluates to P.

The range for y can therefore be set to 0–1 by incorporating the constant Q in the function as follows

$$y_i= Q+\left[\left(E-e_i\right)/E\right]/\left(p_i/P-q_i/Q\right).$$

This achieves the desired solution.

A Loose End When $p = P$

There is a final issue to deal with. Interestingly, the value of y_i is undefined when p_i is *exactly* equal to P. This is because the term $(p_i/P\ -q_i/Q)$ will then be 0 and the same also will be true of the term $[\left(E-e_i\right)/E]$. Thus the expression $\left[\left(E-e_i\right)/E\right]/\left(p_i/P-q_i/Q\right)$ will be undefined because it involves division by zero. As a practical matter, *exact* equality of p_i and P is very rare in conventional empirical analyses of residential segregation in urban areas. Nevertheless, it is a logical possibility that it can occur in empirical studies of segregation and it is certainly likely to occur in methodological analyses and simulation studies. So it is necessary to establish a procedure for handling this situation.

The procedure I adopt is the following: when p_i is exactly P, assign a value for y based on the limiting values of y obtained by taking values of p_i that are arbitrarily close to P, but are just short of reaching exactly P. For example, the value of y can be established in this way by averaging the two values of y obtained using $p_i = P - 0.0000001$ and $p_i = P + 0.0000001$. The two values of y will be exceedingly close; so close in fact that a graph of the y-p relationship will appear as a smooth, continuous function in which y rises monotonically as p ranges from 0 to 1 with only an arbitrarily small "break" in the line at the exact point where $p_i = P$. The procedure suggested here would simply fill in this one point on the line. I offer this as a reasonable, practical strategy to follow until a better alternative is identified.

Appendix F: Establishing the Scaling Function $y = f(p)$ Needed to Cast the Hutchens' Square Root Index (R) as a Difference of Group Means on Scaled Pairwise Contact

In this appendix I establish the scaling function $y = f(p)$ that accomplishes the goal of scoring residential outcomes (y) from area group proportions (p) such that the scores for y fall over the range 0–1 and yield the value of the Hutchens Square Root Index (R) as a difference of means on y for the two groups in the segregation comparison. The result is that, in the example of using R to assess White-Black segregation, $R = Y_W - Y_B$ where Y_W and Y_B are the group means for Whites and Blacks, respectively, on individual residential outcomes (y) scored from the value of the area group proportion (p) for the areas in which the individuals reside.

The scaling function $y = f(p)$ that places R in the differences of group means framework is developed below. Discussion of this function in the main body of this monograph notes that y is a nonlinear transformation of p that changes p from its original or "natural" metric to a new metric that exaggerates group differences on p over portions of the lower and upper ranges of p (i.e., roughly $p < 0.25$ and $p > 0.75$) and compresses group differences on p over middle portions of the range of p (i.e., roughly $0.30 < p < 0.70$). The scaling function for R is very similar in shape and behavior to the scaling function for the Theil Entropy index (H). The main difference is that the nonlinearity in the scaling function for R is more pronounced; that is, it departs from linearity in the same basic manner as the scaling function for H, but the magnitude (amplitude) of the departure from linearity is consistently larger.

To establish the function $y = f(p)$, start with the desired equivalence

$$Y_W - Y_B = R.$$

Next replace R with an expression adapted from the formula for R given in Hutchens (2001, 2004).

$$\mathrm{Y_W} - \mathrm{Y_B} = 1 - \Sigma\ \sqrt{(w_i/W)(b_i/B)}.$$

Next replace $\mathrm{Y_W}$ and $\mathrm{Y_B}$ with the terms of their computing formulas as follows

$$1/\mathrm{W}\cdot\Sigma\,\mathrm{w_i y_i} - 1/\mathrm{B}\cdot\Sigma\,\mathrm{b_i y_i} = 1-\Sigma\sqrt{(w_i/W)(b_i/B)}.$$

Then replace W and B with expressions based on T, P, and Q. Similarly, replace $\mathrm{w_i}$ and $\mathrm{b_i}$ with expressions based on $\mathrm{t_i}$, $\mathrm{p_i}$, and $\mathrm{q_i}$ to obtain

$$1/\mathrm{PT}\cdot\Sigma\,\mathrm{p_i t_i y_i} - 1/\mathrm{QT}\cdot\Sigma\,\mathrm{q_i t_i y_i} = 1-\Sigma\sqrt{(p_i t_i/PT)(q_i t_i/QT)}.$$

Then rearrange terms as follows. First, on the right side isolate $(\mathrm{t_i}^2 / \mathrm{T}^2)$ inside the radical

$$1/\mathrm{PT}\cdot\Sigma\,\mathrm{p_i t_i y_i} - 1/\mathrm{QT}\cdot\Sigma\,\mathrm{q_i t_i y_i} = 1-\Sigma\sqrt{(t_i^2/T^2)(p_i\,/\,P)(q_i\,/\,Q)}.$$

Then move $(\mathrm{t_i}^2 / \mathrm{T}^2)$ outside of the radical as $(\mathrm{t_i}\ /\mathrm{T})$ and then restate it as $\mathrm{t_i}\ (1/\mathrm{T})$ to obtain

$$1/\mathrm{PT}\cdot\Sigma\,\mathrm{p_i t_i y_i} - 1/\mathrm{QT}\cdot\Sigma\,\mathrm{q_i t_i y_i} = 1-\Sigma(\mathrm{t_i}\ /\mathrm{T})\sqrt{(p_i/P)(q_i/Q)}$$

Restate $\sqrt{(p_i/P)(q_i/Q)}$ as $\sqrt{p_i q_i/PQ}$ to obtain

$$1/\mathrm{PT}\cdot\Sigma\,\mathrm{p_i t_i y_i} - 1/\mathrm{QT}\cdot\Sigma\,\mathrm{q_i t_i y_i} = 1-\Sigma\,\mathrm{t_i}\,(1/\mathrm{T})\sqrt{p_i q_i/PQ}.$$

On the left side move P and Q inside the summations

$$1/\mathrm{T}\cdot\Sigma(\mathrm{p_i}/\mathrm{P})\mathrm{t_i y_i} - 1/1\cdot\Sigma(\mathrm{q_i}/\mathrm{Q})\mathrm{t_i y_i} = 1-\Sigma(\mathrm{t_i}/\mathrm{T})\cdot\sqrt{p_i q_i/PQ}$$

On the right side replace 1 with the equivalent expression $\Sigma\ \mathrm{t_i}\ (1/\mathrm{T})$ and replace $(\mathrm{t_i}\ /\mathrm{T})$ with $\mathrm{t_i}\ (1/\mathrm{T})$

$$1/\mathrm{T}\cdot\Sigma(\mathrm{p_i}/\mathrm{P})\mathrm{t_i y_i} - 1/\mathrm{T}\cdot\Sigma(\mathrm{q_i}/\mathrm{Q})\mathrm{t_i y_i} = \Sigma\,\mathrm{t_i}\,(1/\mathrm{T}) - \Sigma\,\mathrm{t_i}\,(1/\mathrm{T})\cdot\sqrt{p_i q_i/PQ}.$$

Next reorganize on both sides

$$1/\mathrm{T}\cdot\left[\Sigma(\mathrm{p_i}/\mathrm{P})\mathrm{t_i y_i} - \Sigma(\mathrm{q_i}/\mathrm{Q})\mathrm{t_i y_i}\right] = \Sigma\,\mathrm{t_i}\left(1/\mathrm{T} - 1/\mathrm{T}\cdot\sqrt{p_i q_i/PQ}\right).$$

Next multiply both sides by T as follows

$$\Sigma(\mathrm{p_i/P})\mathrm{t_iy_i} - \Sigma(\mathrm{q_i/Q})\mathrm{t_iy_i} = \mathrm{T}\cdot\left[\,\Sigma\,\mathrm{t_i}\left(1/\mathrm{T} - 1/\mathrm{T}\sqrt{p_iq_i/PQ}\right)\right].$$

Then move T inside the summation on the right side to obtain

$$\Sigma(\mathrm{p_i/P})\mathrm{t_iy_i} - \Sigma(\mathrm{q_i/Q})\mathrm{t_iy_i} = \Sigma\,\mathrm{t_i}\left(1-\sqrt{p_iq_i/PQ}\,\right).$$

Next reorganize terms on the left side.

$$\Sigma\,\mathrm{t_iy_i}\left[(\mathrm{p_i/P})-(\mathrm{q_i/Q})\right] = \Sigma\,\mathrm{t_i}\left(1-\sqrt{p_iq_i/PQ}\right).$$

Then divide both sides by $[(\mathrm{p_i/P})-(\mathrm{q_i/Q})]$

$$\Sigma\,\mathrm{t_iy_i} = \Sigma\,\mathrm{t_i}\left(1-\sqrt{p_iq_i/PQ}\,\right)\Big/(\mathrm{p_i/P}-\mathrm{q_i/Q}).$$

From the last expression, it is clear that

$$\mathrm{y_i} = \left(1-\sqrt{p_iq_i/PQ}\,\right)\Big/(\mathrm{p_i/P}-\mathrm{q_i/Q}).$$

Adjusting the Range to 0–1

An additional adjustment is required. Under the last expression, the scores for y will yield R as a difference of group means. However, the scores for y will not fall in the range 0-1 as desired. Instead, values of $\mathrm{y_i}$ will range from –Q to P as $\mathrm{p_i}$ varies from its minimum value of 0 to its maximum value of 1. That is, the expression

$$\mathrm{y_i} = \left(1-\sqrt{p_iq_i/PQ}\right)\Big/(\mathrm{p_i/P}-\mathrm{q_i/Q})$$

yields $-$Q when $\mathrm{p_i}$ is 0 and P when $\mathrm{p_i}$ is 1. Accordingly, the range for y can be set to 0–1 by incorporating the constant Q in the function as follows

$$\mathrm{y_i} = \mathrm{Q}+\left(1-\sqrt{p_iq_i/PQ}\right)\Big/(\mathrm{p_i/P}-\mathrm{q_i/Q})$$

A Loose End When $p = P$

One final matter requires attention. It is that $\mathrm{y_i}$ is undefined when $\mathrm{p_i}$ is *exactly* equal to P because the term $[(\mathrm{p_i/P}) - (\mathrm{q_i/Q})]$ will then be 0. Thus the expression

$$\left(1-\sqrt{p_i q_i/PQ}\right)\Big/\left(\mathrm{p_i/P-q_i/Q}\right)$$

will be undefined because it will involve division by zero. As a practical matter, *exact* equality of $\mathrm{p_i}$ and P is very rare in conventional empirical analyses of residential segregation in urban areas. Nevertheless, it is a logical possibility in empirical studies and it is especially likely to occur in methodological analyses and simulation studies. So it is necessary to establish a procedure for handling this situation.

The option I adopt is as follows: when $\mathrm{p_i}$ is exactly P, assign a value for y based on the limiting values of y obtained by taking values of $\mathrm{p_i}$ that are arbitrarily close to P, but are not exactly P. For example, the value of y can be established in this way by averaging the two values of y obtained using $\mathrm{p_i = P - 0.0000001}$ and $\mathrm{p_i = P + 0.0000001}$. The two values of y will be exceedingly close; so close in fact that a graph of the y-p relationship will be a smooth, continuous function in which y rises monotonically as p ranges from 0 to 1 with only an arbitrarily small "break" in the line at the exact point where $\mathrm{p_i = P}$. The procedure suggested here simply fills in this one point on the line. I offer this as a reasonable, practical strategy to follow until a better solution is identified. When this approach is adopted, an interesting regularity is observed; the value of y always converges on 0.50 when $\mathrm{p_i}$ is set arbitrarily close to P.

An Observation

There is another interesting regularity in the y–p relationship. It is that y is always equal to Q when $\mathrm{p_i = Q}$. The basis for this regularity is that the expression $\sqrt{p_i q_i/PQ}$ takes the value of 1 when $\mathrm{p_i = Q}$. Accordingly, the expression

$$\left(1-\sqrt{p_i q_i/PQ}\right)\Big/\left(\mathrm{p_i/P-q_i/Q}\right)$$

takes the value of 0, yielding the result of $\mathrm{y = Q}$. The one exception is when Q is 0.5. In that situation, P also is 0.5 and y is undefined as just described above. However, the above procedure of substituting 0.5 for y when $\mathrm{p_i = P}$ also produces a result consistent with the regularity that $\mathrm{y = Q}$ when $\mathrm{p_i = Q}$.

References

Becker, H. J., James M., & Thomas, G. (1978). The measurement of segregation: The dissimilarity index and Coleman's segregation index compared. In the Proceedings of the social statistics section of the American Statistical Association (pp. 349–353). Washington, DC: American Statistical Association.

Bell, W. (1954). A probability model for the measurement of ecological segregation. *Social Forces, 32*, 357–364.

Blalock, H. M., Jr. (1979). *Social statistics*. McGraw-Hill (Revised Second Edition).

Coleman, J. S., Kelly, S. D., & Moore, J. A. (1975). *Trends in school segregation, 1968–1973*. Washington, DC: The Urban Institute.

Duncan, O. D., & Duncan, B. (1955). A methodological analysis of segregation indices. *American Sociological Review, 20*, 210–217.

Fossett, M. A., & Seibert, M. T. (1997). *Long time coming: Racial inequality in the nonmetropolitan south, 1940–1990*. Boulder: West view Press.

Fossett, M., & South, S. J. (1983). The measurement of inter-group income inequality: A conceptual review. *Social Forces, 61*, 855–871.

Hutchens, R. (2001). Numerical measures of segregation and their desirable properties. *Mathematical Social Sciences, 42*, 13–29.

Hutchens, R. (2004). One measure of segregation. *International Economic Review, 45*, 555–578.

James, D., & Taeuber, K. (1985). Measures of segregation. *Sociological Methodology, 13*, 1–32.

Lieberson, S. (1976). Rank-sum comparisons between groups. *Sociological Methodology, 3*, 276–291.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*(379–423), 623–656.

Stearns, L. B., & Logan, J. (1986). Measuring segregation: Three dimensions, three measures. *Urban Affairs Quarterly, 22*, 124–150.

Theil, H. (1972). *Statistical decomposition analysis*. Amsterdam: North-Holland.

Theil, H., & Finizza, A. J. (1971). A note on the measurement of racial integration of schools by means of informational concepts. *Journal of Mathematical Sociology, 1*, 187–194.

White, M. J. (1986). Segregation and diversity: Measures of population distribution. *Population Index, 65*, 198–221.

Zoloth, B. S. (1976). Alternative measures of school segregation. *Land Economics, 52*, 278–298.