Chapter 2
Alternative Formulas for Selected Indices

The values of all popular indices of uneven distribution can be obtained using a variety of mathematically equivalent computing formulas. For a given index some formulas are more familiar and widely used than others, but no single formula can be declared sacred or best for all purposes. The many alternatives can be confusing to those who are new to segregation research. But their availability benefits researchers by providing a variety of options from which to choose to best serve the needs of a particular study. The relevant considerations can include factors such as efficiency of computation, ease of explaining the index to broad audiences, relevance for establishing appealing substantive interpretations, capacity for enabling practical tasks such as decomposition analysis or the calculation of spatial versions of index scores, and utility for pinpointing technical issues in segregation measurement. Researchers may choose a particular formula specifically to serve the needs of a given study. Or they may use a formula based on familiarity and habit. But in one crucial sense the choice is unimportant as all valid formulas can be used interchangeably without affecting the results of individual index scores, research findings, and substantive conclusions.

To specialists well-versed in the literature on segregation measurement these are not surprising observations. Nevertheless, I raise the point because many researchers and most consumers of segregation research understand the quantitative underpinnings of segregation index scores based primarily on a handful of popular computing formulas. This is not a problem in itself. But problems can arise when lack of familiarity with mathematically equivalent alternatives makes individuals resistant to insights and interpretations that can be gained by drawing on alternative formulations of a particular index. This leads me to suggest that, while some formulas for popular indices of uneven distribution are better known and more widely used, it can be useful to consider other, less well known alternatives. In this chapter I discuss three classes of formulas. The formulas in the first group, which includes some well-known formulas that are very widely used in empirical research, focus attention on outcomes for areas and provide little insight into the relationship

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_2

$$G = 100 \cdot (\Sigma X_{i-1} Y_i - \Sigma X_i Y_{i-1}) \quad \text{(Duncan and Duncan 1955)}$$

$$D = 100 \cdot \tfrac{1}{2} \Sigma \left| (n_{1i}/N_1) - (n_{2i}/N_2) \right| \quad \text{(Duncan and Duncan 1955)}$$

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

Fig. 2.1 Examples of selected area-based computing formulas for indices of uneven distribution (Notes: N_1 and N_2 denote city-wide population counts for the two groups in the comparison; $T = N_1 + N_2$; i denotes area; n_{1i} and n_{2i} denote the area counts for the two groups in the segregation comparison; and X_i and Y_i denote the cumulative proportions of groups 1 and 2, respectively, over areas ranked from low to high on p_i obtained from $n_{1i}/(n_{1i}+n_{2i})$. A summary of notation used is given in Appendices)

between residential segregation and residential outcomes for individuals. The formulas in the second group establish that indices of uneven distribution are connected to the residential outcomes of individuals, but they not provide a basis for gaining insight into how residential outcomes differ across groups. The formulas in the third group go one step further and establish that indices of uneven distribution can be cast in ways that reveal how segregation is specifically connected to group differences on individual-levels residential outcomes associated with neighborhood racial composition.

Many, perhaps most, readers will have given little thought to how indices of uneven distribution are linked to individual residential outcomes. This would not be surprising as this aspect of indices of uneven distribution has not been emphasized in the literature on segregation measurement. It also is not obvious from inspecting the most widely used computing formulas for popular indices. Alternative formulas that do highlight the property tend not to be well known in addition to being infrequently used. In view of this, I use this chapter to briefly introduce formulas that highlight individual residential outcomes and contrast them with standard computing formulas. To streamline presentation, I offer minimal commentary here on the derivations of the new formulas that are introduced in this chapter. For those who are interested, I provide derivations and more detailed discussion of related technical issues as Appendices. In Chaps. 3, 4, and 5 in the body of the monograph I provide general discussions of the new formulas introduced here and then review their benefits for segregation measurement and analysis throughout the remainder of the study.

I begin by introducing computing formulas for three indices of uneven distribution that have very close relations to the segregation curve; namely, the gini index (G), the dissimilarity or delta index (D), and the Hutchens square root index (R). The formulas are given in Fig. 2.1. The formulas for G and D are likely to be familiar to many readers as they are widely used in segregation studies. In no small part this is because these formulas were introduced in Duncan and Duncan (1955), a landmark methodological study that served as the definitive guide to segregation measurement for three decades. In addition, they have continued to remain popular

because they are convenient computing formulas that are relatively easy to implement in empirical analyses. The formula for R was introduced more recently (Hutchens 2001) but I include it with the formulas for D and G because all three measures have close relations to the segregation curve and, as I document later in Chap. 6, all three are highly correlated in empirical applications. G and D are better known to sociologists. But R has technical properties that make it an attractive index to consider if one is committed to using a measure with close relations to the segregation curve.

The point I make about these three formulas is that they focus attention on outcomes for areas, not outcomes for individuals. The formulas adopt this orientation in part because it is efficient for computing index scores from area tabulations – a fact of non-trivial practical import in the early era of segregation research when Duncan and Duncan's study first appeared. In addition, these formulas fit comfortably with approaches to thinking about segregation that have an aggregate-level focus and frame the assessment of even distribution from the point of view of whether or not the racial composition of *areas* or neighborhoods matches the racial composition of the city as a whole. I note, however, that something important is left mysterious and obscure in these formulas. It is the residential outcomes that the individuals residing in these areas experience and how these outcomes may or may not vary systematically for the two groups in the segregation comparison.

The formulas for G and D given here are probably the two most widely applied computing formulas for measuring residential segregation. They also are likely to be the first two computing formulas students of segregation research learn. The fact that these formulas provide little to no basis for drawing insights about how segregation is connected to residential outcomes for individuals speaks volumes about the state of the literature on segregation measurement.

Figure 2.2 provides alternative formulas for G, D, and R and adds in similar formulas for two additional indexes, the Theil entropy index (H) and the separation index (S) (also known as eta squared [η^2] and the variance ratio). With the exception of the formula for R, these computing formulas also are likely to be familiar to many readers because they have been featured in many important methodological studies (e.g., Duncan and Duncan 1955; Zoloth 1976; James and Taeuber 1985; White 1986; Massey and Denton 1988). They, or close variations on them, are widely used in segregation studies. In no small part this is because they are convenient computing formulas that are relatively easy to implement in empirical analyses.

The formulas Fig. 2.2 have a key feature in common. Each formula incorporates the term "t_i" in the core calculations leading to the index value. This term represents the combined population of the two groups in the comparison residing in the i'th area in the city. The calculations involving this term are cumulated over all areas and at some point are divided by "T," the combined city-wide total populations of the two groups. Based on this construction, the index score can be understood as an average value for a quantitative result assessed for all individuals in the segregation comparison.

The point I want to make about these formulas is that the quantitative result computed for individuals can be viewed as an individual-level residential outcome or

$G = 100 \cdot (1/2T^2PQ) \cdot \Sigma\, t_i\, t_j \cdot \left| p_i - p_j \right|$ (James and Taeuber 1985:5)

$D = 100 \cdot (1/2TPQ) \cdot \Sigma\, t_i \cdot |p_i - P|$ (James and Taeuber 1985:6)

$R = 100 \cdot \left[1 - (1/T) \cdot \Sigma\, t_i \cdot \sqrt{p_i q_i / PQ}\right]$ (Appendix F, this monograph)

$H = 100 \cdot \Sigma\, t_i \cdot [(E - E_i)/ET]$ (Massey and Denton 1988:285)

$S = 100 \cdot (1.0 - [(\Sigma\, t_i \cdot p_i q_i)/TPQ])$ (Zoloth 1976:282) or
$100 \cdot (1/TPQ) \cdot \Sigma\, t_i (p_i - P)^2$ (James and Taeuber 1985:6)

Fig. 2.2 Examples of area-based computing formulas for indices of uneven distribution that implicitly feature overall averages on individual-level residential outcomes (Notes: N_1 and N_2 denote city-wide population counts for the two groups in the comparison; $T = N_1 + N_2$; $P = N_1/T$; $Q = N_2/T$; i denotes area; n_1 and n_2 denote the area counts for the two groups in the segregation comparison; $t = n_1 + n_2$; $p_i = n_{1i}/t_i$; $q_i = n_{2i}/t_i$; X_i and Y_i denote the cumulative proportions of groups 1 and 2, respectively, over areas ranked from low to high on p_i; and E denotes entropy for the city overall given by $E = P \cdot \mathrm{Log}_2(1/P) + Q \cdot \mathrm{Log}_2(1/Q)$ and E_i denotes entropy for area i given by $E_i = p_i \cdot \mathrm{Log}_2(1/p_i) + q_i \cdot \mathrm{Log}_2(1/q_i)$. A summary of notation is given in the Appendices)

residential attainment. I emphasize this point with the formulas listed in Fig. 2.3. These are alternative, mathematically equivalent versions of the formulas given in Fig. 2.2. The only difference is that the formulas have been rearranged to highlight and clarify how each index can be understood as an overall average of residential outcome scores (y) for individuals. A more detailed discussion of these formulas are given in the Appendices. Here I limit my comments to noting that the residential outcome terms (y) can be characterized as registering the degree to which the racial composition in the area the individual resides in departs from the racial composition of the city. In the case of G, D, H, and the first formula for S, the calculation of the departure score involves a city-specific constant that "scales" results so the final index score will fall in the range 0–1.

These formulations show that, if one chooses to do so, all popular measures of uneven distribution can be expressed in terms of individual residential outcomes. While this option has been available for most measures for many decades, mathematical expressions of this form have not been as widely used and discussed as the standard computing formulas. One reason for this is that formulating indices of uneven distribution as overall population averages on residential outcomes does not provide any significant practical advantages. Another reason is that these formulations do not support substantive interpretations that are viewed as useful and compelling for the study of segregation. Most studies that measure uneven distribution are motivated by the assumption that it ultimately carries important implications for group differences in residential distributions and residential outcomes. Casting uneven distribution as an overall average for residential outcomes, while a viable mathematical option, does not speak directly to a substantive interest focused on group differences in residential distributions and residential outcomes. Nevertheless, these formulations are relevant for my purposes because they make it clear that all

Index	Averaging Scores for y Over All Individuals	Scores Assigned to Individuals Based on Scaling Function $y_k = f(p_i)$
G =	$100 \cdot (1/T) \cdot \Sigma y_k$	$y_k = \Sigma \mid p_k - p_m \mid / 2TPQ$
D =	$100 \cdot (1/T) \cdot \Sigma y_k$	$y_k = \lvert p_i - P \rvert / 2PQ$
R =	$100 \cdot [1 - (1/T) \cdot \Sigma y_k]$	$y_k = \sqrt{p_i q_i / PQ}$
H =	$100 \cdot (1/T) \cdot \Sigma y_k$	$y_k = (E - E_i)/E$
S =	$100 \cdot (1/T) \cdot \Sigma y_k$	$y_k = (p_i - P)^2/PQ$ or, alternatively,
	$100 \cdot [1 - (1/T) \cdot \Sigma y_k]$	$y_k = p_i q_i / PQ$

Fig. 2.3 Formulas explicitly casting values of indices of uneven distribution as overall population averages on individual residential outcomes (y) (Notes: k and m index individual households; p_i denotes the pair-wise area proportion for the reference group in the i'th area; p_k denotes the value of p_i for the k'th household and p_m denotes the value of p_i for the m'th individual; See notes to Figs. 2.1 and 2.2 for other terms)

indices of uneven distribution have definite relations to residential outcomes for individuals.

Thinking about this led me to raise two questions that are central to this study. They are "Can indices of uneven distribution be formulated in a way that provides direct insights regarding group differences in residential outcomes?" and, if so, "How specifically do indices of uneven distribution register group differences on neighborhood residential outcomes?" The formulas presented in Fig. 2.4 address these questions. The formulas given here cast popular indices of uneven distribution as differences of means on individual residential outcomes (y) that are scored on the basis of the pairwise group proportion (p) for the area of residence. These expressions are new to this monograph and have not been presented previously in the literature on segregation measurement.

These formulas play a crucial role in this study; they constitute the mathematical basis for what I term the "difference of means" framework for segregation measurement. Accordingly, I review these formulas in more detail in Chap. 3 and I also provide additional technical discussions and derivations as Appendices. I conclude this short chapter with a few additional comments. This chapter establishes the point that all popular indices of uneven distribution can be given in a variety of mathematically equivalent formulations. Some are convenient for computing; some support attractive substantive interpretations; and some reveal how segregation is connected to residential outcomes for individuals and how these may differ across groups. All can be used to obtain correct values for index scores and thus they all are interchangeable for that narrow purpose. The new formulas introduced in Fig. 2.4 definitely can be used for this purpose. But that is not their main claim to fame. Their value to segregation research is that they provide unique advantages for segregation

	Difference of Group Means on y	Residential Outcome Scores (y) Assigned to Individuals Based on $y_i = f(p_i)$
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 difference of group means solution is available but $A = 2R - R^2$ for the "symmetric" version of A (i.e., A when $\alpha = \beta = 0.5$).	
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. 2.4 Formulas casting values indices of uneven distribution as differences of group means ($\overline{Y}_1 - \overline{Y}_2$) on individual residential outcomes (y) (Notes:· $\overline{Y}_1$ and $\overline{Y}_2$ are group averages given by $\overline{Y}_1 = \left(1 / N_1\right)\Sigma y_i$ and $\overline{Y}_2 = \left(1 / N_2\right)\Sigma y_i$ with i denoting individuals in the relevant group p_i denotes the pairwise area proportion for the reference group (p_i) in the area where the i'th individual resides and y_i is the residential outcome score generated by the index-specific scoring function $f(p_i)$. See notes to Figs. 2.1 and 2.2 for other terms)

measurement and new options for segregation analysis. They do so by placing all popular indices of uneven distribution in a common framework wherein all indices are given as group differences of means on individual residential outcomes (y) that are scored from the pairwise racial composition (p) of the area in which the individual resides. This framework provides a new basis for understanding, interpreting, and comparing familiar indices. It also opens the door to innovations in segregation measurement and analysis. I explore these possibilities in more detail in the remaining chapters of this monograph starting next with an overview to the "difference of means" framework.

References

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

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

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

Massey, D. S., & Denton, N. A. (1988). The dimensions of residential segregation. *Social Forces, 67*, 281–309.
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.