Chapter 15
New Options for Understanding and Dealing with Index Bias

In this chapter I introduce a new approach for addressing the problem of index bias at the point of measurement. Specifically, I introduce new formulations of popular indices of uneven distribution that are free of bias and take expected values of zero when individuals and households are randomly assigned to residential locations. I accomplish this task by drawing on the difference of means formulations of segregation indices introduced in earlier chapters to first identify and then eliminate the root source of bias in standard versions of popular indices of uneven distribution. The crucial insight from the difference of means formulation is that the values for all popular indices of uneven distribution can be seen as resting on person-specific scores for pairwise group contact (p). Close consideration reveals that the source of index bias is found in these group contact scores. Happily, a surprisingly simple refinement in the calculation of these scores eliminates index bias.

I review the root problem and its solution in more detail in the body of this chapter but offer a brief preview the essence of the problem and the solution here. To begin, recall that the difference of means framework establishes that all popular indices of uneven distribution can be formulated in terms of group differences in scaled residential exposure or contact. More specifically, the score for a particular index of uneven distribution can be obtained by calculating the difference of group means on individual residential outcomes (y) scored using an index-specific scaling function $y = f(p)$. The input to the scaling function, "p", is the individual's level of pairwise contact with the reference group in the comparison. The value of p is calculated from the area population counts for the two groups in the segregation comparison based on $p_i = n_{1i} / (n_{1i} + n_{2i})$. This approach to calculating the value of p introduces inherent upward bias in group differences on scores for p and also group differences on scores of y.

The source of bias is simple; the count terms (i.e., n_{1i} and n_{2i}) used in the calculation of group contact (p_i) include the individual in question. The score for contact thus combines two components of contact – *contact with self* and *contact with neighbors*. For any individual the component of contact that derives from contact

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_15

with neighbors can vary widely; it can range from no (0 %) contact with the reference group to only (100 %) contact with the reference group. In principle, this component of contact can be random for any individual regardless of group membership. Thus, under random assignment the expected value of this component of contact will be the same for every individual regardless of group membership and expected group differences will be zero (0). In contrast, the component of contact that derives from self-contact cannot be randomly assigned; it is fixed and invariant for each individual. Contact with self distorts group comparisons on contact because this component of contact inherently differs by race. Specifically, self-contact makes the assessed value of contact (p) *intrinsically higher* for members of the reference group and *intrinsically lower* for members of the comparison group. This is the source of bias in indices of uneven distribution.

This can be understood intuitively by considering the situation where residential assignments are random. The expected representation of the reference group among neighbors will obviously be same for all individuals and for both groups. But when self-contact is added in, the distribution of values on p necessarily shifts up for members of the reference group and necessarily shifts down for members of the comparison group. Index scores are computed from the difference of groups means on scaled contact (y) scored from simple pairwise contact (p). Since all of the index-specific scaling functions (i.e., $y = f(p)$) score y as a positive, monotonic function of p, the expected distribution of y will necessarily be higher for the reference group than for the comparison group. As a result, standard versions of indices of uneven distribution are biased upward; that is, their expected values under random assignment (E[•]) are positive.

I eliminate index bias in indices of uneven distribution by making a simple refinement to the contact calculation for individuals; *I assess contact using counts for neighbors instead of area population.* For purposes of discussion, I designate the revised version of contact as p′. This modification removes the fixed contribution of self-contact from the calculation of group contact scores for individuals. Intuitively, the expected representation of the reference group among neighbors is the same for all individuals under random assignment regardless of group membership. As a result, the expected distribution of values on contact with neighbors (p′) is the same for both groups. It follows necessarily that the same is true for the expected distribution of scaled contact (y′) scored from p′. Accordingly, the expected value of the group difference of means on scaled contact (y′) also is zero under random assignment. Thus, indices of uneven distribution calculated in this way are unbiased. Below I develop this conclusion more carefully. In Chap. 16 I report results of empirical analyses demonstrating that indices of uneven distribution computed using this relatively simple refinement take an expected value of zero under random assignment.

15.1 The Source of the Initial Insight

I should give credit where credit is due and note that a study by Laurie and Jaggi (2003) set me on the path to discovering a general strategy for developing unbiased versions of all popular indices of uneven distribution. Laurie and Jaggi used a Schelling-style agent simulation model to produce model-generated residential patterns in a virtual city.[1] As is common in agent models they assessed segregation at very small spatial scales. For purposes of the discussion here I consider the example of a city with simple housing grid that is divided into small "blocks" based on 3×3 square sections that contain 9 households.[2] Ordinarily, segregation assessed at this fine-grained spatial resolution would be subject to extremely high levels of index bias. For example, in a city with an 80/20 White-Black group ratio the value of E[D] would be 37.9 and the value of E[S] would be 11.1. Laurie and Jaggi (2003) measured segregation using an index of their own construction which they termed the "ensemble averaged, von Neumann segregation coefficient." They designated their measure as "S" but I term it "LJ" here to credit them and also to avoid confusion with using S to designate the separation index. Lauri and Jaggi claimed their index had an expected value of zero under random distribution; that is $E[LJ] = 0$. Initially I was skeptical of the claim. But I examined the behavior of their index in detail and discovered the claim was valid; Laurie and Jaggi's LJ index was indeed "unbiased." That is, over repeated trials of randomly generated residential distributions the distribution of values for scores on the LJ index will have a mean of zero.

Intrigued by this property and its potential benefits for measuring segregation in agent-models, I examined the formula for their index more closely to see how it related to more well-known indices of uneven distribution (Fossett 2007). I found the formula yielded the average over all individuals of a "scaled" score on same-group contact. For each individual the scaled score is obtained by first taking the difference between the observed proportion same-group among the individual's neighbors from the expected proportion based on the group's representation in the population and then expressing this result as a proportion of the maximum possible deviation under complete segregation. Putting this in notation more familiar to demographers and sociologists, scores for White households (agents) were given by $(p_i - P)/(1 - P)$ where P is proportion White in the population of agents and p_i is

[1] Laurie and Jaggi (2003) is one of many recent studies using Schelling-style agent simulation models – computer-implemented elaborations of the influential agent model of segregation dynamics first introduced in Schelling (1971).

[2] Laurie and Jaggi actually used a smaller, spatially delimited "von Neumann" or "rook's" neighborhood which consists of the 4 neighboring households who share sides with a focal household in a housing grid. I use the 3×3 "bounded" neighborhood to correspond better with practices in sociological segregation studies. All findings I note in this discussion also apply to spatially delimited neighborhoods of any spatial scale. But I defer detailed discussion of this topic for another time.

proportion White for the individual's neighbors.[3] Similarly, scores for Black households (agents) were given by $(q_i - Q)/(1-Q)$ where Q is proportion Black in the population of agents and q_i is proportion Black for the individual's neighbors. The sum of these scores is then divided by T, the total number of households (agents), to obtain the overall average. The resulting expression (dropping subscripts for convenience of presentation) is

$$LJ = (1/T)\cdot\left[\Sigma(p-P)/(1-P)+\Sigma(q-Q)/(1-Q)\right].$$

Interestingly, I found the separate averages for Whites and Blacks calculated as shown below also gave the same result. That is,

$$LJ = (1/W)\cdot\Sigma(p-P)/(1-P) = (1/B)\cdot\Sigma(q-Q)/(1-Q).$$

These expressions can be restated as follows

$$LJ = (\Sigma p/W-P)/(1-P) = (\Sigma q/B-Q)/(1-Q).$$

This expression reveals a close correspondence between LJ and Bell's (1954) revised index of isolation (I_R). Bell's I_R expresses a group's average for same-group contact as a proportion of its possible logical range. For Whites and Blacks, respectively, I_R would be given as

$$I_R = (P_{WW} - P)/(1-P), \text{and}$$

$$I_R = (P_{BB} - Q)/(1-Q)$$

where: $P_{WW} = (1/W)\cdot\Sigma(w_i \cdot p_i)$; $P_{BB} = (1/B)\cdot\Sigma(b_i \cdot q_i)$; $P = (W/T)$; $Q = (B/T)$; W, B, and T are the city totals for the White, Black, and Total populations, respectively; w_i, b_i, and t_i, are the counts for White, Black and Total population in area i; and p_i and q_i are area proportion White and Black, respectively, based on w_i/t_i and b_i/t_i.

The contact expressions P_{WW} and P_{BB} can be restated as $\Sigma(w_i \cdot p_i)/W$ and $\Sigma(b_i \cdot q_i)/B$, respectively. If the calculations are expressed from the point of view of individuals, as in Lauri and Jaggi, they can be given as Σp/W and Σq/B. Thus, I_R for Whites and Blacks will take the same form given above for LJ. Thus,

$$I_R = (\Sigma p/W-P)/(1-P), \text{ and}$$

[3] To clarify terms in this discussion, city-level terms are given as follows: W and B are totals for Whites and Blacks, respectively, $T = W + B$, $P = W/T$, and $Q = B/T$. For each individual, w and b are the number of White and Black *neighbors* in the relevant neighborhood, $t = w + b$, and p and q are proportion White and Black, respectively, based on $p = w/t$ and $q = b/t$.

$$I_R = (\Sigma q / B - Q) / (1 - Q)$$

As shown here the two measures – LJ and I_R – appear to be equivalent, but there is an important difference between them that causes I_R and LJ to exhibit fundamentally different behavior. The difference in behavior is that Bell's I_R will manifest positive bias (i.e., $E[I_R] > 0$) while Laurie and Jaggi's LJ will be unbiased (i.e., $E[LJ] = 0$). The difference in behavior traces to one crucial difference between the calculations for the two indices. It is the difference in how the values of p and q are calculated for LJ and I_R. For Bell's I_R the calculation of contact terms follows the standard methodological practice in sociological segregation studies; the contact terms p and q are calculated using count terms for the full *area population*. Significantly, this calculation *includes* the focal household in the count terms that appear in the numerator and the denominator of the contact calculations. In contrast, for Laurie and Jaggi's LJ the calculation of contact terms p and q is based on a different procedure; it uses count terms for the focal household's *neighbors.* Thus, the approach Laurie and Jaggi use *excludes* the focal household from the count terms used in the calculations. To clarify, the contact scores used in calculating I_R and LJ differ as follows.

For I_R, $p = w / t$ and $q = b / t$.

For LJ, $p' = (w - 1) / (t - 1)$ and $q' = (b - 1) / (t - 1)$.

I use the prime symbol to differentiate contact based on neighbors from contact based on area population.

Closely comparing the design and behavior of the two measures led me to draw several conclusions. One is that, when focusing on a two group comparison, the LJ index can be described as an unbiased version of I_R. Another is that the only difference between the standard (biased) and unbiased versions of I_R is how contact is calculated. Specifically, self-contact is eliminated in the unbiased LJ version and this is accomplished by the simple exercise of excluding the focal household from the count terms that appear in the numerator and denominator of the contact calculations. This revealed that bias in I_R traces to a single source – the impact of incorporating self-contact into the calculation of group contact scores for individuals. It also revealed that bias could be eliminated by following Laurie and Jaggi's example and making the simple adjustment of computing group contact for individuals based on count terms for *neighbors* instead of count terms for *area population.* When this adjustment is implemented, values of I_R take an average value of zero when calculated over repeated trials for random residential distributions.

15.2 Building on the Initial Insight

Based on these intriguing findings, I focused on the question of whether this measurement strategy could be adapted in a general way for application with measures of uneven distribution. I focused first on the separation index (S) as a natural first choice because it is equivalent to Bell's revised index of isolation (I_R) in the special case where the city population consists of only two groups (James and Taeuber 1985; Stearns and Logan 1986; White 1986).[4] In light of this it is straightforward to describe Laurie and Jaggi's LJ index as an unbiased version of the separation index (S). Thus, Laurie and Jaggi deserve credit for establishing the core strategy for developing an unbiased version of S.

Initially I was frustrated in applying this insight to other indices of uneven distribution. The crucial insight of the strategy is to eliminate bias by eliminating the impact of self-contact from group contact calculations. But the best known computing formulas for indices of uneven distribution do not provide an obvious opportunity for acting on this insight because they do not yield index scores as group differences in average contact outcomes for individuals. As one example, James and Taeuber (1985: 6) give the following widely used computing formula for calculating the value of separation index

$$S = 1/NPQ \cdot \Sigma t_i \left(p_i - P\right)^2 .$$

This formula is efficient for computing values of S. But it does not give the value of S as a group difference in average contact scores for individuals. Moreover, I found that implementing the p_i adjustment used by Laurie and Jaggi in this formula did not yield an unbiased version of S with the desirable properties of the version established by Laurie and Jaggi.

I then struck on a second key insight. It is that eliminating bias from index scores first requires that the index be formulated as a difference of means on residential outcomes scored from pairwise contact. This isolates the impact of group differences in self-contact separately by group so its role can be eliminated. This prompted me to search for a formulation of the separation index that (a) would highlight the role of average group contact outcomes for individuals and (b) could be used as a template for deriving similar formulations for other popular indices of uneven distribution.

Appendices outline a derivation I that achieved this goal by expressing the separation index (S) as a group difference of means on contact with the reference group in the comparison.[5] I review a generic formulation in the additional material but give the result here using the example of White-Black segregation with Whites being

[4] That is, one can describe the separation index (S) as a special case of Bell's Revised Index of Isolation (I_R) computed using only pairwise population counts.

[5] Later I found a similar derivation had been reported much earlier in a little known methodological paper by Becker et al. (1978).

designated as the reference group. Thus, S is the White-Black difference in average contact with Whites based on

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

where $P_{WW} = (1/W) \cdot \Sigma (w_i \cdot p_i)$, and $P_{BW} = (1/B) \cdot \Sigma (b_i \cdot p_i)$, with "W" and "B" designating total population for the reference group (Whites) and the comparison group (Blacks), respectively, w_i and b_i indicating area counts for the two groups, and $p_i = w_i / (w_i + b_i)$ indicating pairwise contact with the reference group for individuals residing in area "i".

Refining the contact calculations to eliminate the role of self-contact, leads to the unbiased version of S given as

$$S' = P'_{WW} - P'_{BW}$$

where P'_{WW} and P'_{BW} are contact expressions based on counts for neighbors instead of area population. They are obtained as follows. $P_{WW} = (1/W) \cdot \Sigma (w_i \cdot p'_i)$, and $P_{BW} = (1/B) \cdot \Sigma (b_i \cdot p'_i)$, with p_i' being calculated from $(w_i - 1)/(w_i + b_i - 1)$ for Whites and from $(w_i - 0)/(w_i + b_i - 1)$ for Blacks.

15.3 A More Detailed Exposition of Bias in the Separation Index

I now review the issue of index bias for the separation index (S) in more detail. I continue with the example of White-Black segregation and for simplicity consider a situation where the city in question is not small, consists of only Whites and Blacks, and is divided into areas of constant size in terms of area population (t).[6] I start with the question of "What can be expected when households are distributed randomly across housing units in all areas of the city?" For any household, White or Black, the expected contact with Whites is assessed using counts for *neighbors.* Normally I designate this as p′ but for the current discussion I also sometimes designate it as p_N using the subscript "N" to indicate "computed for neighbors." The expected value of this calculation is essentially equal to proportion White in the city (i.e., $E[P'_{WW}] = E[P'_{BW}] = P = W/(W+B)$).[7]

Intuitively, this is easy to understand. When a household's neighbors are obtained by a random draw from a large city population, the expected proportion Whites for the neighbors will be the city proportion White ($E[p'] = P$). Note that the expected

[6] The assumption that the city is not small assures that an individual household has a negligible impact on the city-wide group proportion for the reference group (P).

[7] For ease of presentation, I ignore the impact of the focal household's contribution to P for the city as a whole. In most empirical applications, the impact is negligible.

result is essentially the same for all households whether White or Black. More exactly, there is a very slight difference in expected value associated with the contribution the focal household makes to the combined total of Whites and Blacks and how this varies with the race of the focal household. In a very small city the P and Q results might differ slightly by race if one calculated P′ as $(W-1)/(W+B-1)$ for Whites and $(W-0)/(W+B-1)$ for Blacks. In larger cities this potential difference becomes negligible and I ignore it here for convenience of exposition.

The results for expected contact in local areas can be quite different when contact with Whites (p) is assessed using counts for *area population* instead of counts for *neighbors*. Expected contact with Whites (E[p]) will now reflect the weighted average of two contributions. The first contribution is the household's contact with White neighbors (p_N). As noted in the previous paragraph, this reflects a random draw of Whites and Blacks and its expected value is equal to P for both Whites and Blacks. The second contribution is the household's self-contact with Whites (p_S). The value of self-contact will be 1 for White households and 0 for Black households so the contribution of self-contact to contact with Whites in the area population (p) varies systematically by race. The relative contribution of the two components of contact depends on the value of area (pairwise) population size (t). Contact with Whites based on area population can be given by

$$p = p_N \cdot (t-1)/t + p_S \cdot (1/t) \cdot$$

where t is area population and $t-1$ is the number of neighbors a household has.

Under random distribution, the expected value of the term $p_N \cdot (t-1)/t$ is the same for every household in the city. But the term $p_S \cdot (1/t)$ is systematically different for Whites and Blacks. Specifically, p_S is 0/t for Blacks and 1/t for Whites. This causes the expected value of the White-Black difference in mean contact with Whites to differ by 1/t.

To further clarify, I examine expected contact separately by race. A White household's expected number of White neighbors under random assignment is given by the household's number of neighbors $(t-1)$ multiplied by expected contact with Whites for neighbors (p_N) which as noted above is $E[P'_{WW}] = P = W/(W+B)$. Unsurprisingly, the White household's expected self-contact with Whites in the area population (p_S) is 1. As a result the expectation for White contact with Whites in the standard contact formulation based on area population (i.e., $E[P_{WW}]$) can be given as follows.

$$E[P_{WW}] = E[P'_{WW}] \cdot ((t-1)/t) + 1.0 \cdot (1/t)$$

A Black household's expected number of White neighbors under random assignment is the same as that expected for a White household. It is given by the household's number of neighbors $(t-1)$ multiplied by expected contact with Whites for neighbors (p_N) which as noted above is $E[P'_{BW}] = P = W/(W+B)$. Unsurprisingly, the Black household's expected self-contact with Whites in the area population (p_S)

is 0. As a result the expected value for Black contact with Whites in the standard contact formulation based on area population (i.e., $E[P_{BW}]$) can be given as follows.

$$E\left[P_{BW}\right]=E\left[P'_{BW}\right]\cdot\left((t-1)/t\right)+0.0\cdot(1/t)$$

Because $E\left[P'_{WW}\right]=E\left[P'_{BW}\right]=P$*, it is now becomes clear that upward bias in the separation index (S) traces solely to role of self-contact in the group contact calculations for the standard formula for the index.*

In the difference of means formulation $S=P_{WW}-P_{BW}$ (given here in pairwise P* contact notation) and the expected value of S is given by the expected value of its components. That is, $E[S]=E\left[P_{WW}\right]-E\left[P_{BW}\right]$. This can be evaluated as follows.

$$E[S]=E\left[P_{WW}\right]-E\left[P_{BW}\right]$$

$$E[S]=\left[\left((t-1)/t\right)\cdot E\left[P'_{WW}\right]+(1/t)\cdot 1.0\right]-\left[\left((t-1)/t\right)\cdot E\left[P'_{BW}\right]+(1/t)\cdot 0.0\right]$$

$$E[S]=[\left((t-1)/t\right)\cdot E\left[P'_{WW}\right]-\left[\left((t-1)/t\right)\cdot E\left[P'_{BW}\right]+\left[(1/t)\cdot 1\right]-(1/t)\cdot 0\right]$$

$$E[S]=\left[\left((t-1)/t\right)\cdot P_{W}-\left((t-1)/t\right)\cdot P_{W}\right]+\left[(1/t)\cdot 1\right]-(1/t)\cdot 0]$$

$$E[S]=(1/t)\cdot 1-(1/t)\cdot 0$$

$$E[S]=1/t$$

Note that this result is identical to the expected value for S previously established and reported by Winship (1977: 1064).

Now consider the expected value for the separation index when contact for individuals is assessed using counts for neighbors instead of counts for area population.[8]

$$E\left[S'\right]=E\left[P'_{WW}\right]-E\left[P'_{BW}\right]$$

$$E\left[S'\right]=P-P$$

$$E\left[S'\right]=0$$

[8] Again, this assumes city size is sufficiently large that an individual household's contribution to P is negligible.

This establishes that an unbiased version of the separation index (i.e., S′ with $E\left[S'\right]=0$ *) can be obtained by eliminating the role of self-contact when assessing each individual's contact with the reference group.*

15.4 Situating This Result and Its Implications in the Difference of Means Framework

I now recast the results for S just presented in the notation of the more general difference of means framework. In that framework the standard formula for S is

$$S = 100\cdot\left(\bar{Y}_1 - \bar{Y}_2\right) = \left(1/W\right)\cdot\Sigma\left(w_i\cdot y_i\right) - \left(1/B\right)\cdot\Sigma\left(b_i\cdot y_i\right).$$

When computing S by this formula, values of y_i are set according to the index-specific scaling function $y = f(p)$. In the case of S, the scaling function is the identity function and thus $y_i = p_i$. Accordingly, the contact formula for S White-Black segregation given in the preceding section

$$S = P_{WW} - P_{BW} = \left(1/W\right)\cdot\Sigma\left(w_i\cdot p_i\right) - \left(1/B\right)\cdot\Sigma\left(b_i\cdot p_i\right)$$

can be converted into the difference of means formula for S by simply substituting y_i for p_i.

I introduce the unbiased version of the separation index (S′) first for two reasons. One is that, as mentioned earlier, it was the first index for which I was able to establish an unbiased version. The second is that the nature of bias for S is especially straightforward and easy to explain. But S is not a special case among indices of uneven distribution. The core strategy of revising the formula to remove the contribution of self-contact can be applied to any index of uneven distribution that can be placed in the difference of means framework.

In standard index calculations group contact is assessed using area population counts and thus reflects the weighted average of two components. The first component registers contact with *neighbors*. This expected value of this component of contact is the same for all individuals and groups in the comparison and so does not contribute to index bias. The second component registers self-contact which is fixed for every individual and differs systematically by group. This introduces bias by systematically inflating contact scores for members of the reference group and reducing contact scores for members of the comparison group. Eliminating the second component from contact calculations yields unbiased group means on contact scores and this results in an unbiased index score.

To summarize, the following two important conclusions apply to all popular indices of uneven distribution – including G, D, A, R, and H – that can be place in the difference of means framework.

- bias in standard index formulations traces to calculating group contact (p_i) for households based on area population counts, and
- unbiased versions of the index can be obtained by calculating group contact based on counts for neighbors.

I now briefly review how these conclusions generalize and apply to other popular and widely used indices of uneven distribution.

15.4.1 Expected Distributions of p′ and y′ Under Random Assignment

When households are randomly assigned to areas, the expected distribution of raw contact scores calculated using counts for *neighbors* (hereafter designated p_i') will be the same for both Whites and Blacks. As a result, expected values for group means on scaled exposure (y_i') scored based on *any* index-specific scaling of "raw" contact among neighbors (p_i') will be the same for both Whites and Blacks (i.e., $E\left[Y_W'\right] = E\left[Y_B'\right]$).

This can be established as follows. The expected distribution of values for raw contact with the reference group (p_i') calculated using counts for neighbors will be given by the binomial probability distribution for a given number of neighbors. This expected distribution will be the same regardless of whether the focal household for this set of neighbors is White or Black. Thus, the expected distribution of p_i' will be the same for Whites and Blacks. Values of contact with the reference group (p_i') determine residential outcome scores (y_i'). So the expected distribution of contact scores (p_i') directly determines the expected distribution of residential outcomes scores (y_i'). This also will be the same for Whites and Blacks. The expected distribution of residential outcomes (y_i') determines the expected mean on scaled contact (Y′) and this also will be the same for Whites and Blacks. Because the expected means on scaled contact are the same for Whites and Blacks (i.e., $Y_W' = Y_B'$), the expected group difference of means (i.e., $Y_W' = Y_B'$), difference under random assignment is zero. This leads to the following general conclusion.

> *Scores for indices computed as a difference of means in scaled contact with the reference group calculated for neighbors (instead of area population) will be unbiased. That is, the expected value of index scores under random assignment will be zero (0.0).*

15.5 Reviewing a Simple Example in Detail

It is instructive to review a simple example in some detail to show how expected group means on residential outcomes (y) differ depending on whether an individual's contact with the reference group (p) is assessed using counts for neighbors or counts for area population. For purposes of illustration I consider the example of a

Table 15.1 Calculations to obtain values of D and S for White-Black segregation from differences of group means on residential outcomes (y) based on contact with Whites for area population and among neighbors under random distribution

Count of Whites	Whites p (×100)	Blacks p (×100)	Share of Whites	Share of Blacks	Whites y_D (×100)	Blacks y_D (×100)	Whites y_S (×100)	Blacks y_S (×100)
Among neighbors								
0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1–11	–	–	–	–	–	–	–	–
12	60.00	60.00	0.04	0.04	0.00	0.00	60.00	60.00
13	65.00	65.00	0.20	0.20	0.00	0.00	65.00	65.00
14	70.00	70.00	0.89	0.89	0.00	0.00	70.00	70.00
15	75.00	75.00	3.19	3.19	0.00	0.00	75.00	75.00
16	80.00	80.00	8.98	8.98	0.00	0.00	80.00	80.00
17	85.00	85.00	19.01	19.01	0.00	0.00	85.00	85.00
18	90.00	90.00	28.52	28.52	100.00	100.00	90.00	90.00
19	95.00	95.00	27.02	27.02	100.00	100.00	95.00	95.00
20	100.00	100.00	12.16	12.16	100.00	100.00	100.00	100.00
Sum or mean			100.00	100.00	67.69	67.69	90.00	90.00
For area population								
0	N/A	0.00	0.00	0.00	N/A	0.00	N/A	0.00
1–11	–	–	–	–	–	–	–	–
12	57.14	57.14	0.01	0.04	0.00	0.00	57.14	57.14
13	61.90	61.90	0.04	0.20	0.00	0.00	61.90	61.90
14	66.67	66.67	0.20	0.89	0.00	0.00	66.67	66.67
15	71.43	71.43	0.89	3.19	0.00	0.00	71.43	71.43
16	76.19	76.19	3.19	8.98	0.00	0.00	76.19	76.19
17	80.95	80.95	8.98	19.01	0.00	0.00	80.95	80.95
18	85.71	85.71	19.01	28.52	0.00	0.00	85.71	85.71
19	90.48	90.48	28.52	27.02	100.00	100.00	90.48	90.48
20	95.24	95.24	27.02	12.16	100.00	100.00	95.24	95.24
21	100.00	N/A	12.16	N/A	100.00	N/A	100.00	N/A
Sum or mean			100.00	100.00	67.69	39.17	90.48	85.71

Notes: "N/A" indicates the combination does not occur. "–" indicates outcomes are omitted because their frequency is negligible

hypothetical city where the population consists of only Whites and Blacks, proportion White for the city (P) is equal to 0.90, and area size (t_i) is equal to 21 households.[9] Table 15.1 presents the expected distributions for contact scores (p) and index-specific residential outcomes scores (y) for the dissimilarity index (D) and the

[9] The number of households is substantially higher than would be found in typical census blocks but substantially lower than would be found in typical census block groups.

Table 15.2 Calculations to obtain values of R and H for White-Black segregation from differences of group means on residential outcomes based on contact with Whites for area population and among neighbors under random distribution

Count of Whites	Whites p (×100)	Blacks p (×100)	Share of Whites	Share of Blacks	Whites y_R (×100)	Blacks y_R (×100)	Whites y_H (×100)	Blacks y_H (×100)
Among neighbors								
0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1–11	–	–	–	–	–	–	–	–
12	60.00	60.00	0.04	0.04	28.99	28.99	42.11	42.11
13	65.00	65.00	0.20	0.20	31.24	31.24	45.70	45.70
14	70.00	70.00	0.89	0.89	33.74	33.74	49.56	49.56
15	75.00	75.00	3.19	3.19	36.60	36.60	53.79	53.79
16	80.00	80.00	8.98	8.98	40.00	40.00	58.54	58.54
17	85.00	85.00	19.01	19.01	44.24	44.24	64.06	64.06
18	90.00	90.00	28.52	28.52	50.00	50.00	70.83	70.83
19	95.00	95.00	27.02	27.02	59.23	59.23	80.08	80.08
20	100.00	100.00	12.16	12.16	100.00	100.00	100.00	100.00
Sum or mean			100.00	100.00	55.96	55.96	73.69	73.69
For area population								
0	N/A	0.00	0.00	0.00	N/A	0.00	N/A	0.00
1–11	–	–	–	–	–	–	–	–
12	57.14	57.14	0.01	0.04	27.79	27.79	40.15	40.15
13	61.90	61.90	0.04	0.20	29.82	29.82	43.45	43.45
14	66.67	66.67	0.20	0.89	32.04	32.04	46.95	46.95
15	71.43	71.43	0.89	3.19	34.51	34.51	50.73	50.73
16	76.19	76.19	3.19	8.98	37.35	37.35	54.87	54.87
17	80.95	80.95	8.98	19.01	40.73	40.73	59.52	59.52
18	85.71	85.71	19.01	28.52	44.95	44.95	64.93	64.93
19	90.48	90.48	28.52	27.02	50.68	50.68	71.57	71.57
20	95.24	95.24	27.02	12.16	59.85	59.85	80.63	80.63
21	100.00	N/A	12.16	N/A	100.00	N/A	100.00	N/A
Sum or mean			100.00	100.00	56.56	46.34	74.35	66.04

Notes: "N/A" indicates the combination does not occur. "–" indicates outcomes are omitted because their frequency is negligible

separation index (S) under random residential distributions based on a binomial probability model. Table 15.2 presents similar results for the Hutchens square root index (R) and the Theil entropy-based index (H). The first panel in each table gives the results when households' contact with Whites is assessed using counts for neighbors. The second panel in each table gives the parallel results when households'

contact with Whites is assessed using counts for area population in the standard way.

I first review the results in Table 15.1. The first column in the first panel of the table lists the possible counts for Whites among neighbors. The areas in the example have 21 total households so every household has exactly 20 neighbors, a situation that would be common when measuring segregation using block-level data. Except for the outcome of 0, which warrants separate comment, the outcomes for counts of White neighbors below 12 are omitted from the listing because their occurrence under random distribution is quantitatively negligible. The values of proportion White among neighbors (p′) is given separately for Whites and Blacks in the next two columns. Note that proportion White among neighbors is the same for both Whites and Blacks under all possible combinations. The share – that is, the proportion – of households in the group expected to experience each of the possible levels of contact under random distribution is given separately for Whites and Blacks in the next two columns. Note that group shares at every outcome are the same for both White and Black households. Scores of residential outcomes y′ scored from p′ using in computing the dissimilarity index (D) under the difference of means calculation approach are reported separately for Whites and Blacks in the next two columns. Scores of residential outcomes (y′) relevant for computing the separation index (S) are reported separately for Whites and Blacks in the last two columns. The results for the expected group means on index-specific residential outcomes are given in the bottom row of the panel. These are obtained by summing the products of group shares and residential outcomes scores (y′).

Table 15.2 continues the exercise and has the same structure as Table 15.1. The only difference is that it provides information on the residential outcomes (y′) that are used in computing the Hutchens square root index (R) and the Theil entropy index (H).

The results for the analysis in the first panels in Tables 15.1 and 15.2 are easy to summarize. For all four indices – D, S, R, and H, Whites and Blacks both experience all possible outcomes on p′ and both groups identical expected distributions across possible outcomes on number of White neighbors. Accordingly, they have identical expected values for the means on the unbiased version of the residential outcome scores (y′) that determine each segregation index score. Consequently, the expected values of D′, S′, R′, and H′ all are zero (0.0). For example, proportion White among neighbors equals the city-wide proportion (0.90) when the count of White neighbors is 18, 19, or 20. Residential outcomes (y′) relevant for calculating D are scored 1.0 in these cases and 0.0 in all other cases. Column 4 shows that 67.69 % of Whites experience this residential outcome. Column 5 shows that the same is true for Blacks. Accordingly, the expected mean for the 0–1 scoring of y scored for D is 0.6769 for both Whites and Blacks (values shown in the final row of columns 6 and 7). This result shows that Whites and Blacks are equally likely to reside in areas where their contact with White neighbors equals or exceeds the proportion White in the city as a whole. As a result, the expected value of D′ is 0.0 (i.e., $E\left[D'\right] = \left(E\left[Y'_W\right] - E\left[Y'_B\right]\right) = \left(0.6769 - 0.6769\right)$).

The group means reported in columns 8 and 9 show that Whites and Blacks also experience identical average levels of contact with Whites neighbors; spe-

cifically, on average 90.0 % of their neighbors are White, a level of contact matching the representation of Whites in the city population overall. So the expected value of S′ also is 0.0 (i.e., $D\left[S'\right] = \left(E\left[Y'_W\right] - E\left[Y'_B\right]\right) = \left(0.9000 - 0.9000\right)$). Similar results are seen when residential outcomes are scored as relevant for computing the Hutchens square root index (R′) (i.e., $E\left[R'\right] = 0.0 = \left(E\left[Y'_W\right] - E\left[Y'_B\right]\right) = \left(0.5596 - 0.5596\right)$) and the Theil entropy index (H′) (i.e., $E\left[H'\right] = 0.0 = \left(E\left[Y'_W\right] - E\left[Y'_B\right]\right) = \left(0.7369 - 0.7369\right)$).

These results are easy to summarize. When neighbors are a random draw, Whites and Blacks have identical probability distributions for experiencing different levels of unbiased contact with White neighbors (p′). It then follows that Whites and Blacks also have identical group means on residential outcomes (y′) scored from unbiased contact with White neighbors (p′).

I now review the results in the second panel of Tables 15.1 and 15.2 where contact with Whites is computed in the standard way based on counts for area population. The results here play out much differently. The key change producing the differences is that counts in the numerator and denominator of the calculation of proportion White (p) now include the focal household. Accordingly, the value for a household's contact with Whites (p) based on area population reflect a weighted average of the household's contact with Whites for neighbors (p′) and the household's self-contact with Whites designated here by p_S which is $1 = \left(1/1\right)$ for White households and $0 = \left(0/1\right)$ for Black households. The relevant expression is

$$p = p' \cdot \left(20/21\right) + p_S \cdot \left(1/21\right)$$

The distribution of values for contact with Whites among neighbors (p′) remains the same as before. This means that all changes in contact with Whites in the lower panel trace to the impact of self-contact with Whites (p_S) which is systematically different for Whites and Blacks.

To see the implications it is useful to consider how the results change for a household with 18 White neighbors, the case that in this example has important implications for the expected value of the dissimilarity index. For both White and Black households who have 18 White neighbors the value of contact with Whites among neighbors (p′) is 0.90 and results in a value of $y' = 1$ when residential outcomes (y′) are scores as relevant for the dissimilarity index (D). The results change when contact with Whites is based on area population (p). For a White household the value of contact with Whites based on area population (p) is given by

$$\begin{aligned} p &= p' \cdot \left(20/21\right) + p_S \cdot \left(1/21\right) \\ &= \cdot \left(18/20\right) \cdot \left(20/21\right) + \left(1/1\right) \cdot \left(1/21\right) \cdot \\ &= \left(0.90 \cdot 0.9524\right) + 0.0476 \\ &= 0.8571 + 0.0476 \\ &= 0.9048. \end{aligned}$$

For a Black household the value of p is given by

$$\begin{aligned} p &= p' \cdot (20/21) + p_S \cdot (1/21) \\ &= (18/20) \cdot (20/21) + (0/1) \cdot (1/21) \cdot \\ &= (0.9524 \cdot 0.90) + 0.0 \\ &= 0.8571 + 0.0 \\ &= 0.8571. \end{aligned}$$

The White and Black households have identical contact with Whites among neighbors and accordingly in the upper panel are scored identically on the residential outcome ($y' = 1$) relevant for computing D′. But in the lower panel the residential outcome (y) relevant for computing D is scored 1 for the White household – based on $0.9048 \geq 0.90$ – and 0 for the Black household – based on $0.8571 < 0.90$.

The expected proportion of households that have 18 White neighbors is 0.2852 for both Whites and Blacks. The difference in how these households are scored on scaled contact with Whites in the upper and lower panels contributes to determining the level of bias in D. Whites are scored the same in both the upper and lower panels; $y' = y = 1$. But Blacks are scored differently in the upper and lower panels; $y' = 1$ in the upper panel and $y = 0$ in the lower panel. This difference reduces the expected Black mean on scaled contact with Whites from $E[Y'_B] = 0.6769$ based on neighbors in the upper panel to $E[Y_B] = 0.3917$ based on area population in the lower panel. In contrast, the expected White mean on scaled contact with Whites is the same – $E[Y'_W] = E[Y_W] = 0.6769$ – under both calculations. Thus, the expected value of D changes from 0.0 when contact with Whites (p′) is based on neighbors ($E[D'] = E(Y'_W] - E(Y'_B] = 0.6769 - 0.6769 = 0.0$) to 0.2852 when contact with Whites (p) is based on area population ($E[D] = E(Y_W] - E(Y_B] = 0.6769 - 0.3917 = 0.2852$).

Scaling to 100 in keeping with convention, the "bias" in the standard version of the index of dissimilarity (D) under random distribution is 28.52. The parallel calculations for the separation index (S) ($E[S] = E(Y_W] - E(Y_B] = 0.9048 - 0.8571 = 0.0477$) indicate that bias in the standard version is 4.77. The interested reader can confirm that these values for E[D] and E[S] are equal to values of E[D] and E[S] obtained using analytic formulas given in Winship (1977).

This example reveals in detail how bias enters into the picture and distorts scores for standard versions of indices of uneven distribution. The example also documents how the simple refinement of assessing group contact based on neighbors instead of area population eliminates index bias for all indices of uneven distribution that can be placed in the differences of means formulation. The basis for this welcome result is easy to summarize. When self-contact is eliminated from that calculation, the two groups in the comparison will have identical expected distributions for the number of neighbors from the reference group and the number of neighbors from the comparison group. It then follows that expected group means on residential outcomes

(y′) scored of the distribution of unbiased contact values (p′) will be identical for both groups.

15.5.1 Additional Reflections on Results Presented in Tables 15.1 and 15.2

The analysis presented in Tables 15.1 and 15.2 clarifies how index bias originates in the role of self-contact. The results provide an intuitive basis for understanding why bias is greater when effective area size (ENS) is small. It is because self-contact will have a bigger impact on assessments of an individual's contact with the reference group when area counts are small as they are in this example. If the same exercise were repeated with area population size set to 5,001 instead of 21, the resulting magnitude of index bias would be much smaller. Alternatively, if the exercise were repeated with area counts of 9 (equivalent to a "Queen's" neighborhood of eight adjacent neighbors plus the focal household), the magnitude of index bias would be even larger.

Reflecting on the difference between unbiased contact (p′) and standard contact (p) also yields additional insight into why the expected level of bias varies from index to index. The role of self-contact in standard calculations of contact is to shift the distribution of values of p up for the reference group and down for the comparison group. These shifts in p are then translated into impacts on scaled contact (y) based on the index-specific scaling function $y = f(p)$. I established earlier that the scaling functions for G, D, R, and H are nonlinear. The nonlinearity has implications for bias. Specifically, *bias at the level of group differences on raw contact (p) will translate into larger group differences in scaled contact (y) when the scaling function is nonlinear and the magnitude of bias is greater when the scaling function is more strongly nonlinear*. This provides a succinct explanation for why levels of bias are higher for G and D compared to R and H and why the level of bias is lowest for S. The scaling function $y = f(p)$ for S is linear; so bias impacting the value of p is carried forward unchanged. The scaling functions for G and D depart from linearity the most; so bias impacting p is "amplified" to a greater degree when values of y are assigned for these indices. The scaling functions for R and H involve milder nonlinearity; so, while bias impacting p also is amplified when values of y are assigned, the resulting distortion is not as dramatic.

Finally, this also provides an explanation for why bias in S does not vary with group size, but bias in the other measures, and especially in G and D, does vary with group size. The reason is that the nonlinear scaling functions for G and D measures become more strongly nonlinear when groups are unequal in size. This means that the role of nonlinearity in exaggerating group differences in y scored from p is magnified for these measures when groups are more imbalanced in size.

15.6 Summary

This chapter reviews how the difference of means formulation of indices of uneven distribution leads to new insights about the nature of index bias and makes it possible to address index bias at the point of measurement. The insight is that, when segregation is cast as group differences in means on scaled group contact, bias can be traced to a relatively simple source; namely, the role of self-contact which inherently and unsurprisingly differs by race. Eliminating self-contact from index calculations by assessing group contact based on neighbors instead of area population eliminates this inherent source of bias in index scores. The chapter shows that resulting "unbiased" versions of unbiased indices are attractive for many reasons. They are attractive on formal grounds because analysis based on binomial probability models shows that they have expected values of zero under random assignment. They are attractive because the index refinements are easy to explain; for any individual group contact can be a random draw when computed using neighbors but it is always inherently biased when computed using area population that includes the individual. Finally, the unbiased versions of indices introduced here are attractive because they allow researchers to use familiar indices and apply familiar substantive interpretations as well as new interpretations.

The next chapter presents evidence on another aspect of the unbiased versions of indices of uneven distribution introduced here; their behavior over varying circumstances of study design. It uses simulation methodology to generated residential distributions over a wide range of circumstances and shows that the unbiased versions of popular indices introduced in this chapter behave as desired in circumstances where bias renders scores for standard versions of the indices untrustworthy and potentially misleading. It also shows that unbiased indices are attractive because they near-exactly replicate the behavior of standard versions of indices in situations where bias is negligible and they yield clearly superior assessments of segregation in situations where the impact of bias on standard versions of indices is non-negligible.

References

Becker, H. J., McPartland, J., & 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.

Fossett, M. A. (2007). Measuring segregation in simulation studies: Conceptual and practical considerations. SimSeg technical paper. Report for NIH Grants R43HD38199 (Simulating Residential Segregation Dynamics: Phase I) and R44HD038199 (Simulating Residential Segregation Dynamics: Phase II). (An extensive revision of a paper originally presented at the

Annual Meetings of the American Sociological Association, Philadelphia, Pennsylvania, August 2005.).
James, D., & Taeuber, K. (1985). Measures of segregation. *Sociological Methodology, 13*, 1–32.
Laurie, A. J., & Jaggi, N. K. (2003). Role of 'Vision' in neighbourhood racial segregation: A variant of the Schelling segregation model. *Urban Studies, 40*, 2687–2704.
Schelling, T. C. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology 1*, 143–186.
Stearns, L. B., & Logan, J. (1986). Measuring segregation: Three dimensions, three measures. *Urban Affairs Quarterly, 22*, 124–150.
White, M. J. (1986). Segregation and diversity: Measures of population distribution. *Population Index, 65*, 198–221.
Winship, C. (1977). A re-evaluation of indices of residential segregation. *Social Forces, 55*, 1058–1066.