INSTITUTE FOR RESEARCH ON POVERTY DISCUSSION PAPERS

201-74

A NOTE ON THE DECOMPOSITION OF INDEXES OF DISSIMILARITY

Hal H. Winsborough

UNIVERSITY OF WISCONSIN-MADISON

A NOTE ON THE DECOMPOSITION OF INDEXES OF DISSIMILARITY

Hal H. Winsborough

May 1974

The paper is one of a series of "Studies in Racial Segregation"

This research was supported in part by funds granted to the Institute for Research on Poverty at the University of Wisconsin by the Office of Economic Opportunity pursuant to the Economic Opportunity Act of 1964. The opinions expressed are those of the author.

ABSTRACT

This note describes a method for decomposing the most commonly used index of residential segregation. The decomposition addresses the following kind of question: How much of the observed Black-White residential segregation is attributable to compositional (say, income) differences between the races; how much to segregation within, say, income classes; and how much is due to the joint and unallocatable effects of composition and composition-specific segregation?

The purpose of this note is to describe and illustrate a method for decomposing indexes of dissimilarity. The kind of question this decomposition addresses is as follows: Given an index of dissimilarity used, say, as a segregation index, how much of the observed value is attributable to compositional differences between the two groups (for example, income differences), how much to within-compositional-category segregation, and how much due to joint and unallocatable effects of composition and composition-specific segregation?

Let P_b be a column vector where p_{bi} indicates the proportion of all Blacks living in census tract i. Let P_w be the comparable vector of proportion of Whites over tracts. The operation $P_b - P_w$ yields the differences over which the index of dissimilarity is computed. The index itself is computed as one-half the sum of the absolute values of this column of figures. We can accomplish this computation by constructing a row vector of length, say K, the number of tracts in the city. Each element of this vector has an absolute value of one-half and sign identical to that of the corresponding element of $P_b - P_w$. Call this vector δ. Then the index of dissimilarity is given by:

$$\Delta = \delta(P_b - P_w)$$

where δ is 1 by K and P_b and P_w are K by 1.

Now, observe that the proportionate distribution of all Blacks over tracts, P_b, can be found from the income-specific proportion of Blacks over tracts and the distribution of Blacks over income categories as follows: Let R_b be a matrix having tracts for rows and income categories for columns. Assume this matrix is proportioned over tracts. Thus r_{bij} is the proportion of all Blacks who are in income category j living in tract i.

Let I_b be a column vector displaying the proportionate distribution of all Blacks over the income categories. Thus i_{bi} indicates the proportion of Blacks for income category i. With these definitions we can see that:

$$P_b = R_b I_b$$

where R_b is K by m and I_b is m by 1, m being the number of income categories, and similarly

$$P_w = R_w I_w.$$

Thus:

$$\Delta = \delta(P_b - P_w) = \delta(R_b I_b - R_w I_w).$$

A Kitagawa-type[1] decomposition of the term on the right yields:

$$\delta(R_b I_b - R_w I_w) = \delta[R_w(I_b - I_w)] + \delta[(R_b - R_w)I_w] + \delta[(R_b - R_w)(I_b - I_w)].$$

The first term on the right indicates the amount of segregation expected if each race had its own income distribution but the income-specific residential distribution of the White population. Thus, this quantity can be taken to indicate the amount of segregation attributable to composition alone. This component is closely related to the index of dissimilarity one would obtain if he computed "expected" proportionate distributions, as in indirect standardization, for both races (using the rates of the White population as standard), and then calculated an index of dissimilarity between them. Indeed, the vector of differences in tract-specific proportions for such a procedure would be identical to that displayed within the square brackets of the first component. In the component, however, this vector of differences is premultiplied by the vector δ which is constructed from the tract-specific signs of the difference in <u>observed</u> proportions. In computing the index of dissimilarity between "expected" proportions, on the other hand, one would effectively premultiply the vector of differences by a different vector, say δ', whose tract-specific signs are determined by the sign of the

difference in expected proportions. Of course, if δ and δ' are identical, the index over "expected" proportions and the component will be identical.

The second term indicates the segregation expected if each race had its own income-specific residential distribution but the income composition of the White population. This value is similar to the segregation index between tract distributions which have been directly standardized for income composition and can be taken to indicate the amount of segregation attributable to composition-specific differences in tract distribution. Again the component and the index on standardized proportions will differ if the signs of the tract-specific differences in the standardized distribution are different from those in the observed distributions.

The third component is not uniquely assignable to either composition or composition-specific distribution, but is rather the difference in composition evaluated over the differences in composition-specific distributions. If a straightforward direct standardization is performed, the analog to this component is treated as a part of composition and hence dispensed with in a directly standardized index. If indirect standardization is performed, the analog of this component is treated as a part of distribution and retained in the indirectly standardized rate.

An Example

Consider a city having five tracts with the following distribution of its Black and White population:

Tract	(P_b) Black	(P_w) White	(P_b-P_w) Difference	δ
	1	2	3	4
1	.10	.32	-.22	$-\frac{1}{2}$
2	.12	.20	-.08	$-\frac{1}{2}$
3	.20	.20	+.00	$+\frac{1}{2}$
4	.20	.14	+.06	$+\frac{1}{2}$
5	.38	.14	+.24	$+\frac{1}{2}$
Total	1.00	1.00	0.00	

As can be seen by treating $\underset{\sim}{\delta}$ as a row vector and premultiplying it $-\delta$ by column 3, the segregation index in this city is .30.

Now, let us suppose one is interested only in the income composition over two values, high and low. The following table gives the income-specific distribution of the races over tracts.

Tracts	(R_b) Black		(R_w) White	
	High	Low	High	Low
1	.10	.10	.40	.20
2	.20	.10	.20	.20
3	.20	.20	.20	.20
4	.20	.20	.10	.20
5	.30	.40	.10	.20
Total	1.00	1.00	1.00	1.00

The following table gives the income composition of the races in our city.

	(I_b) Black	(I_w) White
High	.20	.60
Low	.80	.40

We can confirm that $P_b = R_b I_b$ by computing

$$(.10)(.20) + (.10)(.80) = .10$$

$$(.20)(.20) + (.10)(.80) = .12$$

$$(.20)(.20) + (.20)(.80) = .20$$

$$(.20)(.20) + (.20)(.80) = .20$$

$$(.30)(.20) + (.40)(.80) = .38$$

and that $P_w = R_w I_w$ by computing

$$(.40)(.60) + (.20)(.40) = .32$$

$$(.20)(.60) + (.20)(.40) = .20$$

$$(.20)(.60) + (.20)(.40) = .20$$

$$(.10)(.60) + (.20)(.40) = .14$$

$$(.10)(.60) + (.20)(.40) = .14$$

To compute the first component we need $(I_b - I_w)$ which is given by:

High	-.40
Low	+.40

Premultiplication of this vector by R_w yields:

$$(.40)(-.40) + (.20)(.40) = -.08$$

$$(.20)(-.40) + (.20)(.40) = .00$$

$$(.20)(-.40) + (.20)(.40) = .00$$

$$(.10)(-.40) + (.20)(.40) = .04$$

$$(.10)(-.40) + (.20)(.40) = .04$$

Premultiplication of the result vector by δ yields:

$$(-\tfrac{1}{2})(-.08) + (-\tfrac{1}{2})(.00) + (\tfrac{1}{2})(.00) + (\tfrac{1}{2})(.04) + (\tfrac{1}{2})(.04) = .08$$

To compute the second component we need $(R_b - R_w)$ which is given by:

	High	Low
1	-.30	-.10
2	.00	-.10
3	.00	.00
4	.10	.00
5	.20	.20

Postmultiplication of this matrix by I_w yields:

$$(-.30)(.60) + (-.10)(.40) = -.22$$

$$(.00)\ (.60) + (-.10)(.40) = -.04$$

$$(.00)\ (.60) + (.00)\ (.40) = .00$$

$$(.10)\ (.60) + (.00)\ (.40) = .06$$

$$(.20)\ (.60) + (.20)\ (.40) = .20$$

Premultiplying by δ yields a value of .26 for this component.

To compute the third component we compute $(R_b - R_w)(I_b - I_w)$ as follows:

$$(-.30)(-.40) + (-.10)(.40) = .08$$

$$(.00)\ (-.40) + (-.10)(.40) = -.04$$

$$(.00)\ (-.40) + (.00)\ (.40) = .00$$

$$(.10)\ (-.40) + (.00)\ (.40) = -.04$$

$$(.20)\ (-.40) + (.20)\ (.40) = -.00$$

Premultiplying the result vector by δ yields:

$$(-\tfrac{1}{2})(.08) + (-\tfrac{1}{2})(-.04) + (\tfrac{1}{2})(.00) + (\tfrac{1}{2})(-.04) + (.00)(\tfrac{1}{2}) = -.04$$

Thus, in this example, the compositional component has a value of .08, the composition-specific distributional component has a value of .26, and the unallocatable "interaction" component has a value of -.04. In this

imaginary city, then, residential segregation has little to do with racial difference in income composition. Most of the "action" is attributable to segregation of high income Whites from high income Blacks and low income Whites from low income Blacks.

Let us end this note with two observations about this decomposition method.

The first observation is simply to reiterate Duncan's point that the decomposition of algebraic identities is not necessarily the same thing as "causal" decomposition.

> One egregious error must, however, be avoided: that of treating components and causes on the same footing. By this route, one can arrive at the meaningless result that net migration is a more important "cause" of population growth than is change in manufacturing output. One must take strong exception to a causal scheme constructed on the premise, 'If both demographic and economic variables help explain metropolitan growth, then we may gain understanding of growth processes by lumping the two together.' [George L. Wilber, "Growth of Metropolitan Areas in the South," Social Forces, XLII (May, 1964), p. 491.] On the contrary, "understanding" would seem to require a clear distinction between demographic components of growth and economic causes which may affect growth via one or another of its components.[2]

The second observation pertains to the advantages of a "complete" decomposition, where data availability permits, as opposed to a simple direct or indirect standardization. Given complete access to the data, many demographers might well feel that a direct standardization would, by itself, suffice. In contemplating such a procedure, however, a question arises about which population should be regarded as the "standard." Had we chosen to regard the Black population as "standard," our second component -- the one analogous to a direct standardization -- would have been $\delta[(R_b-R_w)I_b]$. A change in the standard population would have made a difference in this

component of:

$$\delta[(R_b-R_w)I_b] - \delta[(R_b-R_w)I_w]$$

or $$\delta(R_b-R_w)(I_b-I_w)$$

i.e., exactly the amount of the third, interaction, component.

If we consider the component analogous to the "expected value" in the case of indirect standardization, (i.e., the first of our above components), and compute its change under a shift in which population is regarded as "standard" we find the difference is:

$$\delta[R_b(I_b-I_w)] - \delta[R_w(I_b-I_w)]$$

or again

$$\delta(R_b-R_w)(I_b-I_w)$$

i.e., the amount of the third component.

Thus, the third component tells one the degree to which the first two components are sensitive to substitution of the alternative population as the "standard." If one believes that these two populations satisfactorily "bound" the range of reasonableness, then perhaps the best procedure would be to use the range of the two estimates of each component as an interval estimate. Thus, from the above example we might say that the compositional component has a value between .04 and .08 while the composition-specific distributional component has a value between .22 and .26. In her two-components method, Kitagawa chooses to make a point estimate in the middle of this range. However one might choose a midpoint, it is clear that the size of the "range" is an important measure of the sensitivity of the decomposition to changes in the choice of a standard population and hence very much worth the bother.

FOOTNOTES

[1]For a description of these components see Evelyn M. Kitagawa, "Components of a Difference Between Two Rates," JASA L (December, 1955), pp. 1168-1194. See also H. H. Winsborough and Peter Dickinson, "Components of Negro-White Income Differences," Proceedings of the American Statistical Association, Social Statistics Section, 1971, p. 6-8.

[2]Otis Dudley Duncan, "Path Analysis: Sociological Examples," A.J.S., Volume 72, No. 1 (July 1966), p. 10.