The Stata Journal (2022)
22, Number 3, pp. 1–36 DOI: !!

Computing decomposable multigroup indices of segregation

Daniel Guinea-Martin
Universidad de Málaga
Málaga, Spain
daniel.guinea@uma.es

Ricardo Mora
Universidad Carlos III de Madrid
Madrid, Spain
ricmora@eco.uc3m.es

Abstract. Eight multigroup segregation indices are decomposable into a between and a within term. They are two versions of 1) the mutual information index, 2) the symmetric Atkinson index, 3) the relative diversity index, and 4) Theil's H index. In this article, we present the command `dseg`, which obtains all of them. It contributes to the stock of segregation commands in Stata by 1) implementing the decomposition in a single call, 2) providing the weights and local indices used in the computation of the within term, 3) facilitating the deployment of the decomposability properties of the eight indices in complex scenarios that demand tailor-made solutions, and 4) leveraging sample data with bootstrapping and approximate randomization tests. We analyze 2017 census data of public schools in the United States to illustrate the use of `dseg`. The subject topic is school racial segregation.

Keywords: st00!!, dseg, Atkinson, decomposability, multigroup, mutual information, race, relative diversity, Theil's H, schools, segregation

1 Introduction

A classic concern for social scientists is the degree of association between membership in a group—defined by, say, race–ethnicity, gender, or religion—and assignment to an organizational unit—such as an occupation, a neighborhood, or a school. Most indices of segregation measure differences in the proportions of each group in organizational units (hereafter "units"). If each group is represented in each unit as it is in the overall population, indices report no segregation. As the groups' proportions in the units depart from perfect representation, indices become positive. When groups are perfectly split into distinct units, indices reach their maximums.

Traditional indices of segregation, such as Duncan and Duncan's dissimilarity index (DI hereafter; see Duncan and Duncan [1955]) or the Gini index (Flückiger and Silber 1999) model this basic setup appropriately when 1) group belonging is dichotomous, for example, men versus women or whites versus blacks; and 2) the data lack any multilevel structure. However, these indices are inappropriate when conditions 1 and 2 do not hold, that is, when there are more than two groups and segregation has a multilevel structure. In this article, we introduce the command `dseg` (Mora 2014) for computing all known indices of segregation that can be used when this is the case.

Think of multiracial segregation in schools located within districts. Races are simply one of many sources of identity and affiliation (Akerlof and Kranton 2010) that are not necessarily binary. Native language, income, or religion are other possible sources of nonbinary identity that may be consequential to segregation and that violate condition 1. Another source of complexity is that the dimensions of segregation are often multilayered, which is a situation that violates condition 2. For example, in the United States, students who self-classify to a race category other than non-Hispanic white belong to a "minority". In this context, it is naturally appealing to decompose the overall measure of multigroup segregation and determine what chunk of it is due to what aggregation layer (Reardon, Yun, and Eitle 2000); one layer distinguishes non-Hispanic whites ("whites" for short hereafter) from minorities, and the other compares a variety of minority groups among themselves. Multigroup indices that satisfy a *group decomposability* property can partition the original groups into supergroups and perform a decomposition of the index into, on one hand, a between term that captures the segregation among the supergroups and, on the other hand, a within term that captures the contribution made by the groups.

On other occasions, segregation stems from different levels of social organization (Mora and Ruiz-Castillo 2011). School racial segregation is a telling illustration. Consider two cities, X and Y, with K districts and N schools each. Schools in X and Y may be completely segregated, even though the segregating mechanism may arise from different racial compositions 1) in their K districts, 2) in their N schools, or 3) at both levels. To see this, suppose that all districts in city X share the city's overall racial mix. However, schools are completely segregated. Instead, in city Y districts are completely segregated to begin with. Moreover, because in each of its K districts there are students from only one race, this occurs in every single school in town. In city X, one can confidently state that the segregation captured by the indices of racial segregation in schools is effectively race segregation that arises in schools. By contrast, race segregation in the schools of city Y mirrors the racial composition of its districts. Hence, in this city the measure of race segregation in schools confounds segregation in schools with segregation in districts.

To address this and similar situations, we find looking at the property of *unit decomposability* is useful. It allows partitioning units into distinct sets or clusters—based on, for example, districts' boundaries, schools' ownership, or religion—and identifying the contributions to overall segregation made by, on one hand, clusters and, on the other, the final organizational units. To return to our example, we see this means that measures of segregation that satisfy unit decomposability can discriminate between the tendency of racial groups to be found in different 1) districts on one hand and 2) schools on the other.

In this article, we introduce the command `dseg` for computing eight indices of multigroup segregation. As explained in detail in section 3.3, these are 1) the mutual information index, M, and a normalized version, NM, that rescales segregation as a proportion of maximum segregation (we refer to this as "weak normalization"; see Mora and Ruiz-Castillo [2011]); 2) one version of the symmetric Atkinson's index, A, that is group decomposable and one that is unit decomposable; 3) one version of the relative

diversity index, R, that is group decomposable and one that is unit decomposable; and 4) one version of Theil's H index, H, that is group decomposable and one that is unit decomposable. These are the only multigroup segregation indices that satisfy the group decomposability property, the unit decomposability property, or both.

Finally, there might be research settings that call for decomposing a segregation measure along two or more levels in the group and unit dimensions. Joining together the two examples above, suppose we are interested in the segregation in schools among minorities, controlling for the segregation that 1) there is between minorities and whites and that 2) arises from the varying racial mix of school districts, rather than from schools themselves. Then we should conduct the analysis with M because it is the only index that is decomposable along both the group and unit dimensions.

The rest of the article is set out as follows. First, we review the existing commands for calculating multigroup segregation indices. Then, we put forward two notions of segregation and four additive decomposability properties. Based on these, we define the eight segregation indices that `dseg` computes. Afterward, we introduce the `dseg` command itself. Its use is illustrated with an analysis of the U.S. census of the 45,277,593 students enrolled in the 93,443 public schools of the country in 2017. The presentation of `dseg` is organized in three levels: basic, intermediate, and advanced. In it, we progressively show how `dseg` facilitates the deployment of the decomposability properties of the eight indices in complex scenarios that demand tailor-made analyses. Moreover, we emphasize two novelties that `dseg` brings to the stock of segregation commands in Stata: it 1) provides the weights and local indices for each of the clusters or supergroups in the decomposition and 2) tackles the problems caused by small sample sizes with bootstrapping and approximate randomization tests.

2 Existing commands for measuring decomposable or multigroup segregation

Four community-contributed commands compute measures of segregation in settings with only two groups. Commands `duncan` and `duncan2` (Jann 2004) compute the DI using individual-level data. Package `dissim` (Cox 1999) calculates DI from aggregate data. In a multigroup situation, the three commands calculate the DI for all pairwise comparisons of groups. However, they do not provide a unique measure of multigroup segregation. Moreover, the DI is neither unit nor group decomposable. Command `hutchens` (Jenkins 2006) computes the unit-decomposable version of A in the two-group case, also known as the square root index H (not to be confused with Theil's H). For any partition of the units into clusters, the square root index H is unit decomposable. Although `hutchens` provides the decomposition for any partition into clusters, it does not give the weights and local indices for each of the clusters in the decomposition.

Regarding multigroup indices, two community-contributed commands are available. `seg` (Reardon and Townsend 1999) calculates eight multigroup diversity and segregation indices. None of them are group decomposable. Moreover, even though `seg` includes the unit-decomposable versions of Theil's H and R, it does not compute the decompo-

sition of the units into clusters. The command `dicseg` (Gradín 2011) calculates several dichotomous indices of segregation, including the DI, Gini, and applications of M and the unit-decomposable version of Theil's H to the dichotomous case.

Moreover, `localseg` (Gradín 2011) computes measures of what Alonso-Villar and del Río (2010) define as "local segregation". In a study including multiple groups, these measures gauge the level of isolation of any given group.

3 Two notions of segregation, four additive decomposability properties, and eight segregation indices

3.1 Two notions of segregation

Imagine a situation where each individual is assigned to one organizational unit n, where $n \in \{1, \ldots, N\}$, and belongs to one group g, where $g \in \{1, \ldots, G\}$. As James and Taeuber (1985, 24) note, this bivariate distribution of two discrete random variables is typically represented in two-way tables in which the rows correspond to the n units and the columns to the g groups, given that often $N > G$.

In this scenario, many classic articles in the field start by posing the question, What is the meaning of stating that there is more or less "segregation"? (James and Taeuber 1985; Flückiger and Silber 1999; Frankel and Volij 2011; Massey and Denton 1988). The first step toward an answer requires choosing what type of frequencies to use. Chakravarty and Silber (1994) propose measures of segregation that are functions of counts or absolute frequencies. However, the great majority of indices, including the ones computed by `dseg`, are built upon the so-called relative view of segregation, that is, one that is based on proportions or relative frequencies.

Given this baseline, let us introduce more notation. Lowercase p denotes a proportion, and uppercase P denotes a collection of proportions. Moreover, let p_{ng} represent the joint proportion of individuals who are in unit n and belong to group g. These are sometimes called "cell proportions". Then $P_{\text{unit, group}}$ refers to the joint distribution of the discrete random variables *unit* and *group*, indexed by the mute variables n and g, respectively: $P_{\text{unit, group}} = \{p_{11}, p_{12}, \ldots, p_{ng}, \ldots, p_{NG}\}$, where $\sum_{n=1}^{N} \sum_{g=1}^{G} p_{ng} = 1$. Indices of segregation that are only a function of $P_{\text{unit, group}}$ satisfy the property known as size or scale invariance (James and Taeuber 1985; Frankel and Volij 2011): multiplication of absolute frequencies by a constant affects neither the proportions nor, therefore, these indices.

In addition, let $p_{n\bullet} = \sum_{g=1}^{G} p_{ng}$ be the proportion of individuals in unit n that is obtained by summing the cell proportions in different columns. Hence, P_{unit} is the array containing the overall or marginal distribution of individuals across units, regardless of the value of their group: $P_{\text{unit}} = \{p_{1\bullet}, p_{2\bullet}, \ldots, p_{n\bullet}, \ldots, p_{N\bullet}\}$. Let $p_{\bullet g} = \sum_{n=1}^{N} p_{ng}$ be the proportion of individuals who belong to group g. Equivalently, P_{group} is the array made of the overall or marginal distribution of individuals across groups, regardless of the value of their unit: $P_{\text{group}} = \{p_{\bullet 1}, p_{\bullet 2}, \ldots, p_{\bullet g}, \ldots, p_{\bullet G}\}$.

Finally, let $P_{\text{group|unit}} = P_{\text{unit, group}}/P_{\text{unit}}$ be the array of group shares in each unit n. Given the tabular arrangement mentioned earlier, there are N such row collections, each containing G columns. Likewise, let $P_{\text{unit|group}} = P_{\text{unit, group}}/P_{\text{group}}$ be the array of unit proportions conditional on group g. There are G such arrays, each having N conditional proportions.

Most indices, including the popular DI in the two-group case, measure the extent of differences between $P_{\text{group|unit}}$ and P_{group}. They address the question of how the group shares in the units ($P_{\text{group|unit}}$) differ from the group shares in the overall population (P_{group}). Otherwise put, To what extent does the group mixture in the units diverge from the group composition of the population under study? As James and Taeuber (1985) put it, such indices of segregation are "distributional" or "dispersion" measures around the marginal distribution of groups, P_{group}. We label this "axis of measurement" (Denton and Massey 1988) the $P_{\text{group|unit}}$ notion of segregation. In this article, we express it in natural language with the expression "group segregation in units", for example, "race segregation in schools".

Nevertheless, just as logically, one could develop an alternative "axis of measurement" for gauging dispersion around the marginal distribution of units, P_{unit}. An index lying on this axis would measure the extent of differences between $P_{\text{unit|group}}$ and P_{unit}. We label it the $P_{\text{unit|group}}$ notion of segregation. Now the question posed is, How do the unit shares of each group differ from the unit shares in the overall population? Otherwise put, To what extent does the groups' distribution across units diverge from the analogous distribution of the overall population?[1] In natural language, we convey this notion with the expression "unit segregation by group", as in "school segregation by race".

3.2 Four additive decomposability properties

Suppose that there is a partition of the set of N organizational units $\{1, 2, \ldots, N\}$ into a set of K major organizational units or clusters $k \in \{1, \ldots, K\}$ with $K < N$. Let $T_{\text{group},k}$ be the collection of the groups' absolute frequencies in each cluster k. Following Frankel and Volij (2011), we state that a segregation index Ψ is weakly unit decomposable (WUD) if and only if

$$\Psi = \Psi^{\mathcal{K}} + \sum_{k=1}^{K} \omega\left(T_{\text{group},k}\right) \Psi^{\mathcal{U}}(k) \tag{1}$$

where $\Psi^{\mathcal{K}}$ is called the "between term" and results from computing the index using the K clusters as organizational units, $\Psi^{\mathcal{U}}(k)$ is the local index of segregation in the organizational units of cluster k, and $\omega\left(T_{\text{group},k}\right) > 0$ is a weighting factor that is a function of the groups' sizes in each cluster. The so-called within term of the decomposition, $\sum_{k=1}^{K} \omega\left(T_{\text{group},k}\right) \Psi^{\mathcal{U}}(k)$, is a linear combination of the k local indices that there are,

1. In their study of residential segregation with two groups only, Massey and Denton (1988) consider five alternative views or dimensions of segregation. The first two views, evenness and exposure, are related to the $P_{\text{group|unit}}$ and $P_{\text{unit|group}}$ notions of segregation. The last three, concentration, clustering, and centralization, relate to spatial data.

one for each cluster. The linear combination is said to be convex when the weights add up to unity: $\sum_{k=1}^{K} \omega\left(T_{\text{group},k}\right) = 1$. Then the index is strongly unit decomposable (SUD). If segregation changes by the same amount Δ in all clusters, an index that is SUD has the desirable property of changing its within term by Δ (see Mora and Ruiz-Castillo [2011] for a more detailed discussion of the advantages of SUD over WUD indices in the context of entropy-based indices).

Unit-decomposable indices identify different sources of segregation when the units can be organized into a multilevel structure. Returning to the example of race segregation in the schools of a city with K districts, we note the between term $\Psi^{\mathcal{K}}$ captures district racial segregation: segregation that arises from the different race mix of districts (which corresponds to the $P_{\text{group|unit}}$ notion of segregation explained earlier or, in words, "race segregation in districts"); or, viewed otherwise, from the varying distribution of races across districts ($P_{\text{unit|group}}$ notion of segregation or, in words, "district segregation by race"). The within term $\sum_{k=1}^{K} \omega\left(T_{\text{group},k}\right) \Psi^{\mathcal{U}}(k)$ measures school racial segregation per se, controlling for district racial segregation. From the standpoint of the $P_{\text{group|unit}}$ notion of segregation, it captures the reduction in race segregation in schools that would occur if the group shares in all the schools of any district k were made equal to the district's overall group shares. From the standpoint of the $P_{\text{unit|group}}$ notion of segregation, it captures the reduction in school segregation by race that would occur if the distribution of each group across the schools of any district k were made equal to the school distribution of the overall student population of the district.

Likewise, it is possible to define the properties of weak group decomposability and strong group decomposability in the context of a partition of the groups into L supergroups $l \in \{1, \ldots, L\}$ with $L < G$. For example, following Reardon, Yun, and Eitle (2000), we might be interested in weighing the part of race segregation in schools that is fostered by the segregation of whites from minority students compared with the segregation generated by school differences among the miscellaneous minority groups (Asians, blacks, Hispanics, etc.). Then whites and minority students are two supergroups, and a segregation index Ψ is weakly group decomposable (WGD) if and only if

$$\Psi = \Psi^{\mathcal{L}} + \sum_{l=1}^{L} \omega\left(T_{\text{unit},l}\right) \Psi^{\mathcal{G}}(l) \tag{2}$$

where $\Psi^{\mathcal{L}}$ is the segregation computed among the L supergroups (the so-called between term); $\Psi^{\mathcal{G}}(l)$ is the segregation among the groups making up each supergroup l (which amounts to 0 in the case of supergroups composed of only one group, as is the case of the white supergroup); and $\omega\left(T_{\text{unit},l}\right) > 0$ is a weighting factor that is a function of the units' sizes for each group, $T_{\text{unit},l}$. The so-called within term of the decomposition is the term $\sum_{l=1}^{L} \omega\left(T_{\text{unit},l}\right) \Psi^{\mathcal{G}}(l)$; that is, it is a linear combination of the indices defined for each supergroup. If all the weights in the decomposition add up to unity, that is, $\sum_{l=1}^{L} \omega\left(T_{\text{unit},l}\right) = 1$, then the index is strongly group decomposable (SGD).

3.3 Eight segregation indices

The eight indices computed by `dseg` are multigroup and satisfy at least one decomposability property. To our knowledge, there does not exist any other index of segregation that is both multigroup and additively decomposable. Six of the eight indices have been proposed before. These are 1) the mutual information index using, without loss of generality, natural logarithms, M; 2) its weak normalization, NM; 3) the symmetric Atkinson index A that is based on the $P_{\text{unit|group}}$ notion of segregation, $A_{\text{unit|group}}$; 4) a version of Theil's H index that is a normalization of M based on the $P_{\text{group|unit}}$ notion of segregation, $H_{\text{group|unit}}$; 5) a version of Theil's H index that is a normalization of M based on the $P_{\text{unit|group}}$ notion of segregation, $H_{\text{unit|group}}$; and 6) the relative diversity index R that is based on the $P_{\text{group|unit}}$ notion of segregation, $R_{\text{group|unit}}$.

We define in this article for the first time 1) the symmetric Atkinson index A that is based on the $P_{\text{group|unit}}$ notion of segregation, $A_{\text{group|unit}}$; and 2) the relative diversity index R that is based on the $P_{\text{unit|group}}$ notion of segregation, $R_{\text{unit|group}}$. Table 1 presents formulas and the decomposability properties of the eight indices. Frankel and Volij (2011) characterize axiomatically M and $A_{\text{unit|group}}$. Reardon and Firebaugh (2002) study the properties of $H_{\text{group|unit}}$ and $R_{\text{group|unit}}$ (as well as those of other multigroup indices that are not decomposable). Mora and Ruiz-Castillo (2011) discuss the decomposability and normalization properties of all the entropy-based multigroup indices: M, NM, $H_{\text{group|unit}}$, and $H_{\text{unit|group}}$.

Table 1. Additive decomposable multigroup indices of segregation

Index	Notion of segregation	Formula	Range	Decomposable properties	Weights	Original citation	Stata commands
M	$P_{\text{group}\|\text{unit}}$, $P_{\text{unit}\|\text{group}}$	$\sum_{n=1}^{N} p_{n\bullet}\left\{H(P_{\text{group}}) - H(P_{\text{group}\|n})\right\} = \sum_{g=1}^{G} p_{\bullet g}\left\{H(P_{\text{unit}}) - H(P_{\text{unit}\|g})\right\}$	$[0, \log_b(\min)]$, $\min \equiv \min\{N, G\}$	SUD, SGD	$\omega(T_{\text{group},k}) = p_{k\bullet}$, $\omega(T_{\text{unit},l}) = p_{\bullet l}$	Theil and Finizza (1971)	`dseg, dicseg`†
NM	$P_{\text{group}\|\text{unit}}$, $P_{\text{unit}\|\text{group}}$	$\frac{M}{\log_b(\min\{N,G\})}$	$[0,1]$	SUD if $G \leq N$ SGD if $N \leq G$	$\omega(T_{\text{group},k}) = p_{k\bullet}$ $\omega(T_{\text{unit},l}) = p_{\bullet l}$	Mora and Ruiz-Castillo (2011)	`dseg`
$H_{\text{group}\|\text{unit}}$	$P_{\text{group}\|\text{unit}}$	$\sum_{n=1}^{N} p_{n\bullet}\frac{\{H(P_{\text{group}}) - H(P_{\text{group}\|n})\}}{H(P_{\text{group}})}$	$[0,1]$	WUD	$\omega(T_{\text{group},k}) = p_{k\bullet} \times \frac{H(P_{\text{group}\|k})}{H(P_{\text{group}})}$	Theil and Finizza (1971)	`dseg, seg, dicseg`†
$H_{\text{unit}\|\text{group}}$	$P_{\text{unit}\|\text{group}}$	$\sum_{g=1}^{G} p_{\bullet g}\frac{\{H(P_{\text{unit}}) - H(P_{\text{unit}\|g})\}}{H(P_{\text{unit}})}$	$[0,1]$	WGD	$\omega(T_{\text{unit},l}) = p_{\bullet l} \times \frac{H(P_{\text{unit}\|l})}{H(P_{\text{unit}})}$	Mora and Ruiz-Castillo (2011)	`dseg`
$R_{\text{group}\|\text{unit}}$	$P_{\text{group}\|\text{unit}}$	$\sum_{n=1}^{N} p_{n\bullet}\frac{\{\text{I}(P_{\text{group}}) - \text{I}(P_{\text{group}\|n})\}}{\text{I}(P_{\text{group}})}$	$[0,1]$	WUD	$\omega(T_{\text{group},k}) = p_{k\bullet} \times \frac{\text{I}(P_{\text{group}\|k})}{\text{I}(P_{\text{group}})}$	Carlson (1992)	`dseg, seg`
$R_{\text{group}\|\text{unit}}$	$P_{\text{unit}\|\text{group}}$	$\sum_{g=1}^{G} p_{\bullet g}\frac{\{\text{I}(P_{\text{unit}}) - \text{I}(P_{\text{unit}\|g})\}}{\text{I}(P_{\text{unit}})}$	$[0,1]$	WGD	$\omega(T_{\text{unit},l}) = p_{\bullet l} \times \frac{\text{I}(P_{\text{unit}\|l})}{\text{I}(P_{\text{unit}})}$	Here	`dseg`
$A_{\text{group}\|\text{unit}}$	$P_{\text{group}\|\text{unit}}$	$1 - \sum_{g=1}^{G}\prod_{n=1}^{N}\left(p_{g\|n}\right)^{1/N}$	$[0,1]$	WGD	$\omega(T_{\text{unit},l}) = \prod_{n=1}^{N}\left(p_{l\|n}\right)^{1/N}$	Here	`dseg`
$A_{\text{unit}\|\text{group}}$	$P_{\text{unit}\|\text{group}}$	$1 - \sum_{n=1}^{N}\prod_{g=1}^{G}\left(p_{n\|g}\right)^{1/G}$	$[0,1]$	WUD	$\omega(T_{\text{group},k}) = \prod_{g=1}^{G}\left(p_{k\|g}\right)^{1/G}$	Frankel and Volij (2011)	`dseg, hutchens`†

NOTE: M stands for the mutual information index; NM stands for its normalization as a proportion of its maximum value; $H_{\text{group}|\text{unit}}$ is Theil's H based on the $P_{\text{group}|\text{unit}}$ notion of segregation; $H_{\text{unit}|\text{group}}$ is Theil's H based on the $P_{\text{unit}|\text{group}}$ notion of segregation; $R_{\text{group}|\text{unit}}$ is the relative diversity index R that is based on the $P_{\text{group}|\text{unit}}$ notion of segregation. $R_{\text{unit}|\text{group}}$ is the R index that is based on the $P_{\text{unit}|\text{group}}$ notion of segregation. $A_{\text{unit}|\text{group}}$ is the symmetric Atkinson A that is based on the $P_{\text{unit}|\text{group}}$ notion of segregation; $A_{\text{group}|\text{unit}}$ is the A index that is based on the $P_{\text{group}|\text{unit}}$ notion of segregation. Superscript † denotes commands that apply to the dichotomous case. Subscripts $n \in \{1,\ldots,N\}$ and $k \in \{1,\ldots,K\}$ refer to units and clusters, respectively. Subscripts $g \in \{1,\ldots,G\}$ and $l \in \{1,\ldots,L\}$ refer to groups and supergroups, respectively. The entropy function $H(P_X)$ is defined as $H(P_X) = -\sum_{x=1}^{S} p_x \log_b(p_x)$; the diversity function $\text{I}(P_X)$ is defined as $\text{I}(P_X) = \sum_{x=1}^{S} p_x(1-p_x)$, where $p_x \equiv \Pr(X = x)$, $\forall x \in \{1,\ldots,S\}$, and $P_X \equiv \{p_1, p_2, \ldots, p_S\}$. $H(P_X)$ can be defined using logarithms to any base. Base two logarithms express the unit of information in "bits"; natural logarithms express it in "nats". "Weights" refers to weights $\omega(T_{\text{group},k})$ in (1) for units decompositions and to weights $\omega(T_{\text{unit},l})$ in (2) for group decompositions.

4 Introducing the dseg command: An application to U.S. school racial segregation in 2017

As mentioned in the introduction, decomposable indices are valuable for multigroup and multilevel studies of segregation. In this section, we substantiate this claim and showcase the use of `dseg` by applying it to the measurement of school racial segregation in the United States.

There are three reasons for choosing this example. First, it illustrates unit and group decompositions that are of import to classic yet contemporary scientific and public debates in the United States (Coleman et al. 1966) and elsewhere (Casey 2016). Second, these debates motivated the development of two of the segregation indices, M and H (Theil and Finizza 1971), that can be computed with `dseg`. Additionally, they relate to the topics of income and class segregation in schools (Gutiérrez, Jerrim, and Torres 2020) and racial residential segregation (Duncan and Duncan 1955; Massey and Denton 1994). Possibly only the study of occupational segregation, initiated by Gross (1968), takes a similarly prominent place in the academic and public conversation on segregation. The third reason for explaining the use of `dseg` with the topic of school segregation in the United States is that results are reproducible thanks to publicly available data.

4.1 The data: A census of the U.S. student enrollment body in public schools

We use data from the 2017 Common Core of Data (CCD) Local Education Agency Universe Survey, which is publicly available from the National Center for Education Statistics.[2] The CCD are a census of U.S. public schools. They record 1) all the schools and agencies (and their locations) providing free and public elementary and secondary education in the United States and its jurisdictions; and 2) the sex and race of all students enrolled in them.

The original 2017 CCD includes 95,219 schools or agencies. There are 93,443 schools with students in elementary and secondary education. The data are originally aggregated by sex, race, grade, school, school district, and state. The variable `student_count` contains the count of students in each cell. In our analyses, we leave aside sex and grades. In total, there are 45,277,593 students. They are grouped into the seven categories of the original string variable `race_ethnicity`. We generate a numeric version of this variable named `race` to show `dseg`'s ability to handle both character and numeric variables.

2. Accessed on June 10, 2020, at https://nces.ed.gov/ccd/files.asp#Fiscal:2,LevelId:7,SchoolYearId:32,Page:1.

```
. use ccd2017_sjdseg
. tabulate race_ethnicity [fweight=student_count], sort missing
```

Race or Ethnicity	Freq.	Percent	Cum.
White	21,675,558	47.87	47.87
Hispanic/Latino	12,059,119	26.63	74.51
Black or African American	6,856,017	15.14	89.65
Asian	2,366,659	5.23	94.88
Two or more races	1,710,347	3.78	98.65
American Indian/Alaska Native	442,643	0.98	99.63
Native Hawaiian/Other Pacific Islander	167,250	0.37	100.00
Total	45,277,593	100.00	

In the ensuing analyses, we decompose various measures of segregation by geographical level. The original CCD identifies the 16,768 school districts and 51 states (including Washington, DC) where each school is located in. The corresponding variables are the following:

- `district` for the district identifier;
- `state` for the state identifier;
- `schid` and `school` for the school identifier (the first variable corresponds to the original variable in string format; the second is a numeric version that we generate).

In some examples, we also use the core-based statistical area (CBSA) of each school. CBSA is a geographical classification by the U.S. Office of Management and Budget that splits the country into 934 areas. These are either 1) micropolitan areas of 10,000 to 50,000 people, 2) larger metropolitan areas, or 3) rural areas. Only 2,502,368 students are enrolled in schools located in rural areas. Although CBSA information is not available in the original dataset, the National Center for Education Statistics facilitates each school's CBSA (variable `cbsa`) in a separate file,[3] and both sets can be merged thanks to the school identification code (variable `schid`).

4.2 Speaking dseg: Command syntax

We explain the use of the command `dseg` via the study of school racial segregation in the United States with the 2017 CCD. In Stata, `dseg` implements the computation of the only eight multigroup segregation indices that also satisfy decomposability properties. They gauge the association between individuals' group of belonging and a set of organizational units (schools, school districts, CBSAs, and states in our example).

The command `dseg` contributes to the stock of commands on segregation by 1) computing decompositions directly and 2) providing the elements that make up the within terms of the decomposition: $\omega\left(T_{\text{group},k}\right)$ and $\Psi^{\mathcal{U}}(k)$ for clusters k in unit partitions and $\omega\left(T_{\text{unit},l}\right)$ and $\Psi^{\mathcal{G}}(l)$ for supergroups l in group partitions.

3. Accessed on June 10, 2020, at https://nces.ed.gov/programs/edge/Geographic/SchoolLocations.

Its syntax is the following:

`dseg` *index varlist1* [*if*] [*in*] [*weight*], `given(`*varlist2*`)` [`addindex(`*namelist*`)` `by(`*varlist*`)` `within(`*varlist*[`, components`]`)` `missing fast` `bootstraps(`#[`,` *opt*]`)` `random(`#`)` `rseed(`#`)` `clear` `saving(`*filename*[`,` *opt*]`)` `prefix(`*name*`)` `nolist` `format(`%*fmt*`)`]

`fweights` are allowed; see [U] **11.1.6 weight**.

At its simplest, `dseg` requires only that the user specify which index to compute, what notion of segregation to follow, the groups, and the units. The specification for *index* requires an index name: `atkinson`, `diversity`, `mutual`, `n_mutual` (for *NM*), `theil`, `alt_atkinson`, `alt_diversity`, or `alt_theil`. The user chooses the understanding of segregation built around $P_{\text{group|unit}}$ by identifying 1) the groups of interest in the *varlist1* and 2) the organizational units along which groups segregate from each other in the *varlist2*, that is, within the required option `given()`.

However, with `dseg` it is easy to transpose the axis of measurement and follow instead the $P_{\text{unit|group}}$ notion of segregation. It suffices that the user lists unit-related variables in *varlist1* and group-related variables in *varlist2*. Recall that, as explained earlier, only the M index is invariant to the notion chosen.

varlist1 and *varlist2* accept string and numeric variables. For variable lists of length two or larger, `dseg` internally generates a temporal variable with the Cartesian product of the categories of the variables listed. For example, if in *varlist1* we include a numeric variable `race` categorizing observations into seven racial groups and a string variable `sex` distinguishing boys from girls, `dseg` temporarily creates an internal, auxiliary, and numeric variable with the $7 \times 2 = 14$ categories that the combination of `race` and `sex` produces. As shown later, this internal data manipulation ensures that the index decomposition is implemented correctly in all contexts.

With this minimum information—an index name, a notion of segregation, a list of group variables, and another list of unit variables—`dseg` returns the desired index. The command accepts the standard Stata `if` and `in` qualifiers for, respectively, logical conditions and observation ranges. In aggregated data, frequency weights indicate the number of duplicated observations.

4.3 Options

`given(`*varlist2*`)` specifies the units in groups-given-units indices and the groups in units-given-groups indices. `given()` is required.

`addindex(`*namelist*`)` is a list of additional indices to be computed. Any of the eight indices that `dseg` computes can be included in *namelist*. The notion of segregation is set with *varlist1* and *varlist2*. Any index following the alternative notion can be computed by adding the prefix `alt_` to the index name: `alt_atkinson`, `alt_diversity`, or `alt_theil`.

`by(`*varlist*`)` identifies the subsamples over which the index is to be calculated. It is useful for computing the same index and its decomposition for different years, countries, etc. For example, with `by(state)`, `dseg` outputs 51 indices. With `by(cbsa)`, `dseg` outputs 934 indices. With `by(state cbsa)`, `dseg` outputs 1,055 indices. The reason why is that `dseg` considers all existing combinations in the data of the variables' categories. In this example, there are 867 CBSAs that belong to a single state, and 57 CBSAs are split into 2 states, 7 into 3 states, and 2 into 4 states. Finally, the residual category of the `cbsa` variable for rural area is found in 45 states. Hence, the $867 + 57 \times 2 + 7 \times 3 + 2 \times 4 + 45 = 1055$ categories in the geographical classification of CBSAs and states, and for each of them `dseg` with the option `by(cbsa state)` computes one index.

`within(`*varlist*`[ , components ])` specifies the clusters or supergroups that partition the units or groups, thereby defining the decomposition of the index into a between and a within term. For the unit-decomposable index $A_{\text{unit|group}}$, it identifies the clusters defined as the combinations of *varlist1* and *varlist*. For the group-decomposable index $A_{\text{group|unit}}$, it identifies the supergroups defined as the combinations of *varlist1* and *varlist*. For the unit-decomposable indices $H_{\text{group|unit}}$ and $R_{\text{group|unit}}$, it identifies the clusters as the combinations of *varlist2* and *varlist*. For the group-decomposable indices $H_{\text{unit|group}}$ and $R_{\text{unit|group}}$, it identifies the supergroups defined as the combinations of *varlist2* and *varlist*. Finally, for M, which is both SUD and SGD, it identifies the clusters or supergroups defined by the combinations of *varlist2* and *varlist*. In summary, by specifying this option, `dseg` computes either the decomposition in (1) or the decomposition in (2). It all depends on the index chosen and the variables included in *varlist1* and *varlist2*. For example, there are schools (units) within districts (clusters). For Theil's H, setting `race` in *varlist1*, `school` in *varlist2*, and `district` in `within()` decomposes the unit-decomposable index of segregation $H_{\text{group|unit}}$ into a between term that measures race segregation in districts and a within term that captures race segregation in schools within districts. To carry out an analogous unit decomposition with the unit-decomposable symmetric Atkinson index $A_{\text{unit|group}}$, we should include `school` in *varlist1*, `race` in *varlist2*, and `district` in `within()`. With the `components` suboption, `dseg` additionally provides the weights and the local segregation indices from either (1) or (2).

`missing` treats missing values as valid values. By default, `dseg` assumes missing values are instances of random incomplete information. Hence, all observations with missing values in at least one of *varlist1*, *varlist2*, the variables in `by()`, or the variables in `within()` are dropped before the calculations are made. The `missing` option reverts this behavior and forces `dseg` to interpret missing values as categories. For example, missing observations in `race` would be interpreted as representing students from the same "missing race".

`fast` uses the community-contributed command `ftools` (Correia 2016) to speed up computing time with big data. Install `ftools` with `ssc install ftools`. This option can only be used with numeric variables.

`bootstraps(`#[`, `*opt*`])` sets the number of bootstrap samples. This option invokes the `bsample` command to generate a bootstrap sample with replacement with the same number of observations as the original sample. Bootstrap options are passed through by *opt*; see `help bsample`. The `bootstraps()` option cannot be used with weights or simultaneously with `random()`. It results in a new dataset with index values for each of the # bootstrap samples. In the new dataset, the bootstrap samples are identified with variable `bsn`, the bootstrap sample number. The new dataset includes an additional observation with the indices calculated with the original dataset. This observation is identified with `bsn` = 0. The new dataset replaces the current dataset in memory when the `clear` option is used. It is saved when the `saving()` option is used. If none of these two options are used, `dseg` stops and displays an error message before doing the bootstrap.

`random(`#`)` computes the index with # samples simulated under the assumption of no segregation as suggested by Boisso et al. (1994). Each simulated sample is obtained after randomly reshuffling values of *varlist1*. Otherwise, `random()` closely follows the behavior of the option `bootstraps()`: 1) it cannot be directly used with weighted data or simultaneously with `bootstraps()`; 2) the output is a new dataset that includes index values for all simulated samples; and 3) the new dataset must replace the current dataset or be saved.

`rseed(`#`)` sets the seed for the random-number generator.

`clear` replaces data in memory with data containing index values. During execution, the command `dseg` internally creates a temporary dataset with the results. By default, `dseg` lists the index values and stores them in return matrix `r(S)` (in addition to the number of observations, the command name, the names of the indices, and the notion of segregation defined by *varlist1* and *varlist2*). There are three cases in which matrix `r(S)` is not returned: 1) when either `bootstraps()` or `random()` is used, 2) when the resulting matrix is too large, and 3) in the presence of a string variable in the option `by()` or if `components` is used as in `within(`*varlist*`, components)`. In cases 2 and 3, a warning message is displayed. It may be useful to have the index values in a Stata dataset. We can achieve this with the options `clear` and `saving()`. The `clear` option replaces the data in memory with the index values. The structure of the replacing dataset depends on what additional options are included in the call. At its simplest, the new dataset contains only one observation with the index value stored in one variable. Adding the `by()` option enlarges the new dataset to contain as many observations as categories that are defined by the combination of variables in the `by()` option. For example, in the analysis of race segregation in schools, the options `clear` and `by(state)` generate a new dataset with 51 observations (1 for each state) and 1 variable that stores the index values. Option `within(district)` adds two additional variables to this dataset, one with the between term and another with the within term of the decomposition. However, if we use `within(district,`

`components`), the new dataset has 16,768 observations, that is, the number of school districts, and 2 new variables: 1 for the weight and another for the local index of each school district. The overall index and the between and within terms are constant across the districts of the same state because these variables contain the index decomposition for the whole state.

`saving(`*filename*[`,` *opt*]`)` saves the data file *filename* with the index values. Saving options are passed through by *opt*; see `help save`. The structure of the saved dataset follows the same conventions described above for the `clear` option.

`prefix(`*name*`)` attaches *name* in front of the default name for each index, each between and within term, and each local weight in the new dataset. This option is useful in the presence of conflicting names in the new dataset. There are four types of variable names in the new dataset: 1) the indices' names; 2) the variable names in the list of `by(`*varlist*`)` if `by(`*varlist*`)` is used; 3) the variable names in the list of `within(`*varlist*`, components)` if `components` is used; and 4) `bsn` and `ssn` if `bootstraps()` or `random()` is used. For the index versions implicit in the notion of segregation defined by *varlist1* and *varlist2*, the default names are `A`, `H`, `M`, and `R`. For the alternative versions of `A`, `H`, and `R`, the names are `AltA`, `AltH`, and `AltR`, respectively. For the normalized mutual information, the default name is `NM`. When the `within()` option is used, postfixes `_B` and `_W` are added after the default name for the between and the within term, respectively. If the `components` suboption is used, postfixes `_w` and `_l` are used to name the weights and the local segregation indices. In case of a conflicting name, `dseg` stops and issues an error message.

`nolist` prevents the extended output display. This behavior may be desirable when the `by()` or the `within()` option is used because, by default, `dseg` lists all index values. Option `nolist` suppresses this listing but keeps all other output messages from `dseg`. Thus, `nolist` is preferable to using `quietly` with `dseg` when the user wishes to read output messages other than index values. When `nolist` is used together with `clear`, a description of the new dataset is displayed. Options `bootstraps()` and `random()` call the option `nolist` implicitly.

`format(%`*fmt*`)` sets the output format of the index. The default is `format(%9.4f)`; see `help format`.

4.4 Basic usage of dseg

The simplest call to `dseg` specifies an index name and a notion of segregation. For example, if we transform the aggregated 2017 CCD into an individual-level dataset where each row is a student ($N = 45{,}277{,}593$), we can ask for the standard Theil's H, $H_{\text{group|unit}}$, to measure race segregation in schools (with string variables `race_ethnicity` and `schid`) as follows:

```
. preserve
. expand student_count
(44,367,820 observations created)
. dseg theil race_ethnicity, given(schid)
   Decomposable Multigroup Segregation Indexes
   Differences in race_ethnicity given schid
     Index: Theil's H
         H
    0.3505
```

We can use the `addindex()` option to also compute, in one call, the other four indices that follow the same notion of segregation. Moreover, with the `fast` option, the computation of the index speeds up by using the community-contributed command `ftools`, which accepts numeric variables only (`race` and `school`). Finally, we use the `format()` option to obtain a more precise display of results.

```
. dseg theil race, given(school) addindex(mutual n_mutual diversity atkinson)
> format(%7.6f) fast
   Decomposable Multigroup Segregation Indexes
   Differences in race given school
     Indexes:
     Theil's H, Mutual Information, Normalized Mutual Information,
     Relative Diversity, Symmetric Atkinson
          H          M          NM          R          A
   0.350479   0.467817   0.240410   0.351159   1.000000
```

These results provide alternative quantitative answers to the following question: To what extent does the race share of a random public school differ from the racial composition of the whole student enrollment body of the U.S. public schooling system? The answers are measures of race segregation in schools.

Strictly speaking, these segregation indices cannot be compared with each other. Entropy-based indices are nonetheless related. The value of `NM` simply indicates that `M` is $0.4678/\{\log(7)\} \times 100 = 24\%$ of its maximum. The ratio between `M` and `H`, or $H_{\text{group|unit}}$, is the entropy of race, $0.467817/0.350479 \approx 1.33$. Also, note the value of 1 for `A`, or $A_{\text{group|unit}}$, which, recall, is introduced in this article for the first time. In this context, such an index proves to be a poor choice: whenever a race group is absent from one school, it contributes with its maximum $(1/G)$ to segregation. Because no race is present in every single school, the index reaches its maximum value of 1.

If we revert the placement of the variables and write instead `school, given(race)`, then `dseg` computes the five indices that follow the $P_{\text{school|race}}$ notion of segregation. We do so in the next example, which also showcases the use of the frequency weights (`student_count`) available in the 2017 CCD aggregated data. With such data, the `fast` option is no longer necessary.

```
. restore
. dseg theil school [fw=student_count], given(race)
> addindex(mutual n_mutual diversity atkinson) format(%9.6f)
   Decomposable Multigroup Segregation Indexes
   Differences in school given race
     Indexes:
     Theil's H, Mutual Information, Normalized Mutual Information,
     Relative Diversity, Symmetric Atkinson
           H          M         NM          R          A
    0.042076   0.467817   0.240410   0.000024   0.735506
```

These indices give a quantitative answer to the following question: To what extent does the distribution across schools of each race group differ from the school distribution of all the students taken together? Their answers are measures of school segregation by race. Moreover, with the exception of the M index and its normalized version NM, which follow the two notions of segregation simultaneously, their values differ from those whose computation follow the $P_{\text{race|school}}$ notion.

We appreciate that no two values are equal. Note the comparatively large value of $A_{\text{unit|group}}$, the comparatively small values of $H_{\text{unit|group}}$, and especially $R_{\text{unit|group}}$. The large value of $A_{\text{unit|group}}$ (0.735506) reflects that the seven racial categories in the 2017 CCD are simultaneously present in only 22.83% of U.S. schools. This index would obtain a value of 1, as $A_{\text{group|unit}}$ does, were we to ignore these most racially diverse schools. In short, both Atkinson indices are sensitive to zeros, a concern that Frankel and Volij (2011) originally raised with regard to the standard $A_{\text{unit|group}}$.

As for $H_{\text{unit|group}}$ and $R_{\text{unit|group}}$, their lower value with respect to $H_{\text{group|unit}}$ and $R_{\text{group|unit}}$ reflects that they are normalized by the entropy and diversity functions for schools, rather than by the entropy and diversity functions for races, which are unsurprisingly smaller.

All indices obtained so far are direct measures of segregation. However, the object of study, school racial segregation, has a multilevel nature; this entails that alternative sources of segregation may mix up in the segregation that the direct measures capture.

4.5 Intermediate usage: One-way unit decompositions

Often, policymakers and academics debate the relative merits of local- or district-level policies for ameliorating school racial segregation. Rivkin (1994) argues that local policies can achieve little because they are capped by the upper bound set by race segregation in school districts. Allison (1978) and Reardon and Firebaugh (2002) warn of the dangers of using nondecomposable indices in this context. One such danger is the construction of statistical artifacts leading to misleading results. For example, if we were to directly measure school racial segregation at the national level with a nondecomposable index, we may record that it grows over time. Nevertheless, we may also simultaneously find that the school racial segregation of all districts declines. Such an outcome is possible whenever increases in district racial segregation are sufficiently large.

To address this type of concern, we partition the units of analysis (schools) into clusters (districts). In this sort of setting, where units are nested within clusters, the two levels have a hierarchical relationship and, using the Mora and Ruiz-Castillo (2003) expression, indices can be decomposed only one way: from the segregation fueled by the broader clusters to the segregation stemming from the final organizational units. In the terms of our example, with decomposable indices, we can evaluate the impact of policies targeted to race segregation in schools per se, net of the segregation brought about by the geographical location of the schools, be it districtwise or otherwise.

Using (1) on page 5 as a template, let school districts be the clusters and schools the units. Next we present `dseg`'s syntax for computing the three terms in the equation for the following $P_{\text{group|unit}}$ indices: M (which is SUD), NM (which is SUD in this example because $G < N$), and $H_{\text{group|unit}}$ (which is WUD).

```
. dseg mutual race [fw=student_count], given(school)
> addindex(n_mutual theil) within(district)
   Decomposable Multigroup Segregation Indexes
   Differences in race given school
     Indexes:
     Mutual Information, Normalized Mutual Information, Theil's H
   Between/Within district decomposition
         M      M_B      M_W       NM     NM_B     NM_W        H      H_B      H_W
    0.4678   0.3836   0.0842   0.2404   0.1971   0.0433   0.3505   0.2874   0.0631
```

In this decomposition, the total of each index is equal to the direct measurement of race segregation in the units already computed (see the output on page 15). As mentioned above, the point of the decomposition is to assess what amount of the original direct measurement originates from 1) the broader level within which units are nested (captured by the between term, stored in the variables with the `_B` suffix) and 2) the units per se (captured by the within term, stored in the variables with the `_W` suffix). As fractions of the overall index, the between and within components are equivalent because NM and $H_{\text{group|unit}}$ are normalizations of M: $38.36/46.78 = 19.71/24.04 = 28.74/35.05 = 0.82$ and $8.42/46.78 = 4.33/24.04 = 6.31/35.05 = 0.18$. In other words, only 18% of the value produced by the naive measurement of school racial segregation can be unambiguously attributed to the racial segregation in schools. Otherwise put, the majority of the initial measurement captures race segregation in districts.

We can also obtain the decomposition for the $P_{\text{unit|group}}$ indices that are unit decomposable: M and $A_{\text{unit|group}}$ (which is WUD).

```
. dseg mutual race [fw=student_count], given(school)
> addindex(alt_atkinson) within(district)
   Decomposable Multigroup Segregation Indexes
   Differences in race given school
     Index: Mutual Information
   Differences in school given race
     Index: Symmetric Atkinson
   Between/Within district decomposition
         M      M_B      M_W     AltA   AltA_B   AltA_W
    0.4678   0.3836   0.0842   0.7355   0.5006   0.2349
```

Two things are worth noting. First, $A_{\text{unit|group}}$ is Atkinson's unit-decomposable index. Given that we set the $P_{\text{group|unit}}$ notion of segregation by choosing `race` in *varlist1* and `school` in *varlist2*, we need to use `alt_atkinson` in option `addindex()` to compute it. Second, the within-term amounts to almost 32% of the total, a greater share arguably due to this index's sensitivity to zeros.

4.6 Intermediate usage: Two-way unit decompositions

In contrast to the one-way setting described above, there are occasions when the variables defining the clusters and the units can interchange their roles because they have a nonhierarchical relationship. Then the index can be decomposed in two ways, depending on which variable plays what role. The variables `state` and `cbsa` provide a case in point. We can first partition schools into CBSAs; then we can further partition CBSAs into states. In this partitioning sequence, CBSAs are the units that lie within or inside states, the clusters. However, recall that, as explained in the data section, some CBSAs include multiple states. One example is the Washington, DC Metropolitan Area, which includes the federal district of Washington, DC, and parts of the states of Maryland, Virginia, and West Virginia. Hence, it is possible to first partition schools into states and then states into CBSAs. In this partitioning sequence, states are the units that lie within CBSAs, the clusters.

When the sources of segregation on the unit dimension can interchange roles, the unit space over which segregation is measured comprises all combinations of clusters and units. In our example, the overall mensurable concept of segregation is not segregation in either the 867 CBSAs or the 51 states, depending on which way we perform the decomposition. Instead, in both decompositions the final units are the 1,055 areas that are defined by their CBSA and state in the 2017 CCD.

A user who was aware of this nuance in the data could manually generate a variable for all CBSA and state combinations. The user could then call `dseg`, or any other suitable Stata command, on the new variable. With `dseg`, the user could simply enter the original variables in *varlist2*, and the command would internally and automatically combine their categories. For $H_{\text{group|unit}}$:

```
. dseg theil race [fw=student_count], given(state cbsa)
Decomposable Multigroup Segregation Indexes

Differences in race given state cbsa
  Index: Theil's H

      H
 0.1695
```

Moreover, the command `dseg` with the `within()` option internally computes the Cartesian product of the categories of the variables 1) in *varlist2* and 2) in the `within()` option, that is, defining the clusters. Hence, the decomposition where CBSAs are the units that lie within states can be implemented as follows:

```
. dseg theil race [fw=student_count], given(cbsa) within(state)
   Decomposable Multigroup Segregation Indexes
   Differences in race given cbsa
     Index: Theil's H
   Between/Within state decomposition
         H      H_B      H_W
    0.1695   0.1128   0.0568
```

Denoting states by t, we can write the three terms in the decomposition carried out by `dseg` above as follows from left to right:

$$H_{\text{race|CBSA}\times\text{state}} = H_{\text{race|state}} + \sum_t w_t H_{\text{race|CBSA}}(t)$$

$H_{\text{race|CBSA}\times\text{state}}$ stands for the segregation construct being measured, race segregation in states and CBSAs. Of its value (0.1695), only 0.0568, or $0.0568/0.1695 \times 100 = 33.51\%$ can be attributed to race segregation in CBSAs, net of the effect of states.

What proportion of $H_{\text{race|CBSA}\times\text{state}}$ can be attributed to race segregation in states, net of CBSAs? To answer this question, (3) shows the alternative within-CBSAs decomposition:

$$H_{\text{race|CBSA}\times\text{state}} = H_{\text{race|CBSA}} + \sum_c w_c H_{\text{race|state}}(c) \tag{3}$$

Implementing (3) in `dseg` is simply a matter of specifying `state` in *varlist2* and `cbsa` as the clusters within which race segregation in states is measured:

```
. dseg theil race [fw=student_count], given(state) within(cbsa)
   Decomposable Multigroup Segregation Indexes
   Differences in race given state
     Index: Theil's H
   Between/Within cbsa decomposition
         H      H_B      H_W
    0.1695   0.1574   0.0121
```

The between term in this decomposition, 0.1574, captures only race segregation in CBSAs and underestimates the real amount of total segregation that there is in the 1,055 areas that are defined by their CBSA and state. The within term, 0.0121, captures race segregation in states net of CBSAs. It stands in stark contrast to the value of 0.1128 for race segregation in states reported as term `H_B` in the decomposition where states are the clusters.

Moreover, we can now conclude that, net of CBSAs, states generate only $100 \times (0.0121/0.1695) = 7.14\%$ of $H_{\text{race|CBSA}\times\text{state}}$. Interestingly, the sum of the net contributions of states and CBSAs does not equal the value of $H_{\text{race|CBSA}\times\text{state}}$. This topic is discussed in subsection 4.9.3.

4.7 Intermediate usage: One-way group decompositions

Instead of partitioning units into clusters, suppose we are interested in a partition of G demographic groups into L supergroups. For example, we could partition the seven

races in the 2017 CCD into whites and minority students. In this way, we could inquire into what is the contribution to overall school segregation by race of 1) the segregation between white and minority students and 2) the segregation among minorities. With the 2017 CCD, four indices can be decomposed along these lines: $A_{\text{group|unit}}$, $H_{\text{unit|group}}$, $R_{\text{unit|group}}$ (because they are WGD), and the M index (because it is SGD). For brevity, we focus on the last in the example below.

We have shown that the value of the M index of school segregation by race is 0.467817. Let us denote by 1) $M_{\text{school}}^{\text{minority versus white}}$ the school segregation between minority and white students and by 2) $M_{\text{school}}^{\text{minorities}}$ the school segregation among the six minority groups. Note that in this example, there is one supergroup that consists of only one group, whites. By necessity, school segregation by race among whites is zero. Hence, we can do without this case in the ensuing notation of the within term in (4):

$$M = M_{\text{school}}^{\text{minority versus white}} + p_{\bullet\text{minority}} M_{\text{school}}^{\text{minorities}} \tag{4}$$

$p_{\bullet\text{minority}}$ is the minority share in the population. After we create a new dummy variable for minority race group (`mrg`), the following use of `dseg` will produce the intended output: *varlist1* includes the variables that define the units (`school`), *varlist2* consists of the variables that define the groups (`race`), and the `within()` option contains the supergroup variable (`mrg`).

```
. recode race (1/6=1) (7=2), generate(mrg)
(829207 differences between race and mrg)
. dseg mutual school [fw=student_count], given(race) within(mrg)
   Decomposable Multigroup Segregation Indexes
   Differences in school given race
     Index: Mutual Information
   Between/Within mrg decomposition
         M      M_B      M_W
    0.4678   0.2372   0.2306
```

Interestingly, only about half (0.2372) of school racial segregation (0.4678) comes down to the segregation of whites from minority students. The other half (0.2306) originates from segregation among the races in the minority category.

Note that using the command below does not yield the desired outcome:

```
. dseg mutual race [fw=student_count], given(school) within(mrg)
   Decomposable Multigroup Segregation Indexes
   Differences in race given school
     Index: Mutual Information
   Between/Within mrg decomposition
         M      M_B      M_W
    0.9229   0.6922   0.2306
```

This syntax creates a unit space, that is, organizational units, made of all the combinations of `school` and `mrg`: it splits schools into minority-only and white-only schools and measures race segregation in those two types of schools. Given that the population shares of whites and minorities are close to 50% and we are using the mutual information index, by construction the between term reports almost maximum segregation

between whites and minorities: 0.6922 versus $\log(2) \approx 0.6931$. The within term remains unchanged, 0.2306, because it still measures segregation among the races in the minority category only.

4.8 The suboption components: Your way to weights and local indices

As mentioned already, with the `within()` option, the user specifies the clusters or supergroups within which the user wishes to measure segregation. One of the novelties that the `dseg` command brings to the stock of segregation commands in Stata is the reporting of the components (that is, the weights and the indices) that make up the within term. The suboption `components` asks `dseg` to report the weighting factor and the index of segregation for each cluster or supergroup. The following example illustrates its use with $A_{\text{unit|group}}$, Atkinson's unit-decomposable index. We focus on the first five states (by alphabetical order) to simplify the illustration. The goal is to show how unit-decomposable indices measure school segregation by race controlling for states. With this example, we also illustrate the options for 1) replacing the current dataset in memory with a new dataset storing the new variables (`clear`) and 2) not displaying results on the screen (`nolist`).

```
. preserve
. keep if (state==1) | (state==2) | (state==4) | (state==5) | (state==6)
(754,866 observations deleted)
. dseg atkinson school [fw=student_count], given(race)
> within(state, components) clear nolist
```

(*output omitted*)

With this syntax, `dseg` replaces the original dataset with a new one that contains as many observations as there are clusters, that is, the five states of this example.

The data are automatically sorted by the cluster-defining variable (`state` in the example). Next we display the resulting five observations in the current file:

```
. list, abbreviate(20)
```

	state	A	A_B	A_W	A_w	A_l
1.	Alabama	0.6381	0.1738	0.4643	0.0471	0.7793
2.	Alaska	0.6381	0.1738	0.4643	0.0220	0.3858
3.	Arizona	0.6381	0.1738	0.4643	0.1150	0.4991
4.	Arkansas	0.6381	0.1738	0.4643	0.0412	0.7597
5.	California	0.6381	0.1738	0.4643	0.6009	0.5499

The variables `A`, `A_B`, and `A_W` are constants. They correspond to the total amount of school segregation by race (`A`); the between-state term or state segregation by race (`A_B`), which is a measure of how differently the seven races distribute across states; and the within-state term (`A_W`), which is the states' weighted average of school segregation by race. The `components` suboption additionally generates two variables. First, the

variable `A_w` with the weights for the weighted average in the within term. Second, the variable `A_l`, with the local index of school segregation by race in each state. Following (2), the weighted average of `A_l` using `A_w` as weights results in `A_W`:

$$0.4643 = 0.0471 \times 0.7793 + 0.0220 \times 0.3858 + 0.1150 \times 0.4991 + 0.0412 \times 0.7597 + 0.6009 \times 0.5499$$

4.9 Advanced usage: The fine points

Thus far, we have covered single calls to the command `dseg` with an increasing level of complexity. Nevertheless, with this command, we can also design a strategy of multiple calls to achieve an assortment of results that may deepen the analysis of segregation.

4.9.1 Chained decomposition along the unit dimension

Imagine we wished to resume Rivkin's (1994) argument as follows: analogous to segregation in school districts that caps the amount of segregation that can possibly exist in schools per se, it is reasonable to expect in turn that the differential distribution of races across states caps segregation in school districts.

With decomposable indices, we can control for the effect of multiple levels of geographical aggregation on school racial segregation, our ultimate object of interest. Using the relative diversity $R_{\text{group|unit}}$ index as an example, we can decompose race segregation in schools, controlling for districts and states, into 1) a between-state term, STATE, that measures the level of segregation in states; 2) a within-state term, DISTRICT, that gauges the average level of segregation in school districts that is not mixed up with segregation in states; and 3) a within-districts term, SCHOOL, that gauges the average level of segregation in schools that is mixed up with neither segregation in states nor segregation in school districts:

$$R_{\text{race|school}} = \text{STATE} + \text{DISTRICT} + \text{SCHOOL} \tag{5}$$

To reach decomposition (5), and denoting states by t and districts by d, consider the following one-way unit decompositions of the $R_{\text{race|school}}$ and $R_{\text{race|district}}$ indices:

$$R_{\text{race|school}} = R_{\text{race|district}} + \sum_d w_d R_{\text{race|school}}(d) \tag{6}$$

$$R_{\text{race|district}} = R_{\text{race|state}} + \sum_t w_t R_{\text{race|district}}(t) \tag{7}$$

$R_{\text{race|state}}$ captures race segregation in states, $R_{\text{race|school}}(t)$ captures race segregation in the schools of state t, and $R_{\text{race|school}}(d)$ captures race segregation in the schools of district d. Equation (5) follows from replacing $R_{\text{race|district}}$ in (6) with the right term in (7). Then we observe that 1) STATE is equivalent to $R_{\text{race|state}}$; 2) DISTRICT is equivalent to $\sum_t w_t R_{\text{race|district}}(t) = R_{\text{race|district}} - R_{\text{race|state}}$; and 3) SCHOOL is equivalent to $\sum_d w_d R_{\text{race|school}}(d) = R_{\text{race|school}} - R_{\text{race|district}}$.

We know the values of all the terms in (5) with two calls to `dseg`. $R_{\text{race|school}}$ is the total in the second call; STATE is the between term in the first call; DISTRICT is the within term in the first call; finally, SCHOOL is the within term in the second call:

```
. restore
. dseg diversity race [fw=student_count], given(district) within(state)
   Decomposable Multigroup Segregation Indexes
   Differences in race given district
     Index: Relative Diversity
   Between/Within state decomposition
         R      R_B      R_W
    0.2882   0.1087   0.1795
. dseg diversity race [fw=student_count], given(school) within(district)
   Decomposable Multigroup Segregation Indexes
   Differences in race given school
     Index: Relative Diversity
   Between/Within district decomposition
         R      R_B      R_W
    0.3512   0.2882   0.0630
```

Hence,

$$R_{\text{race|school}} = \text{STATE} + \text{DISTRICT} + \text{SCHOOL} \tag{8}$$
$$0.3512 = 0.1087 + \quad 0.1795 \quad + 0.0630$$

In words, the value of school race segregation in states is 0.1087, but it is $\{(0.1795/0.1087) - 1\} \times 100 = 65.13\%$ larger in districts. Finally, once we control for the effect of states and districts, race segregation in schools per se accounts for only $(0.0630/0.3512) \times 100 = 17.94\%$ of the measurement.

4.9.2 Chained decomposition along the unit and group dimensions

Let us revisit the example in subsection 4.7. There we studied the segregation of whites from minorities and among minorities. Let us suppose further that we wanted to control for the differential race shares in states and in school districts, as in the previous subsection 4.9.1. For this task, M is the only instrument in the toolbox because it is additively decomposable in partitions of units and groups. It takes three steps to accomplish this goal.

The first two steps replicate (8). With the M index,

$$\begin{aligned} M &= \text{STATE} + \text{DISTRICT} + \text{SCHOOL} \\ &= M^{\text{race}}_{\text{state}} + \sum_t p_{t\bullet} M^{\text{race}}_{\text{district}}(t) + \sum_d p_{d\bullet} M^{\text{race}}_{\text{school}}(d) \end{aligned} \tag{9}$$

The term $M^{\text{race}}_{\text{school}}(d)$ corresponds to the M index of race segregation in the schools of district d. As such, it can be decomposed as in (4) into a between term that measures segregation between minority and white students and a within term that captures school segregation that arises among the minorities. This is the third step:

$$M^{\text{race}}_{\text{school}}(d) = M^{\text{minority versus white}}_{\text{school}}(d) + p_{\text{minority}}(d) \times M^{\text{minorities}}_{\text{school}}(d) \tag{10}$$

Now we join (9) and (10):

$$\begin{aligned} M &= M^{\text{race}}_{\text{state}} + \sum_t p_{t\bullet} M^{\text{race}}_{\text{district}}(t) + \sum_d p_{d\bullet} M^{\text{minority versus white}}_{\text{school}}(d) + \sum_d p_{d\bullet} M^{\text{minorities}}_{\text{school}}(d) \\ &= \text{STATE} + \quad \text{DISTRICT} \quad + \quad \text{MINORITY VERSUS WHITE} \quad + \quad \text{MINORITIES} \end{aligned} \tag{11}$$

The novelty with respect to (8) lies in the last two terms. The term MINORITY VERSUS WHITE captures minority–white segregation in schools, controlling for districts. The term MINORITIES identifies the contribution to race segregation in schools that comes from the segregation among minorities, controlling for districts. With three calls to the `dseg` command, we follow these three steps in Stata.

```
. dseg mutual race [fw=student_count], given(district) nolist
> within(state) saving(step1,replace)
```

(*output omitted*)

```
. dseg mutual race [fw=student_count], given(school) prefix(step2) nolist
> within(district, components) saving(step2,replace)
```

(*output omitted*)

```
. dseg mutual school [fw=student_count], given(race) prefix(step3) nolist
> within(mrg) by(district) clear
```

(*output omitted*)

The `within(state)` option of `dseg` in the first call identifies the first two terms on the right hand side of (9). The `saving(step1, replace)` option stores in a new dataset (named `step1.dta`) the following variables: 1) `M` stands for the sum of STATE and DISTRICT in (11); 2) `M_B` for STATE; and 3) `M_W` for DISTRICT. The resulting `step1.dta` has one observation with national aggregates.

The second call identifies M and, thanks to the `within(districts, components)` option, also $p_{d\bullet}$ in (11). The `prefix(step2)` option avoids naming conflicts with variables generated in the other two calls. The `saving(step2, replace)` option stores five variables in `step2.dta`. However, we need only two of the five to continue obtaining the values of the terms in (11): `step2M`, which stands for M in (11), and `step2M_w` for $p_{d\bullet}$. `step2.dta` has 16,768 observations, one for each district. Variable `step2M` is constant because it is the nationwide index.

Thanks to the `by(district)` option, the third call to `dseg` generates the three terms in (10) from left to right in every district. The `clear` option replaces the original dataset with the new dataset that has 16,768 observations. Next, we merge the current dataset with `step2.dta`, which also includes one observation per district. We can perform a one-to-one merge directly because `dseg` automatically sorts these datasets by `district` when it creates them.

```
. merge 1:1 district using step2.dta
```
(*output omitted*)

For each district, we calculate the value of the term $p_{d\bullet}M_{\text{school}}^{\text{minority versus white}}(d)$ by multiplying `step2M_w` and `step3M_B`. Likewise, we obtain the value of the term $p_{d\bullet}M_{\text{school}}^{\text{minorities}}(d)$ with the product of `step2M_w` and `step3M_W`.

```
. generate MINORITIES=step2M_w * step3M_W
. generate MINORITY_WHITE=step2M_w * step3M_B
```

Next, we sum these two products over all districts with the `collapse` command and its `sum` option. With this operation, we obtain the last two terms in (11). Recall that the variable `step2M` has the value of M in (11). We keep this constant by using the `mean` option in the `collapse` command.

```
. collapse (sum) MINORITIES MINORITY_WHITE (mean) M=step2M
```

Finally, we merge the only observation in the current data with `step1.dta`, where the values of STATE and DISTRICT are stored. We thereby obtain all the terms in (11), which we rename suitably for displaying the results:

```
. merge 1:1 _n using step1.dta
```
(*output omitted*)
```
. rename M_B STATE
. rename M_W DISTRICT
. list M STATE DISTRICT MINORITY_WHITE MINORITIES, abbreviate(15)
```

	M	STATE	DISTRICT	MINORITY_WHITE	MINORITIES
1.	0.4678	0.1505	0.2331	.0385773	.0456264

Racial segregation in states and districts accounts for around $(0.1505+0.2331)/0.4678\times 100 = 82\%$ of race segregation in schools. The contribution to school racial segregation of segregation among minorities only, controlling for the segregation that arises between minorities and whites, and for the segregation due to states and districts, is 0.0456264 or $0.0456/0.4678 \times 100 = 9.75\%$. Moreover, in subsection 4.7 we found that school segregation among minorities accounts for about half the segregation fueled by the seven race groups. We can now conclude that this result holds after controlling for states and districts: $0.0456/(0.0386 + 0.0456) = 0.54$.

This is only an example. The M index can be decomposed in as many levels as deemed useful. By levels, we mean any arbitrary combination of partitions in the

unit and group dimensions. Then the analyst may use `dseg` to implement the desired multilevel decomposition with only a few command lines.

4.9.3 The interaction term

For brevity, in this section we consider only two-way unit decompositions in unit-decomposable indices. However, its arguments apply to two-way group decompositions and to two-way decompositions that combine a unit and a group partition. Let discrete variables d_1 and d_2 be two unit-level characteristics. By partitioning the units along the levels of d_1, we identify a within term that captures the part of the overall segregation that is exclusively due to d_2: the within term would become zero in the hypothetical case that all segregation within each cluster was eliminated. Moreover, vice versa: the partition defined by the levels of d_2 identifies the segregation exclusively due to d_1.

With the two decompositions, the segregating force of each factor in exclusivity can be fairly appreciated because the influence of the other factor is controlled for. This we have repeated on numerous occasions in the article. However, thus far we have not mentioned that the common metric upon which we weigh the relative standing of each factor allows assessing whether the two factors interact, that is, whether there is some part of the segregation they jointly produce that cannot be unambiguously attributed to any of the two factors by itself. The interaction term I can be defined as the segregation that is not unambiguously attributable to either d_1 or d_2. Mathematically, denoting by 1) Ψ the overall segregation that is jointly produced by the two factors, 2) Ψ^{d_1} the segregation exclusively due to d_1, and 3) Ψ^{d_2} the segregation exclusively due to d_2, then I is equal to whatever remains of the overall segregation once we subtract from it all that is uniquely attributable to each factor:

$$I = \Psi - \left(\Psi^{d_1} + \Psi^{d_2}\right)$$

To illustrate the role that the interaction term may play in the case of unit decompositions, let us compare the exclusive contributions given by the within terms of the two decompositions in subsection 4.6: 0.0568 versus 0.0121; that is, the racial mix of CBSAs yields almost five times more segregation than it does in states. In percentages, CBSAs account, per se, for $100 \times (0.0568/0.1695) = 33.51\%$ of the overall segregation that there is in CBSAs and states; instead, states generate only $100 \times (0.0121/0.1695) = 7.14\%$ of it.

Recall we denote overall race segregation in CBSAs and states by $H_{\text{race|CBSA}\times\text{state}}$ in subsection 4.6. Now we denote the exclusive contribution of states to race segregation in states and CBSAs by $H^{\mathcal{T}} = \sum_c w_c H_{\text{race|state}}(c)$ and the exclusive contribution of CBSAs to race segregation in states and CBSAs by $H^{\mathcal{C}} = \sum_t w_t H_{\text{race|CBSA}}(t)$; the interaction term I is

$$\begin{aligned} I &= H_{\text{race|CBSA}\times\text{state}} - \left\{H^{\mathcal{T}} + H^{\mathcal{C}}\right\} \\ &= 0.1695 - (0.0121 + 0.0568) \\ &= 0.1006 \end{aligned}$$

The interaction accounts for $100 \times (0.1006/0.1695) \approx 59.35\%$ of the race segregation in CBSAs and states. Substantively, this large chunk of race segregation in this instance of the $H_{\text{group|unit}}$ index stems from race differences in the student enrollment body in the geographical units defined by the combination of a CBSA category (including rural) and a state. Such racial differences are greater, in absolute and relative terms, than the ones observed when the units are defined by either the CBSA classification within states or state boundaries within CBSAs.

The interaction term can be negative. For example, Guinea-Martin, Mora, and Ruiz-Castillo (2015) report with the M index a small but negative interaction term that emerges from the joint effect of race and sex on occupational segregation. It reflects that each variable, controlling for the other, generates a more informative distribution than when it is measured directly: sex is more informative about someone's occupation when race is controlled for than when we observe only the distribution of women and men in occupations. Likewise, race is more informative when sex is controlled for. Why? The two factors are mixed up or confounded, and they pull the occupational distribution in opposite directions. Not controlling for one of them waters down the impact of the other.

4.9.4 Survey data: Bootstrapping and simulation techniques

Bias correction and confidence limits. Thus far, we have used census data that cover the population of interest comprehensively. Consequently, proportions p_{ng} can be interpreted as probabilities $\Pr(\text{unit} = n, \text{group} = g)$. For sample data, as the number of observations increases and the experiment of observing individuals' unit n and group g is continued indefinitely, the sample frequencies converge in probability to the corresponding probabilities: $p_{ng} \xrightarrow{p} \Pr(\text{unit} = n, \text{group} = g)$ (Bulmer 1979).

By contrast, survey-based measurements of segregation are finite-sample estimates and, therefore, biased and subject to sample variability (Deutsch, Flückiger, and Silber 1994; Herranz, Mora, and Ruiz-Castillo 2005). Bootstrap methods can help estimate bias and basic bootstrap confidence intervals for segregation indices (Ransom 2000; Allen et al. 2015). Command `dseg` with the option `bootstraps()` implements the nonparametric type of bootstrap. It draws random samples with replacement from the data. It then creates a new dataset with index values calculated on each bootstrap sample.

Next we simulate samples with the 2017 CCD census of the student enrollment body in U.S. public schools. With these samples and the `bootstraps()` option, we show below that the seriousness of the bias in the estimation of segregation measures varies with sample size. We also illustrate the computation of bias-corrected indices and basic bootstrap confidence intervals after using `dseg`.

In some contexts, we may find surveys sampled on some of the key variables in the analysis. For example, school surveys are common. If the sampled schools are representative of all the schools, then a cluster bootstrap may be appropriate and can be implemented in `dseg` with the `bootstraps()` option and its `cluster()` suboption as in `bootstraps(#, cluster(`*varlist*`))`.

However, in the following illustration, we survey students rather than schools. This replicates a common situation where the analysis is based on survey data. For instance, gender segregation in occupations is usually measured with labor force surveys of individuals sampled on households or firms but not on occupations. In this scenario, typically individuals from all groups and units of interest are sampled. To keep computation times short, we sample students enrolled in the 1,334 public schools of Alabama, which happens to be the first state when these are ordered alphabetically. Further assume that our goal is estimating with M the level of race segregation in the 138 school districts of Alabama. To achieve it, we randomly sample the students. Survey samples and weights are seldom designed to accurately estimate the joint distribution of key variables for the analyst, such as school districts and races in our example. Let us assume that survey weights are provided to estimate, say, the number of girls and boys in Alabama.

In the 2017 CCD, there are 671,939 students in Alabama. School districts range in size from 245 to 50,028 students, with an average of approximately 4,869. To illustrate the basic algorithm, let us consider the bias that arises from a 10% random sample of the population and 500 bootstrap replications. To sample, we first expand the 2017 CCD:

```
. drop if student_count==0 | student_count>=.
  (output omitted)
. expand student_count
  (output omitted)
. sample 10
  (output omitted)
. generate survey_count=1
. collapse (sum) survey_count, by(sex race district)
```

The first command line above ensures that the expanded data contain no observations with missing or nonpositive values in `student_count`. The third line randomly samples 10% of the observations. The last line aggregates the data again along the levels of the relevant variables for the weighting scheme (`sex`) and for the ensuing analysis (`race` and `district`).

Next, we generate frequency weights (variable `weights`). They allow the estimation of the number of girls and boys in the Alabamian public school system with a 10% sample of its population. Let `weights.dta` (sorted by `sex`) contain the variable `target` with the total count of students in the population by sex. The weights are equal to the population number of girls and boys per survey observation, rounded to the closest integer.

```
. egen survey=sum(survey_count), by(sex) // # of girls and boys in the survey
. sort sex
. merge m:1 sex using weights.dta
  (output omitted)
. generate weights=round(survey_count*(target/survey))
```

Before calling `dseg`, we expand the sample data with these weights, or else the option `bootstraps()` will not work.

```
. expand weights
. dseg mutual race, given(district) bootstraps(500) saving("boots.dta")
  (output omitted)
```

The new data file `boots.dta` has 501 observations and includes two variables: 1) `bsn` identifies the bootstrap sample (`bsn==0` refers to the original survey sample); 2) `M` is the index value. Data are automatically sorted by `bsn`. Hence, `M[1]`, the first observation in `boots.dta`, corresponds to the M index computed with the original survey sample.

A bootstrap estimate of the expectation of the M index for the survey sample is the average of M in the 500 bootstrap samples: $500^{-1} \times \sum_{b=2}^{501}$ `M[`b`]`. Following Davison and Hinkley (1997), we can estimate the bias in `M[1]` as

$$\widehat{\text{Bias}} = \frac{1}{500} \times \sum_{b=2}^{501} \texttt{M[}b\texttt{]} \; - \; \texttt{M[1]}$$

The bootstrap bias-corrected estimate of M is then obtained as

$$\begin{aligned} \widehat{M}_{bc} &= \texttt{M[1]} - \widehat{\text{Bias}} \\ &= 2 \times \texttt{M[1]} - \frac{1}{500} \times \sum_{b=2}^{501} \texttt{M[}b\texttt{]} \end{aligned} \tag{12}$$

The sample variance of the 500 M values obtained from the bootstrap samples is a consistent estimator of the variance of `M[1]`. Unfortunately, the variance is of limited value because the normal approximation of `M[1]`'s distribution is likely to be poor and so is that of any segregation index computed with a survey sample. Asymptotic results are obtained under the assumption that the number of observations goes to ∞ (Allen et al. 2015). Thus, bootstrap quantile estimates likely are a better strategy for obtaining confidence intervals, provided that the number of bootstrap replications is large enough. The basic bootstrap 90% upper confidence limit for $\widehat{M}_{bc}$ is

$$\widehat{M}^{90} = \widehat{M}_{bc} - (\; \texttt{M[}c\texttt{]} \; - \; \texttt{M[1]})$$

where c is the observation number marking the first decile in the ascendingly ordered distribution of $\{\texttt{M[}b\texttt{]}\}_{b=2}^{501}$.

With these samples and the `bootstraps()` option, we argue below that the seriousness of the bias in the estimation of segregation measures varies with sample size. We

also illustrate the computation of bias-corrected indices and basic bootstrap confidence intervals after using `dseg`.

To show the effect of sample size on the estimation of segregation measures with survey data, we draw one sample for each sampling level set at 5 percentage-point intervals in the range from 5 to 95% of the population. Panel (a) in figure 1 shows the results of this analysis. Unsurprisingly, the survey-based index value is systematically larger than both the population-based index and the bias-corrected estimate. However, bias correction is not satisfactory. It is insufficient for small sample sizes and unnecessary for large sample sizes. Moreover, the true segregation level is not within the confidence interval for sample sizes smaller than 15%.

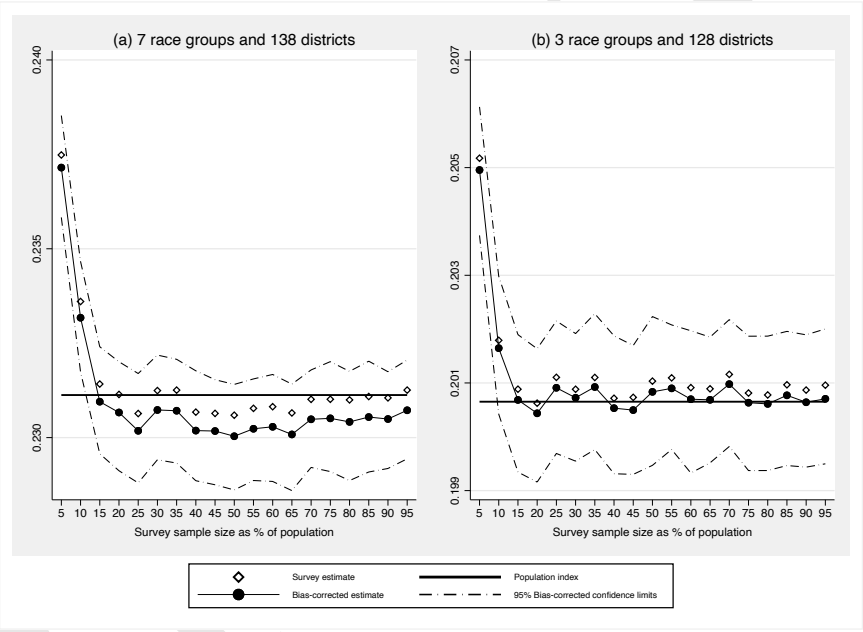

Figure 1. Bias and variance in sample-based segregation indices: Race segregation in the school districts of Alabama.
NOTE: Bias correction and basic bootstrap confidence limits obtained with 500 bootstrap replications (random sampling with replacement). The Population index is the M index obtained with the complete 2017 CCD for Alabama: 0.2311 with 7 race groups and 138 districts; 0.2007 with 3 race groups and 128 districts. The Survey estimate is obtained from a simulated random sample with weights based on the ratio of the population number of boys and girls to the corresponding survey figures. The Bias-corrected estimate is the index obtained by applying (12).

What is best to do, then? A simple strategy would trade off the less ambitious facets of our research goals for a more robust analysis. For example, in Alabama, almost 90% of students are either white (54.6%) or black (32.8%). Correspondingly,

our analyses by race can be based on a partition of three supergroups: white, black, and other. Trading the original classification with seven categories for its simplified ternary makes the population-based index of race segregation in schools fall by only $(0.2311 - 0.2035)/0.2311 \times 100 = 12\%$.

Nevertheless, small districts are a potential source of small-cell bias. Merging those with fewer than 1,000 students with adjoining districts of a similar racial profile reduces their number from 138 to 128.[4] Altogether, aggregating by both race groups and districts lowers the population-based index by $(0.2311 - 0.2007)/0.2311 \times 100 = 13\%$, compared with its original level.

The aggregation of race groups and districts allows us to obtain acceptable estimates with sample sizes of 10% of the population [see panel (b) in figure 1]. The lesson here is that decomposable indices are useful for analyzing survey data that may give rise to small-cell problems. (See Herranz, Mora, and Ruiz-Castillo [2005] for an application of this approach to gender segregation in occupations and industries with the M index.)

Estimation of random segregation. Suppose that $\widehat{\Psi}$ is an estimate of index Ψ that is obtained with a small sample of the population. $\widehat{\Psi}$ can be positive even if $\Psi = 0$ because of integer constraints (each individual must be uniquely allocated to one unit) and sample variation in small units. To discard this possibility, Boisso et al. (1994) propose calculating the value of the segregation index for repeated samples $j = 1, \ldots, J$ under the null hypothesis of no segregation, $\widehat{\Psi}^*_{(j)}$. They also propose carrying out an approximate randomization test. Generally, randomization tests are used to test the null hypothesis that one set of variables is unrelated to another. Random shuffling of one of the sets ensures that there is no relation between the two sets. Hence, in the presence of a true relation, the value of the statistic for the unshuffled data should be unusual (Noreen 1989).

The distribution of the index under the null can be estimated with the empirical distribution function obtained from resampling. The test implies the computation of the p-value as $(1 + \#\{\widehat{\Psi}^*_{(j)} \geq \widehat{\Psi}\})/(J + 1)$, where $\#\{A\}$ means the number of times the event A occurs (Davison and Hinkley 1997).

Implementing this approximate randomization test with the `dseg` command requires individual-level data and writing a few lines of code. Next is an example with 999 replications based on the 10% Alabamian sample data that we created earlier:

```
. expand weights
  (output omitted)
. dseg mutual race, given(district) random(999) clear
  (output omitted)
```

4. We merge Sheffield City with Tuscumbia City; Midfield City with Fairfield City; Greene County with Sumter County; Linden City with Marengo County; Acceleration Day And Evening Academy Charter Agency with Mobile County; Lanet City with Chambers County; Barbour County with Bullock County; Daleville City with Ozark city; and Elba City with Enterprise City and Coffee County.

```
. generate count=sum(M>=M[1]) in 2/l
```
(*output omitted*)
```
. generate pvalue=(1+count)/_N
```
(*output omitted*)
```
. list pvalue in 2, clean noobs
    pvalue
      .001
```

The null hypothesis of no segregation is rejected if the p-value is less than or equal to the specified rejection level for the test. In this example, p-value $= 1/1000$ because the segregation index of the unshuffled data is never smaller than the index in any of the 999 null samples. Hence, the null is rejected at conventional significance levels.

5 Conclusions

In this article, we presented the community-contributed command `dseg`. The command computes eight indices of segregation that are multigroup and decomposable. These are useful qualities when segregation stems from multiple groupings and levels of organization in the units. Six of the eight indices follow one of two notions on segregation. The mutual information index and its weak normalization follow both. The most common notion in the literature compares the group mix in each unit with the overall mix. The other notion compares the group distribution across units with the overall distribution. The syntax of `dseg` aids the user in being explicit about what notion the index chosen follows and, correspondingly, about the actual meaning of "segregation" in their research.

The command `dseg` also provides the following advantages to users. First, it computes one- and two-way decompositions intuitively, directly, and securely. Second, it can give the weights and local indices that are used in the calculation of the within term of the decomposition. These quantities can then be used in chained decompositions. Third, it accepts frequency weights for working with aggregate or sample data. Fourth, it helps deal with small sample-size problems in the estimation of the indices through bootstrapping and approximate randomization tests.

The findings from our illustration of the usage of `dseg` with U.S. school enrollment data show that decomposable multigroup indices are a resourceful tool for analyzing segregation with a multilevel organization of groups or units, or both. For example, we find that 1) the contribution to race segregation in schools of segregation among minorities accounts for a sizable portion of the total, and this fact remains after controlling for the differential race shares in states and districts; and 2) only a small share of school racial segregation can be unambiguously attributed to the segregation of races in schools, net of state and district.

6 Acknowledgments

We thank Stephen P. Jenkins as well as an anonymous reviewer for comments that helped improve the article. We also acknowledge the financial support of Spain's Ministry of Science (MCIN/AEI/10.13039/501100011033, project number PID2019-108576RB-I00).

7 Programs and supplemental materials

To install a snapshot of the corresponding software files as they existed at the time of publication of this article, type

```
. net sj 22-3
. net install st00!!     (to install program files, if available)
. net get st00!!         (to install ancillary files, if available)
```

8 References

Akerlof, G. A., and R. E. Kranton. 2010. *Identity Economics: How Our Identities Shape Our Work, Wages, and Well-Being*. Princeton, NJ: Princeton University Press.

Allen, R., S. Burgess, R. Davidson, and F. Windmeijer. 2015. More reliable inference for the dissimilarity index of segregation. *Econometrics Journal* 18: 40–66. https://doi.org/10.1111/ectj.12039.

Allison, P. D. 1978. Measures of inequality. *American Sociological Review* 43: 865–880. https://doi.org/10.2307/2094626.

Alonso-Villar, O., and C. del Río. 2010. Local versus overall segregation measures. *Mathematical Social Sciences* 60: 30–38. https://doi.org/10.1016/j.mathsocsci.2010.03.002.

Boisso, D., K. Hayes, J. Hirschberg, and J. Silber. 1994. Occupational segregation in the multidimensional case: Decomposition and tests of significance. *Journal of Econometrics* 61: 161–171. https://doi.org/10.1016/0304-4076(94)90082-5.

Bulmer, M. G. 1979. *Principles of Statistics*. 2nd ed. Mineolo, NY: Dover.

Carlson, S. M. 1992. Trends in race/sex occupational inequality: Conceptual and measurement issues. *Social Problems* 39: 268–290. https://doi.org/10.2307/3096962.

Casey, D. L. 2016. *The Casey Review: A Review into Opportunity and Integration*. London: Department for Communities and Local Government.

Chakravarty, S. R., and J. Silber. 1994. Employment segregation indices: An axiomatic characterization. In *Models and Measurement of Welfare and Inequality*, ed. W. Eichhorn, 912–920. Berlin: Springer. https://doi.org/10.1007/978-3-642-79037-9_48.

Coleman, J. S., E. Q. Campbell, C. J. Hobson, J. McPartland, A. M. Mood, F. D. Weinfeld, and R. L. York. 1966. *Equality of Educational Opportunity*. Washington, DC: U.S. Government Printing Office.

Correia, S. 2016. ftools: Stata module to provide alternatives to common Stata commands optimized for large datasets. Statistical Software Components S458213, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s458213.html.

Cox, N. J. 1999. dissim: Stata module to calculate dissimilarity index. Statistical Software Components S365901, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s365901.html.

Davison, A. C., and D. V. Hinkley. 1997. *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.

Denton, N. A., and D. S. Massey. 1988. Residential segregation of Blacks, Hispanics, and Asians by socioeconomic status and generation. *Social Science Quarterly* 69: 797–817.

Deutsch, J., Y. Flückiger, and J. Silber. 1994. Measuring occupational segregation: Summary statistics and the impact of classification errors and aggregation. *Journal of Econometrics* 61: 133–146. https://doi.org/10.1016/0304-4076(94)90080-9.

Duncan, O. D., and B. Duncan. 1955. A methodological analysis of segregation indexes. *American Sociological Review* 20: 210–217. https://doi.org/10.2307/2088328.

Flückiger, Y., and J. Silber. 1999. *The Measurement of Segregation in the Labor Force*. Heidelberg: Physica-Verlag.

Frankel, D. M., and O. Volij. 2011. Measuring school segregation. *Journal of Economic Theory* 146: 1–38. https://doi.org/10.1016/j.jet.2010.10.008.

Gradín, C. 2011. segregation: Stata module to compute segregation indices. Statistical Software Components S457266, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s457266.html.

Gross, E. 1968. Plus ça change...? The sexual structure of occupations over time. *Social Problems* 16: 198–208. https://doi.org/10.2307/800005.

Guinea-Martin, D., R. Mora, and J. Ruiz-Castillo. 2015. The joint effect of ethnicity and gender on occupational segregation. An approach based on the Mutual Information Index. *Social Science Research* 49: 167–178. https://doi.org/10.1016/j.ssresearch.2014.08.007.

Gutiérrez, G., J. Jerrim, and R. Torres. 2020. School segregation across the world: Has any progress been made in reducing the separation of the rich from the poor? *Journal of Economic Inequality* 18: 157–179. https://doi.org/10.1007/s10888-019-09437-3.

Herranz, N., R. Mora, and J. Ruiz-Castillo. 2005. An algorithm to reduce the occupational space in gender segregation studies. *Journal of Applied Econometrics* 20: 25–37. https://doi.org/10.1002/jae.829.

James, D. R., and K. E. Taeuber. 1985. Measures of segregation. *Sociological Methodology* 15: 1–32. https://doi.org/10.2307/270845.

Jann, B. 2004. duncan: Stata module to calculate dissimilarity index. Statistical Software Components S447202, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s447202.html.

Jenkins, S. P. 2006. hutchens: Stata module to calculate the Hutchens 'square root' segregation index with optional decompositions by subgroup. Statistical Software Components S456601, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s456601.html.

Massey, D. S., and N. A. Denton. 1988. The dimensions of residential segregation. *Social Forces* 67: 281–315. https://doi.org/10.2307/2579183.

———. 1994. *American Apartheid: Segregation and the Making of the Underclass*. Cambridge, MA: Harvard University Press.

Mora, R. 2014. dseg: Stata module to compute decomposable multigroup segregation indexes. Statistical Software Components S457839, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s457839.html.

Mora, R., and J. Ruiz-Castillo. 2003. Additively decomposable segregation indexes. The case of gender segregation by occupations and human capital levels in Spain. *Journal of Economic Inequality* 1: 147–179. https://doi.org/10.1023/A:1026198429377.

———. 2011. Entropy-based segregation indices. *Sociological Methodology* 41: 159–194. https://doi.org/10.1111/j.1467-9531.2011.01237.x.

Noreen, E. W. 1989. *Computer-Intensive Methods for Testing Hypotheses: An Introduction*. New York: Wiley.

Ransom, M. R. 2000. Sampling distributions of segregation indexes. *Sociological Methods and Research* 28: 454–475. https://doi.org/10.1177/0049124100028004003.

Reardon, S. F., and G. Firebaugh. 2002. Measures of multigroup segregation. *Sociological Methodology* 32: 33–67. https://doi.org/10.1111/1467-9531.00110.

Reardon, S. F., and J. B. Townsend. 1999. seg: Stata module to compute multiple-group diversity and segregation indices. Statistical Software Components S375001, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s375001.html.

Reardon, S. F., J. T. Yun, and T. M. Eitle. 2000. The changing structure of school segregation: Measurement and evidence of multiracial metropolitan-area school segregation, 1989–1995. *Demography* 37: 351–364. https://doi.org/10.2307/2648047.

Rivkin, S. G. 1994. Residential segregation and school integration. *Sociology of Education* 67: 279–292. https://doi.org/10.2307/2112817.

Theil, H., and A. J. Finizza. 1971. A note on the measurement of racial integration of schools by means of informational concepts. *Journal of Mathematical Sociology* 1: 187–193. https://doi.org/10.1080/0022250X.1971.9989795.

About the authors

Daniel Guinea-Martin is an assistant professor in the Department of Sociology at Universidad de Málaga (UMA), Spain. He specializes in research methods and inequality.

Ricardo Mora is an associate professor in the Department of Economics at Universidad Carlos III de Madrid. Mora develops and applies quantitative methods in the fields of economics and sociology.