Measuring Segregation Using Stata: The Two-group Case

Carlos Gradín[1]

Universidade de Vigo

This version: February, 17, 2014

This document presents the formulae of the indices reported by the *dicseg.ado* command (version 2, February 2014) included in the SEGREGATION Stata Module. For an application of some of these indices, see Gradín (2013).

For the multigroup case, use the command *localseg.ado*.

1. Notation

Let us consider a population of N individuals (e.g. workers) distributed across $T > 1$ units (e.g. occupations), with $N = \sum_{j=1}^{T} n_j$>0; $n_j \geq 0$ being the total number of workers in the jth occupation, $j = \{1, \dots T\}$.

Let us also consider an exhaustive partition of this population into 2 groups (e.g. men and women; white and nonwhite):

$$n = (n^1, n^2) = (n_1^1, \dots, n_T^1, n_1^1, \dots, n_T^2).$$

Each group has size $N^i = \sum_{j=1}^{T} n_j^i > 0$, where $n_j^i \geq 0$ is the number of members of the ith group ($i = 1, 2$) in jth occupation, with $N = N^1 + N^2$.

2. The segregation curve

A partial ordering of distributions can be obtained by comparing the corresponding segregation curves, which correspond to the Lorenz curves in the measurement of inequality.

[1] Facultade de CC. Económicas, Universidade de Vigo, Lagoas-Marcosende, s/n, Vigo, Galicia, Spain. Email: cgradin@uvigo.es.

The first known application of this curve in segregation appeared in Duncan and Duncan (1955), while Hutchens (1991) provides its rationale by showing its connection with inequality measurement.

Let us consider the total distribution of employment n^{+} and that of each group n^{+i}, $i = \{1, 2\}$, in which occupations are ordered in increasing values of n_j^2/n_j (or equivalently, n_j^2/n_j^1 if $n_j^1 > 0$ for all j). Each value of the segregation curve $S_j(n^{+1}, n^{+2})$ represents for each occupation the cumulative proportion of one group $\sum_{s=1}^{j} n_s^{+1}/N^1$ on the abscissa and the cumulative proportion of the other group $\sum_{s=1}^{j} n_s^{+2}/N^2$ on the ordinate.

Indices *G*, *GE*, and *A* defined below are consistent with the ranking from the segregation curves.

3. Segregation indices

Most segregation indices are borrowed from the analysis of inequality (for the relationship between measuring inequality and segregation, see Silber, 1989 or Hutchens, 1991).

The *dissimilarity* index *D* is the equivalent to the Relative Mean Deviation (Pietra index) in inequality analysis (Jahn, Schmid and Schrag, 1947; popularized by Duncan and Duncan, 1955):

$$D(n^1, n^2) = \frac{1}{2}\sum_{j=1}^{T}\left|\frac{n_j^2}{N^2} - \frac{n_j^1}{N^1}\right|.$$

Karmel and MacLachlan (1988) KM is a transformation of *D:*

$$KM(n^1, n^2) = 2\,\frac{N^1}{N}\frac{N^2}{N}D(n^1, n^2).$$

The *Gini* index *G*. First introduced to the measurement of segregation by Jahn, Schmid and Schrag (1947). It is the double of the area between the diagonal and the segregation curve.

When the distribution is ordered (n^{+}), *G* can be expressed as follows (Hutchens, 1991):

$$G(n^1,n^2) = 1 - \sum_{j=1}^{T} \frac{n_j^{+2}}{N^2}\left(\frac{n_j^{+1}}{N^1} + 2\sum_{s=j+1}^{T}\frac{n_s^{+1}}{N^1}\right).$$

D, *KM* or *G* are always finite.

D and *G* are bounded between 0 (minimum segregation when both distributions are identical, $\frac{n_j^1}{N^1} = \frac{n_j^2}{N^2}\ \forall j$) and 1 (maximum segregation when there is no overlapping between both distributions, $n_j^1 n_j^2 = 0\ \forall j$). The *KM* index takes values between 0 and $2\frac{N^1}{N}\frac{N^2}{N}$.

D, *KM*, and *G* are symmetric in types, segregation does not change after swapping the groups (e.g. $G(n^1,n^2) = G(n^2,n^1)$).

The *Generalized Entropy family* of indices, defined for any real number α, allows comparing segregation with different levels of sensitiveness to segregation that occurs at different points of the distribution n^2. (See, for example, Hutchens, 1991, 2001, 2004).

$$GE(n^1,n^2;\alpha) = \begin{cases} \frac{1}{\alpha(\alpha-1)}\sum_{j=1}^{T}\frac{n_j^1}{N^1}\left[\left(\frac{n_j^2/N^2}{n_j^1/N^1}\right)^{\alpha} - 1\right], & \alpha \neq 0,1 \\ \sum_{j=1}^{T}\frac{n_j^1}{N^1}\, ln\left(\frac{n_j^1/N^1}{n_j^2/N^2}\right), & \alpha = 0. \\ \sum_{j=1}^{T}\frac{n_j^2}{N^2}\, ln\left(\frac{n_j^2/N^2}{n_j^1/N^1}\right), & \alpha = 1 \end{cases}$$

This family embraces as particular cases the Mean Log Deviation for $\alpha = 0$, a scalar transformation of the Hutchens' (2001, 2004) square root *H* for α =.5, the Theil index *T* for α =1, and half the squared Coefficient of Variation for α =2.

$$H(n^1,n^2) = \frac{1}{4} GE(n^1,n^2;\alpha = .5) = 1 - \sum_{j=1}^{T} \sqrt{\frac{n_j^1}{N^1}\frac{n_j^2}{N^2}}.$$

$$CV(n^1,n^2) = \sqrt{\sum_{j=1}^{T} \frac{n_j^1}{N^1}\left[\left(\frac{n_j^2/N^2}{n_j^1/N^1}\right)^2 - 1\right]}.$$

$$GE(n^1,n^2;\alpha = 2) = \frac{1}{2} CV^2(n^1,n^2).$$

$GE(n^1,n^2;\alpha)$ is not generally symmetric in types (except if $\alpha = .5$ or *H*). However, it is easy to check that $GE(n^1,n^2;\alpha) = GE(n^2,n^1;1-\alpha)$.

Unlike *D*, *KM* or *G*, GE is not generally finite unless we explicitly delimit occupations such that $n_j^1 > 0, n_j^2 > 0$ for all *j*. GE is always finite (Hutchens, 20014) only for $GE(n^1,n^2;0 < \alpha < 1)$. If for some *j*, $n_j^2 = 0$, then GE is infinite for $\alpha \leq 0$; if $n_j^1 = 0$, then GE is infinite for $\alpha \geq 1$.[2]

Also unlike *G* or *D*, GE is generally not bounded between 0 and 1. In the case of minimum segregation, $GE(n^1,n^2;\alpha) = 0$. In the case of maximum segregation, $GE(n^1,n^2;\alpha \leq 0 \; or \; \alpha \geq 1)$ is infinite, as mentioned above, while $GE(n^1,n^2;0 < \alpha < 1)$=$\frac{1}{\alpha(1-\alpha)}$ ≥ 4. Thus, we need to multiply $GE(n^c,n^r;0 < \alpha < 1)$ by $\alpha(1-\alpha)$ if we want the maximum to be 1 (such as in the squared root).

Atkinson Index. Similarly, we can define the Atkinson's (1970) index in the context of segregation.

[2] If for some *j*, $n_j^1 = 0$, we use the fact that $\lim_{x \to 0} x ln(x) = 0$ to compute $GE(n^1,n^2;\alpha = 0)$. The same with $GE(n^1,n^2;\alpha = 1)$ if $n_j^2 = 0$ for some *j*.

$$A(n^1,n^2;\varepsilon)=\begin{cases}1-\left[\sum_{j=1}^{T}\frac{n_j^1}{N^1}\left(\frac{n_j^2/N^2}{n_j^1/N^1}\right)^{1-\varepsilon}\right]^{\frac{1}{1-\varepsilon}}=1-\left[\sum_{j=1}^{T}\left(\frac{n_j^1}{N^1}\right)^{\varepsilon}\left(\frac{n_j^2}{N^2}\right)^{1-\varepsilon}\right]^{\frac{1}{1-\varepsilon}}, & \varepsilon>0,\varepsilon\neq 1\\ 1-\frac{\prod_{j=1}^{T}\left(n_j^2/n_j^1\right)^{n_j^1/N^1}}{N^2/N^1}, & \varepsilon=1\end{cases}.$$

Where ε measures the sensitivity to segregation. Like *GE*, A index might be infinite in some cases. If some $n_j^2=0$, then, $A(n^1,n^2;\varepsilon>1)$ is infinite.[3]

$A(n^1,n^2;\varepsilon)$ (with $n_j^2>0$ for all *j*) is bounded between 0 and 1, and is ordinally equivalent to $GE(n^1,n^2;\alpha)$ for $\varepsilon=1-\alpha>0$.

A is not symmetric in types, unless $\varepsilon=.5$. However, it can be checked that:[4]

$$[1-A(n^1,n^2;\varepsilon)]^{1-\varepsilon}=[1-A(n^2,n^1;1-\varepsilon)]^{\varepsilon}\ \forall\varepsilon\ .$$

Mutual Information Index (Theil and Finizza, 1971).

$$T(n^1,n^2)=\frac{N^1}{N}\log\left(\frac{N}{N^1}\right)+\frac{N^2}{N}\log\left(\frac{N}{N^2}\right)-\sum_{j=1}^{T}\frac{n_j}{N}\left(\frac{n_j^1}{n_j}\log\left(\frac{n_j}{n_j^1}\right)+\frac{n_j^2}{n_j}\log\left(\frac{n_j}{n_j^2}\right)\right).$$

We report log in base 2, so the index is bounded between 0 and 1 (Mora and Ruiz-Castillo, 2003).

---

[3] In this case $A(n^1,n^2;\varepsilon=1)=1$. Note also that if some $n_j^1=0$, then in this case the $\lim_{n_j^1\to 0}\left(n_j^2/n_j^1\right)^{n_j^1/N^1}=1$.

[4] In particular: $=0\rightarrow A(n^1,n^2;0)=0\rightarrow[1-A(n^1,n^2;0)]^1=1=[1-A(n^2,n^1;1)]^0$, and $\varepsilon=1\rightarrow[1-A(n^1,n^2;1)]^0=1=[1-A(n^2,n^1;0)]^1$.

References

Atkinson, Anthony B. (1970), On the Measurement of Inequality, Journal of Economic Theory, 2: 244-263.

Duncan, Otis D. and Bervely Duncan, (1955), A Methodological Analysis of Segregation Indexes, *American Sociological Review*, 20(2): 210-217.

Gradín, Carlos (2013), Conditional occupational segregation of minorities in the U.S., Journal of Economic Inequality, 11(4): 473-493.

Jahn, Julius, Calvin F. Schmid, and Clarence Schrag (1947), The Measurement of Ecological Segregation, *American Sociological Review*, 12(3): 293-303.

Hutchens, Robert M. (1991), “Segregation Curves, Lorenz Curves, and Inequality in the Distribution of People across Occupations”, Mathematical Social Sciences, 21: 31-51.

Hutchens, Robert M. (2004), One Measure of Segregation, *International Economic Review*, 45(2): 555-578.

Hutchens, Robert M. (2004), “One Measure of Segregation”, International Economic Review 45(2): 555-578.

Karmel, T. and Maclachlan M. (1988), Occupational Sex Segregation - Increasing or Decreasing, Economic Record, 64, 187-195.

Theil, Henri and Anthony J. Finizza (1971), A Note on the Measurement of Racial Integration of Schools by Means of Informational Concepts, *Journal of Mathematical Sociology*, 1: 187-194

Silber, Jacques (1989), On the Measurement of Employment Segregation, *Economic Letters*, 30: 237-243.

Mora, Ricardo and Javier 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.