CMPO

THE CENTRE FOR MARKET AND PUBLIC ORGANISATION

More Reliable Inference for Segregation Indices

Rebecca Allen, Simon Burgess and Frank Windmeijer

April 2009

Working Paper No. 09/216

Centre for Market and Public Organisation
Bristol Institute of Public Affairs
University of Bristol
2 Priory Road
Bristol BS8 1TX
http://www.bristol.ac.uk/cmpo/

Tel: (0117) 33 10799
Fax: (0117) 33 10705
E-mail: cmpo-office@bristol.ac.uk

ISSN 1473-625X

University of BRISTOL

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

The Leverhulme Trust

CMPO Working Paper Series No. 09/216

More Reliable Inference for Segregation Indices

Rebecca Allen[1], Simon Burgess[2]
and
Frank Windmeijer[2, 3]

[1] Institute of Education, University of London
[2] CMPO, University of Bristol
[3] CSE, Cemmap, IFS London

April 2009

Abstract

The most widely used measure of segregation is the dissimilarity index, *D*. It is now well understood that this measure also reflects randomness in the allocation of individuals to units; that is, it measures deviations from evenness not deviations from randomness. This leads to potentially large values of the segregation index when unit sizes and/or minority proportions are small, even if there is no underlying systematic segregation. Our response to this is to produce an adjustment to the index, based on an underlying statistical model. We specify the assignment problem in a very general way, with differences in conditional assignment probabilities underlying the resulting segregation. From this we derive a likelihood ratio test for the presence of any systematic segregation and a bootstrap bias adjustment to the dissimilarity index. We further develop the asymptotic distribution theory for testing hypotheses concerning the magnitude of the segregation index and show that use of bootstrap methods can improve the size and power properties of test procedures considerably. We illustrate these methods by comparing dissimilarity indices across school districts in England to measure social segregation.

Keywords: segregation, dissimilarity index, bootstrap methods, hypothesis testing

JEL Classification: C12, C13, C15, C46, I21

Electronic version: www.bristol.ac.uk/cmpo/publications/papers/2009/wp216.pdf

Address for Correspondence
CMPO, Bristol Institute of Public Affairs
University of Bristol
2 Priory Road
Bristol
BS8 1TX
Simon.burgess@bristol.ac.uk
www.bristol.ac.uk/cmpo/

CMPO is jointly funded by the Leverhulme Trust and the ESRC

1. Introduction

Segregation remains a major topic of research in a number of contexts such as neighbourhoods, workplaces and schools. Researchers study segregation by poverty status, by gender and by ethnicity among other characteristics. Almost always, these studies are comparative in some way: for example, arguing that ethnic segregation in neighbourhoods is higher in one city than another, or that gender segregation by occupation has changed over time. There is often also an implicit or explicit causal model in mind, and the difference in segregation is associated with some behavioural process. However, the inferential framework for segregation indices is under-developed, limiting the progress that can be made. This paper proposes an approach to strengthen this framework.

It is central to our approach to think of segregation as the outcome of a process of assignment. This includes the assignment of people to neighbourhoods, workers to jobs, or pupils to schools. In general, this allocation is likely to be the result of the inter-locking decisions of different agents rather than a dictator model. This perspective offers a number of advantages. First, it ties the outcome to a set of processes that can be analysed and estimated. Second, it makes it clear that the observed outcome is one of a set of possible outcomes, and so naturally leads on to a framework for statistical inference. Third, the connection with the underlying processes makes explicit that it is this systematic or behaviour-based segregation that is the object of interest in terms of analysing the causes of segregation.

There is a large literature concerning the measurement of segregation, with a number of indices in use, all with differing properties. The most widely used measure of segregation is the dissimilarity index, D, defined below (Duncan and Duncan, 1955). It is now widely understood this measure also reflects randomness in the allocation of individuals to units; that is, it measures deviations from evenness not deviations from randomness. Furthermore, the impact of randomness on D depends on the nature of the context (made precise below). This makes difficult one of the prime tasks of the measurement of segregation – to make statements on true differences in segregation between cities, school districts, industries or time periods. For example, the overall proportion of the minority group influences this because a very small minority group is more likely to be unevenly distributed across units by chance, compared to a larger minority group. This problem is particularly acute with small unit sizes. This is easy to see in the following example. Consider a large population, half male and half female. Suppose they are assigned to work in two very large firms. A random assignment process would produce an outcome close to a 50:50 gender split in each firm and an estimated D of about zero. However, if they were allocated to many firms of size 2, then a random assignment procedure would lead to many all-female firms, many all-male firms and many mixed firms and a high value for *D*. The high value reflects a strong deviation from evenness despite pure randomness. Others have noted the problem of small unit size in the measurement of segregation, see e.g. Carrington and Troske (1997). They proposed an adjustment to segregation indices that has since been used by researchers measuring workplace segregation where small units are particularly likely (e.g. Hellerstein and Neumark, 2008) and school segregation (e.g. Söderström and Uusitalo, 2005).

Comparing segregation across areas or time, small unit bias should be of concern to researchers for two reasons. First, the size of the bias will differ across comparison areas, potentially leading to an incorrect ranking of levels of segregation across areas. Second, the presence of small unit bias makes a correlation between measured segregation index values and a potentially causal variable, say X, difficult to interpret. It will impact on the estimated effect of X on measured segregation, even if the parameters of the problem (unit size, minority fraction and population) do not vary across areas. More challengingly, it is likely that the bias as a function of these parameters will be correlated with X, making the true relationship between X and D difficult to identify.

In this paper we propose an inferential framework for the canonical segregation measure, *D*, based on an underlying statistical model. This setup is related to, but different from, that used by Ransom (2000). He derives (asymptotic) inference procedures for D by specifying the sampling variation of a multinomial distribution. We specify the assignment problem in a very general way, and set out the difference in assignment probabilities that underlies the resulting segregation; this is Section 2. From this we derive a likelihood test for the presence of any systematic segregation and a bootstrap bias adjustment to the standard *D* in sections 3 and 4. Following Ransom (2000), we further develop the asymptotic distribution theory for testing hypotheses concerning the magnitude of the segregation index and show that use of bootstrap methods can improve the size and power properties of test procedures considerably; this is in section 5. In section 6 we illustrate the methods in an example of social segregation in schools in England. Section 7 concludes.

2. Statistical Framework

Underlying an assignment of individuals to units is an allocation process. This might be purely random, or it may be influenced by the actions of agents, including those whose allocation we are studying, as well as others. This systematic allocation process will in general reflect the preferences and constraints of both the individual (such as preferences for racial composition of neighbourhood or ability to pay for houses in a particular neighbourhood) and of the unit to match with particular individuals (such as a firm's desire for highly educated workers or school admissions procedures that favour children of particular religious denomination). Typically the research question is about characterising segregation arising from this behaviour.

Our notation is as follows. There are units $j = 1,...,J$ nested within an area. Individuals $i = 1,...,n$ either have, or do not have, a characteristic measurable on a dichotomous scale, $c = \{0,1\}$. This could be black ethnicity, female or poverty status. The number of individuals in the area with status $c = 1$ is denoted n^1, and n^0 denotes the number of individuals with status $c = 0$. Individuals are assigned to units and we observe the resulting allocations, n_j^1 individuals in unit *j* having status *c* = 1 and n_j^0 individuals in unit *j* having status *c* = 0. The total number of individuals in unit *j* is $n_j = n_j^1 + n_j^0$.

There are many indices used to measure segregation (see Duncan and Duncan, 1955, Massey and Denton, 1988, and White, 1986 for an overview). The formula for each provides an implicit definition of segregation. Massey and Denton (1988) characterise segregation along five dimensions: evenness (dissimilarity), exposure (isolation), concentration (the amount of physical space occupied by the minority group), clustering (the extent to which minority neighbourhoods abut one another), and centralisation (proximity to the centre of the city). Throughout this paper we use the index of dissimilarity (denoted *D*), the most popular unevenness index in the literature. However, our analysis can be extended to other unevenness segregation indices.

The formula for the index of dissimilarity D in the area, which is bounded by 0 (no segregation) and 1, is given by (see Duncan and Duncan, 1955)[1]:

$$D = \frac{1}{2}\sum_{j=1}^{J}\left|\frac{n_j^1}{n^1} - \frac{n_j^0}{n^0}\right|.$$

The basis for an allocation procedure is a set of conditional probabilities that assigns an individual i to unit j, given the individual's status c:

$$p_j^a \equiv P\left(unit = j \mid c = a\right), \quad j = 1,...,J; a = 0,1.$$

We define systematic segregation as being present when

$$\exists\, j: \quad p_j^1 \neq p_j^0.$$

We can see the relationship between D and the conditional probabilities of the underlying allocation process by noting that the fraction n_j^1 / n^1 and n_j^0 / n^0 are estimates of these conditional probabilities:

$$\hat{p}_j^0 = \frac{n_j^0}{n^0};\ \hat{p}_j^1 = \frac{n_j^1}{n^1},$$

and therefore the index of dissimilarity is equal to

$$D = \frac{1}{2}\sum_{j=1}^{J}\left|\hat{p}_j^1 - \hat{p}_j^0\right|.$$

Formalising the allocation process, an area population of n individuals with a given proportion $p = n^1 / n$ with status $c = 1$, is allocated to J units according to the population conditional probability rules. Each individual is allocated independently, for $c = 1$ individuals according to the probabilities $p_j^1,\ j = 1,...,J$, and for $c = 0$ individuals according to the probabilities $p_j^0,\ j = 1,...,J$. The outcomes of this process are the allocations n_j^1 and n_j^0. Clearly, unit sizes are not fixed in this setup as

[1] D measures the share of either group that must be removed, without replacement, to achieve zero segregation (Cortese et al., 1976; Massey and Denton, 1988). It can be shown to be equal to the maximum distance between the line of equality and a segregation curve that sorts units by p_j, then plots the cumulative share of $c = 1$ individuals against the cumulative share of $c = 0$ individuals (Duncan and Duncan, 1955).

they are equal to $n_j = n_j^1 + n_j^0$ and therefore determined by the stochastic allocation. The expected unit sizes are given by

$$E\left(n_j\right) = n^1 p_j^1 + n^0 p_j^0 .$$

We can now interpret the index of dissimilarity as an estimator for the population quantity

$$D_{pop} = \frac{1}{2}\sum_{j=1}^{J}\left|p_j^1 - p_j^0\right| .$$

It is clear that $D_{pop} = 0$ if $p_j^1 = p_j^0$ for all $j = 1,..., J$.

From the allocation process described above, we can estimate the conditional probabilities by maximum likelihood. As the allocations are two independent multinomial distributions the log-likelihood function, given the observed allocations is given by

$$\log L = \log\left(\frac{n^1!}{n_1^1!...n_J^1!}\right) + \log\left(\frac{n^0!}{n_1^0!...n_J^0!}\right) + \sum\nolimits_{j=1}^{J} n_j^1 \log\left(p_j^1\right) + \sum\nolimits_{j=1}^{J} n_j^0 \log\left(p_j^0\right),$$

Clearly, the maximum likelihood estimates are given by $\hat{p}_j^1 = \frac{n_j^1}{n^1}$ and $\hat{p}_j^0 = \frac{n_j^0}{n^0}$, $j = 1,.., J$, i.e. exactly the same as the estimates entering *D*.

Ransom (2000) proposed the use of the following statistical model for a random sample of size n:

$$P\left(n_1^0, n_2^0, ..., n_J^0, n_1^1, n_2^1, ..., n_J^1; \pi_{jc}\right) = n!\prod_{j=1}^{J}\prod_{c=0}^{1}\frac{\left(\pi_{jc}\right)^{n_j^c}}{n_j^c!}$$

where π_{jc} is the joint probability of observing an individual with status c and in unit j in the sample, i.e. $\pi_{ja} = P\left(unit = j, c = a\right)$. Mora and Ruiz-Castillo (2007), and references therein, consider a similar setup for an information index of multi-group segregation. Ramsom (2000, p. 458) notes that this model is not appropriate when the population is observed as then the π_{jc} are known. The parameters π_{jc} are not those that enter the segregation index D_{pop}, which are the conditional probabilities $p_j^c = P\left(unit = j \mid c\right) = \pi_{jc} / \sum\nolimits_{s=1}^{J}\pi_{sc}$.

Our model is applicable even when we observe the complete, finite population, but randomness is achieved by the random allocation process to units. Our statistical model is for a finite population of size $n = n^0 + n^1$, with parameters p_j^c, $j = 1,...,J$, $c = 0,1$, and is given by

$$P\left(n_1^0, n_2^0, ..., n_J^0, n_1^1, n_2^1, ..., n_J^1; n^0, n^1, p_j^c\right) = \prod_{j=1}^{J}\prod_{c=0}^{1} n^c!\frac{\left(p_j^c\right)^{n_j^c}}{n_j^c!}.$$

In the remainder of the paper we will focus on this particular model. A different model applies where unit sizes n_j are assumed fixed, in addition to our assumptions that the population size *n* and minority fraction *p* are fixed. In this case, the allocation mechanism is determined by the conditional probabilities $P(c = a \mid unit = j)$. As

$$P(unit = j \mid c = a) = P(unit = j)P(c = a \mid unit = j)/P(c = a)$$

D_{pop} can equivalently be written as

$$\begin{aligned} D_{pop} &= \frac{1}{2}\sum_{j=1}^{J} P(unit = j)\left|\frac{P(c=1 \mid unit = j)}{P(c=1)} - \frac{1 - P(c=1 \mid unit = j)}{1 - P(c=1)}\right| \\ &= \frac{1}{2}\sum_{j=1}^{J}\frac{n_j}{n}\left|\frac{P(c=1 \mid unit = j)}{P(c=1)} - \frac{1 - P(c=1 \mid unit = j)}{1 - P(c=1)}\right| \end{aligned}$$

Finally, if instead of the full population we obtain a random sample from the population, D will still be an estimator of D_{pop}, in both cases of random or fixed unit sizes.

2.1 Bias

As D is an estimator for D_{pop}, we define the bias of D as

$$bias = E(D) - D_{pop},$$

where the expectation is taken over the independent multinomial distributions with probabilities p_j^c, $j = 1,...,J; c = 0,1$ for given population size n and minority proportion p:

$$E(D) = \frac{1}{2}\sum\nolimits_{\left\{n_1^0,...,n_J^0\right\}}\sum\nolimits_{\left\{n_1^1,...,n_J^1\right\}}\left(\left(\sum_{j=1}^{J}\left|\frac{n_j^1}{n^1} - \frac{n_j^0}{n^0}\right|\right)\prod_{j=1}^{J}\prod_{c=0}^{1} n^c!\frac{\left(p_j^c\right)^{n_j^c}}{n_j^c!}\right)$$

The value of $E(D)$ is a function of the underlying conditional probabilities, summarised by D_{pop}, and of unevenness generated by the randomness of the allocation process. As has been well documented in the literature (see e.g. Carrington and Troske (1997)), D can be severely upward biased when unit sizes are small and allocation is 'random', meaning that there is no systematic segregation, $p_j^1 = p_j^0$ for all j, and hence $D_{pop} = 0$. For small number of units J and small unit sizes, we can calculate the expected value of D analytically. The figure below graphs the bias $E(D) - D_{pop}$ for $J = 4$, $n = \{20, 40, 60\}$, $p = 0.1$ and for various values of D_{pop}. These values of D_{pop} are obtained by setting the p_j^c according to a scheme discussed in Section 5 below. The expected unit sizes are the same for the 4 units, i.e. 5 when $n = 20$, 10 when $n = 40$ and 15 when $n = 60$.

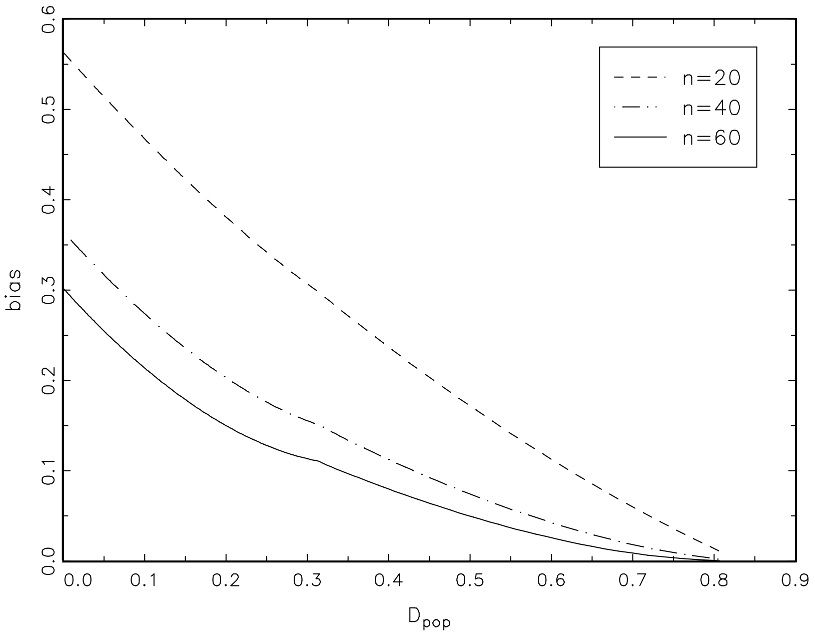

Figure 1. Bias $E(D) - D_{pop}$, $J = 4$, $p = 0.1$, equal expected unit sizes

The small-unit bias is apparent in the figure. When expected unit sizes are equal to 5, $E(D)$ is equal to 0.56 when $D_{pop} = 0$. The graph also shows that the bias is a decreasing function of increasing systematic segregation (D_{pop}) and a decreasing function of expected unit size.

3. Bootstrap Bias Correction

The purpose of our adjustment to D is to reduce the upward bias on the estimate of D_{pop}, as highlighted in Figure 1. Our proposal is to use a bootstrap type bias correction, as described in e.g. Hall (1992) and Davison and Hinkley (1997). Given an observed allocation, a new sample is generated with the same sample size n and minority proportion p, but using the observed conditional probabilities $\hat{p}_j^1 = n_j^1 / n^1$ and $\hat{p}_j^0 = n_j^0 / n^0$ for the allocation process. The value for D in this bootstrap sample is denoted D_b. Repeating this B times, we can calculate

$$\bar{D}_b = \frac{1}{B}\sum_{b=1}^{B} D_b \,.$$

The population value of the segregation measure in the bootstrap sample is D itself, and so a measure of the bias of D is given by $\bar{D}_b - D$. A bootstrap bias corrected estimate of D_{pop} is then obtained as

$$D_{bc} = D - \left(\bar{D}_b - D\right) = 2D - \bar{D}_b \,.$$

This type of bias correction works well if the bias is constant for different values of D_{pop}. This is clearly not the case here, as the biases as displayed in Figure 1 are much larger for smaller values of D. This bias correction is therefore not expected to work well for small unit sizes combined with small values of D_{pop}. We show in the next sections that this bootstrap procedure reduces enough of the bias to make inferences about levels of segregation, provided unit sizes are not too small. Where unit sizes are very small, we show in section 4 that the observed level of segregation can rarely statistically be distinguished from evenness. Thus, we suggest that in these cases the data is inappropriate for making inferences about segregation.

3.1 Monte Carlo Simulations

This section evaluates the performance of the bootstrap bias adjustment for estimating levels of segregation. To do this we follow Duncan and Duncan's (1955) approach of generating a level of unevenness between no segregation and complete segregation using a single parameter, $0 \le q < 1$. This parameter maps a set of parabolic segregation curves via the formula:[2]

[2] Although this set of segregation curves cannot represent all distributions of segregation, it is a sufficient set to examine different levels of systematic segregation for the purposes of this paper.

$$P\left(unit \le j \mid c=1\right)=\frac{\left(1-q\right)P\left(unit \le j \mid c=0\right)}{1-q\cdot P\left(unit \le j \mid c=0\right)}$$

This formula, combined with the constraint of equal expected unit sizes, fixes the conditional allocation probabilities for both groups. An allocation is then generated by assigning n^1 and n^0 individuals to the J units using these calculated conditional probabilities.

For each D, $\bar{D}_b$ is calculated from 100 bootstrap samples. This process is repeated 1,000 times for each n, p and D_{pop} combinations over the following parameter space:

- Number of units, *J*, is fixed at 50;
- Unit sizes n_j are equal in expectation, with expected unit size varying from 6 to 200;
- Proportion of $c=1$ individuals, *p*, varies from 0.01 to 0.5;
- Systematic segregation generator, *q*, varies from 0 to 0.99.

The biases of D and D_{bc} are presented in Table 1. It shows that where the minority proportion is very small tiny (e.g. $p=0.05$), unit sizes are small (e.g. $E\left(n_j\right)=10$) and systematic segregation is very low (e.g. $D_{pop}=0.056$), observed segregation incorrectly suggests a highly segregating process underlies the allocation, $D=0.55+0.056=0.606$, and the bootstrap correction does little to correct this bias $D_{bc}=0.43+0.056=0.486$. At the other extreme, where the minority proportion is large (e.g. $p=0.3$), unit sizes are large (e.g. $n=200$) and systematic segregation is high (e.g. $D_{pop}=0.818$), no correction is needed because the expected value of observed segregation is not different from D_{pop}. However, in much social science data, the phenomenon of interest tends to have moderate (D_{pop} around 0.1 to 0.4) rather than very high levels of segregation. In this range, the proposed bootstrap correction tends to work well and is necessary, provided that p and $E\left(n_j\right)$ are not both simultaneously very small. For example, when the minority proportion is 10%

Table 1: Bias of D and D_{bc} for $J = 50$ and combinations of p, $E\left(n_j\right)$ and D_{pop}.

		D_{pop}							
p	$E\left(n_j\right)$	0		0.056		0.127		0.225	
		$D-D_{pop}$	$D_{bc}-D_{pop}$	$D-D_{pop}$	$D_{bc}-D_{pop}$	$D-D_{pop}$	$D_{bc}-D_{pop}$	$D-D_{pop}$	$D_{bc}-D_{pop}$
0.01	**6**	0.94	0.92	0.89	0.87	0.81	0.80	0.72	0.70
	10	0.90	0.87	0.85	0.82	0.78	0.75	0.68	0.65
	20	0.82	0.76	0.76	0.70	0.69	0.63	0.60	0.54
	30	0.74	0.65	0.68	0.60	0.61	0.53	0.52	0.44
	40	0.67	0.56	0.61	0.51	0.55	0.44	0.46	0.36
	50	0.60	0.48	0.55	0.43	0.48	0.37	0.40	0.29
	100	0.38	0.22	0.33	0.17	0.27	0.12	0.20	0.063
	200	0.28	0.16	0.22	0.11	0.17	0.064	0.12	0.025
0.05	**6**	0.74	0.65	0.68	0.60	0.61	0.53	0.52	0.45
	10	0.60	0.48	0.55	0.43	0.48	0.36	0.40	0.29
	20	0.30	0.24	0.35	0.19	0.29	0.14	0.22	0.079
	30	0.33	0.21	0.28	0.16	0.22	0.10	0.16	0.056
	40	0.29	0.17	0.24	0.13	0.18	0.073	0.13	0.032
	50	0.26	0.16	0.21	0.11	0.16	0.061	0.10	0.024
	100	0.18	0.11	0.13	0.062	0.089	0.026	0.054	0.005
	200	0.13	0.074	0.082	0.033	0.048	0.007	0.027	-0.000
0.10	**6**	0.55	0.41	0.49	0.36	0.43	0.30	0.35	0.23
	10	0.41	0.26	0.36	0.21	0.30	0.15	0.23	0.094
	20	0.29	0.18	0.24	0.13	0.19	0.079	0.13	0.037
	30	0.24	0.14	0.19	0.095	0.14	0.050	0.091	0.016
	40	0.21	0.12	0.16	0.077	0.11	0.037	0.071	0.010
	50	0.19	0.11	0.14	0.065	0.093	0.027	0.057	0.006
	100	0.13	0.078	0.085	0.034	0.051	0.008	0.029	0.000
	200	0.093	0.055	0.051	0.016	0.027	0.002	0.015	0.000
0.30	**6**	0.35	0.22	0.30	0.17	0.24	0.12	0.18	0.066
	10	0.27	0.16	0.22	0.12	0.17	0.068	0.12	0.028
	20	0.19	0.11	0.14	0.067	0.10	0.03	0.061	0.006
	30	0.16	0.092	0.11	0.048	0.070	0.016	0.041	0.002
	40	0.14	0.08	0.089	0.036	0.053	0.009	0.030	0.000
	50	0.12	0.071	0.076	0.029	0.044	0.006	0.024	-0.001
	100	0.086	0.051	0.045	0.014	0.023	0.001	0.013	0.000
	200	0.061	0.036	0.025	0.005	0.011	0.000	0.006	0.000
0.50	**6**	0.32	0.19	0.26	0.14	0.21	0.091	0.15	0.048
	10	0.25	0.15	0.20	0.098	0.15	0.055	0.098	0.020
	20	0.18	0.10	0.13	0.058	0.086	0.024	0.051	0.004
	30	0.14	0.083	0.097	0.041	0.059	0.012	0.034	0.001
	40	0.12	0.072	0.079	0.030	0.046	0.007	0.025	0.000
	50	0.11	0.07	0.07	0.024	0.037	0.004	0.020	-0.001
	100	0.079	0.046	0.039	0.011	0.019	0.001	0.011	0.001
	200	0.056	0.033	0.021	0.003	0.009	-0.000	0.005	-0.000

Notes: Mean bias reported for 1000 replications. Number of bootstrap replications 100.

Table 1 continued

		D_{pop}							
p	$E(n_j)$	**0.292**		**0.382**		**0.634**		**0.818**	
		$D-D_{pop}$	$D_{bc}-D_{pop}$	$D-D_{pop}$	$D_{bc}-D_{pop}$	$D-D_{pop}$	$D_{bc}-D_{pop}$	$D-D_{pop}$	$D_{bc}-D_{pop}$
0.01	**6**	0.65	0.63	0.56	0.54	0.31	0.30	0.15	0.14
	10	0.61	0.58	0.53	0.50	0.29	0.27	0.14	0.13
	20	0.53	0.48	0.45	0.40	0.24	0.21	0.11	0.10
	30	0.46	0.39	0.39	0.32	0.20	0.16	0.095	0.076
	40	0.40	0.31	0.33	0.25	0.18	0.13	0.081	0.059
	50	0.35	0.24	0.28	0.19	0.15	0.097	0.069	0.045
	100	0.16	0.035	0.13	0.014	0.058	-0.007	0.025	-0.006
	200	0.094	0.011	0.069	-0.000	0.029	-0.007	0.011	-0.006
0.05	**6**	0.47	0.39	0.39	0.32	0.21	0.17	0.099	0.079
	10	0.34	0.24	0.29	0.19	0.15	0.096	0.070	0.045
	20	0.18	0.056	0.14	0.030	0.066	0.003	0.030	0.000
	30	0.13	0.032	0.10	0.014	0.043	-0.004	0.019	-0.004
	40	0.10	0.018	0.075	0.005	0.032	-0.005	0.013	-0.004
	50	0.083	0.013	0.061	0.004	0.026	-0.003	0.012	-0.001
	100	0.040	-0.000	0.029	-0.002	0.012	-0.002	0.005	-0.001
	200	0.020	-0.001	0.014	-0.001	0.0057	-0.001	0.002	-0.000
0.10	**6**	0.31	0.19	0.25	0.15	0.13	0.069	0.061	0.031
	10	0.19	0.065	0.15	0.040	0.069	0.006	0.031	0.000
	20	0.11	0.021	0.079	0.007	0.034	-0.003	0.015	-0.003
	30	0.070	0.006	0.051	0.000	0.022	-0.003	0.010	-0.001
	40	0.054	0.003	0.04	-0.001	0.016	-0.002	0.007	-0.001
	50	0.043	0.001	0.031	-0.002	0.013	-0.001	0.005	-0.001
	100	0.021	-0.001	0.015	-0.001	0.006	-0.000	0.003	-0.000
	200	0.011	0.000	0.008	0.000	0.003	-0.000	0.001	-0.000
0.30	**6**	0.14	0.043	0.11	0.022	0.050	0.001	0.023	-0.001
	10	0.091	0.014	0.067	0.003	0.028	-0.003	0.012	-0.002
	20	0.047	0.002	0.034	0.000	0.014	-0.001	0.006	-0.001
	30	0.031	0.000	0.022	-0.001	0.009	0.000	0.004	-0.001
	40	0.022	-0.002	0.015	-0.002	0.006	-0.002	0.002	-0.001
	50	0.018	-0.001	0.013	-0.001	0.005	-0.001	0.002	-0.001
	100	0.001	0.000	0.006	-0.000	0.003	0.000	0.002	-0.000
	200	0.005	0.000	0.003	0.000	0.001	-0.000	0.000	-0.000
0.50	**6**	0.12	0.028	0.092	0.011	0.041	-0.002	0.018	-0.002
	10	0.075	0.008	0.055	0.000	0.021	-0.005	0.009	-0.003
	20	0.038	0.000	0.028	0.000	0.011	-0.001	0.004	-0.001
	30	0.025	-0.001	0.019	-0.001	0.008	0.000	0.003	-0.001
	40	0.019	0.000	0.013	-0.001	0.005	-0.001	0.001	-0.001
	50	0.015	0.000	0.010	-0.001	0.004	-0.001	0.002	-0.000
	100	0.008	0.001	0.006	0.000	0.002	0.000	0.001	-0.000
	200	0.004	-0.000	0.003	-0.000	0.001	-0.000	-0.000	-0.000

Notes: Mean bias reported for 1000 replications. Number of bootstrap replications 100.

and unit sizes are expected to be 30, if underlying segregation is 0.225, the observed index of segregation would be upward biased by 0.091 whereas the bootstrap correction would successfully reduce this bias to just 0.016.

Figure 2 illustrates the pattern of results for an expected unit size of 30. For a reasonably large minority proportion of 20% and above, D_{bc} succeeds in removing most of the bias of D, provided underlying segregation levels are not very low. Once the size of the minority proportion falls below 7.5%, the bias correction is poor, except where D_{pop} is high.

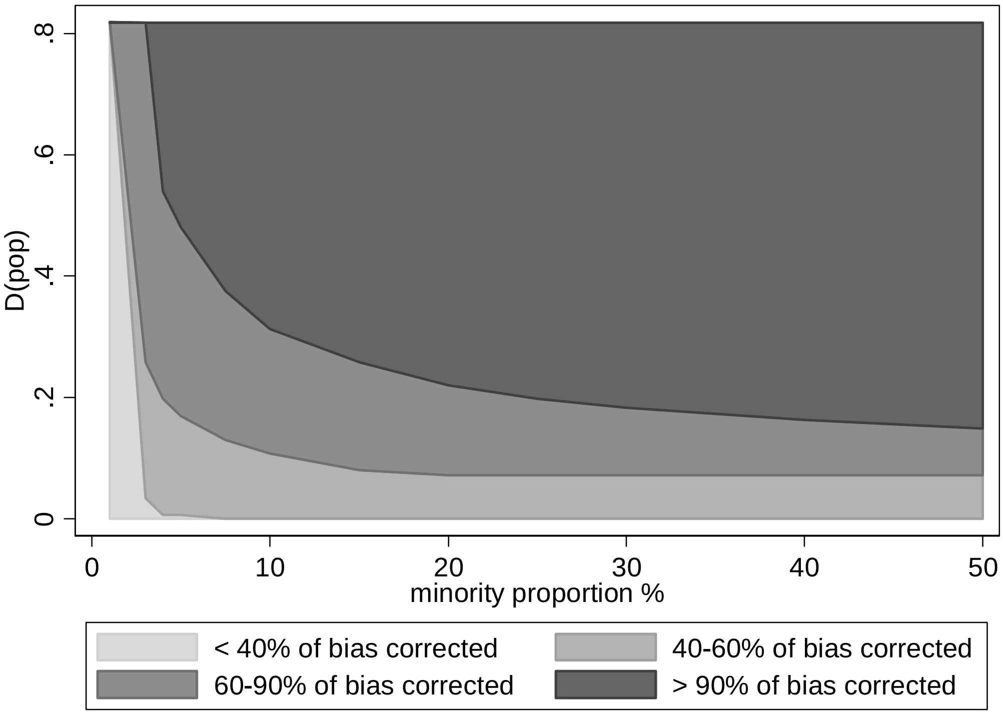

Figure 2. Performance of bootstrap bias correction for $E\left(n_j\right) = 30$

4. Tests of no systematic segregation

To complement this bootstrap bias correction, we provide a test for no systematic segregation. We consider two alternative methods to test whether we can reject the hypothesis that the level of segregation observed was generated by randomness alone, $D_{pop} = 0$. It is common in the literature to run a randomisation procedure to generate the distribution of D under the null of no systematic segregation (see e.g. Boisso et al., 1994), and D is compared to this distribution. Here, we generate the distribution of D under the null of no systematic segregation by creating B samples generated using the restricted conditional probabilities $\hat{p}_j^0 = \hat{p}_j^1 = \hat{p}_j = \left(n_j^0 + n_j^1\right)/n$ and calculating D in each sample, which we denote D^*. The null hypothesis $H_0 : D_{pop} = 0$ is then rejected at level α if $\frac{1}{B}\sum_{b=1}^{B} 1\left(D_b^* > D\right) < \alpha$, where $1(.)$ is the indicator function

Alternatively, following the statistical model developed in Section 2, we can employ a likelihood ratio test for the hypothesis

$$H_0 : p_j^0 = p_j^1 = p_j \ \forall\, j,$$

which is given by

$$LR = -2\left(\sum\nolimits_{j=1}^{J} n_j \log\left(\hat{p}_j\right) - \sum\nolimits_{j=1}^{J} n_j^0 \log\left(\hat{p}_j^0\right) - \sum\nolimits_{j=1}^{J} n_j^1 \log\left(\hat{p}_j^1\right)\right),$$

and which follows an asymptotic χ_{J-1}^2 distribution. This asymptotic distribution is for large n and fixed J, and therefore for large unit sizes. For large J and/or small unit sizes, the asymptotic approximation can be expected to be poor, as we originally found in our simulation results discussed below. We therefore also utilise a bootstrap procedure to improve the size properties of the test. Let LR^* be the value of the likelihood ratio test in a sample generated from $\hat{p}_j^0 = \hat{p}_j^1 = \hat{p}_j = \left(n_j^0 + n_j^1\right)/n$. Then the null hypothesis of no systematic segregation is rejected at level α if $\frac{1}{B}\sum_{b=1}^{B} 1\left(LR_b^* > LR\right) < \alpha$.

Table 2 presents the test results for $J = 50$ and $E\left(n_j\right) = 30$, for various values of D_{pop} and minority proportions p. The size and power properties of the two tests are virtually identical. They have good size properties for all minority proportions p. The tests fail to reject the null for small values of D_{pop} combined with small minority proportions p, exactly the circumstances where the bootstrap bias correction does not remove much of the bias of D, as indicated in Figure 2. Clearly, any calculation of

D and D_{bc} should be accompanied by the D^* and/or bootstrapped LR tests. If these tests fail to reject, no further inference should be pursued.

Table 2. Rejection frequencies of D randomisation and Likelihood Ratio tests, $J = 50$, $E\left(n_j\right) = 30$, level $\alpha = 0.05$.

p	Test	D_{pop}							
		0	**0.056**	**0.127**	**0.225**	**0.292**	**0.382**	**0.634**	**0.818**
0.01	D^*	0.062	0.069	0.086	0.136	0.196	0.356	0.96	1.000
	LR	0.056	0.059	0.076	0.124	0.188	0.360	0.97	1.000
0.05	D^*	0.068	0.073	0.162	0.527	0.849	0.991	1.000	1.000
	LR	0.056	0.073	0.161	0.538	0.861	0.995	1.000	1.000
0.15	D^*	0.053	0.100	0.429	0.984	1.000	1.000	1.000	1.000
	LR	0.044	0.084	0.416	0.984	1.000	1.000	1.000	1.000
0.30	D^*	0.046	0.141	0.735	1.000	1.000	1.000	1.000	1.000
	LR	0.045	0.136	0.740	1.000	1.000	1.000	1.000	1.000
0.50	D^*	0.050	0.160	0.812	1.000	1.000	1.000	1.000	1.000
	LR	0.049	0.161	0.828	1.000	1.000	1.000	1.000	1.000

5. Inference

Having established that the bootstrap bias correction works well for a large part of the parameter space, the next step is to develop reliable inference procedures such as 95% confidence intervals and Wald test statistics for equivalence of segregation in different areas. We start by deriving the asymptotic distribution of D given our statistical framework, following the procedures as developed in Ransom (2000).

The estimated conditional probabilities $\hat{p}_j^c$, for $c = \{0,1\}$, are asymptotically normally distributed, as

$$\sqrt{n^c}\begin{pmatrix} \hat{p}_1^c - p_1^c \\ \hat{p}_2^c - p_2^c \\ \vdots \\ \hat{p}_J^c - p_J^c \end{pmatrix} \xrightarrow{d} N\left(0, \begin{bmatrix} p_1^c\left(1-p_1^c\right) & -p_1^c p_2^c & \cdots & -p_1^c p_J^c \\ -p_1^c p_2^c & p_2^c\left(1-p_2^c\right) & & -p_2^c p_J^c \\ \vdots & & \ddots & \vdots \\ -p_1^c p_J^c & -p_2^c p_J^c & \cdots & p_J^c\left(1-p_J^c\right) \end{bmatrix}\right) \equiv N\left(0, \Omega^c\right).$$

As $n^1 = pn$ and $n^0 = (1-p)n$, the limiting distribution of D can then be obtained via the delta method:

$$\sqrt{n}\left(D - D_{pop}\right) \xrightarrow{d} N\left(0, \lambda'\left(p^{-1}\Omega^1 + (1-p)^{-1}\Omega^0\right)\lambda\right)$$

where λ is a J-vector with rth element $\lambda_r = sign\left(p_r^1 - p_r^0\right)/2$, where $sign(q) = 1$ if $q > 0$ and $sign(q) = -1$ if $q < 0$.[3] This follows from

$$\frac{\partial D_{pop}}{\partial p_r^1} = \frac{\partial}{\partial p_r^1}\frac{1}{2}\sum_{j=1}^{J}\left|p_j^1 - p_j^0\right| = sign\left(p_r^1 - p_r^0\right)/2;$$
$$\frac{\partial D_{pop}}{\partial p_r^0} = \frac{\partial}{\partial p_r^0}\frac{1}{2}\sum_{j=1}^{J}\left|p_j^1 - p_j^0\right| = -sign\left(p_r^1 - p_r^0\right)/2.$$

Clearly, this derivation is only valid when $p_r^1 \neq p_r^0$.

The asymptotic distribution of D is then given by

$$D \overset{a}{\sim} N\left(D_{pop}, n^{-1}\lambda'\left(p^{-1}\Omega^1 + (1-p)^{-1}\Omega^0\right)\lambda\right),$$

or, equivalently,

$$D \overset{a}{\sim} N\left(D_{pop}, \lambda'\left(\Omega^1/n^1 + \Omega^0/n^0\right)\lambda\right)$$

which can form the basis for constructing confidence intervals and Wald test statistics for hypotheses of the form $H_0 : D_{pop} = \delta$. Denoting $\hat{\lambda}$ and $\hat{\Omega}^c$ the estimated counterparts of λ and Ω^c substituting the observed fractions $\hat{p}_j^c$ for p_j^c, the Wald test is then computed as

$$W = \frac{(D-\delta)^2}{\hat{\lambda}'\left(\hat{\Omega}^1/n^1 + \hat{\Omega}^0/n^0\right)\hat{\lambda}}$$

and converges in distribution to a χ_1^2 distributed random variable under the null.

Clearly, we don’t expect this approximation to work well when δ, group sizes and/or minority proportions are small, if only due to the upward bias of D as established in the previous sections. However, the Wald test W is asymptotically pivotal in the sense

[3] Although Ω^c is singular because $\sum_j p_j^c = 1$, exactly the same results are obtained by redefining D as a function of $2(J-1)$ probabilities only.

that its limiting distribution is not a function of nuisance parameters. We can therefore use bootstrap p-values which may result in an improvement of the finite sample behaviour of the test (see Hall (1992) and Davison and Hinkley (1997)). Denoting the Wald statistic in the *b*-th bootstrap sample as W_b, calculated as

$$W_b = \frac{\left(D_b - D\right)^2}{\hat{\lambda}_b' \left(\hat{\Omega}_b^1 / n^1 + \hat{\Omega}_b^0 / n^0\right) \hat{\lambda}_b}$$

the bootstrap p-value is then given by $\frac{1}{B}\sum_{b=1}^{B} 1\left(W_b > W\right)$.

This bootstrap procedure is equivalent to a symmetric two-tailed test for the t-statistic. Let τ denote the t-test

$$\tau = \frac{D - \delta}{\sqrt{\hat{\lambda}' \left(\hat{\Omega}^1 / n^1 + \hat{\Omega}^0 / n^0\right) \hat{\lambda}}},$$

then a test that does not assume symmetry can be based on the equal-tail bootstrap p-value

$$2 \min\left(\frac{1}{B}\sum_{b=1}^{B} 1\left(\tau_b < \tau\right), \frac{1}{B}\sum_{b=1}^{B} 1\left(\tau_b > \tau\right)\right).$$

Alternatively, we can base the inference directly on the bootstrap bias corrected estimator of D_{pop}. In order to estimate the variance of the bias corrected estimator, we perform a double bootstrap procedure. For every bootstrap sample we generate another set of bootstrap samples, enabling us to generate a bootstrap estimate of the variance of D_{bc}. Denoting this estimate $V\hat{a}r_b\left(D_{bc}\right)$, the Wald test statistic is then calculated as

$$W_{bc} = \frac{\left(D_{bc} - \delta\right)^2}{V\hat{a}r_b\left(D_{bc}\right)}$$

and this is again compared to the χ_1^2 distribution.

Figure 3 shows p-value plots for testing the true hypothesis $H_0 : D_{pop} = 0.2922$, for $E\left(n_j\right) = 30$, $J = 50$ and $p = 0.3$. The Wald test based on the asymptotic normal distribution of D and using the χ_1^2 critical values is denoted *W*, whereas the Wald test using the bootstrap critical values is denoted W_{pb}. The test based on the equal tail bootstrap p-value for the t-test is denoted T_{pb}. The Wald statistic using the double bootstrap variance estimate for the bias corrected estimator is denoted W_{bc}. The results shown are for 10,000 Monte Carlo replications. Per replication 599 bootstrap samples are drawn for the calculation of D_{bc} and the bootstrap distribution of the Wald test. Per bootstrap sample we draw a further 50 double bootstrap samples for the calculation of $V\hat{a}r_b\left(D_{bc}\right)$.

The mean of D is equal to 0.323, whereas that of D_{bc} is equal to 0.292. There is therefore a 10% upward bias in D, but D_{bc} is unbiased. The standard deviation of D is equal to 0.023, that of D_{bc} is equal to 0.027, and their root mean squared errors are given by 0.039 and 0.027 respectively. As is clear from Figure 3, the asymptotic Wald test, W, using the χ_1^2 critical values does not have good size properties. It rejects the true null too often, for example at 5% nominal size, it rejects the null in 18.5% of the replications. In contrast, using the p-values from the bootstrap distribution of the Wald statistic improves the size behaviour considerably. At the 5% level, the rejection frequency is now reduced to 7.3%. Using the equal-tailed bootstrap p-values for the t-test also improves on the size performance of the asymptotic Wald statistic, but it performs less well than W_{pb}. However, the best size performance in this case is obtained by using W_{bc} with χ_1^2 critical values. At the 5% level, it only rejects 5.4% of the time.

Figure 4 shows the p-value plot for a similar design, but now for smaller expected group sizes $E\left(N_j\right)=20$ and a smaller minority proportion, $p=0.10$. The bias of D in this case is 0.106, or 36%, whereas that of D_{bc} is 0.020, or 6.5%. The standard deviation of D is equal to 0.037, that of D_{bc} is equal to 0.048, and their root mean squared errors are given by 0.111 and 0.051 respectively

The size distortions of all test statistics are now more severe. The asymptotic Wald test is severely size distorted, with a 68% rejection rate at the 5% level. The Wald and asymmetric t-test using the bootstrapped p-values behave best, with their size properties being very similar. At the 5% level, the rejection frequencies for these tests are 9.4% and 9.5% respectively. W_{bc} has only a slightly worse size performance than these two bootstrap tests, it rejects the true null 10.7% of the time at the 5% level.

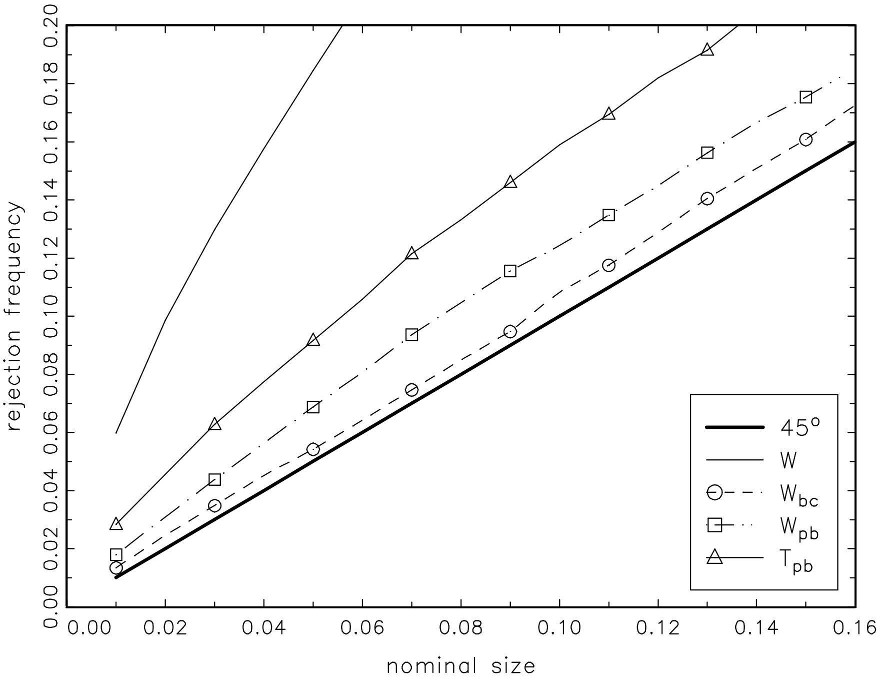

Figure 3. P-value plot, $H_0 : D_{pop} = 0.292$, $E\left(n_j\right) = 30$, $J = 50$, $p = 0.30$.

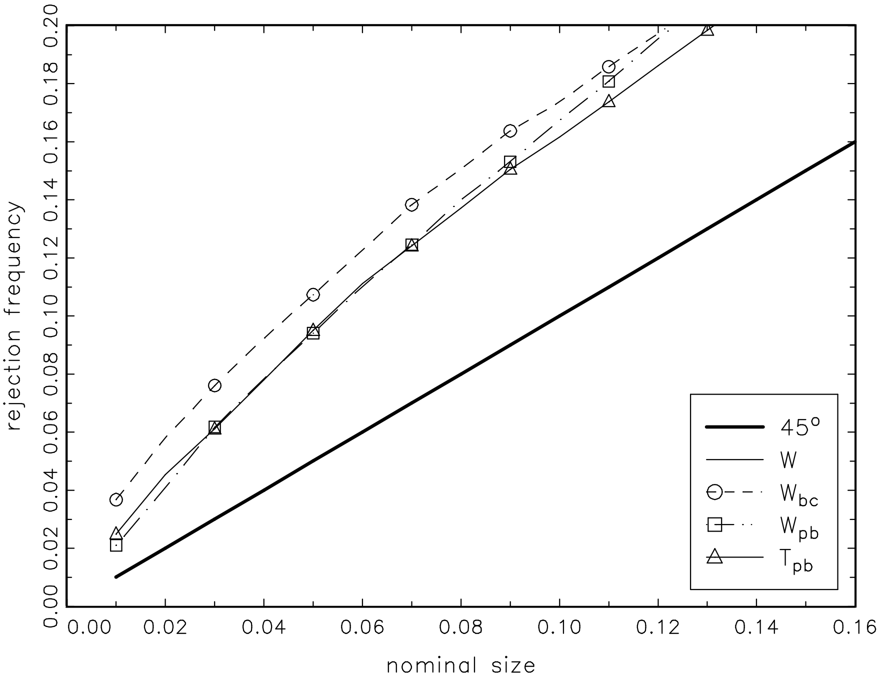

Figure 4. P-value plot, $H_0 : D_{pop} = 0.292$, $E\left(n_j\right) = 20$, $J = 50$, $p = 0.10$.

Clearly, in general, inference can only be based on these latter three tests when the sample size, D_{pop} and/or the minority proportion are small, although as the figures show, some size distortions occur also for these tests.

There is a one-to-one correspondence between the p-value plots as depicted in figures 3 and 4 and the coverage properties of the confidence intervals associated with the particular test statistics. Using the normal approximation, $(1-\alpha)\%$ confidence intervals associated with the asymptotic Wald and W_{bc} tests are constructed as

$$D - z_{(1-\alpha/2)}\sqrt{V\hat{a}r(D)} < D_{pop} < D + z_{(1-\alpha/2)}\sqrt{V\hat{a}r(D)}$$

and

$$D_{bc} - z_{(1-\alpha/2)}\sqrt{V\hat{a}r_b(D_{bc})} < D_{pop} < D_{bc} + z_{(1-\alpha/2)}\sqrt{V\hat{a}r_b(D_{bc})}$$

respectively, where $z_{(1-\alpha/2)}$ is the $100*(1-\alpha/2)$ percentile of the normal distribution. For the bootstrap Wald test the associated confidence interval is given by

$$D - \sqrt{w^*_{(1-\alpha)}V\hat{a}r(D)} < D_{pop} < D + \sqrt{w^*_{(1-\alpha)}V\hat{a}r(D)},$$

where $w^*_{(1-\alpha)}$ is the $100*(1-\alpha)$ percentile of the distribution of the bootstrap replications W_b. The equal-tailed bootstrap t-test has the corresponding confidence interval given by

$$D - \tau^*_{(1-\alpha/2)}\sqrt{V\hat{a}r(D)} < D_{pop} < D + \tau^*_{(\alpha/2)}\sqrt{V\hat{a}r(D)},$$

where $\tau^*_{(1-\alpha/2)}$ and $\tau^*_{(\alpha/2)}$ are the $100*(1-\alpha/2)$ and $100*(\alpha/2)$ percentiles of the distribution of the bootstrap replication τ_b.

For the example with $E(n_j) = 20$ and $p = 0.10$ as described above, the observed rejection frequencies of 68%, 10.7%, 9.4% and 9.5% for the W, W_{bc}, W_{pb} and T_{pb} tests respectively translate into coverage probabilities of 32%, 89.3%, 91.6% and 91.5% of the associated 95% confidence intervals. Given the upward bias of D this leads to an interesting observation concerning the confidence interval based on the bootstrap Wald test W_{pb}. As the size and associated coverage properties of this test are reasonably good, but as the confidence interval is symmetric around the upward biased D, this suggest that the W_{pb} based confidence interval will be quite large.

Table 3. Average lower and upper limits of 95% confidence intervals

Test	Lower limit	Upper limit
W_{bc}	0.227	0.395
W_{pb}	0.226	0.569
T_{pb}	0.209	0.376

Table 3 shows the averages of the lower and upper limits of the 95% confidence intervals based associated with W_{bc}, W_{pb} and T_{pb} respectively. This confirms that the W_{pb} based confidence interval is on average indeed much larger than those based on W_{bc} and T_{pb}. Whereas the lower limit is quite similar to those of the other two confidence intervals, its upper limit is much higher, as expected due to the symmetry around the upward biased D. Clearly, W_{pb} can therefore have poor power properties when D has substantial bias.

A researcher will in general be interested in determining whether segregation has changed significantly within an area over time, or whether segregation in one area is significantly different from that in another, similar or perhaps neighbouring area. We consider the performances of the test statistics for comparing the two hypothetical areas for which the results were simulated above. Area 1 has $J = 50$, $E\left(n_j\right) = 30$ and $p = 0.30$, whereas Area 2 has $J = 50$, $E\left(n_j\right) = 20$ and $p = 0.10$. To study the size properties of the tests for the null hypothesis

$$H_0 : D_{pop,1} = D_{pop,2}$$

we set the two area population segregation measures $D_{pop,1} = D_{pop,2} = 0.2922$ as before. Given the area specific conditional allocation probabilities, the allocations in the areas are determined independently and therefore the Wald test

$$W = \frac{\left(D_1 - D_2\right)^2}{V\hat{a}r\left(D_1\right) + V\hat{a}r\left(D_2\right)}$$

is asymptotically χ_1^2 distributed. The Wald test based on the bootstrap bias corrected estimates is defined as

$$W_{bc} = \frac{\left(D_{bc,1} - D_{bc,2}\right)^2}{V\hat{a}r_b\left(D_{bc,1}\right) + V\hat{a}r_b\left(D_{bc,2}\right)},$$

whereas the bootstrap p-values for the W_{pb} test are based on the distribution of the bootstrap replications of

$$W_b = \frac{\left(D_{b,1} - D_{b,2} - \left(D_1 - D_2\right)\right)^2}{V\hat{a}r\left(D_{b,1}\right) + V\hat{a}r\left(D_{b,2}\right)},$$

where $D_{b,1}$ and $D_{b,2}$ are calculated from independent bootstrap replications. The bootstrap p-values for the T_{pb} test are obtained in an equivalent way.

Figure 5 depicts the p-value plots for the true null of equal population segregation measures D_{pop} in the two areas. The asymptotic Wald test again over-rejects substantially, 27.3% at the 5% level. The W_{bc} test displays the best size properties in this case, followed by T_{pb} and then W_{pb}. The rejection probabilities for these tests at the 5% level are 8.8%, 9.2% and 10.9% respectively.

We next turn to evaluate the power properties of these tests when the two population segregation measures $D_{pop,1}$ and $D_{pop,2}$ are not equal. We keep $D_{pop,2}$ equal to 0.2922, but increase $D_{pop,1}$ to 0.3819. As discussed above, because D_2 is substantially biased upwards, we expect the W_{pb} test to have low power. This is confirmed by the p-value plots in Figure 6. The standard Wald test has power below nominal size, but especially the bootstrap based Wald test W_{pb} fails to reject the null of equal segregation completely. In contrast, both W_{bc} and T_{pb} show reasonable power properties, with T_{pb} having most power to detect this deviation from the null, although it has not been size adjusted. The p-value plots, not shown here, for the true null that $D_{pop,1} - D_{pop,2} = 0.0897$ are very similar to those in Figure 5. Clearly, these results combined together show that for simple hypothesis testing W_{bc} and T_{pb} are the test procedures with reasonably good size and power properties in the settings we considered.

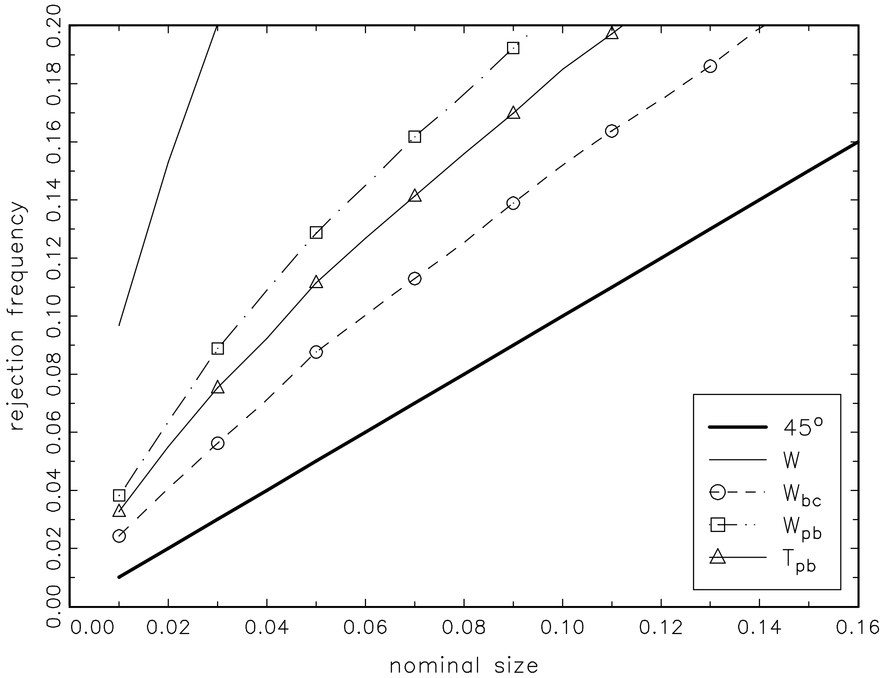

Figure 5. P-value plot, $H_0 : D_{pop,1} = D_{pop,2}$, size properties

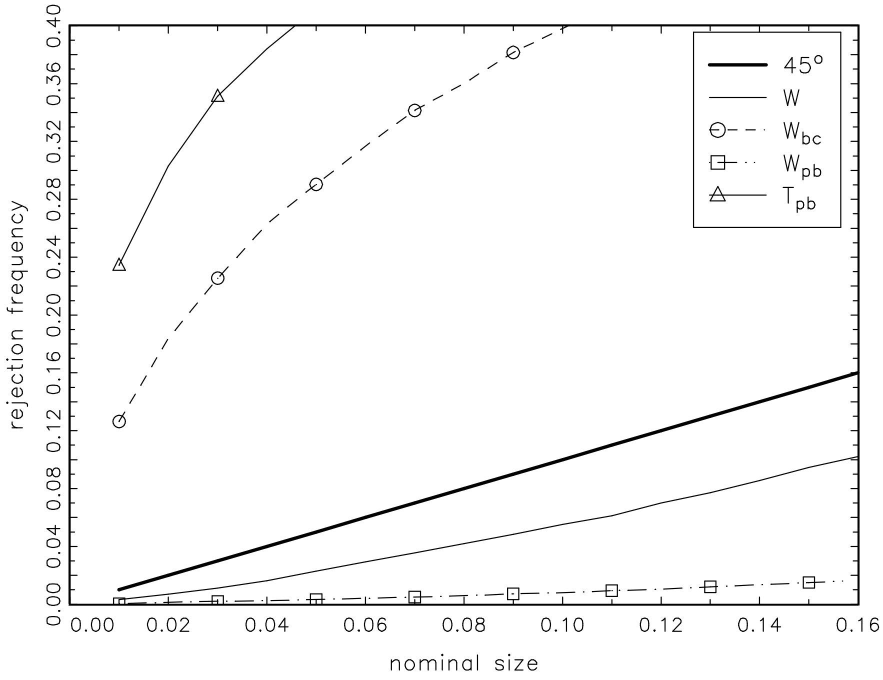

Figure 6. P-value plot, $H_0 : D_{pop,1} = D_{pop,2}$, power properties

6. Social segregation in schools

In this section we illustrate our inference procedures with an empirical application relating to social segregation in primary schools in England. The dichotomous measure is an indicator of poverty based on eligibility for free school meals (FSM). This context is useful as it naturally produces small unit sizes, and shows a range of minority proportions and overall populations across different Local Authorities (LAs). We use administrative data collected by the Department for Children, Families and Schools and made available to researchers as part of the National Pupil Database on pupils aged 10/11 in English primary schools in 2006. Measurement of school segregation in using this dataset has been carried out by many researchers (e.g. Allen and Vignoles, 2006; Burgess et al., 2006; Gibbons and Telhaj, 2006). Using the tools developed above, we can assess whether the small unit sizes and/or small minority populations lead to incorrect inferences about differences in segregation across areas.

We provide two cases. First, we compare two similar pairs of LAs, showing that quite small differences in their characteristics imply different outcomes of inference; these are North-East Lincolnshire and North Lincolnshire, and Blackburn and Oldham. Second, we compare all the different LAs in inner-city London, and consider which pair-wise comparisons yield significant differences. Table 4 shows the descriptive statistics and the dissimilarity indices of the LAs.

North-East Lincolnshire and North Lincolnshire have the same number of pupils, 2005 and 2011 respectively, but differ in the number of schools, 46 vs. 57 respectively, and therefore the average cohort size, and also differ in the percentages of children eligible for free-school meals, 21% vs. 13%. The dissimilarity index for North-East Lincolnshire is 0.43, higher than that of North Lincolnshire, which has an index of 0.36. Blackburn and Oldham differ rather more in size, but have closer average unit sizes, and slightly higher percentages of children eligible for free-school meals.

Table 4: Key parameters of primary schools across English local authorities

LA name	Number of pupils	Number of schools	Average cohort size	% FSM	*D*
North-East Lincolnshire	2005	46	44	21	0.43
North Lincolnshire	2011	57	35	13	0.36
Blackburn	2105	51	41	26	0.34
Oldham	2990	86	35	21	0.47
Camden	1394	41	34	42	0.23
Greenwich	2666	66	40	36	0.29
Hackney	2194	54	41	43	0.22
Hammersmith & Fulham	1177	39	30	45	0.30
Islington	1845	48	38	41	0.26
Kensinton & Chelsea	881	27	33	36	0.32
Lambeth	2428	60	40	40	0.24
Lewisham	2833	70	40	29	0.30
Southwark	2929	72	41	36	0.21
Tower Hamlets	2703	68	40	61	0.20
Wandsworth	2124	60	35	27	0.29
Westminster	1336	39	34	39	0.33

Are the school allocations in North-East Lincolnshire more segregated than those in North Lincolnshire? Table 5 shows that the observed D marginally overstates the level of segregation in each local authority, but the bootstrap correction to D does not alter the ranking. The table further presents the various test procedures and confidence intervals as described in the previous section. Here we generate 999 bootstrap samples with a further 100 samples for the double bootstrap variance estimate of D_{bc}. The LR test for no systematic segregation is clearly rejected for both LAs, with both bootstrap p-values equal to 0. The rejection of the null of equal segregation in North-East Lincolnshire and North Lincolnshire depends on the test statistics employed. Using the preferred test statistics W_{bc} and T_{pb} we reject the null of equal segregation in the two LAs at the 5% and 1% level respectively.

Table 6 shows the test statistics for Blackburn and Oldham. In this example, we can reject, with a high degree of confidence, the null of equal segregation in these areas. This greater confidence than in the Lincolnshire example is possible, despite similar segregation levels, because the local authorities are slightly larger and the minority proportion is higher.

Table 5. Bias corrected dissimilarity indices, confidence intervals and test statistics for North-East and North Lincolnshire

	North-East Lincolnshire	**North Lincolnshire**
D	0.433	0.364
D_{bc}	0.419	0.322
LR-test, boot. p-value	0	0
CI-W	[0.386-0.481]	[0.306-0.421]
CI-W_{bc}	[0.369-0.469]	[0.264-0.379]
CI-W_{pb}	[0.377-0.490]	[0.271-0.456]
CI-T_{pb}	[0.370-0.462]	[0.261-0.373]
	$H_0 : D_{pop,NEL} = D_{pop,NL}$, p-values	
W	0.067	
W_{bc}	0.011	
W_{pb}	0.121	
T_{pb}	0.004	

Table 6. Bias corrected dissimilarity indices, confidence intervals and test statistics for Blackburn and Oldham

	Blackburn	**Oldham**
D	0.342	0.472
D_{bc}	0.325	0.454
LR-test, boot. p-value	0	0
CI-W_{bc}	[0.282-0.368]	[0.418-0.490]
CI-T_{pb}	[0.287-0.360]	[0.420-0.486]
	$H_0 : D_{pop,Blackburn} = D_{pop,Oldham}$, p-values	
W_{bc}	0.000	
T_{pb}	0	

For our second illustration, Table 7 compares observed and bootstrap corrected segregation levels across the 12 local authorities in Inner London. The bootstrap correction makes little differences to the ranking of segregation levels, with just Wandsworth and Greenwich switching positions. The test statistics show that the LAs can be approximately divided into three groups, with possible multiple membership, where the tests do not reject the null of equal segregation. These groups are: Tower Hamlets, Southwark and Hackney, with the lowest level of segregation; Hackney, Camden and Lambeth, with medium level of segregation; and Wandsworth, Greenwich, Hammersmith & Fulham, Lewisham, Kensington & Chelsea and

Westminster with the highest level of segregation. Islington is a medium segregation LA with some overlap with the group of highest segregation LAs.

7. Conclusions

To make statements about the true underlying degree of segregation, or understand the processes causing segregation, it is desirable to measure the level of systematic segregation. However, where minority proportions and unit sizes are small, the level of segregation observed by researchers in their data is known to be significantly greater than systematic segregation. Furthermore, because the size of the bias of observed segregation over systematic segregation is known to be a function of minority proportion, unit sizes and systematic segregation, differences in any of these parameters between areas or over time may lead to incorrect inferences.

In this paper we have proposed and tested a bootstrap procedure for adjusting segregation indices for this bias. Our bootstrap correction works well provided both the minority proportion and unit size are not very small. Where very small minority proportions and unit sizes render our correction useless, we show that levels of segregation are often not statistically distinguishable from zero. We have developed and tested our statistical framework using the index of dissimilarity, D, but it can, in principle, be applied to other segregation indices.

From our statistical framework we have developed inference tests for a null of no systematic segregation; a null of equal segregation in two areas; and establishing confidence intervals for levels of systematic segregation. In tests using unit sizes, minority proportions and underlying segregation levels similar to those encountered by social scientists, the Wald statistic using the double bootstrap variance estimate for the bias corrected estimator (W_{bc}) and the test based on the equal tail bootstrap p-value for the t-test (T_{pb}) are found to perform best. The methods proposed in this paper provide a framework for more reliable inference as to levels of segregation, which will aid the further investigation of the causes of segregation.

Table 7: Bias corrected dissimilarity indices, confidence intervals and test statistics for Inner London

	D	D_{bc}	LR (p)	CI-W_{bc} CI-T_{pb}	Sout	Hack	Camd	Lamb	Isli	Wand	Gree	Hamm	Lewi	Kens	West
Tower	.197	.161	0	[.123-.199]	.764	.291	.289	.088	.032	.001	.000	.001	.000	.000	.000
Hamlets				[.127-.193]	.695	.176	.208	.036	.012	.000	.000	.000	.000	.002	.000
Southwark	.206	.169	0	[.132-.206]		.435	.300	.151	.058	.003	.001	.001	.000	.001	.000
				[.137-.200]		.330	.412	.096	.042	.000	.000	.000	.000	.000	.000
Hackney	.219	.191	0	[.151-.231]			.888	.542	.267	.030	.013	.013	.002	.007	.001
				[.156-.224]			.881	.476	.208	.014	.000	.000	.000	.004	.002
Camden	.231	.196	0	[.144-.247]				.698	.398	.077	.044	.035	.011	.017	.004
				[.152-.239]				.633	.322	.046	.020	.016	.004	.010	.000
Lambeth	.240	.208	0	[.170-.247]					.581	.106	.056	.047	.011	.022	.005
				[.175-.241]					.484	.052	.020	.020	.002	.010	.000
Islington	.257	.225	0	[.181-.269]						.327	.220	.157	.071	.077	.027
				[.184-262]						.272	.178	.086	.028	.062	.018
Wandsworth	.290	.255	0	[.213-.298]							.821	.582	.430	.311	.169
				[.219-.291]							.833	.519	.380	.262	.126
Greenwich	.286	.262	0	[.223-.300]								.701	.547	.381	.216
				[.226-.297]								.641	.511	.312	.148
Hammersmith	.303	.274	0	[.222-.327]									.910	.627	.450
& Fulham				[.226-.318]									.899	.579	.358
Lewisham	.304	.278	0	[.241-.315]										.656	.449
				[.247-.310]										.641	.402
Kensington &	.317	.295	0	[.231-.358]											.843
Chelsea				[.232-.350]											.833
Westminster	.328	.303	0	[.250-.357]											
				[.253-.352]											

References

Allen, R. and Vignoles, A. (2007), What Should an Index of School Segregation Measure?, CEE Discussion Papers 0060, Centre for the Economics of Education, LSE.

Boisso, D., Hayes, K., Hirschberg, J. and Silber, J. (1994), Occupational Segregation in the Multidimensional Case, *Journal of Econometrics*, 61, 161-171.

Burgess, S., McConnell, B., Propper, C. and Wilson, D. (2006), The Impact of School Choice on Sorting by Ability and Socio-economic Factors in English Secondary Education, in: L. Woessmann and P. Peterson (eds), *Schools and the Equal Opportunity Problem*, MIT Press, Cambridge.

Carrington, W. J. and Troske, K. R. (1997), 'On Measuring Segregation in Samples with Small Units, *Journal of Business and Economic Statistics*, 15 , 402-409.

Cortese, C. F., Falk, R. F. and Cohen, J. K. (1976), Further Considerations on the Methodological Analysis of Segregation Indices, *American Sociological Review*, 41, 630-637.

Davison, A.C. and Hinkley, D.V. (1997), *Bootstrap Methods and their Applications*, Cambridge University Press, Cambridge.

Duncan, O. and Duncan, B. (1955), A Methodological Analysis of Segregation Indexes, *American Sociological Review*, 20, 210-217.

Gibbons, S. and Telhaj, S. (2006), Are Schools Drifting Apart? Intake Stratification in English Secondary Schools, CEEDP 64, Centre for the Economics of Education, LSE.

Hall, P. (1992), *The Bootstrap and Edgeworth Expansion*, Springer-Verlag, New York.

Hellerstein, J. and Neumark, D. (2008), Workplace Segregation in the United States: Race, Ethnicity and Skill, *Review of Economics and Statistics*, 90, 459-477.

Massey, D.S. and Denton, N.A. (1988), The Dimensions of Residential Segregation *Social Forces*, 67, 281-315.

Mora, R. and Ruiz-Castillo, J. (2007), The Statistical Properties of the Mutual Information Index of Multigroup Segregation, Working Paper 07-74, Department of Economics, Universidad Carlos III de Madrid.

Ransom, M. R. (2000), Sampling Distributions of Segregation Indexes, *Sociological Methods & Research*, 28, 454-475.

Söderström, M. and Uusitalo, R. (2005), School Choice and Segregation: Evidence from an Admission Reform, IFAU Working Paper 2005:7.

White, M. J. (1986), Segregation and Diversity Measures in Population Distribution, *Population Index*, 52, 198-221.