AMERICAN SOCIOLOGICAL ASSOCIATION

Sociological Methodology
Volume 42, 365–389
© American Sociological Association 2012
DOI: 10.1177/0081175012452207
http://sm.sagepub.com
SAGE

10

A Nonparametric Index of Stratification

Xiang Zhou[1]

Abstract

The author presents a nonparametric approach to measuring stratification that highlights the distinction between stratification and inequality. Using pairwise comparison of ranks, the author develops an index of stratification that gauges the overall degree to which population subgroups occupy distinct strata with respect to a hierarchical outcome. This new index possesses a number of desirable properties that are not satisfied by existing measures of stratification. The overall index can be decomposed as a weighted average of pair-specific indices of stratification, which capture the extent of separation between any two particular groups. Besides, this index can be easily extended to measure conditional stratification through control of a third variable. In addition, the author builds a parallel between stratification and inequality in their measurement by developing a general formula of which the index of stratification and the Gini index of inequality can be considered as two special cases. Finally, this new approach is applied to depict the temporal trends of wage stratification by gender, race, and educational attainment over the past three decades in the United States.

Keywords

Gini, inequality, pairwise comparison, rank correlation, stratification

1. INTRODUCTION

Among the most deeply embedded notions in sociology are inequality and stratification. They have been traditionally used in different settings and from different perspectives. In particular, inequality refers usually to a *state* in which economic or social resources are unevenly distributed across individuals or between population

[1] University of Michigan, Ann Arbor, MI, USA

Corresponding Author:
Xiang Zhou, University of Michigan, 426 Thompson Street, 2067 ISR, Ann Arbor, MI 48109, USA
Email: xiangzh@umich.edu

subgroups. Although economists have discussed inequality almost exclusively in the context of income and wealth, sociologists have extended the concept to embrace variations in other domains of social life, such as educational attainment and health status. Stratification, as a sociological construct, is frequently used to emphasize the *process* by which a society is divided into a number of hierarchically arranged groups. The stratification hierarchy, as Max Weber argued, can be based on three distinct dimensions: economic condition, social status, and political power. Over the past half century, sociologists have increasingly used socioeconomic status, a combined measure of income and education, in their studies of stratification.

Nevertheless, the boundary between inequality and stratification in their use is hardly visible in today's substantive research. Empirical researchers sometimes analyze patterns and determinants of inequality (particularly between-group differences) under the name of stratification, as though they were exchangeable terms (e.g., Greenhalgh 1985; Hagan 1990; Kao and Thompson 2003; Lenski 1984; C. E. Ross and Bird 1994; P. Ross 1981). More recently, sociologists have attempted to reify the concept of stratification by comparing some kind of between-group variation with the total variation in a particular hierarchy, notably earnings. For example, Kim and Sakamoto (2008) used occupation R^2 and the decomposition of the Theil index to determine the impact of occupation structure on the rise of wage inequality in the United States. Mouw and Kalleberg (2010) reevaluated the role of occupations in explaining the increase in wage inequality using the standard decomposition of the variance of log wages. Meanwhile, Liao (2006, 2008) proposed a measure of stratification on the basis of the relative size of the between-class Gini coefficient of inequality. This class of methods is conducive to disentangling stratification from inequality but has two drawbacks. First, it depends on the specific measure of variation and is thus unable to provide a unified index for assessing stratification. More important, it lacks a key feature that is crucial to quantify the amount of stratification. To distinguish stratification from inequality, an ideal measure of stratification should be independent of the level of inequality. Unfortunately, a ratio of between-group inequality to total inequality rarely satisfies this property and may confound changes in inequality with changes in stratification.

In fact, as far back as two decades ago, Yitzhaki and Lerman (1991) observed this confusion:

> Articles using the term stratification continue to appear, but they typically attempt to determine the impact of the social and economic determinants of inequality (usually the variance) of such outcomes as earnings, incomes, occupational level or education. While these analyses may contribute to our understanding of average effects of, say, education on income levels, they provide no basis for making a distinction between stratification and inequality. (p. 313)

In the same article, Yitzhaki and Lerman developed an index of stratification for gauging the extent to which population subgroups overlap with respect to a hierarchical measure, such as income. However, their approach to measuring stratification is group specific and fails to provide an index for the whole population. Inspired by their pioneering work, in this article, I aim to propose a new measure that captures the overall degree of stratification for a hierarchical outcome.

The structure of this article is as follows. In section 2, I draw a distinction between inequality and stratification in four aspects: units of comparison, patterns of interest, information requirement, and existing measures. In section 3, I discuss Yitzhaki and Lerman's (1991) measure of stratification and its limitations. In section 4, I introduce a nonparametric index of stratification, which satisfies a number of desirable properties. The overall index, as I show, can be decomposed as a weighted average of pair-specific measures of stratification. Meanwhile, I extend this index to evaluate conditional stratification through control of a third variable. In section 5, I build a parallel between stratification and inequality in their measurement by developing a general formula of which the index of stratification and the Gini coefficient can be considered as two special cases. In section 6, I illustrate this new index by displaying the temporal trends of wage stratification by gender, race, and educational attainment over the past three decades in the United States. Section 7 presents my conclusions.

2. STRATIFICATION AND INEQUALITY

Before addressing the issue of measurement, I draw a conceptual distinction between stratification and the traditional notion of inequality. As mentioned at the beginning, the sociological construct of stratification hinges on the concept of strata, or layers. As Lasswell (1965) put it in *Class and Stratum*, ''Stratification is the process of forming observable layers, or the state of being comprised of layers'' (p. 10). In the context of income, one may consider these layers as different segments of the overall income distribution. These segments could correspond to a specific attribute of the individual, such as gender, race, or education. For example, when referring to *income stratification by race*, race is considered an agent by which the distribution of income is stratified among different racial groups. Therefore, complete stratification of income by race corresponds to the case in which there is no overlap between different racial groups in their ranges of income. This conceptualization makes income stratification fundamentally different from income inequality, which focuses on the variation of income across individuals or between population subgroups. Income stratification by some attribute can be fairly high in a society with a very low income inequality along the same axis, and vice versa.

To see this point, consider two hypothetical populations, A and B, both of which are composed of male and female workers. Suppose that the sex-specific distributions of annual income are given in Figure 1, respectively for population A (left) and population B (right). In population A, the average earnings are $32,000 for men and $30,000 for women, indicating a fairly low level of income inequality between the sexes (measured by either difference or ratio). However, for the same population, there is little overlap between men and women in their ranges of income distribution. This suggests that almost the highest-earning woman makes less than does the lowest-earning man. Therefore, in population A, income stratification by sex is virtually complete, although income inequality between men and women seems

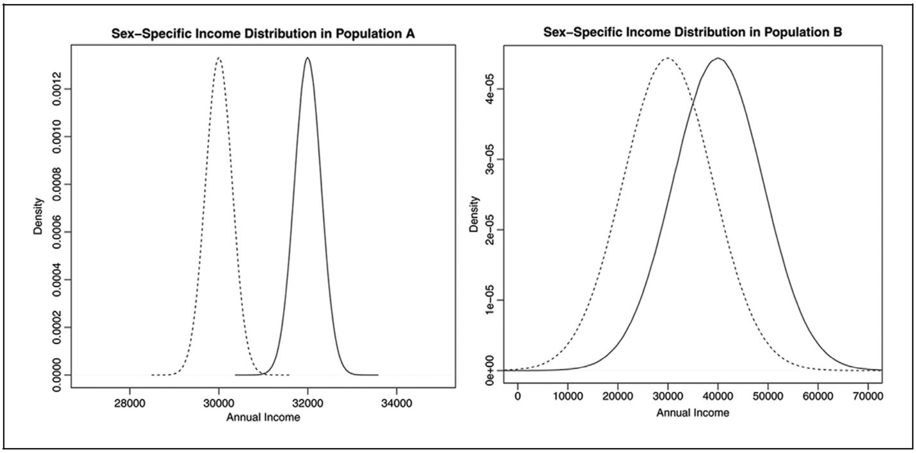

Figure 1. Sex-specific Distributions of Annual Income in Two Hypothetical Populations

reasonably low. Population B shows the opposite end of the spectrum: The average earnings are $40,000 for men and $30,000 for women, revealing a considerably higher gender gap in income than in population A. At the same time, the spread of income is so wide for either men or women that it becomes fairly difficult to predict who earns more between a randomly chosen man and a randomly chosen woman. Hence, compared with population A, population B demonstrates a lower degree of income stratification but a higher level of income inequality between men and women.

The preceding example implies that the demarcation between inequality and stratification consists in the conceptual boundary between *variation* and *segmentation*. To better delineate this boundary, it is important to notice the associated distinction between *level* and *rank*. To assess the magnitude of variation (i.e., inequality), a researcher must obtain the absolute levels of all individual observations. In contrast, only ranks are needed to evaluate the degree of segmentation (i.e., stratification). One can derive the ranks of all observations by ordering their levels, but not vice versa; therefore, ranks are nested in levels with respect to the amount of information they contain. A monotonic transformation of the variate, such as $x \rightarrow x^3$, can alter the distribution of levels and thus the size of inequality; however, it will maintain the extent of stratification by preserving the rank order of individual observations. As a result, a measure of stratification that depends on the absolute level of the variate is liable to contaminate the concept with irrelevant information. I further illustrate this point in section 3.3.

Table 1 compares inequality and stratification in four aspects: units of comparison, patterns of interest, information required for measurement, and existing measures. The first three rows recapitulate the foregoing discussion: Although inequality focuses on the variation of levels, either across individuals or between groups,

Table 1. Distinction between Inequality and Stratification

	Inequality		Stratification
Units of comparison	Individuals	Population subgroups	Population subgroups
Patterns of interest	Across-individual variation	Between-group variation	Between-group segmentation
Information required for measurement	Levels	Levels	Ranks
Existing measures	Gini index, Theil index, etc.	Intergroup gap (difference, ratio)	Between-group proportion of variation, Yitzhaki and Lerman's (1991) index

stratification hinges on the segmentation of ranks between population subgroups. The last row summarizes existing measures that were used to quantify different patterns of interest. In particular, the Gini index and the Theil index are frequently used to measure the overall degree of inequality, either over the whole population or within a subpopulation. Disparity between population subgroups, by contrast, is predominantly gauged by some kind of intergroup gap, such as the difference of two group means. Nonetheless, there is little consensus on the measurement of stratification. As mentioned in the preceding section, sociologists have recently relied on some kind of between-group proportion of variation, such as R^2, whereas the stratification index proposed by Yitzhaki and Lerman (1991) has gained more popularity in economics.

3. YITZHAKI AND LERMAN'S INDEX OF STRATIFICATION

3.1. Definition and Properties

In this section, I briefly introduce the index that Yitzhaki and Lerman (1991) developed for measuring income stratification. First, consider a population of n subjects, which can be classified into g mutually exclusive groups by an individual attribute, such as gender, race, or educational attainment. Let y_{si} be the income of the ith member of the sth group ($1 \leq s \leq g$). Second, I denote by F_s the cumulative distribution of income over members of group s and by F_{ns} the cumulative distribution of income over the entire population excluding group s. Therefore, $F_s(y_{si})$ and $F_{ns}(y_{si})$ are the percentile ranks of observation y_{si} respectively within group s and among all population members except group s. Further, I use $\mathrm{Cov}_s(u, v)$ to designate the covariance between u and v among members of group s only. On the basis of these notations, a group-specific index of stratification is defined as follows:

$$Q_s = \frac{\mathrm{Cov}_s[(F_s - F_{ns}), y]}{\mathrm{Cov}_s(F_s, y)}. \tag{1}$$

In this expression, the numerator is the covariance over group s between the variate, y_{si}, and the difference between its percentile rank in group s, $F_s(y_{si})$, and the percentile rank it would have in the rest of the population, $F_{ns}(y_{si})$. The denominator, as a normalizing factor, is the covariance between the variate and its percentile rank in group s.

Yitzhaki and Lerman (1991) showed that the previous index possesses properties that reflect the extent to which group s forms a distinct stratum in the population. First, it ranges from –1 to 1 and reaches the maximum of 1 if and only if $\text{Cov}_s(F_{ns}, y) = 0$. This condition suggests that there is no variation in $F_{ns}(y)$ across all members of group s. In this case, no members of other groups are in the range of the variate of group s, that is, group s alone occupies a certain stratum of the overall distribution. Second, Q_s decreases as $\text{Cov}_s(F_{ns}, y)$ increases, that is, as more members of other groups enter the range of the variate of group s. Therefore, the more integrated the members of other groups are with those of group s in terms of y, the lower is the value of Q_s. Third, the index equals zero when $F_s = F_{ns}$, that is, when the income distribution over members of group s is identical to that over the entire population. In this case, group s does not occupy a stratum at all.

Furthermore, Q_s takes a negative value when $\text{Cov}_s(F_{ns}, y) > \text{Cov}_s(F_s, y)$, which means that the divergence of $F_{ns}(y)$ is larger than that of $F_s(y)$ among members of group s. This suggests a scenario in which group s is not a homogeneous group but consists of several groups that occupy different ranges of the overall distribution. Finally, Q_s achieves the minimum of –1 when $\text{Cov}_s(F_{ns}, y) = 2\text{Cov}_s(F_s, y)$, corresponding to the case in which group s comprises two strata that are located at the extremes of the overall distribution, with the members of each strata taking identical values of y.

With these properties, the index Q_s adequately characterizes the degree of segmentation in the overall distribution between one group and the rest of the population. In addition, this index satisfies the property of scale invariance. That is, if all the measures are multiplied by a constant c, the index Q_s will be unchanged:

$$Q_s^* = \frac{\text{Cov}_s[(F_s - F_{ns}), cy]}{\text{Cov}_s(F_s, cy)} = \frac{c\text{Cov}_s[(F_s - F_{ns}), y]}{c\text{Cov}_s(F_s, y)} = Q_s.$$

Similarly, Q_s is translation invariant, that is, not changing if all the measures are added by a constant, because

$$Q_s^* = \frac{\text{Cov}_s[(F_s - F_{ns}), y + c]}{\text{Cov}_s(F_s, y + c)} = \frac{\text{Cov}_s[(F_s - F_{ns}), y]}{\text{Cov}_s(F_s, y)} = Q_s.$$

3.2. Decomposition

The preceding index measures a group's stratification with respect to the rest of the population. The Q index could also be calculated for one group in terms of its overlap with another group, which can be formulated as

$$Q_{ts} = \frac{\mathrm{Cov}_s[(F_s - F_t), y]}{\mathrm{Cov}_s(F_s, y)}. \tag{2}$$

As in the case of Q_s, Q_{ts} possesses properties that capture the extent of separation of group s from group t. However, there is an asymmetry between group s and group t in this definition, which implies that Q_{ts} may not be equal to Q_{st}. To see this, consider a population which consists of only two groups: group 1 and group 2. In this population, group 1 forms a distinct stratum in the sense that no observations of group 2 are in the range of the variate of group 1, whereas observations of group 2 are heterogeneous in the sense that half of them are above the maximum of group 1 and half of them are below the minimum of group 1. In this case, it is not hard to show that $Q_{21} = 1$ but that $Q_{12} < 0$.

The group-level index of stratification Q_s can be expressed as a weighted average of Q_{ts} values. In fact, the percentile rank that a member of group s would have in the rest of the population is a weighted average of the percentile ranks that he or she would have in all groups except group s; that is,

$$F_{ns} = \sum_{t \neq s} w_{ts} F_t. \tag{3}$$

Here,

$$w_{ts} = \frac{p_t}{1 - p_s}, \tag{4}$$

where p_s denotes the population share of group s ($1 \leq s \leq g$). Obviously, $\sum_{t \neq s} w_{ts} = 1$. Plugging equation 3 into equation 1 produces

$$Q_s = \frac{\mathrm{Cov}_s[(F_s - \sum_{t \neq s} w_{ts} F_t), y]}{\mathrm{Cov}_s(F_s, y)} \tag{5}$$

$$= \sum_{t \neq s} w_{ts} Q_{ts}. \tag{6}$$

Therefore, the stratification of group s as measured by Q_s is a weighted average of Q_{ts} values, with the weight of Q_{ts} proportional to the size of group t.

3.3. Limitations

Although Yitzhaki and Lerman's (1991) index largely captures the amount of segmentation between population subgroups, it has a number of drawbacks. First and foremost, the index of Q_s capitalizes on the information of both ranks (F_s and F_{ns}) and levels (y). Meanwhile, the discussion in section 2 suggests that only ranks are needed for gauging the concept of stratification. A measure that depends on the absolute level may confound changes in stratification with irrelevant changes of the variate. For example, a monotonic transformation of income, such as $y \rightarrow \log y$, should not change the level of income stratification. Unfortunately, this requirement is not satisfied by the index of Q_s, because in general

$$\frac{\mathrm{Cov}_s[(F_s - F_{ns}), y]}{\mathrm{Cov}_s(F_s, y)} \neq \frac{\mathrm{Cov}_s[(F_s - F_{ns}), \log y]}{\mathrm{Cov}_s(F_s, \log y)}.$$

Second, Yitzhaki and Lerman's (1991) index is group specific and does not measure the degree of stratification for the whole population. In particular, for a population that composes two groups, the stratification index for one group may be dramatically different from that for the other, which complicates the evaluation of the overall degree of stratification. Third, the range of Yitzhaki and Lerman's index is [–1,1], rather than [0,1]. Although the interval of [0,1] reflects precisely the unidimensional concept of stratification, the interpretation of negative values is ambiguous. In fact, taking a negative value corresponds to an interaction effect of stratification and heterogeneity, which is undesirable for a measure of stratification.[1]

4. A NONPARAMETRIC INDEX OF STRATIFICATION

In this section, I develop a new index of stratification on the basis of pairwise comparisons of ranks. This new index aims to capture the overall extent of stratification for the population with a value between 0 and 1. More important, it is fully nonparametric and thus independent of the distribution of levels. In the subsections that follow, I (1) introduce its definition and properties, (2) discuss the decomposition of the overall index of stratification into pair-specific components, and (3) extend this index to measure conditional stratification through control of a third variable.

4.1. Definition and Properties

4.1.1. Two-group case. I begin the introduction of the new index with a two-group case. Consider a population that consists of n_{M} men and n_{F} women. Let $y_{\mathrm{M}i}$ be the income of the ith man and $y_{\mathrm{F}j}$ the income of the jth woman. First, we order all the subjects (including men and women) from the lowest to the highest in terms of income and use $r_{\mathrm{M}i}$ and $r_{\mathrm{F}j}$ to denote the ranks of the ith man and of the jth woman. Then, we calculate the average ranks both for men and women and denote them R_{M} and R_{F}. For convenience, we assume $R_{\mathrm{M}} > R_{\mathrm{F}}$; that is, men rank higher than women on average. Given these notations, we measure the extent of income stratification between men and women by the following quantity:

$$S = \frac{\sum_{i=1}^{n_{\mathrm{M}}} \sum_{j=1}^{n_{\mathrm{F}}} [1(r_{\mathrm{M}_i} > r_{\mathrm{F}_j}) - 1(r_{\mathrm{M}_i} < r_{\mathrm{F}_j})]}{n_{\mathrm{M}} n_{\mathrm{F}}}. \tag{7}$$

The above measure adequately quantifies the amount of income stratification between men and women. On one hand, it takes zero when $\sum_{i=1}^{n_{\mathrm{M}}} \sum_{j=1}^{n_{\mathrm{F}}} 1(r_{\mathrm{M}i} > r_{\mathrm{F}j}) = \sum_{i=1}^{n_{\mathrm{M}}} \sum_{j=1}^{n_{\mathrm{F}}} 1(r_{\mathrm{M}i} < r_{\mathrm{F}j})$, that is, when there is no difference between men and women in their relative positions. In fact, it is not hard to show that $S = 0$ if and only if $R_{\mathrm{M}} = R_{\mathrm{F}}$. On the other hand, the index increases as men become more separated from women in the income ranking and reaches the maximum of 1 when $y_{\mathrm{M}i} > y_{\mathrm{F}j}$ holds true for any pair of man and woman, corresponding to the case of complete stratification.

4.1.2. Multigroup case. To expound the general case of multiple groups, consider a hypothetical population of n subjects, who belong to g mutually exclusive groups. As in section 3.1, let y_{si} be the income of the ith member of the sth group $(1 \leq s \leq g)$. First, we rank all the subjects in order of increasing income and denote by r_{si} the income rank corresponding to the observation of y_{si}. Second, we calculate the average rank for each group so that these groups can be ordered from the lowest average rank to the highest. Then, we use R_s to denote the average rank of the sth group. Thus, r and R provide two sets of income ranks, one for individual observations and the other for population subgroups. On the basis of these notations, we define an index of stratification by the following concordance score between the two sets of ranks:

$$S = \frac{\sum_{s=1}^{g}\sum_{t=1}^{g}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}[1(r_{si} > r_{tj}) - 1(r_{si} < r_{tj})]1(R_s > R_t)}{\sum_{s=1}^{g}\sum_{t=1}^{g}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}1(R_s > R_t)}, \tag{8}$$

where n_s and n_t denote the number of observations in group s and group t, respectively. Taking into account the facts that $1(r_{si} > r_{tj}) = 1(y_{si} > y_{tj})$ and that $\sum_{s=1}^{g}\sum_{t=1}^{g}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}1(R_s > R_t) = \sum_{s=2}^{g}\sum_{t=1}^{s-1}n_s n_t$ (assuming no ties between groups), equation 8 can be rewritten as[2]

$$S = \frac{\sum_{s=1}^{g}\sum_{t=1}^{g}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}[1(y_{si} > y_{tj}) - 1(y_{si} < y_{tj})]1(R_s > R_t)}{\sum_{s=2}^{g}\sum_{t=1}^{s-1}n_s n_t}. \tag{9}$$

From these expressions, we see that the index of S measures essentially the confidence with which we can predict the relative position of two individuals from different groups on the basis of the relative position of the two groups to which they belong. Specifically, if we denote by P_{agree} the probability that the order of two individuals from different groups agrees with the order of their groups, then

$$\begin{aligned} S &= \frac{\sum_{s,t,i,j}1\left(r_{si} > r_{tj}\right)1(R_s > R_t)}{\sum_{s,t,i,j}1(R_s > R_t)} - \frac{\sum_{s,t,i,j}1\left(r_{si} < r_{tj}\right)1(R_s > R_t)}{\sum_{s,t,i,j}1(R_s > R_t)} \\ &= P_{\mathrm{agree}} - \left(1 - P_{\mathrm{agree}}\right) \\ &= 2P_{\mathrm{agree}} - 1. \end{aligned}$$

As a result, we have built a one-to-one correspondence between S and P_{agree}. Rewriting the above relation yields

$$P_{\text{agree}} = \frac{1}{2}(1+S),$$

and

$$1 - P_{\text{agree}} = \frac{1}{2}(1-S).$$

Consider the aforementioned two-group case: If the index of stratification is 0.4, the probability will be $\frac{1+0.4}{2} = 0.7$ that a randomly chosen man earns higher than a randomly chosen woman.

As in the two-group case, the general definition given in equation 8 satisfies a number of properties that mirror the concept of stratification. First, it takes zero when $\sum_{R_s > R_t} \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} 1(y_{si} > y_{tj}) = \sum_{R_s > R_t} \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} 1(y_{si} < y_{tj})$, which means no between-group difference in their relative positions in the overall income distribution. Second, the index of S increases as groups of higher ranks become more separated from groups of lower ranks and attains the maximum of 1 when $y_{si} > y_{tj}$ holds true for all pairs of (s, t) with $R_s > R_t$, corresponding to the case of complete stratification, that is, no overlap of income ranges between different groups.[3] Furthermore, because this approach capitalizes only on the rank order of incomes, the index of S is independent of the magnitude of income inequality, as well as the shape of income distribution. Therefore, it can take any value between 0 and 1 under whatever kind of income distribution.

4.1.3. Advantages of the nonparametric index. Compared with previous measures of stratification, the S index has several advantages. First, because equations 8 and 9 hinge on the pairwise comparison of individual ranks from different groups, this new approach disentangles the measurement of stratification thoroughly from the magnitude of within-group variation. Specifically, S reaches the maximum of 1 as long as the distribution of income is completely segmented between different groups. This property, however, is missing in stratification measures that are based on between-group proportion of variation, which do not attain the maximum of 1 until there is no within-group variation. More important, the S index is invariant under rank-preserving transformations, that is, transformations of the variate that do not alter the rank order of individual observations. Measures based on between-group proportion of variation, as well as Yitzhaki and Lerman's (1991) index, do not satisfy this feature, because both of them are explicitly dependent on the levels of the variate. Finally, in contrast to the group-specific measure proposed by Yitzhaki and Lerman, the S index provides a measure of stratification for the whole population. In Table 2, I compare different measures of stratification in five dimensions. Whereas all three approaches share the common merits of translation independence and scale independence, the S index shows superiority in capturing more crucial aspects of stratification.

Table 2. Comparison of Different Measures for Stratification

Characteristic	Measures Based on Between-group Proportion of Variation	Yitzhaki and Lerman's (1991) Index	*S* Index
Translation independence	Yes (for R^2)	Yes	Yes
Scale independence	Yes (for R^2)	Yes	Yes
Independent of within-group inequality	No	Yes	Yes
Invariant under rank-preserving transformations	No	No	Yes
Measuring stratification for the whole population	Yes	No	Yes

4.2. Decomposition

The index defined previously gauges the overall extent of stratification pertaining to a given grouping scheme. As in the case of Yitzhaki and Lerman's (1991) index of Q_s, S can also be expressed as a weighted average of pair-specific components. To see this, we first exchange the index pairs (s, i) and (t, j) in equation 8 and obtain

$$S = \frac{\sum_{t=1}^{g} \sum_{s=1}^{g} \sum_{j=1}^{n_t} \sum_{i=1}^{n_s} \left[1\left(r_{si} < r_{tj}\right) - 1\left(r_{si} > r_{tj}\right)\right] 1(R_s < R_t)}{\sum_{t=2}^{g} \sum_{s=1}^{t-1} n_t n_s} \tag{10}$$

$$= -\frac{\sum_{s=1}^{g} \sum_{t=1}^{g} \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} \left[1\left(r_{si} > r_{tj}\right) - 1\left(r_{si} < r_{tj}\right)\right] 1(R_s < R_t)}{\sum_{s=2}^{g} \sum_{t=1}^{s-1} n_s n_t}. \tag{11}$$

Then, taking the average of equation 8 and equation 11 gives another expression of S:

$$S = \frac{\sum_{s=1}^{g} \sum_{t=1}^{g} \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} \left[1\left(r_{si} > r_{tj}\right) - 1\left(r_{si} < r_{tj}\right)\right] \left[1(R_s > R_t) - 1(R_s < R_t)\right]}{2 \sum_{s=2}^{g} \sum_{t=1}^{s-1} n_s n_t}. \tag{12}$$

The above formula enables us to decompose the overall stratification index as

$$S = \sum_{u=1}^{g} \sum_{v=1}^{g} w_{uv} S_{uv}, \tag{13}$$

where

$$w_{uv} = \frac{n_u n_v}{2 \sum_{s=2}^{g} \sum_{t=1}^{s-1} n_s n_t}, \tag{14}$$

and

$$S_{uv} = \frac{\sum_{i=1}^{n_u} \sum_{j=1}^{n_v} \left[1\left(r_{ui} > r_{vj}\right) - 1\left(r_{ui} < r_{vj}\right)\right]\left[1(R_u > R_v) - 1(R_u < R_v)\right]}{n_u n_v}. \tag{15}$$

Here, S_{uv} indicates the degree of separation between the uth group and the vth group in their relative positions; in other words, the lower is the overlap between group u and group v in their spreads of income distribution, and the higher is the value of S_{uv}. At the same time, w_{uv} represents the weight of the pair (u, v) in determining the overall level of income stratification. For example, if the entire labor force in the United States is divided into three racial groups, (1) non-Hispanic whites, (2) African Americans, and (3) others, then S_{12} provides a measure for the level of separation between non-Hispanic whites and African Americans in their spreads of income distribution, and $\frac{n_1 n_2}{2(n_1 n_2 + n_1 n_3 + n_2 n_3)}$ gives the weight of the pair (non-Hispanic whites, African Americans) in determining the overall index of income stratification by race.

Note that in this decomposition, group u and group v are symmetric in the component of S_{uv}. Therefore, S_{uv} is equal to S_{vu} for any $u \neq v$. This is in stark contrast to the discrepancy between Q_{st} and Q_{ts} for the index proposed by Yitzhaki and Lerman (1991). Let us reconsider the example in section 3.2, in which members of group 2 surround members of group 1 from both ends. It is not hard to see that $S = S_{12} = S_{21} = 0$, which is consistent with the fact that group 1 and group 2 do not differ in their average positions in the income ranking. Therefore, in this case, the index of S provides a more intuitive result than do Q_1 and Q_2 in evaluating the degree of stratification.

4.3. Conditional Stratification

Social scientists have long been interested in the notion of conditional inequality, for example, wage disparity between blacks and whites after controlling for productivity-related covariates (Cancio, Evans, and Maume 1996). I now extend the S index to embrace the concept of conditional stratification. Suppose we have a ‘‘confounding’’ variable, z, which is associated with both the grouping attribute and the target outcome. For example, occupational sex segregation may account for a substantial portion of wage stratification between men and women. If the confounding variable z is categorical, as in occupation, we can restrict our pairwise comparisons to those pairs that are taking the same value of z. This procedure provides a definition of conditional stratification:

$$S(z) = \frac{\sum_{s=1}^{g} \sum_{t=1}^{g} \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} [1(r_{si} > r_{tj}) - 1(r_{si} < r_{tj})] 1(R_s > R_t, z_{si} = z_{tj})}{\sum_{s=1}^{g} \sum_{t=1}^{g} \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} 1(R_s > R_t, z_{si} = z_{tj})}, \tag{16}$$

where $z_{si} = z_{tj}$ means that observations si and tj belong to the same category of z. Sometimes, the confounding variable for which the researcher aims to control is a continuous measure, such as age; in this case, we can construct a discrete approximation of z by partitioning its range into a few short intervals. Similarly, if there are several confounding variables, we may divide the joint domain of these variables into a number of relatively homogeneous regions, thus creating a synthetic categorical variable, and then apply equation 16.[4]

5. LINKING THE *S* INDEX AND THE GINI COEFFICIENT

In this section, I demonstrate a similarity between stratification and inequality in their measurement. In particular, I show that both the S index of stratification and the Gini index of inequality can be expressed as a weighted average of interpersonal comparisons.

Let us first revisit the Gini index of inequality. Using the same set of notations adopted to introduce Q_s and S, we can express the Gini coefficient as (Dagum 1997; Schwartz and Winship 1980):

$$\text{Gini} = \frac{1}{n^2} \sum_{s=1}^{g} \sum_{t=1}^{g} \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} \frac{|y_{si} - y_{tj}|}{2\bar{y}}, \tag{17}$$

where $\bar{y}$ is the population average of income. Hence, the Gini index is simply the average of $\frac{|y_{si} - y_{tj}|}{2\bar{y}}$ over all pairs of observations.

At the same time, equation 9 indicates that the index of S is indeed the average of $1(y_{si} > y_{tj}) - 1(y_{si} < y_{tj})$ over those pairs of (y_{si}, y_{tj}) for which $R_s > R_t$ or, to put it another way, a weighted average of $1(y_{si} > y_{tj}) - 1(y_{si} < y_{tj})$ over all pairs of observations, with the weight proportional to $1(R_s > R_t)$. Therefore, the index of S and the Gini coefficient can be considered as two members of the following class:

$$\Lambda = \frac{\sum_{s,t,i,j} K(s,t,i,j) w(s,t,i,j)}{\sum_{s,t,i,j} w(s,t,i,j)}, \tag{18}$$

where the kernel function $K(s, t, i, j)$ denotes a measure of comparison between observations si and tj, and $w(s, t, i, j)$ denotes the corresponding weight up to a normalizing factor. Specifically, for the Gini index of inequality, $K(s,t,i,j) = \frac{|y_{si} - y_{tj}|}{2\bar{y}}$, and $w(s, t, i, j) = 1$, whereas for the S index of stratification, $K(s, t, i, j) = 1(y_{si} > y_{tj}) - 1(y_{si} < y_{tj})$, and $w(s, t, i, j) = 1(R_s > R_t)$.

S_{uv}, the pair-specific measure of stratification, can also be incorporated into the general measure of Λ. In fact, equation 15 suggests that S_{uv} is essentially the average of $[1(y_{ui} > y_{vj}) - 1(y_{ui} < y_{vj})][1(R_u > R_v) - 1(R_u < R_v)]$ between observations from group u and observations from group v. Hence, if we define

$$\delta(a,b) = 1(a > b) - 1(a < b),$$

Table 3. Gini, S, S_{uv}, and $S(z)$ Expressed as Different Realizations of Λ

Index	$K(s, t, i, j)$	$w(s,t,i,j)$	Corresponding Concept
Gini	$\frac{\lvert y_{si} - y_{tj} \rvert}{2\bar{y}}$	1	Overall inequality
S	$\delta(y_{si}, y_{tj})$	$1(R_s > R_t)$	Overall stratification
S_{uv}	$\delta(y_{si}, y_{tj})\delta(R_s, R_t)$	$1(s = u, t = v)$	Stratification between group u and group v
$S(z)$	$\delta(y_{si}, y_{tj})$	$1(R_s > R_t, z_{si} = z_{tj})$	Stratification conditional on z

S_{uv} can be written in the form of Λ by setting

$$K(s,t,i,j) = \delta(y_{si}, y_{tj})\delta(R_s, R_t)$$

and

$$w(s,t,i,j) = 1(s = u, t = v).$$

In addition, equation 16 implies that the measure of conditional stratification also becomes a special case of Λ as long as we set $w(s, t, i, j) = 1(R_s > R_t, z_{si} = z_{tj})$ and use the same kernel function for S.

As a consequence, Gini, S, S_{uv}, and $S(z)$ can be considered as different realizations of the general measure of Λ. Table 3 summarizes their corresponding kernels and weights, as well as the concepts that they aim to quantify. The column of kernel functions highlights the distinction between inequality and stratification with regard to information required: Whereas the Gini coefficient consists in the size of absolute distances, the stratification indices hinge on the comparison of relative positions.

6. WAGE STRATIFICATION IN THE UNITED STATES, 1980 TO 2010

In the following, I illustrate the new index of stratification by displaying the temporal trends of wage stratification in the United States. In particular, I estimate the stratification indices of nonmissing weekly wages by gender, race, and education using data from the Merged Outgoing Rotation Groups of the Current Population Survey at seven time points: 1980, 1985, 1990, 1995, 2000, 2005, and 2010. To operationalize race, I divide people into three racial categories: (1) non-Hispanic whites, (2) African Americans (i.e., blacks), and (3) others. To operationalize education, I classify people into four educational groups: (1) no diploma, (2) high school diploma, (3) some college, and (4) college degree.

In Figure 2, I demonstrate the trend of the S index for wage stratification by gender. The solid line shows a dramatic decline of wage stratification between men and women, from 0.51 in 1980 to 0.20 in 2010. Considering that the average income rank for men is consistently higher than that for women, this result suggests that the probability that a randomly selected man earns more than a randomly selected woman has

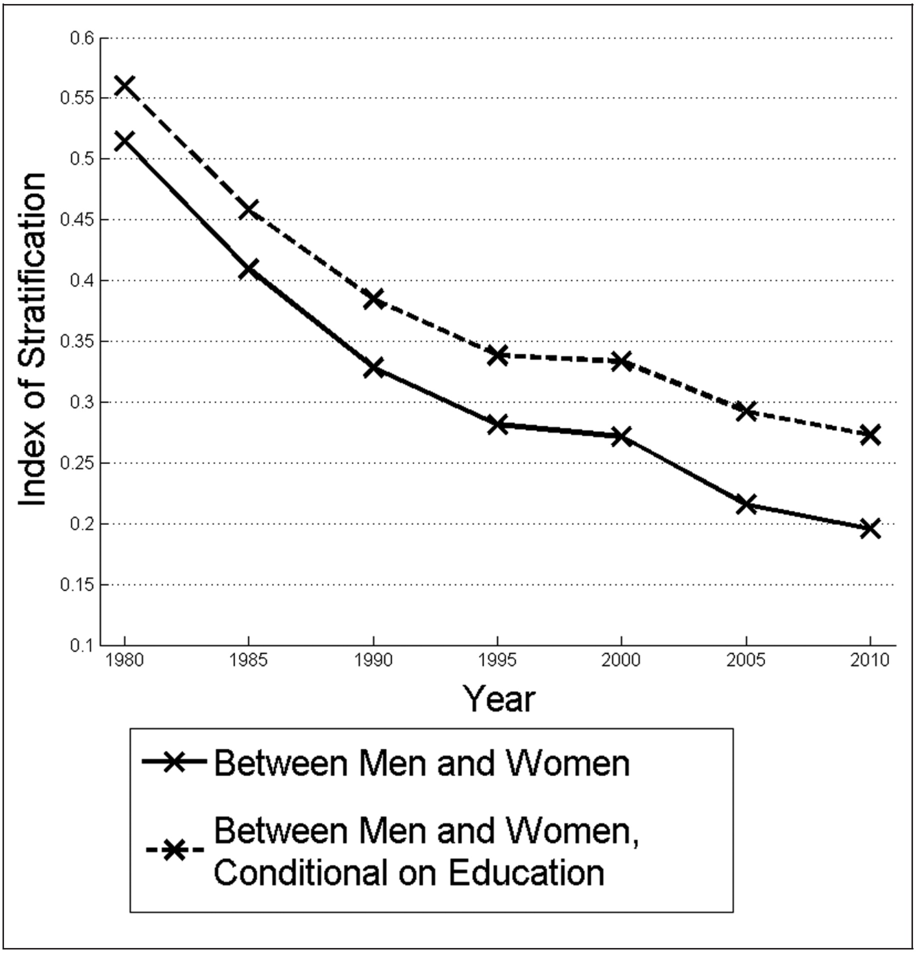

Figure 2. Wage Stratification by Gender in the United States, 1980 to 2010

declined from (1 + 0.51)/2 = 0.76 to (1 + 0.2)/2 = 0.6 over the past three decades. The dashed line depicts the trend of the same index conditional on educational attainment, which results in a higher level of stratification over the entire period. This is because women, especially women who are in the labor market, have surpassed men in educational attainment for three decades (Fischer and Hout 2006).

Figure 3 reveals the dynamics of wage stratification by race in two aspects: (1) across the aforementioned three racial categories (measured by S) and (2) between non-Hispanic whites and African Americans (measured by S_{12}). As in Figure 2, solid lines are used to represent the stratification measure in absolute terms, and dashed lines demonstrate the extent of stratification conditional on education. On one hand, we can see that over the three decades, different racial groups have been less

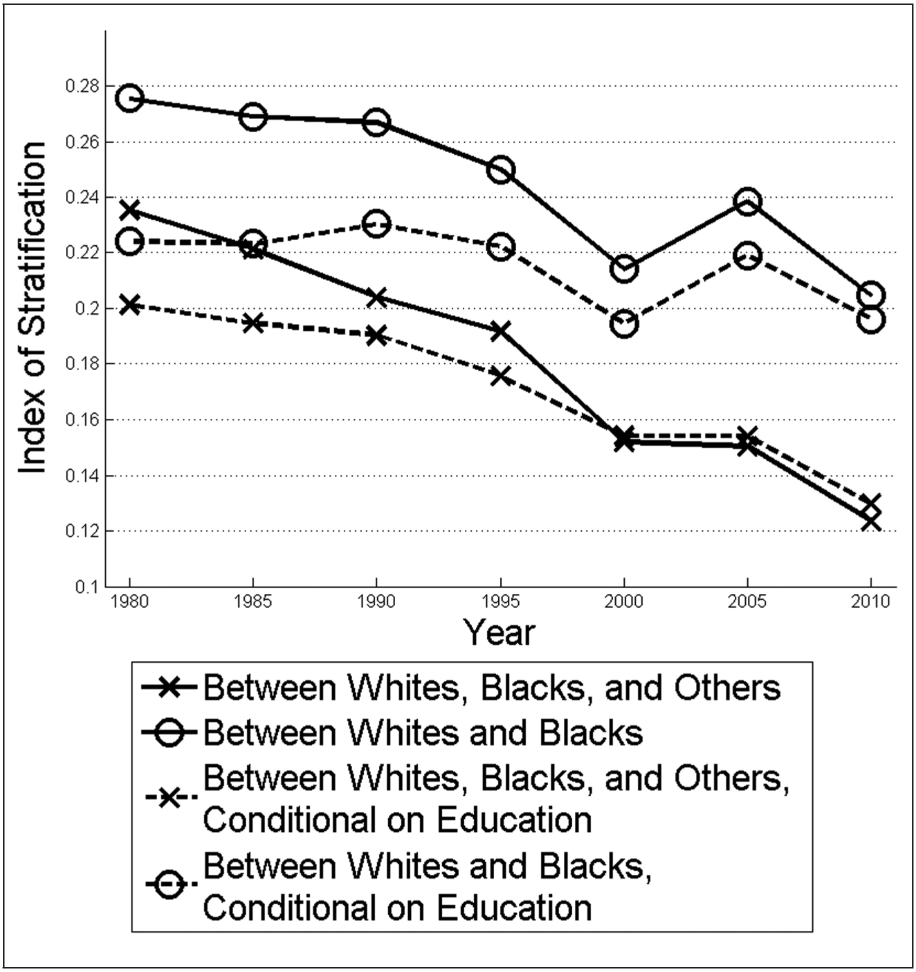

Figure 3. Wage Stratification by Race in the United States, 1980 to 2010

separated in their earnings, with a decline of the stratification index from 0.24 to 0.13. Wage stratification between non-Hispanic whites and blacks is consistently higher than the overall index of S, although S_{12} has also dwindled over the same period, with a noticeable rebound from 2000 to 2005. On the other hand, the dashed lines indicate that wage stratification by race is only slightly reduced after racial disparities in education are taken into account. In particular, wage stratification between non-Hispanic whites and blacks has been increasingly less mediated by their differences in educational attainment.

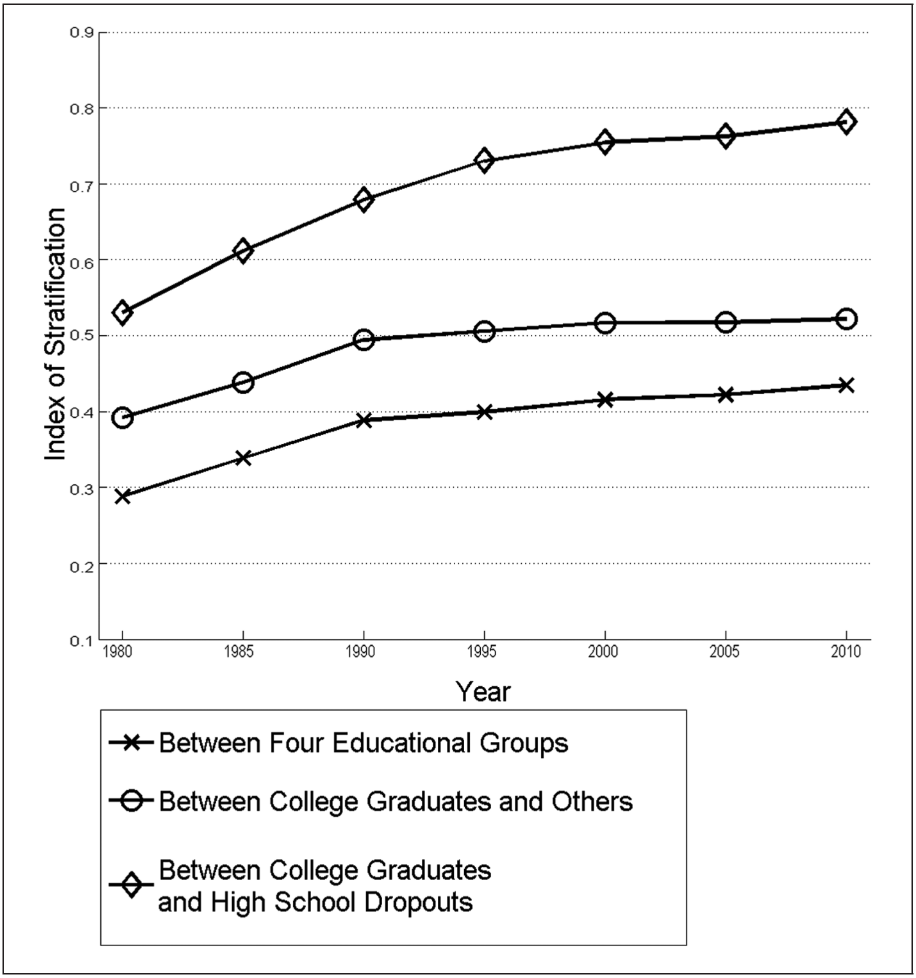

Figure 4. Wage Stratification by Education in the United States, 1980 to 2010

Wage stratification by education, by contrast, has experienced a substantial rise during the three decades. This is clearly manifested in Figure 4, which depicts the trends of the stratification index for three grouping schemes: (1) across the four educational categories defined, (2) between college graduates and others, and (3) between college graduates and high school dropouts. First, wage stratification across the four educational groups has significantly increased, from 0.29 in 1980 to 0.43 in 2010. Then, by focusing on the divide between college graduates and others, we see a higher degree of stratification between college degree holders and the rest, which has also climbed steadily during these years. Finally, when we restrict our attention exclusively to the separation between college graduates and high school dropouts, we obtain an even higher index of stratification (S_{14}), which has exhibited a more rapid growth over the past 30 years, from 0.53 in 1980 to 0.78 in 2010. This implies that the probability that a randomly chosen high school dropout earns more than a randomly chosen college graduate has dropped from (1 – 0.53)/2 = 0.24 to (1 – 0.78)/2 = 0.11.

7. CONCLUDING REMARKS

In this article, I have proposed a new approach to measuring stratification between population subgroups with respect to a quantitative outcome, such as income. This approach provides a nonparametric measure of stratification through a concordance score between two sets of ranks, one from individual values and the other from the ordering of population subgroups. The resulting index is able to capture the overall extent of stratification with a value between 0 and 1. More important, it possesses certain characteristics that highlight the distinction between stratification and inequality. In particular, this index is invariant under rank-preserving transformations of the variate, which is not satisfied by any previous measure of stratification.

The overall index of S can also be decomposed as a weighted average of pair-specific measures of stratification (S_{uv} values), which enable one to assess the degree of separation between two specific groups. Moreover, this index can be easily extended to evaluate conditional stratification through control of a third variable. In addition, I have demonstrated a parallel between stratification and inequality in their measurement by developing a general formula of which S, S_{uv}, $S(z)$, and the Gini index can all be considered special cases.

Finally, I applied the new index to explore the temporal trends of wage stratification by gender, race, and educational attainment over the past three decades in the United States. In sum, the results show that the dividing axis of the U.S. labor market has shifted gradually from gender to education, especially the attainment of a college diploma, over the past three decades.

APPENDIX

MATLAB and R Codes for the S Index

The following MATLAB function provides a general routine for calculating the S index both in absolute terms and conditional on a third variable. It has three arguments: `y`, a column vector of the outcome variable (continuous); `x`, a column vector of the grouping variable (discrete); and `z`, a column vector of the control variable (discrete). To produce the unconditional measure of stratification, set `z = 0`.

```
function [output] = strat(y,x,z)
if z==0
Fy=tiedrank(y);
n=length(y);
df=[y,x,Fy];
```

```
value_x=unique(x);
number_x=length(value_x);
info_x=zeros(number_x,2);
new_df=[];
for i=1:number_x
temp=df(x==value_x(i),:);
sz=size(temp);
info_x(i,1)=sz(1);
info_x(i,2)=mean(temp(:,3));
new_df=[new_df;temp,info_x(i,2).*ones(sz(1),1)];
end
new_df=sortrows(new_df,4);
info_x=sortrows(info_x,2);
cum_n=[0;cumsum(info_x(:,1))];
T=0;
P=0;
for j=1:number_x-1
a=cum_n(j);
b=cum_n(j+1);
for p=a+1:b
for q=b+1:n
P=P+(new_df(q,3)>new_df(p,3));
T=T+1;
end
end
end
output=2*P/T-1;
else
Fy=tiedrank(y);
df=[y,x,z,Fy];
value_x=unique(x);
number_x=length(value_x);
info_x=zeros(number_x,2);
value_z=unique(z);
number_z=length(value_z);
weight_z=zeros(number_z,1);
strat_z=zeros(number_z,1);
new_df=[];
for i=1:number_x
temp=df(x==value_x(i),:);
sz=size(temp);
info_x(i,1)=sz(1);
info_x(i,2)=mean(temp(:,4));
new_df=[new_df;temp,info_x(i,2).*ones(sz(1),1)];
```

```
end
new_df=sortrows(new_df,5);
info_x=sortrows(info_x,2);
for k=1:number_z
cond_new_df=new_df(new_df(:,3)==value_z(k),:);
cond_info_x=zeros(number_x,2);
for i=1:number_x
temp=cond_new_df(cond_new_df(:,2)==value_x(i),:);
sz=size(temp);
cond_info_x(i,1)=sz(1); %number of observations
cond_info_x(i,2)=mean(temp(:,5)); % average rank
end
cond_info_x=sortrows(cond_info_x,2);
cond_cum_n=[0;cumsum(cond_info_x(:,1))];
T=0;
P=0;
for j=1:number_x-1
a=cond_cum_n(j);
b=cond_cum_n(j+1);
for p=a+1:b
for q=b+1:cond_cum_n(number_x+1)
P=P+(cond_new_df(q,4)>cond_new_df(p,4));
T=T+1;
end
end
end
strat_z(k)=2*P/T-1;
weight_z(k)=T;
end
weight_z=weight_z/sum(weight_z);
output=weight_z'*strat_z;
end
```

The following code is an R implementation of the same function:

```
strat=function(y,x,z=0){
if (length(z)==1){
Fy=rank(y,ties.method="average")
n=length(y)
df=data.frame(y,x,Fy)
value.x=unique(x)
number.x=length(value.x)
info.x=array(0,c(number.x,2))
new.df=NULL
```

```
for (i in seq(1,number.x)){
info.x[i,1]=dim(df[x==value.x[i],])[1]
info.x[i,2]=mean(df[x==value.x[i],]$Fy)
temp=cbind(df[x==value.x[i],],info.x[i,2])
new.df=rbind(new.df,temp)
}
new.df=new.df[order(new.df[,4]),]
info.x=info.x[order(info.x[,2]),]
cum_n=c(0,cumsum(info.x[,1]))
T=0
P=0
for (j in seq(1,number.x-1)){
start=cum_n[j]
end=cum_n[j+1]
t1=new.df[seq(start+1,end),1] %*% t(rep(1,n-end))
t2=rep(1,end-start) %*% t(new.df[seq(end+1,n),1])
t3=sign(t2-t1)
P=P+sum(t3)
T=T+(end-start)*(n-end)
}
print(P/T)
}
else{
Fy=rank(y,ties.method="average")
df=data.frame(y,x,z,Fy)
value.x=unique(x)
number.x=length(value.x)
info.x=array(0,c(number.x,2))
value.z=unique(z)
number.z=length(value.z)
weight.z=array(0,c(number.x,1))
strat.z=array(0,c(number.x,1))
new.df=NULL
for (i in seq(1,number.x)){
info.x[i,1]=dim(df[x==value.x[i],])[1]
info.x[i,2]=mean(df[x==value.x[i],]$Fy)
temp=cbind(df[x==value.x[i],],info.x[i,2])
new.df=rbind(new.df,temp)
}
new.df=new.df[order(new.df[,5]),]
info.x=info.x[order(info.x[,2]),]
for (k in seq(1,number.z)){
cond.new.df=new.df[new.df[,3]==value.z[k],]
cond.info.x=array(0,c(number.x,2))
for (i in seq(1,number.x)){
temp=cond.new.df[cond.new.df[,2]==value.x[i],]
cond.info.x[i,1]=dim(temp)[1]
cond.info.x[i,2]=mean(temp[,5])
}
cond.info.x=cond.info.x[order(cond.info.x[,2]),]
cond.cum.n=c(0,cumsum(cond.info.x[,1]))
T=0
P=0
for (j in seq(1,number.x-1)){
start=cond.cum.n[j]
end=cond.cum.n[j+1]
t1=cond.new.df[seq(start+1,end),1] %*% t(rep(1,cond.cum
.n[number.x+1]-end))
t2=rep(1,end-start)%*%t(cond.new.df[seq(end+1,cond.cum
.n[number.x+1]),1])
t3=sign(t2-t1)
P=P+sum(t3)
T=T+ (end-start)*(cond.cum.n[number.x+1]-end)
}
strat.z[k]=P/T
weight.z[k]=T
}
weight.z=weight.z/sum(weight.z)
print(t(weight.z) %*% strat.z)
}
}
```

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Notes

1. This weakness, however, will turn into a blessing if a researcher is interested in detecting within-group heterogeneity as well as measuring stratification.
2. If we consider R also as a ranking of the n subjects (with ties), then the S index is a rescaled version of the Kendall rank correlation coefficient between r and R, for which the denominator is $n(n-1)$ rather than $\sum_{s=2}^{g}\sum_{t=1}^{s-1} n_s n_t$ (Kendall 1955).

3. When the population consists of only two groups, it can be shown that the index of S ranges from 0 to 1. Generally, whether S can take negative values is still an open question to the author. In practice, however, S almost always falls in the interval [0,1].
4. Like other nonparametric methods, the conditional index of stratification is subject to the curse of dimensionality. When there are many confounding variables, simple partitioning of the high-dimensional space will result in a huge number of hypercubes, many of which may contain few or no observations. In this case, more advanced techniques of dimension reduction, such as latent class analysis, could be used to construct a categorical variable for which to control.

References

Cancio, S. A., T. D. Evans, and D. J. Maume. 1996. ‘‘Reconsidering the Declining Signicance of Race: Racial Differences in Early Career of Wages.’’ *American Sociological Review* 61(4):541–56.

Dagum, C. 1997. ‘‘A New Approach to the Decomposition of the Gini Income Inequality Ratio.’’ *Empirical Economics* 22(4):515–31.

Fischer, C. S. and M. Hout. 2006. *Century of Difference*. New York: Russel Sage Foundation.

Greenhalgh, S. 1985. ‘‘Sexual Stratification: The Other Side of ‘Growth with Equity’ in East Asia.’’ *Population and Development Review* 11(2):265–314.

Hagan, J. 1990. ‘‘The Gender Stratification of Income Inequality among Lawyers.’’ *Social Forces* 68(3):835–55.

Kao, G. and J. S. Thompson. 2003. ‘‘Racial and Ethnic Stratification in Educational Achievement and Attainment.’’ *Annual Review of Sociology* 29:417–42.

Kendall, M. G. 1955. *Rank Correlation Methods*. New York: Hafner.

Kim, C. and A. Sakamoto. 2008. ‘‘The Rise of Intra-occupational Wage Inequality in the United States, 1983 to 2002.’’ *American Sociological Review* 73(1):129–57.

Lasswell, T. 1965. *Class and Stratum*. Boston: Houghton Mifflin.

Lenski, G. 1984. ‘‘Income Stratification in the United States: Toward a Revised Model of the System.’’ Pp. 173–205 in *Research in Social Stratification and Mobility*, Vol. 3, edited by D. Treiman and R. Robinson. Greenwich, CT: JAI.

Liao, T. F. 2006. ‘‘Measuring and Analyzing Class Inequality with the Gini Index Informed by Model-based Clustering.’’ *Sociological Methodology* 36(1):201–24.

Liao, T. F. 2008. ‘‘The Gini Unbound: Analyzing Class Inequality with Model-based Clustering.’’ Pp. 201–21 in *Advances on Income Inequality and Concentration Measures*, edited by G. Betti and A. Lemmi. New York: Routledge.

Mouw, T. and A. L. Kalleberg. 2010. ‘‘Occupations and the Structure of Wage Inequality in the United States.’’ *American Sociological Review* 75(3):402–31.

Ross, C. E. and C. E. Bird. 1994. ‘‘Sex Stratification and Health Lifestyle: Consequences for Men’s and Women’s Perceived Health.’’ *Journal of Health and Social Behavior* 35(2):161–78.

Ross, P. 1981. ‘‘Sex Stratification in the Workplace: Male-female Differences in Economic Returns to Occupation.’’ *Social Science Research* 10(3):195–224.

Schwartz, J. and C. Winship. 1980. ‘‘The Welfare Approach to Measuring Inequality.’’ *Sociological Methodology* 11(1):1–36.

Yitzhaki, S. and R. Lerman. 1991. ‘‘Income Stratification and Income Inequality.’’ *Review of Income and Wealth* 37(3):313–29.

Bio

Xiang Zhou is a doctoral student in the Department of Sociology and the Department of Statistics at the University of Michigan. His research focuses on social demography, causal inference, and Chinese studies.