Some Issues on Application of Standardization and Decomposition Analysis

Jichuan Wang*

Children's Research Institute (CRI), Children's National Medical Center (CNMC), The George Washington University School of Medicine, USA

Abstract: This article is to 1) address some issues when multiple populations and/or sample uncertainty are involved in application of standardization and decomposition analysis (SDA); and 2) introduce a computer program that provides opportunity of dealing with such issues.

In the recent issues of AJPH, there are some interesting discussions on the application of a demographic technique – standardization and decomposition analysis (SDA) – for studying temporal trends in low-birthweight (LBW) rates in the U.S. [1-3]. The purpose of this study is to promote further discussion and application of SDA, as well as to shed some light on the issues debated between Schempf & Becker [2] and Yang *et al*. [3].

It is well-known that differences in crude rates between populations are often confounded by population compositions or distributions of confounding factors. SDA can be used not only to adjust the crude rate, but also to decompose the difference in the crude rates between populations into component effects, such as the "rate effect," representing the "real" rate difference, and "factor composition effects," representing the rate difference attributed to compositional differences or distribution differences in specific confounding factors [4-9]. For example, in Yang et al.'s study [1] change in LBW rate in the U.S. population during a given decade was decomposed into a rate effect and a factor effect; the former represents the "real" changes in parity-specific LBW rates, and the latter represents the changes in the observed crude LBW rates that were attributed to the changes in age and parity compositions in the U.S. child-bearing women population during a given decade. Although three factors (i.e., mother's age, birth parity, and ethnicity) were actually involved in the study, mothers' age and birth parity were combined as one factor, and one-factor SDA was conducted by ethnic group for periods 1980-1990 and 1990-2000, separately.

The key dispute in the debate between Schempf & Becker and Yang *et al.* is about the standardization solution. In Yang et al's study [1] rates and factor distributions were standardized based on different populations (e.g., 1980 and 1990), respectively. Schempf & Becker [2] suggest a symmetric solution using the average rate and average factor distribution as the weights.

*Address correspondence to this author at the Children Research Institute (CRI), Children's National Medical Center (CNMC), The George Washington University School of Medicine, 111 Michigan Ave. N.W., Washington, D.C. 20010, USA; Tel: 202-476-2978; E-mail: jiwang@cnmc.org

To better understand the debate between Schempf & Becker [2] and Yang *et al.* [1, 3] I rewrite Equation 2 in Schempf & Becker [2] as following:

$$-R_{1.} = \sum_i \frac{F_{2i} + F_{1i}}{2}\left(R_{2i} - R_{1i}\right) + \sum_i \frac{R_{2i} + R_{1i}}{2}\left(F_{2i} - F_{1i}\right) \quad (1)$$

This is actually one of the formulas described by Kitagawa [4] for SDA with two populations and one factor where $R_{1.}$ and $R_{2.}$ are the crude rates; R_{1i} and R_{2i} are the factor-specific rates (e.g., age-parity-specific LBW rates in Yang et al.'s study [1]); and F_{1i} and F_{2i} are the factor compositions in Populations 1 and 2 (or the same population at different time points), respectively. The difference between the two crude rates, ($R_{1.}$-$R_{2.}$), is decomposed into two components: 1) *rate effect* -- the first term in Equation 1 in which compositions of the confounding factor are standardized across populations; thus the crude rate difference contained in this term represents the difference in factor-specific rates between the two populations. 2) *factor component effect* -- the second term in Equation 1 in which factor-specific rates are standardized; thus, this term represents the crude rate difference that is attributed to the difference in the factor compositions between the two populations. The SDA solution expressed in Equation 1 is called *additive-effect* or *main-effect* solution.

Alternatively, the crude rate difference can be decomposed in the following way [4]:

$$R_{2.} - R_{1.} = \sum_i F_{1i}\left(R_{2i} - R_{1i}\right) + \sum_i R_{1i}\left(F_{2i} - F_{1i}\right) + \sum_i \left(R_{2i} - R_{1i}\right)\left(F_{2i} - F_{1i}\right) \quad (2)$$

where Population 1 is defined as the standard population, and the F_{1i} and R_{1i} are used as weights for the *rate effect* and *factor component effect*, respectively; and the third component is an *interaction effect*, which is due to differences in both factor composition and factor-specific rates. As a matter of fact, the *additive-effect* solution described in Equation 1 and the *interaction-effect* solution described in Equation 2 are equivalent to each other if we integrate half of the third component into the first and second components, respectively.

There may be personal preference in regard to the different SDA solutions. The *additive-effect* solution is proposed

1874-9186/08 2008 Bentham Open

as the meaningful one for general purposes, while the *interaction-effect* solution may be considered for studying temporal changes in outcomes of the same population [4]. The *additive-effect* solution has been generalized by Das Gupta for studying multiple confounding factors, as well as multiple populations [8, 9]

The current study focuses on two issues that are not addressed in the debate between Schempf & Becker and Yang et al, and are often encountered in applications of SDA: 1) how to appropriately conduct SDA when multi-populations are involved; and 2) how to conduct significance testing for component effects in SDA when sample data, rather than population data, are used.

First, when multiple populations (or the same population at multiple time-points) are analyzed, pair-wise comparisons of the populations are inappropriate because the pair-wise results are usually not internally consistent. For example, the standardized rates for each population may not remain the same in different pair-wise comparisons. As such, the difference in the standardized rates between population 1 and population 2 plus the difference between population 2 and population 3 may not be equal to the difference between population 1 and population 3. In other words, the factor effects may not be internally consistent. The correct way for conducting SDA with multiple populations or for the same population at multiple time-points is to conduct all pair-wise comparisons simultaneously adjusting for internal inconsistence. The formulas for such analysis have been developed by Das Gupta [8, 9]. When comparing populations 1 and 2 in the presence of populations 3, 4, …, and K, the standardized rate in population 1 controlling for all other factors but A is calculated as:

$$A_{1.23...K} = \frac{\sum_{j=2}^{K} A_{1.j}}{K-1} + \frac{\sum_{i=2}^{K}\left(\sum_{j\neq 1,i}^{K} A_{i.j} - (K-2)A_{i.1}\right)}{K(K-1)} \quad (3)$$

and the factor effect of A, controlling for all other factors, is:

$$A_{12.3...K} = A_{12} + \frac{\sum_{j=3}^{K}(A_{12} + A_{2j} - A_{1j})}{K} \quad (4)$$

In fact, the formulas shown in Equations 3 and 4 are based on the results of separate pair-wise population comparisons. For example, let $A_{1.2}$ denote the standardized rate in population 1, and A_{12} the factor effect of A, standardizing all other factors but A when populations 1 and 2 are compared without presence of other populations. The corresponding notations are $A_{1.23...K}$ and $A_{12.3...K}$ in the presence of other populations (e.g., populations 3…K). Suppose we have three populations in SDA, Equations 3 and 4 would become:

$$A_{1.23} = \frac{(A_{1.2} + A_{1.3})}{2} + \frac{(A_{2.3} + A_{2.1}) + (A_{3.2} + A_{3.1})}{6} \quad (5)$$

$$A_{12.3} = A_{12} + \frac{(A_{12} + A_{23} - A_{13})}{3} \quad (6)$$

where $A_{1.23}$ and $A_{12.3}$ are the standardized rate in population 1 and factor effect of A, respectively, standardizing all other factors but A, when populations 1 and 2 are compared in presence of population 3. When populations 2 and 3 are compared in presence of population 1, the corresponding standardized rate in population 2 and factor effect of A are $A_{2.31}$ and $A_{23.1}$, respectively. $A_{1.32}$ and $A_{13.2}$ can be readily calculated for comparing populations 1 and 3. As a result, each population will have only one set of standardized rates when standardization is conducted with respect to the same set of factors no matter which population this population is compared with; and factor effects will be internally consistent. For example, difference in the standardized rates between populations 1 and 2 plus the difference between populations 2 and 3 will add up to the difference between populations 1 and 3. Factor effects and standardized rates with respect to other factors can be calculated in the same way. That is, the same formulas apply to other factors regardless of how many factors are involved in the SDA [8, 9].

The second important issue in SDA applications that was not addressed in the debate between Schempf & Becker and Yang *et al.* is that SDA is based on algebraic calculation; as such, traditionally, sample uncertainty is not taken into account in SDA when sample data, rather than population data, are analyzed. As a matter of fact, survey data which are randomly sampled from a target population under study are most often used for data analyses in population studies and many other fields of social sciences. In order to apply SDA to survey data, significance testing for component effects should be considered. Although mathematic derivations of the standard errors of the component effects are possible using delta method, it is cumbersome. A non-parametric method -- bootstrap -- can be readily applied for statistical inference in situations where it is difficult or impossible to derive the standard error of a statistic in the usual way [10-13].

The author has developed a computer program DECOMP, which enables to conduct SDA with multiple populations/samples and conduct significance testing for component effects via bootstrap [14-16]. Both grouped data (contingency table) and individual data can be analyzed in DECOMP. If bootstrapping is desired, individual data must be used. However, DECOMP allows one to convert a grouped data set into an individual data set if the outcome measure in the data is rate, percentage, or proportion. The computer program DECOMP can be downloaded from the author's website (http://www.wright.edu/~jichuan.wang).

SDA has some explicit advantages. First, its results are easy to interpret. Outcome difference/change is decomposed into component effects that are attributed to "real" difference/change and effects of confounding factors; and the relative contributions of all component effects sum up to 100%. These kinds of results are much easier than the statistical model parameter estimates, particularly for policy makers, to understand. Second, SDA has no constraints on the specifi-

cation of relationship (e.g., linearity), the nature of the variables (e.g., random), or the form of variable distributions (e.g., normality), that are usually assumed for statistical analyses. Third, SDA can be applied to study differences/changes in a wide range of outcome measures such as rate, percentage, proportion, ratio, as well as arithmetic mean, among populations/samples [5, 16]. And finally, with an available computer program (e.g., DECOMP), we can conduct significance testing for component effects [14-16]. As such, hypotheses can be tested and results of a SDA study can be generalized from the study sample to the target population.

ACKNOWLEDGMENTS

This research was supported by a grant from the National Institute on Drug Abuse (Grant # 3 R01 DA10099-06S1).

ABBREVIATION

SDA = Standardization & decomposition analysis

REFERENCES

[1] Yang Q, Greenland S, Flanders WD. Associations of maternal age- and parity-related factors with trends in low-birthweight rates: United States, 1980 through 2000. Am J Public Health 2006; 96: 856-61.

[2] Schempf A, Becker S. On the application of decomposition methods. Am J Public Health 2006; 96: 1899.

[3] Yang Q, Greenland S, Flanders WD. Yang *et al.* respond. Am J Public Health 2006; 96: 1899-1901.

[4] Kitagawa EM. Components of a difference between two rates. J Am Stat Assoc 1955; 50: 1168-94.

[5] Kitagawa EM. Standardized comparisons in population research. Demography 1964; 1: 296-315.

[6] Pullum TW. Standardization (World Fertility Survey Technical Bulletins, No. 597), Voorburg, Netherlands: International Statistical Institutes; 1978.

[7] United Nations. The methodology of measuring the impact of family planning programs, manual IX (Population Studies No. 66). New York: U.N.; 1979.

[8] Das Gupta P. Decomposition of the difference between two rates and its consistency when more than two populations are involved. Math Popul Stud 1991; 3: 105-25.

[9] Das Gupta P. Standardization and decomposition of rates: a user's manual. U.S. Bureau of the Census, Current Population Reports (Series P23-186). Washington, D.C.: U.S. Government Printing Office; 1993.

[10] Efron B. Bootstrap methods: another look at the Jackknife. Ann Stat 1979; 7: 1-26.

[11] Efron B. Nonparametric standard errors and confidence intervals (with discussion). Can J Stat 1981; 9: 139-72.

[12] Miller RG. The Jackknife -- a review. Biometrika 1974; 61: 1-15.

[13] Mooney CZ, Duval BD. Bootstrapping: A Nonparametric Approach to Statistical Inference. Newbury Park, CA: Sage Publications; 1993.

[14] Wang J, Rahman A, Siegal HA, Fisher JH. Standardization and decomposition of rates: Useful analytic techniques for behavior and health studies. Behav Res Method Instrum Comput 2000; 32: 357-66.

[15] Wang J. Components of difference in HIV seropositivity rate among injection drug users between low and high HIV prevalence regions. AIDS Behav 2003; 7: 1-8.

[16] Wang J, Carlson RG, Falck RS, Leukefeld C, Booth BM. Multiple sample standardization and decomposition analysis: an application to comparisons of methamphetamine use among rural drug users in three American states. Stat Med 2007; 26: 3612-23 (corrections for printing errors in Tables II and III are available from the author).

Received: July 25, 2008 Revised: November 06, 2008 Accepted: November 14, 2008