On the index of dissimilarity for lack of fit in loglinear and log-multiplicative models

Jouni Kuha*[,a], David Firth[b]

[a] *Department of Statistics, London School of Economics, UK*
[b] *Department of Statistics, University of Warwick, Coventry, UK*

Abstract

The index of dissimilarity, often denoted by Delta, is commonly used, especially in social science and with large datasets, to describe the lack of fit of models for categorical data. In this paper the definition and sampling properties of the index are investigated for general loglinear and log-multiplicative models. It is argued that in some applications a standardized version of the index is appropriate for interpretation. A simple, approximate variance formula is derived for the index, whether standardized or not. A simple bias reduction formula is also given. The accuracy of these formulae and of confidence intervals based upon them is investigated in a simulation study based on large-scale social mobility data.

Key words: Bias reduction, Delta, Iterative proportional fitting, Model selection, Raking, Stratified sampling, Total variation distance

1. Introduction

In the presence of a large amount of informative data, even 'good' models are typically rejected by conventional lack-of-fit tests based on statistical significance. In such cases attention turns to assessment of the extent to which a model's lack of fit is important from the subject-matter point of view. For example, in situations where a statistical model is to be used mainly or entirely as a basis for forecasting, predictive performance will usually be a more important criterion than formal goodness of fit.

In the context of models for categorical data a commonly used statistic is the so-called *index of dissimilarity* or *dissimilarity index* (e.g. Agresti, 2002, pp. 329–330), which aims to quantify lack of model fit by estimating the smallest fraction of the population under study that would need to be re-classified in order to make the fitted model exactly correct. The index of dissimilarity thus has a fairly direct interpretation in terms of the magnitude of departures from the model, and the statistic itself is simple to compute from model residuals (Section 2 below). These appealing properties have led to routine use of the index of dissimilarity for model assessment, especially in social science where the computed statistic is sometimes referred to simply as 'Delta' (e.g. Erola and Moisio 2007; Jonsson et al. 2009; Pfeffer 2008; Uggen and Blackstone 2004; Wells et al. 2003). The Delta statistic is used as a supplement to, rather than a replacement for, model-selection criteria such as those based on the log likelihood. The present work explores the definition and estimation of the index of dissimilarity in a fairly general setting.

Previously the sample index of dissimilarity has been used mainly as a descriptive device, and in particular no measure of the associated uncertainty, such as an estimated standard error, has been available. A primary aim of this paper is to rectify this deficiency. An approximate variance formula is derived in Section 4, based on large-sample arguments. The same large-sample considerations also suggest an approximate bias correction, given in Section 5. These results apply to general parametric models for data collected

*Corresponding address: Department of Statistics, London School of Economics, Houghton St, London WC2A 2AE, UK. Tel: +44 20 7955 6835; fax: +44 20 7955 7005.

Email addresses: j.kuha@lse.ac.uk (Jouni Kuha), d.firth@warwick.ac.uk (David Firth)

Submitted for publication *This version 1 May 2009*

by multinomial or Poisson sampling, including any loglinear and log-multiplicative models; and in keeping with the simplicity of the index itself, the variance and bias formulae are straightforward to compute from quantities made available by standard software for such models. As a preliminary to all of this work, the definition of the index of dissimilarity itself is placed under scrutiny, and it is suggested that often a standardized version of the index will be of most interest (Section 3). The accuracy of the various approximations and the effects of standardization are investigated in Section 6 using simulation, based on data from six countries in a large-scale study of social class mobility.

2. Index of dissimilarity and models for categorical data

Suppose that $\mathbf{Y} = (Y_1, \ldots, Y_K)'$ is a vector of K observed cell counts corresponding to the cells of a possibly multidimensional contingency table with a sample size $N = \sum_{i=1}^{K} Y_i$ and observed cell proportions $\mathbf{p} = (p_1, \ldots, p_K)' = \mathbf{Y}/N$. The corresponding vectors of fitted counts and proportions for some model M of interest estimated from the data are denoted by $\hat{\mathbf{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_K)'$ and $\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \ldots, \hat{\pi}_K)' = \hat{\mathbf{Y}}/N$. The index of dissimilarity for the fitted model is defined as

$$\hat{\Delta} = \frac{\sum_{i=1}^{K} |Y_i - \hat{Y}_i|}{2N} = \frac{\sum_{i=1}^{K} |p_i - \hat{\pi}_i|}{2} = \frac{\hat{\boldsymbol{\delta}}' \hat{\mathbf{e}}}{2} \tag{1}$$

where $\hat{\mathbf{e}} = (\hat{e}_1, \ldots, \hat{e}_K)' = (p_1 - \hat{\pi}_1, \ldots, p_K - \hat{\pi}_K)'$ are the raw residuals, and $\hat{\boldsymbol{\delta}} = (\hat{\delta}_1, \ldots, \hat{\delta}_K)' = (\mathrm{sgn}(\hat{e}_1), \ldots, \mathrm{sgn}(\hat{e}_K))'$ indicate their signs. We assume that the fitted model has the property that $\sum_i \hat{Y}_i = N$, so that the residuals sum to zero and $\hat{\boldsymbol{\pi}}$ are the fitted cell proportions. This will be true, in particular, for hierarchical loglinear and log-multiplicative models fitted by maximum likelihood, which are the main focus of this article. The index $\hat{\Delta}$ can then be interpreted as the smallest proportion of observations in $\mathbf{Y}$ that would need to re-classified to other cells to make all observed cell counts exactly equal to the fitted values.

A version of the index of dissimilarity has a long history in sociology and human geography as a measure of residential and geographic segregation. According to Goodman and Kruskal (1959), the index was first suggested by Gini (1914); other early references include Jahn et al. (1947) and Duncan and Duncan (1955); see also White (1986) for a more recent introduction and further references on segregation indices. Suppose that we want to compare the distributions of two groups, for example blacks and whites, across C locations such as schools or neigborhoods. The populations are completely segregated if no members of the two groups share the same location, and completely unsegregated if the proportions of the two groups are the same in every location. The segregation index of dissimilarity is defined as

$$D = \frac{1}{2} \sum_{j=1}^{C} \left| \frac{Y_{1j}}{N_1} - \frac{Y_{2j}}{N_2} \right| \tag{2}$$

where Y_{ij} is the number of members of group i in location j, and N_i is total number of members of group i in the sample. The index (2) can be interpreted as the general measure (1) where the observed data are one of the rows of the group-by-location table and the 'model' is given by the column proportions in the other row. Alternatively, D is $\hat{\Delta}$ for the independence model for the two-way table, divided by $N^2/(2N_1N_2)$, which is the maximum value that $\hat{\Delta}$ can achieve for the independence model when the row totals are regarded as fixed; for more general models the maximum achievable $\hat{\Delta}$ is usually not available in a closed form.

The statistic $\hat{\Delta}$ estimates a corresponding population quantity. Suppose that the true distribution of $\mathbf{Y}$ given N is multinomial with cell probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)'$, so that $\mathbf{p}$ converges to $\boldsymbol{\pi}$ as N increases. It is assumed that $\pi_i > 0$ for all i. If the table has structural zeros for which the cell probability is known to be zero, these contribute nothing to (1) and can be omitted from $\mathbf{Y}$.

Suppose further that the fitted proportions $\hat{\boldsymbol{\pi}}$ converge to $\boldsymbol{\pi}^* = (\pi_1^*, \ldots, \pi_K^*)'$, which differ from $\boldsymbol{\pi}$ unless model M is the true model. The population value of the index of dissimilarity is then defined as

$$\Delta = \frac{\sum_{i=1}^{K} |\pi_i - \pi_i^*|}{2}, \tag{3}$$

which is consistently estimated by $\hat{\Delta}$. This is the *total variation distance* between the two discrete distributions with probabilities $\boldsymbol{\pi}$ and $\boldsymbol{\pi}^*$ (Feller 1971; the definition there is for 2Δ). It can also be interpreted as $\Delta = \sup_{A \subset \mathcal{K}} |\pi(A) - \pi^*(A)|$ where $\pi(A) = \sum_{i \in A} \pi_i$, $\pi^*(A) = \sum_{i \in A} \pi_i^*$ and $\mathcal{K} = \{1, 2, \ldots, K\}$.

The population value Δ depends on the true probabilities $\boldsymbol{\pi}$ and the model M under consideration, and also on the method used to fit M to obtain $\hat{\mathbf{Y}}$. One implication of the definition is that fitted cell probabilities $\hat{\boldsymbol{\pi}}$, in particular those given by a maximum likelihood fit, do not in general converge to the value $\boldsymbol{\pi}^*$ that would minimize (3) for M; for this it would be necessary to base estimation on minimization of the sum of the absolute values of the residuals. But then the fitted counts would not satisfy $\sum_i \hat{Y}_i = N$, which is important for the interpretation of $\hat{\Delta}$. Another variant would be to consider measures of fit based on squared rather than absolute residuals; these could then be related to the τ_b measure of association of Goodman and Kruskal (1954, 1979) as in Haberman (1982). However, the substantive interpretation of $\hat{\Delta}$ would again be lost. We will not pursue these approaches here, but will concentrate on examining the properties of the commonly-used version of $\hat{\Delta}$ defined in (1).

An equivalent and often more convenient way to describe the distribution of the cell counts $\mathbf{Y}$ is to treat them as independent Poisson random variables with means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)' = N\boldsymbol{\pi}$, subject to the constraint $\sum_i \mu_i = N$ (Fisher, 1925; Birch, 1963). Here the two formulations will be used interchangeably. Even when the sample size is actually random and the Poisson thus the appropriate distribution, a standardization argument given in Section 3 will motivate conditioning on at least N in all cases. Conversely, if the sampling design is such that other margins than just N are fixed, the computations will always be conditional on at least these margins.

3. Standardization

Consider for the moment a loglinear parametrization

$$\log \mu_i = \mathbf{X}_{1i}'\boldsymbol{\theta}_1 + \mathbf{X}_{2i}'\boldsymbol{\theta}_2 \tag{4}$$

for the cell means of the true model. Here $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$ is a $K \times 1$ vector of parameters and $\mathbf{X}_i' = (\mathbf{X}_{1i}', \mathbf{X}_{i2}')$ are the rows of a nonsingular $K \times K$ design matrix $\mathbf{X}_S = [\mathbf{X}_1\ \mathbf{X}_2]$. Without loss of generality we assume that the parameters are subject to *treatment constraints* where parameters corresponding to effects involving the first level of any variable are excluded from $\boldsymbol{\theta}$ and the corresponding columns not included in $\mathbf{X}_S$. The model is saturated, so that no restrictions are imposed on the true cell means. The minimal sufficient statistics for $\boldsymbol{\theta}$ are $\mathbf{S}_S = (\mathbf{S}_1', \mathbf{S}_2')' = \mathbf{X}_S'\mathbf{Y}$. These are cell counts in a set of marginal tables obtained from $\mathbf{Y}$.

The distribution of $\mathbf{Y}$ can be parametrized in terms of the canonical parameters $\boldsymbol{\theta}$ or the expectation parameters $\boldsymbol{\eta} = (\boldsymbol{\eta}_1', \boldsymbol{\eta}_2')' = \mathrm{E}(\mathbf{S}_S)$. A further alternative is a mixed parametrization $\boldsymbol{\psi} = (\boldsymbol{\eta}_1', \boldsymbol{\theta}_2')'$. The relationships between the parametrizations are one-to-one when all μ_i are positive. Furthermore, $\boldsymbol{\eta}_1$ and $\boldsymbol{\theta}_2$ are variation-independent and orthogonal (see, for example, Barndorff-Nielsen and Cox 1994).

The most common class of models for contingency tables are loglinear models of the form

$$\log \mu_i = \mathbf{X}_{1i}'\boldsymbol{\theta}_1, \tag{5}$$

obtained by setting $\boldsymbol{\theta}_2 = \mathbf{0}$ in (4). The model is assumed to be hierarchical, so that for any marginal total included in $\mathbf{S}_1$, all corresponding lower-order margins in $\mathbf{S}_S$ are also included. Suppose that the parameter estimates and fitted values are obtained by maximum likelihood estimation, and denote the estimates of $\boldsymbol{\psi}$ by $\tilde{\boldsymbol{\psi}} = (\tilde{\boldsymbol{\eta}}_1', \tilde{\boldsymbol{\theta}}_2')'$ for the saturated model and $\hat{\boldsymbol{\psi}} = (\hat{\boldsymbol{\eta}}_1', \mathbf{0}')'$ for model (5). The estimating equations guarantee that $\tilde{\boldsymbol{\eta}}_1 = \hat{\boldsymbol{\eta}}_1 = \mathbf{S}_1$, so that both models always fit the observed lower-order margins $\mathbf{S}_1$ perfectly. Given this, it is desirable that any statistic that purports to characterize the lack of fit of the fitted model M should depend only on $\tilde{\boldsymbol{\theta}}_2$ and in particular on how different $\tilde{\boldsymbol{\theta}}_2$ is from zero. The usual index of dissimilarity, however, does not have this property. Both its population value and the estimate $\hat{\Delta}$ depend not only on the associations constrained by the model but also on the margins fitted by it. An illustration of this is given by the following simple example.

Example: Independence model for a 2×2 table
Consider a 2×2 contingency table with true cell probabilities, cell means and observed counts π_{ij}, $\mu_{ij} = N\pi_{ij}$ and Y_{ij} $(i, j = 1, 2)$ respectively. Suppose that the model of interest M is the independence model, for which the fitted counts are $\hat{Y}_{ij} = Y_{i+}Y_{+j}/N$, where the subscript + indicates summation. In this simple case all the important properties of $\hat{\Delta}$ can be examined in closed form. Suppose that the model is expressed in a loglinear form where all parameters corresponding to the first row or column are set to zero. The sufficient statistics are then $\mathbf{S}_1 = (N, Y_{2+}, Y_{+2})'$ and the corresponding expectation parameters $\boldsymbol{\eta}_1 = (\mu_{++}, \mu_{2+}, \mu_{+2})'$. The remaining canonical parameter is the log odds ratio $\theta_2 = \log(\pi_{11}\pi_{22}/\pi_{12}\pi_{21})$, which is set to zero by the independence model.

The fitted probabilities $\hat{\pi}_{ij} = \hat{Y}_{ij}/N$ converge to $\pi^*_{ij} = \pi_{i+}\pi_{+j}$, and the true value of Δ is $\Delta = \sum_{i,j} |\pi_{ij} - \pi^*_{ij}|/2$. In this case Δ simplifies further, because all the residuals have the same absolute value. Thus $\Delta = 2|e|$ where, for example, $e = \pi_{22} - \pi^*_{22}$. By solving for π_{22} in the quadratic equation

$$\exp(\theta_2) = \frac{\pi_{22}(1 - \pi_{2+} - \pi_{+2} + \pi_{22})}{(\pi_{2+} - \pi_{22})(\pi_{+2} - \pi_{22})}, \tag{6}$$

Δ can then be expressed as a function of $\boldsymbol{\eta}_1$ and θ_2. Figure 1 illustrates how Δ depends on the margins $\boldsymbol{\eta}_1$ as well as the odds ratio $\exp(\theta_2)$. For the same value of the odds ratio, the index can take values between zero and some maximum, achieved at $\pi_{2+} = \pi_{+2} = 1/2$.

[Figure 1 around here.]

The dependence of $\hat{\Delta}$ on the margins complicates its interpretation in at least two situations. The first occurs when $\hat{\Delta}$ is computed from a table where some margins have been fixed by the sampling design. The sample index then does not estimate Δ in the true population. Second, values of $\hat{\Delta}$ for the same model M computed for samples from different populations are not directly comparable, because differences in lack of fit are confounded by differences in the marginal distributions.

Both of these problems can be remedied by table standardization. The idea is to create a new table where the lower-order margins fitted by the model are equal to some benchmark values, but the higher-order associations are left unchanged. Specifically, a standardized table with cell proportions $\mathbf{p}_0$ satisfies the conditions

$$\mathbf{X}_0'\mathbf{p}_0 = \mathbf{S}_0 \tag{7}$$

$$\mathbf{W}' \log \mathbf{p}_0 = \mathbf{W}' \log \mathbf{p}, \tag{8}$$

where $\mathbf{S}_0$ is a given set of non-zero marginal proportions and log denotes the element-wise logarithmic operator (see Imrey et al. 1981; Little and Wu 1991). The matrix $\mathbf{X}_0$ is a design matrix corresponding to the margins included in $\mathbf{S}_0$; here we take $\mathbf{X}_0 = \mathbf{X}_1$ so that $N\mathbf{S}_0$ contains the same margins as $\mathbf{S}_1$, but this will be relaxed in the next Section. The matrix $\mathbf{W}$ is an orthocomplement matrix of $\mathbf{X}_0$ (i.e. $\mathbf{W}'\mathbf{X}_0 = \mathbf{0}$ and $[\mathbf{X}_0\ \mathbf{W}]$ is non-singular). For example, if the number of columns in $\mathbf{X}_0 = \mathbf{X}_1$ is K_1, one possible $\mathbf{W}'$ is given by the last $K - K_1$ rows of $\mathbf{X}_S^{-1} = [\mathbf{X}_1\, \mathbf{X}_2]^{-1}$.

Condition (7) guarantees that the standardized table with proportions $\mathbf{p}_0$ matches the benchmark margins $\mathbf{S}_0$, while (8) implies that the higher-order associations $\mathbf{W}' \log \mathbf{p}_0 = \tilde{\boldsymbol{\theta}}_2$ are the same as for the original proportions $\mathbf{p}$. The index of dissimilarity for a model calculated for the standardized table is the standardized statistic $\hat{\Delta}_0 = \hat{\Delta}_0(\mathbf{Y}; \mathbf{S}_0)$, which is a function of both the observed data $\mathbf{Y}$ and the standard margins $\mathbf{S}_0$. The unstandardized value of $\hat{\Delta}$ is equal to $\hat{\Delta}_0(\mathbf{Y}; \mathbf{S}_1/N)$.

The choice of appropriate benchmark margins $\mathbf{S}_0$ depends on the situation. In the stratified sampling case, the table should be standardized to the population margins if these are at least approximately known. In the case of comparing populations, one possibility is to use the margins of one of the samples as a benchmark for the others. An alternative is to use some general standard such as tables which are as uniform as possible. Such margins are, of course, entirely artificial, but they may seem a convenient default choice; this is considered further in Section 6.

The standardized index $\hat{\Delta}_0$ estimates Δ for a population where the true distribution of $\mathbf{Y}_0$ is specified by parameters $(N_0\mathbf{S}_0', \boldsymbol{\theta}_2')'$, where N_0 is the (arbitrary) sample size of the standardized table. When standardization is done to adjust for the sampling design, this is approximately the true Δ in the target population.

In the comparison case, on the other hand, it refers to an artificial population, so only the relative values of $\hat{\Delta}_0$ are of interest. The unstandardized $\hat{\Delta}$ is still required for estimating the true population Δ for each sample.

Table standardization (also known as *raking* the table) can be carried out using standard iterative algorithms for fitting loglinear models (see, for example, Agresti 2002). The first step is to generate a table of counts $\mathbf{Y}_{0*}$ specified by $\boldsymbol{\psi} = (N_0\mathbf{S}^{0\prime}, \boldsymbol{\theta}_2^{*\prime})'$ for some arbitrary $\boldsymbol{\theta}_2^*$. With iterative proportional fitting (IPF), the standardized table $\mathbf{Y}_0$ is then obtained by fitting model (5) to $\mathbf{Y}_{0*}$, with the initial values given by $\mathbf{Y}$ (the algorithm actually uses only the sufficient statistics $N_0\mathbf{S}_0$, but most IPF routines in standard packages expect starting values in the form of a full set of cell counts). If the Newton–Raphson algorithm is used, the Poisson regression model (5) is first fitted to $\mathbf{Y}_{0*}$. The resulting fitted values $\hat{\mathbf{Y}}_0$ correspond to parameter values $\boldsymbol{\psi} = (N_0\mathbf{S}_0', \mathbf{0}')'$. The model is then fitted again, now to $\hat{\mathbf{Y}}_0$ and using $\log Y_i$ as offsets; $\mathbf{Y}_0$ is given as the vector of fitted values.

4. Inference

4.1. Introduction

Statistical inference for the segregation index of dissimilarity D, given by (2), has been considered by, for example, Cortese et al. (1976), Inman and Bradley (1991), Ransom (2000) and Mulekar et al. (2008). Two cases may be of interest here. One is the derivation of the standard error of the estimated D under any population value of it. The other is the distribution of D when the population value is 0. This leads to a goodness-of-fit test for the hypothesis of no segregation in the population. The distributional results are then complicated by the fact that under the null hypothesis, the population D is on the boundary of the parameter space. Such tests are not available for the general $\hat{\Delta}$.

Here we consider the distribution of the general $\hat{\Delta}$, given by (1), when the fitted model M does not hold in the population: the true distribution is the most general, saturated model (4) with $\boldsymbol{\theta}_2$ not constrained to be zero. This is the most relevant case in applications where $\hat{\Delta}$ is of interest, where it is rarely realistic to assume that the model under consideration is exactly correct in the population. Furthermore, better-behaved goodness-of-fit tests for models for contingency tables are already available, so there is little need for one based on $\hat{\Delta}$.

The variance results given below are not limited to the loglinear models (5). In particular, they apply also to log-multiplicative models which aim to simplify interactions using bilinear structures: examples are the well-known Goodman RC and RC(m) association models (see Goodman, 1985) and the *uniform difference* (unidiff) models of Erikson and Goldthorpe (1992) and Xie (1992). The RC and unidiff models are both of the form

$$\log \mu_i = \mathbf{X}_{1i}'\boldsymbol{\theta}_1 + (1 + \mathbf{U}_{1i}'\boldsymbol{\beta}_2)\,(\mathbf{U}_{2i}'\boldsymbol{\beta}_3) \tag{9}$$

where $\mathbf{U}_{i1}'$ and $\boldsymbol{U}_{2i}'$ are rows of two design matrices $\mathbf{U}_1$ and $\mathbf{U}_2$, and $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ are parameter vectors; a general expression which also covers RC(m) models is a straightforward extension.

Another broad class of models for categorical data are latent class models (see e.g. Hagenaars and McCutcheon 2002). Our general variance results hold also for them, but specific expressions are not given here.

The discussion of the previous section suggests that the properties of $\hat{\Delta}$ as an estimator of Δ for a loglinear model should be considered conditionally on $\mathbf{S}_1 = \hat{\boldsymbol{\eta}}_1$. The distribution of $\hat{\Delta}$ given $\mathbf{S}_1$ then depends on the data only through $\tilde{\boldsymbol{\theta}}_2$. Similarly, if any model fits some set of margins $\mathbf{S}$ perfectly, as (9) does for $\mathbf{S} = \mathbf{X}_1'\mathbf{Y}$, inference for $\hat{\Delta}$ should be conditional on those margins. Since they can be obtained with little extra effort, we will also present variance results for $\hat{\Delta}$ (and the standardized $\hat{\Delta}_0$) conditional on less than the full set of fitted margins, including the special case of the unconditional variances. This will also allow us to consider the distributions of differences of $\hat{\Delta}$ for different models, conditional on shared fitted margins.

Our approach is to obtain an estimate for the asymptotic conditional variance of $\hat{\Delta}$ under the saturated model and use it to compute symmetric confidence intervals for Δ. The derivation is based on approximate normality of $\hat{\Delta}$, and the intervals are thus expected to behave best when this assumption is justified. The simulations reported below indicate that the results are satisfactory whenever the sample size is sufficiently

large that model lack-of-fit is clearly detectible, i.e. in precisely the situations where a summary such as Δ is of most interest. The asymptotic approach is to a large extent motivated by the desire to keep the computations as simple as possible. Ideally, a confidence interval for Δ should be, like $\hat{\Delta}$ itself, computable using little more than standard output for the fitted model. Improved variance estimates and confidence intervals could perhaps be obtained using resampling methods or other computationally intensive techniques, but these will not be considered here.

Example: independence model for a 2×2 table (continued)
In the 2×2 example considered in Section 3 the exact distribution of $\hat{\Delta}$ is available. Like the true Δ, the estimate can in this case be expressed as $\hat{\Delta} = 2|\hat{e}|$ where $\hat{e} = (Y_{22} - \hat{Y}_{22})/N = p_{22} - \hat{\pi}_{22}$, say. Since $\hat{Y}_{22}$ is a constant given $\mathbf{S} = \mathbf{S}_1 = (N, Y_{2+}, Y_{+2})'$, the distribution of $\hat{\Delta}$ is determined by the distribution of Y_{22} given both margins of the table. This is an extended hypergeometric distribution (e.g. Breslow and Day 1980, sec. 4.3; Cox and Snell 1989, sec. 2.3). In all but very small samples, it is well approximated by a normal distribution with mean μ_{22} and variance $\sigma_y^2 = (\sum \mu_{ij}^{-1})^{-1}$. Since $\hat{\pi}_{22}$ converges to π_{22}^*, the conditional distribution of $\hat{e}$ is approximately normal with mean $\mu_e = \pi_{22} - \pi_{22}^*$ and variance $\sigma_e^2 = \sigma_y^2/N^2$, estimated by $\hat{e}$ and $\hat{\sigma}_e^2 = N^{-1}(\sum p_{ij}^{-1})^{-1}$ respectively. The distribution of $|\hat{e}|$ is then folded normal (see Johnson et al., 1995, p. 454) with mean and variance

$$\mu_{|e|} = 2\sigma_e\phi(\mu_e/\sigma_e) + \{1 - 2\Phi(-\mu_e/\sigma_e)\}\,\mu_e \tag{10}$$

$$\sigma_{|e|}^2 = \sigma_e^2 - (\mu_{|e|}^2 - \mu_e^2) \tag{11}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are respectively the density and distribution functions of the standard normal distribution.

The variance estimates considered in this section are derived using the basic assumption that the distribution of $\hat{e}$ is such that the sign of $\hat{e}$ is effectively constant. In the 2×2 case this implies that $\sigma_{|e|}^2 \approx \sigma_e^2$ and thus the variance of $\hat{\Delta}$ can be estimated by $\widehat{\text{var}}(\hat{\Delta}) = 4\,\hat{\sigma}_e^2$. From (10) and (11) it can be seen that such an approximation is justified when σ_e is small and the standardized residual μ_e/σ_e is large. The basic assumption is most likely to fail in small samples and when the model fits well, i.e. θ_2 is close to zero, so that μ_e is small.

4.2. Approximate variance: the general case

The same approach can be applied in generality. We will again use asymptotic results for the distribution of $\mathbf{Y}$ given N. Suppose the fitted probabilities $\hat{\boldsymbol{\pi}} = \hat{\boldsymbol{\pi}}(\hat{\boldsymbol{\beta}})$ for a model are functions of the estimate $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$ of a p-dimensional parameter $\boldsymbol{\beta}$. For example, the model for $\boldsymbol{\pi}$ corresponding to the loglinear model (5) is the multinomial logistic model given by

$$\hat{\boldsymbol{\pi}} = \exp(\mathbf{X}\hat{\boldsymbol{\beta}})/\{\mathbf{1}'\exp(\mathbf{X}\hat{\boldsymbol{\beta}})\} \tag{12}$$

where $\exp(\cdot)$ and division denote elementwise operations, $\mathbf{1}$ is a $K \times 1$ vector of ones, and $\mathbf{X}$ is obtained from $\mathbf{X}_1$ by omitting the initial column of ones and $\hat{\boldsymbol{\beta}}$ from $\hat{\boldsymbol{\theta}}_1$ by omitting its first element (the constant term in the loglinear model). Similarly, the model corresponding to (9) is

$$\hat{\boldsymbol{\pi}} = \frac{\exp\{\mathbf{X}\hat{\boldsymbol{\beta}}_1 + (\mathbf{1} + \mathbf{U}_1\hat{\boldsymbol{\beta}}_2) \odot (\mathbf{U}_2\hat{\boldsymbol{\beta}}_3)\}}{\mathbf{1} + \exp\{\mathbf{X}\hat{\boldsymbol{\beta}}_1 + (\mathbf{1} + \mathbf{U}_1\hat{\boldsymbol{\beta}}_2) \odot (\mathbf{U}_2\hat{\boldsymbol{\beta}}_3)\}} \tag{13}$$

where $\odot$ denotes the Hadamard product, and $\mathbf{X}$ is again obtained by omitting the first column of $\mathbf{X}_1$ in (9) and $\hat{\boldsymbol{\beta}}_1$ similarly by omitting the constant term in $\boldsymbol{\theta}_1$. For this model $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1', \hat{\boldsymbol{\beta}}_2', \hat{\boldsymbol{\beta}}_3')'$.

Suppose that the fitted model has the property that always $\mathbf{X}'\hat{\boldsymbol{\pi}} = \mathbf{X}'\mathbf{p} = \mathbf{S}_M$, where $\mathbf{X}$ is chosen so that $\mathbf{S}_M$ is the largest set of nonredundant margins for which this is true (since the model is always conditional on N, it is not included in $\mathbf{S}_M$ and thus the columns of $\mathbf{X}$ do not include $\mathbf{1}$). For example, the matrix $\mathbf{X}$ in both (12) and (13) is of this kind. Our aim is to obtain the variance of $\hat{\Delta}$ conditional on all appropriate subsets $\mathbf{Z}'\mathbf{p} = \mathbf{S} \subseteq \mathbf{S}_M$, including $\mathbf{S}_M$ itself (full conditioning) and an empty set (unconditional variance).

The basic idea of the variance calculation is to use the asymptotic normal distribution of $\mathbf{p}$ and apply the delta method to $\hat{\boldsymbol{\pi}}$ and $\mathbf{S}$, considering them as functions of $\mathbf{p}$. The asymptotic variance matrix of $\mathbf{p}$ given N under multinomial sampling is

$$\mathrm{var}(\mathbf{p}) = \{\mathrm{Diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'\}/N = \mathbf{V}_\pi/N$$

(see, for example, chapter 14 of Agresti (2002) for these and some of the other results used below). This is consistently estimated by $\widehat{\mathrm{var}}(\mathbf{p}) = \{\mathrm{Diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}'\}/N = \mathbf{V}_p/N$. The joint distribution of $\mathbf{p}$ and $\mathbf{S} = \mathbf{Z}'\mathbf{p}$ is also asymptotically normal, and the asymptotic conditional variance matrix of $\mathbf{p}$ given $\mathbf{S}$ is obtained from it as

$$\mathrm{var}(\mathbf{p}|\mathbf{S}) = \{\mathbf{V}_\pi - \mathbf{V}_\pi\mathbf{Z}(\mathbf{Z}'\mathbf{V}_\pi\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{V}_\pi\}/N \tag{14}$$

(note that $\mathrm{var}(\mathbf{S}) = \mathbf{Z}'\mathbf{V}_\pi\mathbf{Z}/N$ is nonsingular when all elements of $\mathbf{S}$ are positive).

Suppose that $\hat{\boldsymbol{\beta}}$ are solutions to p estimating equations $\mathbf{u}(\hat{\boldsymbol{\beta}}, \mathbf{p}) = \mathbf{u} = \mathbf{0}$. Using implicit function differentiation we then obtain

$$\mathbf{A} = \frac{\partial\hat{\boldsymbol{\pi}}}{\partial\mathbf{p}} = \frac{\partial\hat{\boldsymbol{\pi}}}{\partial\hat{\boldsymbol{\beta}}}\,\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\mathbf{p}} = \frac{\partial\hat{\boldsymbol{\pi}}}{\partial\hat{\boldsymbol{\beta}}}\left\{-\left[\frac{\partial\mathbf{u}}{\partial\hat{\boldsymbol{\beta}}}\right]^{-1}\frac{\partial\mathbf{u}}{\partial\mathbf{p}}\right\},$$

where $\partial\hat{\boldsymbol{\pi}}/\partial\mathbf{p}$ denotes the $K \times K$ matrix with elements $\partial\hat{\pi}_i/\partial p_j$, and the other matrices are defined similarly. This exists if $\partial\mathbf{u}/\partial\hat{\boldsymbol{\beta}}$ is non-singular, i.e. when $\hat{\boldsymbol{\beta}}$ itself exists. In particular, for maximum likelihood estimation $\mathbf{u}$ is the score function and $-\partial\mathbf{u}/\partial\hat{\boldsymbol{\beta}}$ is the observed information matrix. For the log-linear/ multinomial logistic model (12), $\partial\hat{\boldsymbol{\pi}}/\partial\hat{\boldsymbol{\beta}} = \{\mathrm{Diag}(\hat{\boldsymbol{\pi}}) - \hat{\boldsymbol{\pi}}\hat{\boldsymbol{\pi}}'\}\mathbf{X} = \mathbf{V}_{\hat{\pi}}\mathbf{X}$, $\mathbf{u}(\hat{\boldsymbol{\beta}}, \mathbf{p}) = \mathbf{X}'(\mathbf{p} - \hat{\boldsymbol{\pi}})$, and $\mathbf{A} = (\mathbf{V}_{\hat{\pi}}\mathbf{X})\,(\mathbf{X}'\boldsymbol{V}_{\hat{\pi}}\mathbf{X})^{-1}\,\mathbf{X}'$. For model (13), $\partial\hat{\boldsymbol{\pi}}/\partial\hat{\boldsymbol{\beta}} = \mathbf{V}_{\hat{\pi}}\mathbf{E}$ where $\mathbf{E} = [\mathbf{X} \;\; \mathrm{Diag}(\mathbf{U}_2\hat{\boldsymbol{\beta}}_3)\mathbf{U}_1 \;\; \{\mathbf{I} + \mathrm{Diag}(\mathbf{U}_1\hat{\boldsymbol{\beta}}_2)\}\mathbf{U}_2]$, $\mathbf{u}(\hat{\boldsymbol{\beta}}, \mathbf{p}) = \mathbf{E}'(\mathbf{p} - \hat{\boldsymbol{\pi}})$, and $\mathbf{A} = (\mathbf{V}_{\hat{\pi}}\,\mathbf{E})\,\mathbf{F}^{-1}\,\mathbf{E}'$ where

$$\mathbf{F} = \mathbf{E}'\mathbf{V}_{\hat{\pi}}\mathbf{E} - \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{U}_1'\mathrm{Diag}(\hat{\mathbf{e}})\mathbf{U}_2 \\ \mathbf{0} & \mathbf{U}_2'\mathrm{Diag}(\hat{\mathbf{e}})\mathbf{U}_1 & \mathbf{0} \end{bmatrix}$$

and $\hat{\mathbf{e}} = \mathbf{p} - \hat{\boldsymbol{\pi}}$.

Defining $\mathbf{P} = (\mathbf{p}', \hat{\boldsymbol{\pi}}')'$, we obtain $\partial\mathbf{P}/\partial\mathbf{p} = [\mathbf{I} \;\; \mathbf{A}']' = \mathbf{B}$. The delta method then shows that the asymptotic variance matrix of $\mathbf{P}$ given $\mathbf{S}$ is $\mathrm{var}(\mathbf{P}|\mathbf{S}) = \mathbf{B}\,\mathrm{var}(\mathbf{p}|\mathbf{S})\,\mathbf{B}'$, and the asymptotic conditional variance matrix of the residuals is

$$\mathrm{var}(\hat{\mathbf{e}}|\mathbf{S}) = \mathrm{var}(\mathbf{p}|\mathbf{S}) + \mathrm{var}(\hat{\boldsymbol{\pi}}|\mathbf{S}) - \mathrm{cov}(\mathbf{p}, \hat{\boldsymbol{\pi}}|\mathbf{S}) - \mathrm{cov}(\mathbf{p}, \hat{\boldsymbol{\pi}}|\mathbf{S})'. \tag{15}$$

For model (12), the formulas simplify if $\mathbf{Z} = \mathbf{X}$, so that we are conditioning on the full margins for the fitted model. Then $\mathrm{var}(\hat{\boldsymbol{\pi}}|\mathbf{S}) = \mathrm{cov}(\mathbf{p}, \hat{\boldsymbol{\pi}}|\mathbf{S}) = \mathbf{0}$ and $\mathrm{var}(\hat{\mathbf{e}}|\mathbf{S}) = \mathrm{var}(\mathbf{p}|\mathbf{S})$. Another special case is the unconditional variance matrix of $\hat{\mathbf{e}}$, which is obtained by replacing $\mathrm{var}(\mathbf{p}|\mathbf{S})$ with $\mathrm{var}(\mathbf{p}) = \mathbf{V}_\pi/N$.

Finally, a consistent estimate of $\mathrm{var}(\hat{\Delta}|\mathbf{S})$ is obtained by estimating $\boldsymbol{\pi}$ by $\mathbf{p}$ in (14), and treating $\hat{\boldsymbol{\delta}}$ as a constant in (1). The variance estimate is given by

$$\widehat{\mathrm{var}}(\hat{\Delta}|\mathbf{S}) = \hat{\boldsymbol{\delta}}'\,\widehat{\mathrm{var}}(\hat{\mathbf{e}}|\mathbf{S})\,\hat{\boldsymbol{\delta}}/4. \tag{16}$$

4.3. Standardized index of dissimilarity

A small modification of these results gives an estimate for the variance of standardized statistic $\hat{\Delta}_0$, conditional on margins $\mathbf{S}$ in the original (unstandardized) table. Here the standardization can only be up to the margins $\mathbf{S}_M$ fitted by the model, so that $\mathbf{X}_{0*}'\mathbf{p} \subseteq \mathbf{S}_M$, where $\mathbf{X}_{0*}$ is obtained from $\mathbf{X}_0$ in (7) by omitting the column of ones. Let $\mathbf{p}_0$ denote the cell proportions in a standardized table satisfying (7)–(8), and $\hat{\boldsymbol{\pi}}_0$ the fitted probabilities for model M fitted to $\mathbf{p}_0$. The standardized index of dissimilarity is then $\hat{\Delta}_0 = \hat{\boldsymbol{\delta}}_0'\hat{\mathbf{e}}_0/2$,

where $\hat{\mathbf{e}}_0 = \mathbf{p}_0 - \hat{\boldsymbol{\pi}}_0$ and $\hat{\boldsymbol{\delta}}_0$ are the residuals and their signs for the model fitted to the standardized table. First, applying implicit function differentiation to (7)–(8) gives

$$\mathbf{C} = \frac{\partial \mathbf{p}_0}{\partial \mathbf{p}} = \mathbf{W}\,(\mathbf{W}'\mathbf{D}_{p0}^{-1}\mathbf{W})^{-1}\,\mathbf{W}'\mathbf{D}_p^{-1}$$

where $\mathbf{D}_{p0}$ and $\mathbf{D}_p$ are diagonal matrices with $\mathbf{p}_0$ and $\mathbf{p}$ respectively on the diagonal (see Imrey et al., 1981; Little and Wu, 1991). Then $\partial\hat{\boldsymbol{\pi}}_0/\partial\mathbf{p} = (\partial\hat{\boldsymbol{\pi}}_0/\partial\mathbf{p}_0)(\partial\mathbf{p}_0/\partial\mathbf{p}) = \mathbf{AC}$, where $\mathbf{A} = \partial\hat{\boldsymbol{\pi}}^0/\partial\mathbf{p}^0$ is of the same form as $\partial\hat{\boldsymbol{\pi}}/\partial\mathbf{p}$ for the same model. Defining $\mathbf{P}_0 = (\mathbf{p}_0', \hat{\boldsymbol{\pi}}_0')'$, $\partial\mathbf{P}_0/\partial\mathbf{p} = [\mathbf{C}'\ (\mathbf{AC})']' = \mathbf{B}$ and $\mathrm{var}(\mathbf{P}_0|\mathbf{S}) = \mathbf{B}\,\mathrm{var}(\mathbf{p}|\mathbf{S})\,\mathbf{B}'$, an estimate of $\mathrm{var}(\hat{\Delta}_0|\mathbf{S})$ is then obtained from (15)–(18) by replacing $\hat{\boldsymbol{\delta}}$ with $\hat{\boldsymbol{\delta}}_0$ and the conditional variances and covariances of $\mathbf{p}$ and $\hat{\boldsymbol{\pi}}$ with those of $\mathbf{p}_0$ and $\hat{\boldsymbol{\pi}}_0$.

4.4. Differences

The derivation can be extended to obtain an estimate for the variance of the difference $\hat{\Delta}_2 - \hat{\Delta}_1 = \{\hat{\boldsymbol{\delta}}_2'\hat{\mathbf{e}}_2 - \hat{\boldsymbol{\delta}}_1'\hat{\mathbf{e}}_1\}/2$, conditional on $\mathbf{S}$, where $\hat{\Delta}_1$ and $\hat{\Delta}_2$ are values of the index of dissimilarity for two (not necessarily nested) models M_1 and M_2, and $\mathbf{S}$ is a set of margins fitted by both models. Then

$$\widehat{\mathrm{var}}(\hat{\Delta}_2 - \hat{\Delta}_1|\mathbf{S}) = \widehat{\mathrm{var}}(\hat{\Delta}_2|\mathbf{S}) + \widehat{\mathrm{var}}(\hat{\Delta}_1|\mathbf{S}) - 2\,\widehat{\mathrm{cov}}(\hat{\Delta}_2, \hat{\Delta}_1|\mathbf{S}) \tag{17}$$

where

$$\begin{aligned}&\widehat{\mathrm{cov}}(\hat{\Delta}_2, \hat{\Delta}_1|\mathbf{S})\\ &\quad= \hat{\boldsymbol{\delta}}_2'\{\widehat{\mathrm{var}}(\mathbf{p}|\mathbf{S}) - \widehat{\mathrm{cov}}(\mathbf{p}, \hat{\boldsymbol{\pi}}_1|\mathbf{S}) - \widehat{\mathrm{cov}}(\mathbf{p}, \hat{\boldsymbol{\pi}}_2|\mathbf{S})' + \widehat{\mathrm{cov}}(\hat{\boldsymbol{\pi}}_2, \hat{\boldsymbol{\pi}}_1|\mathbf{S})\}\hat{\boldsymbol{\delta}}_1/4,\end{aligned} \tag{18}$$

$\widehat{\mathrm{cov}}(\hat{\boldsymbol{\pi}}_2, \hat{\boldsymbol{\pi}}_1|\mathbf{S}) = \mathbf{A}_2\,\widehat{\mathrm{var}}\,(\mathbf{p}|\mathbf{S})\mathbf{A}_1'$ and $\mathbf{A}_i = \partial\hat{\boldsymbol{\pi}}_i/\partial\mathbf{p}$.

5. Bias reduction

When the assumption that the residual signs $\hat{\boldsymbol{\delta}}$ are known constants fails, (10) and (11) show that the variance estimate (16) will overestimate the true variance of $\hat{\Delta}$ for the independence model for a 2×2 table. Those expressions also indicate that although $\hat{\Delta}$ is a consistent estimate of Δ, it will have an upward bias in finite samples. These findings almost always hold also in more general cases. The simulations reported below suggest that the bias is quite persistent and may often be a more serious problem than overestimation of the variance. It is thus of interest to consider modifications of $\hat{\Delta}$ to try to reduce the bias.

Bias correction based on solution of (10) directly is not simple, and neither is it necessarily desirable: some preliminary numerical work, not reported here, indicates that the bias-corrected estimator thus derived may have a severely inflated variance. Instead, we consider reducing the positive bias of each $|\hat{e}_i|$ by applying a simple shrinkage method, producing a new estimator of Δ in the generic form

$$\tilde{\Delta} = \tilde{\boldsymbol{\delta}}'\hat{\mathbf{e}}/2,$$

where $\tilde{\delta}_1, \ldots, \tilde{\delta}_K$ are appropriately defined shrinkage factors, with $\tilde{\delta}_i/\hat{\delta}_i \in (0, 1)$. A possible set of values for the $\{\tilde{\delta}_i\}$ can be obtained heuristically by re-writing (10) as

$$\mu_{|e|} = \mu_e\,[2\phi(\rho_e)/\rho_e + \{1 - 2\Phi(-\rho_e)\}]\,,$$

where $\rho_e = \mu_e/\sigma_e$. This suggests taking $\tilde{\delta}_i = \tilde{\delta}_i^{(1)}$, say, where

$$\tilde{\delta}_i^{(1)} = [2\phi(\hat{\rho}_{ei})/\hat{\rho}_{ei} + \{1 - 2\Phi(-\hat{\rho}_{ei})\}]^{-1} \quad (i = 1, \ldots, K),$$

and $\hat{\rho}_{ei} = \hat{e}_i/\hat{\sigma}_{ei}$ with $\{\hat{\sigma}_{ei}^2\}$ the diagonal elements of $\widehat{\mathrm{var}}(\hat{\mathbf{e}}|\mathbf{S})$ In what follows we in fact use an alternative, slightly stronger shrinkage formula $\tilde{\delta}_i = \tilde{\delta}_i^{(2)}$, say, where

$$\tilde{\delta}_i^{(2)} = 1 - 2\Phi\,(-\hat{\rho}_{ei}) \quad (i = 1, \ldots, K),$$

since it was found that slightly more shrinkage is required than is provided by $\tilde{\boldsymbol{\delta}}^{(1)}$. The two alternatives $\tilde{\boldsymbol{\delta}}^{(1)}$ and $\tilde{\boldsymbol{\delta}}^{(2)}$ both have the desired properties of signs that match those of $\hat{\mathbf{e}}$ and magnitudes in $(0, 1)$; and in fact their magnitudes are not very dissimilar, but $\tilde{\boldsymbol{\delta}}^{(2)}$ is always slightly smaller, with $1 < \tilde{\delta}_i^{(1)}/\tilde{\delta}_i^{(2)} < 1.571$ in all cases.

The estimates $\hat{\Delta}$ and $\tilde{\Delta}$ are asymptotically equivalent to first order. The variance of $\tilde{\Delta}$ can thus be estimated by (16) or alternatively by substituting $\tilde{\boldsymbol{\delta}}$ for $\hat{\boldsymbol{\delta}}$ in (16). In our simulations the latter variance estimator almost always gave confidence intervals which had worse coverage than intervals based on (16), so it is not considered further.

Table 1 shows results of a simulation study of the 2×2 example. Each row of the table is based on 1000 independent tables generated with the given N, π_{2+}, π_{+2}, and θ_2. The overestimation of both Δ and the standard error of $\hat{\Delta}$ is apparent, with the bias in $\hat{\Delta}$ being the more serious in many cases. As expected, the biases are largest when the sample size is small, the marginal distributions uneven or the odds ratio small. In these cases the bias of $\hat{\Delta}$ is reduced by $\tilde{\Delta}$, but there is little difference between the estimators in terms of the coverage of their respective confidence intervals.

[Table 1 around here]

6. Example and simulations

6.1. Cross-national comparisons of social mobility

The examples considered here are based on data on social mobility described by Erikson and Goldthorpe (1992), who combined data from a number of national mobility surveys conducted in the 1970s. The social class variables in each study were recoded to a common class schema to achieve a high degree of cross-national comparability. The data used in the core analyses of Erikson and Goldthorpe consisted of men from nine nations, divided into four ten-year birth cohorts. Each subject was classified into one of five 'classes of origin' (the subject's father's social class) and 'classes of destination' (the subject's own current class). The classes, which are based on occupational characteristics, are labelled white-collar workers, petty bourgeoisie, farm workers, skilled workers, and non-skilled workers.

Because simulations based on the full data set were computationally too demanding, we considered reduced versions of the data. The four cohorts were combined into two (ages 25–44 and 45–64), and only three nations were used in each simulation. Two groups of three countries are considered. The first consists of England, France and Sweden, three industrialized Western European countries which were the subject of a pilot study by Erikson et al. (1979). The second group of countries is Ireland, Hungary and Poland. All of these have large agricultural sectors, but ones that operated under very different conditions during the period when the data were collected.

Each of the two sets of data is a $3 \times 2 \times 5 \times 5$ table. The sample sizes are $26,655$ for the group of England, France and Sweden, and $40,057$ for Ireland, Hungary and Poland. The individual cell counts vary between 1 and 1205 and 2 and 4441 in the two samples respectively, and 14 and 16 of the 150 counts are less than 10. The data are given in an Appendix.

The main substantive interest in this example is in comparing patterns of social fluidity between origin and destination classes in different cohorts and nations. These are operationalized in the statistical models as associations between origin class (O) and destination class (D) given cohort (C) and nation (N). Table 2 shows some such models fitted to the two data sets. The first line of the table gives the model of 'perfect social fluidity', where O and D are conditionally independent given N and C. This model is known to be unrealistic, and is presented only as a baseline. The second model adds an association (two-way interaction) between O and D, and the others allow it to vary between nations or cohorts or both.

The models in Table 2 are loglinear, except for model V. This is the uniform difference (unidiff) model Erikson and Goldthorpe (1992); Xie (1992) for which

$$\log \mu_{ijkl} = \mathbf{x}'_{ijkl}\boldsymbol{\theta}_1 + \delta_{ij}^{NC}\,\gamma_{ikl}^{NOD}.$$

Here μ_{ijkl} denotes the expected count for the cell for nation i, cohort j, origin class k and destination class l, and $\mathbf{x}'_{ijkl}\boldsymbol{\theta}_1$ is identical to the linear predictor for model I. The term γ^{NOD}_{ikl} represents the general pattern of origin–destination association for nation i as in the corresponding terms of model IV, while δ^{NC}_{ij} describes the relative strength of that association for cohort j; to ensure identifiability, a constraint such as $\delta^{NC}_{i1} = 1$ is required. Extending standard notation for loglinear models, the unidiff model can be expressed concisely as $(NCO, NCD, (NC) * (NOD))$ The model is of the general log-multiplicative form (9) where $\mathbf{X}_1$ is the design matrix for model I, $\mathbf{U}_1$ is a 150×3 matrix with columns containing indicator variables for $(i = 1, 2, 3; j = 2)$ and $\mathbf{U}_2$ is a 150×48 matrix obtained from the design matrix of model IV by keeping only those terms corresponding to the $O \times D$ and $N \times O \times D$ interactions. The model fits exactly those margins also fitted by model I, so the standard errors for model V in Table 2 are conditional on these margins.

[Table 2 around here]

Except for the entirely inadequate model I, $\hat{\Delta}$ is fairly small for all of the models in Table 2. Even though the lack of fit of some of them would in such large samples be judged strongly significant by conventional goodness-of-fit tests, $\hat{\Delta}$ indicates that only two per cent or less of the observations would need to be reclassified to make the models fit perfectly. This result, which may seem surprising, is not uncommon in large contingency tables. The values of $\hat{\Delta}$ are estimated with fairly high precision, with the widths of 95% confidence intervals around one percentage point or less for these two data sets.

The estimates in Table 2 suggest, in agreement with earlier published analyses of these data, that cross-national variation in mobility patterns is stronger than differences between birth cohorts, at least for the group of Ireland, Hungary and Poland. The difference in $\tilde{\Delta}$ between models III and IV is 0.97 percentage points for these countries, with a 95 % confidence interval (conditional on the shared margins of model II) of 0.52–1.42, suggesting that model IV gives a better fit, as measured by Δ. For the group of England, France and Sweden, the difference is only 0.08 and its confidence interval easily covers zero. Comparisons of the unstandardized estimates of Δ between the two data sets may, however, be confounded by differences in their marginal distributions. Such differences exist, most obviously in the proportion of farm workers in the destination and particularly origin classes, which is much higher in Ireland, Poland and Hungary. To remove the effects of the margins from the comparisons, standardized estimates of the index of dissimilarity were computed. The initial standard table used here for each of the two groups of nations is such that the three nations concerned in each case are equally represented, both birth cohorts are of the same size, and the class origin-destination margin within each cohort is as estimated from all six nations taken together; standardization for each model is then based on the relevant margins derived from these two standard tables.

The standardized estimates, also shown in Table 2, reinforce the earlier results. Lack of fit is fairly similar for the two data sets for model VI, but appears to be greater in the group of Ireland, Hungary and Poland for at least some of the other models. In particular, standardization makes it clearer that differences between nations are larger in this group than between England, France and Sweden.

6.2. Simulations

The models and data sets reported in Table 2 were used as a basis for a simulation study of the estimators of Δ. For each model, four-way tables were generated conditional on the margins $\mathbf{S}_1$ fitted by the model (the margins of model I for model V) and having $\boldsymbol{\theta}_2$ equal to the $\tilde{\boldsymbol{\theta}}_2$ estimated from the data. The true values of Δ are thus equal to the $\hat{\Delta}$ shown in Table 2.

The tables were generated using a Gibbs sampling approach originally proposed by Forster et al. (1996) for estimating exact p-values, with the minor difference that here the tables were generated under a saturated model rather than a restricted one. The algorithm produces a sequence of tables from the required conditional distribution. These are autocorrelated, so that the effective number of simulated tables and estimates of Δ is smaller than the number of simulations. This was examined by calculating two estimates of the standard error of the simulation mean of $\hat{\Delta}$, a naive estimate which assumes independence between simulations, and an estimate which accounts for the autocorrelations by including a weighted sum of them up to some fixed lag (see Geyer 1992); specifically, the Parzen window function (e.g. Priestley, 1981, p. 443) was used for the weighting. The effective sample sizes were then estimated by dividing the number of simulations by the ratio of these two standard errors. The effective number of simulations for all the loglinear models based on the

original data (first two panels in Table 3) is between 500 and 600. The effective simulation sizes in the last four panels, relating to the standardized index, are typically larger than 600. The simulations were carried out using the OX 2.0 matrix programming language of Doornik (2001), and some auxiliary calculations with S-Plus 2000 for Windows.

[Table 3 around here.]

The first pair of panels of Table 3 show results of simulations based on the original data sets in Table 2. The variances of both $\hat{\Delta}$ and $\tilde{\Delta}$ are well estimated by (16). The bias in $\hat{\Delta}$, however, is apparent in spite of the large sample sizes. The modification $\tilde{\Delta}$ removes much of the bias. Because standard errors are small, the reduction in bias results in dramatically improved coverage probabilities for confidence intervals based on $\tilde{\Delta}$ rather than $\hat{\Delta}$. In general, the standard deviations of $\hat{\Delta}$ are very nearly the same for all of the models.

Simulations for the estimates of Δ standardized as in Table 2 were computed using the unstandardized tables generated for the corresponding models, subsequently standardized to have the relevant margins implied by the two standard 4-way tables described above. A constant 0.5 was added to each cell of each simulated table before standardization, motivated by standard notions of continuity correction: this was found empirically to improve the variance approximation, as well as avoiding potential numerical difficulties in the standardization caused by sampling zeros. The results are shown in the middle pair of panels of Table 3. The main difference from results for the unstandardized index is that the sampling variation in $\hat{\Delta}$ is slightly increased, even taking into account that Δ itself is increased by this particular standardization: that is, its coefficient of variation is also slightly increased. The sampling variation in estimates of the standardized Δ continues to be quite well approximated by the estimated standard error, and the approximate bias reduction formula $\hat{\Delta}$ remains effective in improving the coverage of a nominally 95% confidence interval.

These encouraging findings do not, however, apply to all standardizations of Δ that might be considered. The lower pair of panels in Table 3 show the effects of standardization based on a completely uniform table with $N/150$ observations in each cell, from which all derived margins used for standardization are of course also uniform. As was mentioned in Section 3 above, this might seem a convenient choice of standardization when there is no obvious 'standard' population from a subject-matter viewpoint. However, the principal effects of this standardization are firstly to increase substantially the sampling variation in $\hat{\Delta}$ and $\tilde{\Delta}$, even when measured in terms of the coefficient of variation; and secondly to induce a larger bias than before in the estimated standard errors. Both findings are connected with the fact that the standardization used here, to uniform margins $\mathbf{S}_0$, is rather strong and in particular may inflate small counts in the original tables so that their contribution to $\hat{\Delta}$ increases from negligible to substantial, with the corresponding residuals more frequently having the wrong sign. As a simpler illustration of these effects, consider the independence model for variables A and B for the constructed data shown in Table 4. While standardization to uniform margins reduces $\hat{\Delta}$ in this case, the variance is substantially increased because of the relatively large role played by small counts in the last row of the data; and the corresponding residuals in $\hat{\mathbf{e}}_0$ for the last row of the table are very far from our working assumption that in repeated sampling they are approximately normally distributed with constant sign. As a general rule, standardization which greatly inflates the impact of small counts should be avoided.

[Table 4 around here.]

Even when N is large, a contingency table can be sparse in that it has many small individual cell counts. Further simulations, not reported in detail here, were carried out to examine the behaviour of $\hat{\Delta}$ and estimates of its variance in such cases. The results were similar to those for the simulations reported above. This suggests that sparseness, even in extreme cases, does not have a serious effect on the adequacy of our approximate variance formulas. As argued above, the most problematic cases are those with relatively large residuals whose signs are very uncertain. Small cell counts, on the other hand, are associated with residuals which are either small (and thus contribute little to $\hat{\Delta}$ and its variance) or large but with well-established (negative) signs. They will thus have little influence on the behaviour of the approximate variance of $\hat{\Delta}$.

7. Concluding remarks

The approximate variance formula for $\hat{\Delta}$ or the bias-reduced form $\tilde{\Delta}$ has been found to work well in situations where sample size is sufficiently large that there is clear evidence of model lack-of-fit, which are precisely the situations where the estimation of a summary such as Δ is of most value. With smaller sample sizes, specifically when evidence against the model of interest is not statistically significant at conventional levels, the approximations used are rather crude and may not be so accurate.

We have seen that standardization, while often desirable for reasons of interpretation or comparison, can result in rather poor estimation of Δ if it involves gross inflation of small counts in the original data. Any standardization of Δ for interpretive or comparative purposes should therefore be chosen with caution. The estimated standard error is a useful diagnostic in this regard.

Estimates of Δ and their approximate variances are straightforward to implement in many standard statistical software packages. What is required is that a program should be able to fit the model of interest, extract relevant parts of its output and carry out simple matrix calculations, as well as providing facilities for combining such commands into command macros or functions for convenient use. Apart from the simulations, the calculations reported here were carried out using S-language functions which the authors will make available in fully documented form through the usual archive network.

Acknowledgments

This paper is a substantial development of earlier unpublished work supported by the Economic and Social Research Council under the Analysis of Large and Complex Datasets programme, carried out at Nuffield College Oxford.

Appendix: Data used in Section 6

England, 25–44					Sweden, 25–44					France, 25–44				
612	70	13	170	89	113	9	1	27	16	1258	128	19	400	212
148	76	4	83	51	46	16	2	32	24	469	307	23	285	186
59	25	85	86	80	64	16	30	76	84	388	184	982	539	773
482	134	8	684	360	84	22	3	88	52	573	118	23	815	390
250	66	12	393	316	62	18	3	88	54	363	81	37	541	424
England, 45–64					Sweden, 45–64					France, 45–64				
416	52	6	138	87	56	10	1	13	7	736	146	37	180	102
155	97	11	120	79	32	23	3	22	12	358	344	49	160	131
68	26	99	90	108	64	28	87	70	97	357	198	1205	322	610
469	99	11	628	376	59	10	3	69	31	396	124	33	382	214
207	45	7	311	282	44	11	3	54	42	239	84	31	243	232
Ireland, 25–44					Hungary, 25–44					Poland, 25–44				
73	7	2	20	13	422	8	21	210	121	938	26	45	489	249
36	32	2	15	19	101	19	12	113	85	141	45	28	144	94
52	28	182	39	56	321	30	576	667	970	1272	132	3085	2360	2040
37	9	2	73	30	239	10	21	400	156	755	54	78	1660	609
37	18	9	74	66	192	12	49	434	462	416	35	83	963	545
Ireland, 45–64					Hungary, 45–64					Poland, 45–64				
37	7	4	10	9	261	8	7	44	36	457	22	33	114	83
17	22	6	11	15	176	39	24	149	89	137	56	67	112	78
36	34	292	24	72	459	48	864	497	926	1046	131	4441	1167	1114
20	2	2	35	23	164	14	14	204	73	469	40	119	675	260
31	4	16	26	59	167	11	37	148	209	324	23	105	362	272

Source: Erikson and Goldthorpe (1992), obtained from the authors.

Note: In each 5×5 table for a nation and a birth cohort, rows correspond to class of origin and columns to class of destination, both in the order white-collar workers, petty bourgeoisie, farm workers, skilled workers, and non-skilled workers.

References

Agresti, A., 2002. Categorical Data Analysis, 2nd Edition. Wiley, New York.

Barndorff-Nielsen, O. E., Cox, D. R., 1994. Inference and Asymptotics. Chapman & Hall, London.

Birch, M. W., 1963. Maximum likelihood in three-way contingency tables. Journal of the Royal Statistical Society B 25, 220–233.

Breslow, N. E., Day, N. E., 1980. Statistical Methods in Cancer Research. Volume 1 — The Analysis of Case–control Studies. International Agency for Research on Cancer, Lyon.

Cortese, C. F., Falk, R. F., Cohen, J. K., 1976. Further considerations on the methodological analysis of segregation indices. American Sociological Review 41, 630–637.

Cox, D. R., Snell, E. J., 1989. Analysis of Binary Data, 2nd Edition. Chapman and Hall, London.

Doornik, J. A., 2001. Ox: An Object-oriented Matrix Language, 4th Edition. Timberlake Consultants Press, London.

Duncan, O. D., Duncan, B., 1955. A methodological analysis of segregation indexes. American Sociological Review 20 (2), 210–217.

Erikson, R., Goldthorpe, J. H., 1992. The Constant Flux: A Study of Class Mobility in Industrial Societies. Clarendon Press, Oxford.

Erikson, R., Goldthorpe, J. H., Portocarero, L., 1979. Intergenerational class mobility in three Western European societies: England, France and Sweden. British Journal of Sociology 30, 415–441.

Erola, J., Moisio, P., 2007. Social mobility over three generations in Finland, 1950–2000. European Sociological Review 23, 169–183.

Feller, W., 1971. An Introduction to Probability Theory and its Applications, 2nd Edition. Vol. 2. Wiley, New York.

Fisher, R. A., 1925. Theory of statistical estimation. Proceedings of the Cambridge Philosophical Society 22, 700–725.

Forster, J. J., McDonald, J. W., Smith, P. W. F., 1996. Monte Carlo exact conditional tests for log-linear and logistic models. Journal of the Royal Statistical Society B 58 (2), 445–453.

Geyer, C. J., 1992. Practical Markov chain Monte Carlo (with discussion). Statistical Science 7 (4), 473–511.

Gini, C., 1914. Di una misura della dissomiliglianza tra due gruppi di quantitá e delle sue applicazioni allo studio delle relazioni statistiche. Atti del Reale Instituto Veneto di Scienze, Lettere ed Arti (Series 8) 74 (2), 185–213.

Goodman, L. A., 1985. The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries. Annals of Statistics 13 (1), 10–69.

Goodman, L. A., Kruskal, W. H., 1954. Measures of association for cross classifications. Journal of the American Statistical Association 49, 732–764.

Goodman, L. A., Kruskal, W. H., 1959. Measures of association for cross classifications. II: Further discussion and references. Journal of the American Statistical Association 54, 123–163.

Goodman, L. A., Kruskal, W. H., 1979. Measures of Association for Cross Classifications. Springer-Verlag, New York.

Haberman, S. J., 1982. Analysis of dispersion of multinomial responses. Journal of the American Statistical Association 77 (379), 568–580.

Hagenaars, J. A., McCutcheon, A. L. (Eds.), 2002. Applied Latent Class Analysis. Cambridge University Press, Cambridge.

Imrey, P. B., Koch, G. G., Stokes, M. E., 1981. Categorical data analysis: Some reflections on the log linear model and logistic regression. Part I: Historical and methodological overview. International Statistical Review 49, 265–283.

Inman, H. F., Bradley, Jr., E. L., 1991. Approximations to the mean and variance of the index of dissimilarity in $2 \times C$ tables under a random allocation model. Sociological Methods & Research 20 (2), 242–255.

Jahn, J., Schmidt, C. F., Schrag, C., 1947. The measurement of ecological segregation. American Sociological Review 12 (3), 293–303.

Johnson, N. L., Kotz, S., Balakrishnan, N., 1995. Continuous Univariate Distributions, 2nd Edition. Vol. 2. Wiley, New York.

Jonsson, J. O., Grusky, D. B., , Di Carlo, M., Pollak, R., Brinton, M. C., 2009. Micro-class mobility: Social reproduction in four countries. American Journal of Sociology 114, 977–1036.

Little, R. J. A., Wu, M.-M., 1991. Models for contingency tables with known margins when target and sampled populations differ. Journal of the American Statistical Association 86 (413), 87–95.

Mulekar, M. S., Knutson, J. C., Champanerkar, J. A., 2008. How useful are approximations to mean and variance of the index of dissimilarity. Computational Statistics & Data Analysis 52, 2098–2109.

Pfeffer, F. T., 2008. Persistent inequality in educational attainment and its institutional context. European Sociological Review 24, 543–565.

Priestley, M. B., 1981. Spectral Analysis and Time Series. Academic Press, London.

Ransom, M. R., 2000. Sampling distributions of segregation indexes. Sociological Methods & Research 28, 454–475.

Uggen, C., Blackstone, A., 2004. Sexual harassment as a gendered expression of power. American Sociological Review 69, 64–92.

Wells, T., Sandefur, G. D., Hogan, D. P., 2003. What happens after the high school years among young persons with disabilities. Social Forces 82, 803–832.

White, M. J., 1986. Segregation and diversity measures in population distribution. Population Index 52 (2), 198–221.

Xie, Y., 1992. The log-multiplicative layer effect model for comparing mobility tables. American Sociological Review 57, 380–395.

Table 1: Simulation results for estimates of Δ for an independence model in a 2×2 table, based on 1000 simulated tables. Columns denoted $\hat{\Delta}$ and $\tilde{\Delta}$ are simulation means, 'sd' denotes simulation standard deviations, '$\widehat{\text{se}}$' the simulation mean of estimated standard errors, and 'coverage' the coverage of nominally 95% confidence intervals. All results are in percentage units.

									Coverage	
π_{2+}	π_{+2}	θ_2	Δ	$\hat{\Delta}$	$\tilde{\Delta}$	sd($\hat{\Delta}$)	sd($\tilde{\Delta}$)	$\widehat{\text{se}}(\hat{\Delta})$	$\hat{\Delta}$	$\tilde{\Delta}$
$N = 100$										
0.5	0.5	1.0	12.25	12.05	11.66	4.56	4.99	4.83	96.0	91.9
0.5	0.5	0.1	1.25	4.25	3.16	3.23	3.30	4.97	96.0	96.0
0.5	0.1	1.0	4.24	4.36	3.94	2.47	2.71	2.56	83.8	83.8
0.5	0.1	0.1	0.45	2.41	1.85	1.95	1.97	2.81	86.6	86.6
0.1	0.1	1.0	2.16	2.45	2.05	1.97	1.95	2.34	97.6	97.6
0.1	0.1	0.1	0.17	1.32	1.05	1.28	1.12	1.83	98.3	98.3
$N = 1000$										
0.5	0.5	1.0	12.25	12.21	12.21	1.54	1.54	1.53	95.2	95.2
0.5	0.5	0.1	1.25	1.61	1.27	1.16	1.23	1.58	97.7	97.7
0.5	0.1	1.0	4.24	4.22	4.21	0.84	0.84	0.86	95.1	95.1
0.5	0.1	0.1	0.45	0.80	0.58	0.57	0.58	0.94	98.5	98.5
0.1	0.1	1.0	2.16	2.17	2.15	0.72	0.76	0.73	94.7	94.7
0.1	0.1	0.1	0.17	0.48	0.35	0.39	0.39	0.58	94.9	96.9

Table 2: Fitted models for the social mobility example. Here L^2 denotes the deviance and 'p-val.' its p-value from a χ^2 distribution with 'd.f.' degrees of freedom. All the results for estimates of Δ are in percentage units.

							Standardized		
Model	d.f.	L^2	p-val.	$\hat{\Delta}$	$\tilde{\Delta}$	$\widehat{\text{se}}(\hat{\Delta})$	$\hat{\Delta}$	$\tilde{\Delta}$	$\widehat{\text{se}}(\hat{\Delta})$
ENG, FRA, SWE ($N = 26,655$)									
I = (NCO, NCD)	96	8385	$\ll 0.001$	21.43	21.31	0.28	22.97	22.75	0.41
II = I + OD	80	106	0.030	1.93	1.49	0.25	2.70	2.05	0.41
III = I + COD	64	77	0.125	1.49	1.13	0.24	2.42	1.84	0.37
IV = I + NOD	48	55	0.225	1.41	1.05	0.26	1.53	1.02	0.36
V = I + (NC)*(NOD)	45	43	0.555	1.14	0.81	0.23	1.34	0.83	0.41
VI = I + NOD + COD	32	25	0.789	0.75	0.48	0.22	1.25	0.77	0.42
IRL, HUN, POL ($N = 40,057$)									
I = (NCO, NCD)	96	10650	$\ll 0.001$	20.11	20.05	0.19	23.03	22.78	0.37
II = I + OD	80	269	< 0.001	2.38	2.18	0.17	4.54	4.02	0.39
III = I + COD	64	221	< 0.001	2.09	1.93	0.16	4.23	3.73	0.43
IV = I + NOD	48	70	0.021	1.19	0.96	0.19	2.17	1.64	0.42
V = I + (NC)*(NOD)	45	63	0.040	1.08	0.87	0.18	1.94	1.44	0.40
VI = I + NOD + COD	32	25	0.816	0.66	0.48	0.18	1.42	0.93	0.40

Table 3: Simulations based on the models of Table 2. See Table 1 for a description of the entries.

Model	Δ	$\hat{\Delta}$	$\tilde{\Delta}$	sd($\hat{\Delta}$)	sd($\tilde{\Delta}$)	$\hat{\text{se}}(\hat{\Delta})$	Coverage $\hat{\Delta}$	Coverage $\tilde{\Delta}$
ENG, FRA, SWE ($N = 26,655$)								
I	21.43	21.53	21.43	0.27	0.28	0.27	92.9	93.3
II	1.93	2.48	2.09	0.21	0.22	0.25	38.1	92.3
III	1.49	1.99	1.66	0.19	0.21	0.23	40.5	90.2
IV	1.41	1.90	1.58	0.20	0.22	0.25	52.4	92.6
V	1.14	1.70	1.39	0.21	0.22	0.25	36.4	83.9
VI	0.75	1.21	0.96	0.17	0.19	0.22	47.1	88.1
IRL, HUN, POL ($N = 40,057$)								
I	20.11	20.15	20.09	0.18	0.18	0.19	96.5	94.0
II	2.38	2.59	2.40	0.16	0.17	0.17	81.1	95.9
III	2.09	2.29	2.14	0.16	0.17	0.16	78.3	93.8
IV	1.19	1.43	1.24	0.16	0.17	0.18	75.6	94.7
V	1.08	1.33	1.14	0.15	0.17	0.18	74.1	95.4
VI	0.66	0.88	0.73	0.15	0.16	0.16	75.8	93.0
ENG, FRA, SWE, standardized as in Table 2								
I	22.97	22.90	22.69	0.38	0.39	0.40	96.1	90.3
II	2.70	3.66	3.08	0.33	0.36	0.40	30.1	87.8
III	2.42	3.34	2.80	0.32	0.35	0.38	29.7	84.9
IV	1.53	2.29	1.84	0.29	0.31	0.37	46.3	90.8
V	1.34	2.15	1.69	0.29	0.31	0.38	44.6	89.7
VI	1.25	1.95	1.55	0.31	0.34	0.38	59.7	91.3
IRL, HUN, POL, standardized as in Table 2								
I	23.03	23.18	22.95	0.32	0.33	0.37	95.5	95.2
II	4.54	5.06	4.58	0.38	0.41	0.41	79.3	94.8
III	4.23	4.72	4.27	0.40	0.43	0.42	81.7	94.5
IV	2.17	2.78	2.34	0.31	0.34	0.40	71.3	96.2
V	1.94	2.55	2.13	0.32	0.34	0.38	69.0	93.6
VI	1.42	2.16	1.72	0.33	0.35	0.43	65.4	94.2
ENG, FRA, SWE, standardized to uniform margins								
I	25.80	25.72	25.29	0.57	0.59	0.62	96.8	86.4
II	5.13	7.23	5.87	0.63	0.68	0.90	31.0	93.8
III	4.51	6.61	5.32	0.64	0.69	0.92	32.4	92.6
IV	3.69	5.29	4.29	0.66	0.70	0.90	59.8	95.1
V	2.33	3.53	2.81	0.44	0.46	0.57	45.5	92.3
VI	2.90	4.51	3.56	0.70	0.74	0.96	67.1	94.8
IRL, HUN, POL, standardized to uniform margins								
I	26.94	27.05	26.68	0.62	0.64	0.63	96.1	92.2
II	7.31	8.73	7.53	0.64	0.68	0.82	62.3	97.5
III	6.53	8.08	6.94	0.70	0.74	0.86	60.3	96.4
IV	4.40	5.66	4.65	0.66	0.71	0.84	72.9	97.6
V	3.28	4.09	3.38	0.50	0.54	0.57	74.1	95.4
VI	3.09	4.60	3.62	0.73	0.76	0.98	77.0	98.0

Table 4: A constructed 3×2 table to illustrate the potential influence of small counts when standardized margins differ greatly from those of the original table.

		Unstandardized				Standardized to uniform margins	
		B				B	
		1	2			1	2
	1	6000	3000		1	4004	2002
A	2	3000	6000	A	2	2002	4004
	3	9	9		3	3003	3003
		$\hat{\Delta} \approx 1/6$				$\hat{\Delta}_0 = 1/9$	

Figure 1: True Δ for an independence model for a 2×2 table as a function of the row proportion π_{2+}. The column proportion π_{+2} is 0.5 on the left and 0.05 on the right. The odds ratio $\exp(\theta_2)$ is 5 in both cases.