Statistica Sinica **13**(2003), 655-671

USING AUXILIARY INFORMATION FOR IMPROVING ESTIMATION IN THE NUMBER OF SPECIES PROBLEM

S. Lynne Stokes

Southern Methodist University

Abstract: Researchers from a variety of disciplines have studied the problem of estimating the number of distinct classes in a population, known in statistics as the number of species problem. The topic of this paper is the special case of the problem in which the population is finite, its size (or the sampling rate) is known, and auxiliary information correlated to class size is available from sampled classes. We use this information to improve estimation by linking class size to this information via a loglinear model. The parameters of the model are estimated from the sample using conditional maximum likelihood, where the conditioning event is that the class is observed in the sample. The model is then used to estimate the probability of observation for every sampled class, which is in turn used in a Horvitz-Thompson-like estimator of number of classes. The paper shows that the improvement in estimation over other available estimators can be dramatic, especially if the class sizes vary widely. The performance of the estimator degrades when the model is misspecified, but still competes well with alternative estimators.

Key words and phrases: capture-recapture, Horvitz-Thompson estimator, loglinear model, number of classes.

1. Introduction

The problem of estimating the number of distinct classes in a population, known as the number of species problem, has been studied by researchers from a variety of disciplines outside ecology, including archeology and computer science. A review article by Bunge and Fitzpatrick (1993) reports papers on the topic from as early as 1948. A special case of the problem is addressed in this paper; specifically, how can we estimate the number of classes in a finite population whose size is known? This problem arises in an array of applications other than the classical number of species, including the following.

(a) Data are organized in tables called relations in relational databases. For example, each row of a database, called a record, might represent a telephone call made, and a column, called an attribute, might represent the originating time of the call. Processing complex queries to a database efficiently has become more important as they have grown in size. Knowledge of the number of distinct values of an attribute is vital for determining the most efficient

algorithm for computing a specified output relation (e.g., Chaudhuri, Motwani and Narassayya (1998)) and is therefore important for good database performance.

(b) Lists containing overlapping entries may need to be combined and unduplicated. Examples of these situations are estimation of the number of valid signatures in petitions (Smith-Cayama and Thomas (1999)) and the number of sampling units on combined sampling frames (Deming and Glasser (1959)).

The problem being studied in this paper can be described as follows. A population of size N consists of D mutually disjoint classes of items, N_j denotes the size of the jth class, $N = \sum_{j=1}^{D} N_j$. A sample of items is chosen from this population, and we let f_i denote the number of classes represented exactly i times in the sample and $d = \sum_{i=1}^{N} f_i$ the total number of classes represented in the sample. In this paper, we propose a method for estimating D from this sample. Our innovation is that we allow auxiliary information about the observed classes to be used in the estimation process.

Throughout our development, we assume the sample of items is selected by a Bernoulli sample design with a known sampling rate q. A Bernoulli design is one in which each unit of the population is chosen into the sample independently and with equal probability. (See, for example, Sarndal, Swensson and Wretman (1992), Section 3.2.) We assume this design, rather than the more familiar simple random sampling, for two reasons. First, it is the design actually used for sampling in some database products, which was the motivating application for work on this problem. Secondly, it provides an approximation to a simple random sample (srs) that proves more tractable for this analysis. If a srs of size n is the true design, then we set the sampling rate $q = n/N$ in the development that follows. Thus the assumption of a known sampling rate is equivalent to an assumption of a known population size N.

The long history for the number of species problem has produced many estimators, none of which perform well under all circumstances. First consider the unrealistic special case in which it is known that the class sizes are equal; i.e., that $N_1 = \cdots = N_D = \overline{N} = N/D$. Under Bernoulli sampling with probability of selection q, $d \sim \text{Bin}(D, 1-(1-q)^{\overline{N}})$, so that a method of moments estimator of D is the solution of the equation

$$d = D[1-(1-q)^{\overline{N}}]. \tag{1.1}$$

If q is sufficiently small and N sufficiently large, $(1-q)^{\overline{N}} \approx \exp(-q\overline{N}) \approx \exp(-n/D)$, where n is the number of items selected in the Bernoulli sample. This leads to the following approximation for the estimating equation in (1.1):

$$d = D[1-\exp(-n/D)]. \tag{1.2}$$

A number of authors (e.g., Good (1950), Lewontin and Prout (1956)) have suggested that the solution of (1.2), denoted $\hat{D}_{eq}$, be used as an estimator of D in the case in which multinomial sampling from an equal class-size population is assumed. This estimator performs well when those assumptions are met, but underestimates D when class sizes vary.

Haas and Stokes (1998) proposed a family of estimators for D when class sizes vary, called generalized jackknife estimators. Using this approach, they attempted to correct the bias of d as an estimator of D after approximating it by a function that could be more easily estimated. Their "first-order" estimator is $\hat{D}_{uj1} = d/(1-(1-q)f_1/n)$; their "second-order" estimator is $\hat{D}_{uj2} = (\hat{D}_{uj1}/d)[d - f_1(1-q)\ln(1-q)\hat{\gamma}^2(\hat{D}_{uj1})/q]$, where $\hat{\gamma}^2(\hat{D}_{uj1}) = \max(0, (\hat{D}_{uj1}/n^2)$ $\sum_{i=1}^{n} i(i-1)f_i + \hat{D}_{uj1}/N - 1)$ is an estimator of the square of the coefficient of variation (cv) of the class sizes. Neither of these estimators performs well as cv increases. $\hat{D}_{uj2}$ was found to be improved, however, by using a stabilizing device suggested by Chao, Ma and Yang (1993). First fix $c \geq 1$ and remove any class whose frequency in the sample exceeds c (in the simulations which follow, $c = 50$). Then compute the estimator $\hat{D}_{uj2}$ from the reduced sample and subsequently increment it by the number of large classes removed. We denote this estimator by $\hat{D}_{uj2a}$. Haas and Stokes (1998) recommended it as the best nonparametric non-branching estimator they were able to find.

The population configurations that are especially problematic are those having small average class size $\overline{N}$. Since $D = N/\overline{N}$, a small change in the average class size, which is hard to detect from the sample, can produce a large change in D. For some class size configurations, there is simply too little information available from the sample for estimating D, unless the sampling rate is large.

Because performance of the available nonparametric estimators of D is not entirely satisfactory, researchers have examined alternatives that attempt to augment the information available from the sample in various ways. One approach has been to assume a certain family of distributions for class sizes (e.g., Sichel (1997)). Here we take an alternative approach. We assume that explanatory variables for the size of each observed class are available and that a model linking class size to these variables can be determined. In the database application, the auxiliary information might be the size of the sampled class at some point in the past when the entire database was processed (as it might be once a year, say). A similar type of information for the number of species problem may exist in the form of compilations of counts of "sightings" of species for similar habitats, or for the same habitat in earlier time periods. In the sampling frame application, observable characteristics of the sampling units themselves, such as family size, may be predictive of the number of lists on which the units appear if the lists are administrative records.

The remainder of the paper will proceed as follows. Section 2 presents the model for class size, describes the method for estimating it, and examines an

estimator of D based upon the model. A method for estimating variance is given in Section 3. The results of a simulation study are presented in Section 4. The study examines the performance of the new estimator in both correctly specified and misspecified artificial populations, and in a real population.

2. The Estimator

In this section, we present an estimator for D based on a sample from a finite population of items. The data are composed of the number of sample items falling into, as well as a vector of explanatory variables about, each observed class. The estimation process has two conceptual components, which are summarized and then considered in detail.

(1) The proposed estimator is

$$\hat{D} = \sum_{O_j=1} 1/\hat{\pi}_j, \tag{2.1}$$

where $O_j = 1$ when class j is observed in the sample and $= 0$ otherwise, and $\hat{\pi}_j$ is an estimator of $\pi_j = \Pr[O_j = 1]$, the probability that class j is observed in the sample. If the π_j's were known rather than estimated, $\hat{D}$ could be recognized as the Horvitz-Thompson estimator of D, the size of the population of classes. Haas and Stokes (1998) considered an estimator of the form (2.1) that made use only of the number of sample observations in each observed class, but not any auxiliary information, to estimate π_j. It performed poorly.

(2) The second component of the estimation process is methodology for estimating π_j. This probability is determined by the size of the jth class, which we link by a regression model to auxiliary information about the class. The model must be estimated from the observed classes, which are size-biased, since large classes are more likely to be observed than small classes. This selection bias must be taken into account for estimation.

This method is similar in spirit to Alho (1990) and Huggins (1989), who proposed model-based generalizations of the capture-recapture estimator. Their methods allowed each individual in a population to have its own capture probabilities on each occasion, modeled by logistic regression as functions of characteristics of the individuals, and estimated from the data of ever-captured individuals. In their case, π_j represented the probability that the jth individual was captured on either occasion. Our approach is similar, but leads to a different structure for both the class selection probability π_j and for the underlying regression model. We now describe the assumed structure for our population.

Recall that N_j denotes the size of the jth class in the population. Assume that

$$N_j \sim P^+(\lambda_j), \tag{2.2}$$

$$\lambda_j = \exp(\boldsymbol{x}_j\boldsymbol{\beta}), \quad j = 1, \ldots, D, \tag{2.3}$$

and that the N_j's are mutually independent. P^+ denotes a positive Poisson distribution, $\boldsymbol{x}_j$ is a row of k explanatory variables for the size of class j in the design matrix $\boldsymbol{X}$, and $\boldsymbol{\beta}$ is the $k \times 1$ vector of regression coefficients that must be estimated from the sample. Suppose the resulting population of $N = \sum_{j=1}^{D} N_j$ items is sampled with a Bernoulli design having selection probability q. Then the number of items in class j in the resulting sample, n'_j, has conditional distribution

$$n'_j | N_j \sim Bin(N_j, q) \tag{2.4}$$

for $j = 1, \ldots, D$, and are independent. The marginal distribution of n'_j is

$$\begin{aligned} f_{\lambda_j}(k) &= \Pr[n'_j = k] \\ &= \sum_{r=k}^{\infty} \Pr[n'_j = k | N_j = r] \Pr[N_j = r] \\ &= \begin{cases} e^{-q\lambda_j}(q\lambda_j)^k / k!(1 - e^{-\lambda_j}) & \text{for } k = 1, \ldots \\ (e^{-q\lambda_j} - e^{\lambda_j})/(1 - e^{-\lambda_j}) & \text{for } k = 0. \end{cases} \end{aligned} \tag{2.5}$$

Note, however, that n'_j cannot be observed in the sample since it is truncated when $n'_j = 0$. Instead, the size of the jth observed class, denoted by n_j, has probability function $f_{\lambda_j}(k)/[1 - f_{\lambda_j}(0)]$. (To aid in notation, we assume that the n'_j's have been reordered so that $n'_j > 0$ for $j = 1, \ldots, d$.) From (2.5), one can show, that for $j = 1, \ldots, d$,

$$n_j \sim P^+(q\lambda_j). \tag{2.6}$$

The n_j's are conditionally independent, given the classes observed. Under this model the O_j's, $j = 1, \ldots, D$, are independently distributed as Bernoulli (π_j), and one can show from (2.5) that

$$\pi_j = \Pr[n'_j > 0] = \frac{1 - \exp(-q\lambda_j)}{1 - \exp(-\lambda_j)}. \tag{2.7}$$

Therefore, following (2.1), we propose as an estimator

$$\hat{D} = \sum_{O_j=1} \frac{1 - \exp(-\hat{\lambda}_j)}{1 - \exp(-q\hat{\lambda}_j)}, \tag{2.8}$$

where $\hat{\lambda}_j = \exp(\boldsymbol{x}_j \hat{\boldsymbol{\beta}})$, and $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimator of $\boldsymbol{\beta}$ under (2.6).

From (2.6), the conditional loglikelihood function for the sample of n_j's given the observed classes is

$$\ln L = \sum_{O_j=1} [n_j \ln \lambda_j - \ln(\exp(q\lambda_j - 1))] + \text{ constant},$$

with λ_j defined in (2.3). Thus the likelihood equations can be written as

$$\sum_{O_j=1} \boldsymbol{x}_j n_j = \sum_{O_j=1} \boldsymbol{x}_j E(n_j), \tag{2.9}$$

$$E(n_j) = \frac{q\lambda_j}{1-\exp(-q\lambda_j)}. \tag{2.10}$$

We solve these equations for $\boldsymbol{\beta}$ using Newton's method. The covariance matrix for $\boldsymbol{n} = (n_1, \ldots, n_d)$, denoted by $\boldsymbol{W}$, is diagonal with the truncated Poisson variance as its jth element:

$$\text{Var}\,(n_j) = \frac{q\lambda_j}{1-\exp(-q\lambda_j)}\left[1 - \frac{q\lambda_j \exp(-q\lambda_j)}{1-\exp(-q\lambda_j)}\right]. \tag{2.11}$$

Letting $\boldsymbol{\beta}_0$ denote a starting value, we can solve (2.9) iteratively by computing $\boldsymbol{\beta}_{u+1} = \boldsymbol{\beta}_u + (\boldsymbol{X'W}_u\boldsymbol{X})^{-1}\boldsymbol{X'}\,(\boldsymbol{n} - \boldsymbol{E}_u(\boldsymbol{n}))$, $u = 0, 1, \ldots$, where $\boldsymbol{W}_u$ and $E_u(\boldsymbol{n})$ give the estimated values of $\boldsymbol{W}$ and $E(\boldsymbol{n})$ (from (2.11) and (2.10)) based on the current value $\boldsymbol{\beta}_u$. Iteration continues until convergence occurs. An estimator for the covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\hat{\text{Cov}}\,(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X'\hat{W}X})^{-1}, \tag{2.12}$$

where $\hat{\boldsymbol{W}}$ is the estimator of $\boldsymbol{W}$, obtained by substituting $\hat{\lambda}_j$ for λ_j.

What does this estimator look like when there are no explanatory variables, i.e., when the design matrix is a vector of 1's? In that case the mle for $\lambda_j = \lambda$ is (from (2.9)) the solution of the equation

$$1 - \exp(-q\lambda) = \frac{q\lambda d}{n}, \tag{2.13}$$

and (from (2.8))

$$\hat{D} = d\frac{1-\exp(-\hat{\lambda})}{1-\exp(-q\hat{\lambda})} = \frac{n(1-\exp(-\hat{\lambda}))}{q\hat{\lambda}},$$

where $\hat{\lambda}$ is the solution of (2.13). Since $Nq \approx n$ when N is large, we have

$$\hat{D} \approx \frac{N}{\hat{\lambda}/[1-\exp(-\hat{\lambda})]}. \tag{2.14}$$

The denominator of (2.14), which is an estimator of average class size, must be greater than or equal to 1; thus we are assured that $\hat{D} \leq N$. We now compare $\hat{D}$ to $\hat{D}_{eq}$ defined from (1.2). We reparameterize (2.13) by defining $\lambda = N/x$, and again use the approximation $Nq \approx n$ to obtain the expression $d = x[1 - \exp(-n/x)]$. Note that the solution to this equation is $\hat{D}_{eq}$, so that

$\hat{D}_{eq} = \hat{D}/[1 - \exp(-\hat{\lambda})]$. So if there are no explanatory variables, $\hat{D}$ and $\hat{D}_{eq}$ are nearly identical when $\hat{\lambda}$ is as large as about 4 (when $\hat{D} = (0.98)\hat{D}_{eq}$). As $\hat{\lambda} \to 0$ (meaning that the class sizes, following the positive Poisson distribution, are becoming more uniformly of size 1), the two estimators diverge, with $\hat{D} \to N$ and $\hat{D}_{eq} \to \infty$, since that estimator was developed for the infinite population case.

The estimator $\hat{D}$ can be thought of as a generalization of a post-stratified version of the equal class-size estimator of D, in the same way that Alho's estimator is a generalization of Sekar and Deming's (1949) post-stratified dual system estimator. If one knew that the set of classes could be divided into strata of nearly equal-sized classes based on some observable characteristic of the sampled class, one could estimate D by $\hat{D}_{ps} = \sum_{h=1}^{H} \hat{D}_{eq,h}$, where $\hat{D}_{eq,h}$ is calculated from those classes that fall into stratum h. This would yield approximately the same estimator as $\hat{D}$ if the only explanatory variables were the indicator variables for post-stratum membership.

3. Asymptotic Properties and Variance Estimation

In this section we examine some large sample properties of $\hat{D}$ and develop an estimator for its standard error. Discussion of asymptotic properties of an estimator of parameters of a finite population requires an artificial structure to make the concept of an increasing sample size meaningful. The approach we take here is similar to that of Huggins (1989) in his development of large sample properties of an estimator of population size in a capture-recapture setting.

Consider a sequence $\{U_1, U_2, \ldots\}$, where U_r is a population consisting of D_r classes and $D_r \to \infty$ as $r \to \infty$. We do not assume that the populations are nested ($U_1 \subset U_2 \subset \cdots$) although we could; that is sometimes done for consistency arguments in finite populations. We do assure the similarity of the populations by assuming that the vectors of covariates $\boldsymbol{x}_{rj}$ (for the jth class in population r) are independent and identically distributed within each population, and that they have come from a common underlying distribution for all r. Further we assume that the structure of the class size configuration and the sample drawn from each population is determined as shown in (2.2)−(2.4). We assume that $\boldsymbol{\beta}$ and q remain constant over all r.

The derivative of the loglikelihood function (2.9) can be written for the rth sample as $d \ln L_r / d\boldsymbol{\beta} = S_r(\boldsymbol{\beta}) = \sum_{j=1}^{D_r} \boldsymbol{x}_j O_j (n'_j - q\lambda_j/(1 - e^{-q\lambda_j}))$. (For simplicity of notation, suppress the subscript r for the parameters and random variables associated with each class in population r.) From (2.7) we have $E(O_j) = \pi_j$ and, from (2.5), one can show that $E(O_j n'_j) = E(n'_j) = q\lambda_j/(1 - e^{-\lambda_j})$. Therefore $S_r(\boldsymbol{\beta})$ is the sum of D_r i.i.d. terms, each of which has expectation $\mathbf{0}$. To find the variance-covariance matrix of each term, denoted by $\boldsymbol{i}_\beta$, we calculate

$$E\left\{ \left[O_j \left(n'_j - \frac{q\lambda_{rj}}{1 - e^{-q\lambda_j}} \right) \right]^2 \middle| \boldsymbol{x}_j \right\} = \frac{q\lambda_j}{1 - e^{-\lambda_j}} \left[1 - \frac{q\lambda_j e^{-q\lambda_j}}{1 - e^{-q\lambda_j}} \right] = c_j.$$

Thus

$$\boldsymbol{i}_{\boldsymbol{\beta}} = E[\boldsymbol{x}_j \boldsymbol{x}_j' c_j], \tag{3.1}$$

where the expectation is taken over the distribution of the covariates.

Let $\hat{D}_r(\boldsymbol{\beta}) = \sum_{j=1}^{D_r} O_j/\pi_j$. Though this cannot be considered an estimator since $\boldsymbol{\beta}$ is unknown, it would be the proposed estimator if $\boldsymbol{\beta}$ were replaced by $\hat{\boldsymbol{\beta}}$. We first investigate the properties of $\hat{D}_r(\boldsymbol{\beta})$. Observe that $\hat{D}_r(\boldsymbol{\beta}) - D_r = \sum_{j=1}^{D_r}[(O_j - \pi_j)/\pi_j]$ is the sum of D_r i.i.d. random variables, each with mean 0 and variance

$$v_{\boldsymbol{\beta}} = E\left\{ \text{Var}\left[\frac{O_j}{\pi_j} \middle| \boldsymbol{x}_j \right] \right\} = E\left[\frac{1-\pi_j}{\pi_j} \right]. \tag{3.2}$$

We assume that this expectation exists, which can be assured by requiring that the probability of sampling a class does not become too small. For example, we can require that the distribution of the covariates be bounded. It is easy to show that the covariance between $S_r(\boldsymbol{\beta})$ and $\hat{D}_r(\boldsymbol{\beta})$ is $\mathbf{0}$. Thus from the Central Limit Theorem, as $r \to \infty$,

$$D_r^{-1/2}(S_r'(\boldsymbol{\beta}), \hat{D}_r(\boldsymbol{\beta}) - D_r) \to \text{ Multinormal}(\mathbf{0}, \text{ Diag}(\boldsymbol{i}_{\boldsymbol{\beta}}, v_{\boldsymbol{\beta}})). \tag{3.3}$$

One can also show that $-E[dS_r(\boldsymbol{\beta})/d\boldsymbol{\beta}] = E[S_r(\boldsymbol{\beta})][S_r(\boldsymbol{\beta})]'$, so that the consistency and asymptotic normality of the solution of the likelihood equations for the rth sample ($\hat{\boldsymbol{\beta}}_r$) is assured. Specifically, we have that $\sqrt{D_r}(\hat{\boldsymbol{\beta}}_r - \boldsymbol{\beta})$ is asymptotically multivariate normal with covariance matrix $\boldsymbol{i}_{\boldsymbol{\beta}}^{-1}$. This allows us to determine that as $r \to \infty$,

$$\left(\sqrt{D_r}(\hat{\boldsymbol{\beta}}_r - \boldsymbol{\beta}), D_r^{-1/2}(\hat{D}_r(\boldsymbol{\beta}) - D_r)\right) \to \text{ Multinormal}(\mathbf{0}, \text{ Diag}(\boldsymbol{i}_{\boldsymbol{\beta}}^{-1}, v_{\boldsymbol{\beta}})). \tag{3.4}$$

Now we need to evaluate the behavior of $\hat{D}_r(\boldsymbol{\beta})$ when $\boldsymbol{\beta}$ is replaced by $\hat{\boldsymbol{\beta}}$. To distinguish the two, we denote the latter as $\hat{D}_r(\hat{\boldsymbol{\beta}}_r)$. We can write the first order Taylor approximation as

$$\hat{D}_r(\hat{\boldsymbol{\beta}}_r) - D_r = \hat{D}_r(\boldsymbol{\beta}) - D_r + (\hat{\boldsymbol{\beta}}_r - \boldsymbol{\beta})' \left[\frac{d}{d\boldsymbol{\beta}} \hat{D}_r(\boldsymbol{\beta}) \right]_{\boldsymbol{\beta}^*}, \tag{3.5}$$

where $\boldsymbol{\beta}^*$ lies between $\hat{\boldsymbol{\beta}}_r$ and $\boldsymbol{\beta}$. After some algebra,

$$\begin{aligned} \frac{1}{D_r}\left[\frac{d}{d\boldsymbol{\beta}} \hat{D}_r(\boldsymbol{\beta}) \right] &= -\frac{1}{D_r} \sum_{j=1}^{D_r} \frac{O_j}{\pi_j^2} \frac{d\pi_j}{d\boldsymbol{\beta}} \\ &= \frac{1}{D_r} \sum_{j=1}^{D_r} \frac{O_j \lambda_j \boldsymbol{x}_j}{\pi_j} \left[\frac{\exp(-\lambda_j)}{1-\exp(-\lambda_j)} - \frac{q\exp(-q\lambda_j)}{1-\exp(-1\lambda_j)} \right]. \end{aligned} \tag{3.6}$$

Since this is a sample mean of i.i.d. random variables, it converges in probability (for any value of $\boldsymbol{\beta}$) to a vector we denote by $\boldsymbol{u}$. Thus we can show from (3.5) that, as $r \to \infty$,

$$\frac{1}{\sqrt{D_r}}(\hat{D}_r(\hat{\boldsymbol{\beta}}_r) - D_r) \to \text{ Normal}(0, v_{\boldsymbol{\beta}} + \boldsymbol{u'} \boldsymbol{i}_{\boldsymbol{\beta}}^{-1} \boldsymbol{u}). \tag{3.7}$$

This suggests that for populations in which D is sufficiently large, the standard error of $\hat{D}$ can be obtained by separately estimating the two components of its approximate variance which (from (3.7)) can be written as $V(\hat{D}) \approx V_{\boldsymbol{\beta}} + \boldsymbol{U'} \boldsymbol{I}_{\boldsymbol{\beta}}^{-1} \boldsymbol{U}$, where $V_{\boldsymbol{\beta}} = D v_{\boldsymbol{\beta}}$, $\boldsymbol{U} = D\boldsymbol{u}$ and $\boldsymbol{I}_{\boldsymbol{\beta}} = D\boldsymbol{i}_{\boldsymbol{\beta}}$. Note that if $\boldsymbol{\beta}$ were known and x fixed, we could estimate $V_{\boldsymbol{\beta}}$ unbiasedly by $\sum_{j=1}^{D} O_j(1-\pi_j)/\pi_j^2$. Since we do not know $\boldsymbol{\beta}$, we substitute its estimator in this expression to obtain

$$\hat{V}_{\boldsymbol{\beta}} = \sum_{j=1}^{D} \frac{O_j(1-\hat{\pi}_j)}{\hat{\pi}_j^2}. \tag{3.8}$$

For estimation of the second component of the variance, we observe from (3.6) that a reasonable estimator of $\boldsymbol{U}$ is

$$\hat{\boldsymbol{U}} = \sum_{j=1}^{D} \frac{O_j \hat{\lambda}_j \boldsymbol{x}_j}{\hat{\pi}_j} \left[\frac{\exp(-\hat{\lambda}_j)}{1-\exp(-\hat{\lambda}_j)} - \frac{q\exp(-q\hat{\lambda}_j)}{1-\exp(-q\hat{\lambda}_j)} \right]. \tag{3.9}$$

For estimation of $\boldsymbol{I}_{\boldsymbol{\beta}}$, we use (based on (3.1))

$$\hat{\boldsymbol{I}}_{\boldsymbol{\beta}} = \sum_{j=1}^{D} \frac{O_j \boldsymbol{x}_j \boldsymbol{x}_j' \hat{c}_j}{\hat{\pi}_j} = \sum_{j=1}^{D} O_j \boldsymbol{x}_j \boldsymbol{x}_j' \frac{q\hat{\lambda}_j}{1-e^{-q\hat{\lambda}_j}} \left[1 - \frac{q\hat{\lambda}_j e^{-q\hat{\lambda}_j}}{1-e^{-q\hat{\lambda}_j}} \right].$$

Note that $\hat{\boldsymbol{I}}_{\boldsymbol{\beta}}^{-1}$ can be written as $(\boldsymbol{X'}\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}$, which matches the previously proposed variance estimator for $\hat{\boldsymbol{\beta}}$ shown in (2.12). Therefore, we can write the estimator of the variance of $\hat{D}$ as

$$\hat{V}(\hat{D}) = \hat{V}_{\boldsymbol{\beta}} + \hat{\boldsymbol{U}}'(\boldsymbol{X'}\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}\hat{\boldsymbol{U}}, \tag{3.10}$$

where components are defined in (3.8), (3.9) and (2.12).

4. Simulation Results

A simulation study was conducted to compare the performance of $\hat{D}$ with that of two estimators that do not make use of auxiliary information, $\hat{D}_{eq}$ and $\hat{D}_{uj2a}$. These two were chosen because one or the other would be expected to perform about as well as known estimators for any population configuration.

When the cv of class size is small, we would expect $\hat{D}_{eq}$ to dominate; else the stabilized jackknife would be expected to perform best.

Three experiments were performed. One studied the benefits of the new estimator in its most favorable light, that is, when (2.6) holds. The other two were designed to assess its performance under more realistic conditions. The first simulates estimation when the model is misspecified in a particular way, and the second simulates sampling from a real population.

4.1. Performance of $\hat{D}$ under assumed model

The estimators of D were studied for different populations, generated by manipulating parameters in (2.2)−(2.4). The experiment used three two-level factors to control the characteristics of the population. These factors were: size of D (100 and 1000); coefficient of variation of class size (small ($cv^2 < 0.25$) and large ($cv^2 \approx 1.7$)); average class size (small ($\overline{N} \approx 6$) and large ($\overline{N} \approx 22$)). Eight ($2 \times 2 \times 2$) populations were generated, each having one combination of the three characteristics. Samples were repeatedly selected from each population according to one of several Bernoulli designs.

The simulation was conducted as follows. For each of the eight populations, a set of D independent variables $X_1, \ldots, X_D$ was generated as independent with $X_j \sim$ Uniform $(0, 1)$. Next the D class sizes $N_1, \ldots, N_D$ were generated, where $N_j \sim P^+(\lambda_j)$, with $\ln(\lambda_j) = \beta_0 + \beta_1 x_j$, and $(\beta_0, \beta_1) = (1.6, 0.6)$ (small cv, small $\overline{N}$); (3.1, 0.6) (small cv, large $\overline{N}$); (-1.5, 5) (large cv, small $\overline{N}$); (-0.2, 5) (large cv, large $\overline{N}$). Then 1000 Bernoulli samples were drawn from the generated population with each sampling rate. Sampling rates examined ranged from $q =$ 0.10 (0.05 for $D = 1000$) to $q = 0.50$ ($q = 0.25$ for small cv, large $\overline{N}$ populations). $\hat{D}$, $\hat{D}_{eq}$ and $\hat{D}_{uj2a}$ were computed from each sample. The estimated variance of $\hat{D}$ (3.10) and its nominal 95% confidence interval

$$\hat{D} \pm 1.96\sqrt{\hat{V}(\hat{D})} \tag{4.1}$$

were also computed from each sample.

Table 1 summarizes the results for $D = 1000$ and for $q = 0.05$, 0.10, 0.25 and 0.50. The first entry in the cells labeled by estimator names is the simulated estimate of relative bias, that is, the average of each estimator as a fraction of D:

$$RB = \frac{1}{1000} \sum_{s=1}^{1000} \hat{D}_s / D, \tag{4.2}$$

where $\hat{D}_s$ is the estimate from one of the three estimators on sample replicate s. The second entry in each cell is the simulated estimate of the relative root

mean squared error, defined as the square root of the average squared error of each estimator as a fraction of D:

$$\sqrt{RMSE} = \left[\frac{1}{1000}\sum_{s=1}^{1000}(\hat{D}_s - D)^2\right]^{1/2}/D. \tag{4.3}$$

Estimates of the simulated standard error of the relative bias and relative average squared error were both less than 0.01 for each cell entry. The table also shows the average values of sample size and number of species observed in the samples for the various simulated populations and sampling rates. The table shows that the proposed estimator provides marked improvement over the alternatives considered when cv is large. Surprisingly, the improvement was at least as great when average cell size was large as when it was small. When cv is small, $\hat{D}$ and $\hat{D}_{eq}$ are nearly identical in performance, with $\hat{D}_{uj2a}$ generally slightly worse. The results for the $D = 100$ case show similar relative performance of the estimators but are not displayed.

Table 1. RB(4.2) and $\sqrt{RMSE}$(4.3) of estimators for correctly specified model for population having $D = 1000$.

q	cv	$\overline{N}$ Small				$\overline{N}$ Large			
		$\overline{n}$ $\overline{d}$	$\hat{D}_{eq}$	$\hat{D}_{uj2a}$	$\hat{D}$	$\overline{n}$ $\overline{d}$	$\hat{D}_{eq}$	$\hat{D}_{uj2a}$	$\hat{D}$
0.05	Large	333 237	0.46 0.54	0.52 0.49	1.05 0.18	1195 467	0.52 0.48	0.64 0.36	1.00 0.09
	Small	332 277	0.89 0.14	0.90 0.15	1.03 0.14	1493 780	1.01 0.02	1.00 0.03	1.01 0.03
0.10	Large	666 372	0.51 0.49	0.62 0.39	1.03 0.10	2389 598	0.61 0.39	0.78 0.23	1.00 0.05
	Small	665 481	0.97 0.06	0.92 0.10	1.00 0.06	2987 950	1.00 0.01	1.00 0.01	1.00 0.01
0.25	Large	1666 585	0.63 0.37	0.84 0.17	1.02 0.04	5973 770	0.77 0.23	0.96 0.05	1.00 0.02
	Small	1662 800	0.98 0.02	0.97 0.04	1.00 0.02	7467 999	1.00 0.00	1.00 0.00	1.00 0.00
0.50	Large	3332 778	0.79 0.21	1.05 0.05	1.01 0.02	11945 900	0.90 0.10	1.05 0.05	1.00 0.01
	Small	3324 956	0.99 0.01	0.99 0.01	1.00 0.01	n.a.	n.a.	n.a.	n.a.

n.a. Not available. Simulations were not conducted for this case, since $q = 0.25$ resulted in virtually perfect estimation by all estimators.

Table 2 summarizes the performance of the variance estimator for $\hat{D}$ (3.10) and the confidence interval procedure (4.1) for $D = 100$ and $D = 1000$. The first entry in each cell of this table shows the square root of the relative bias of the estimated variance of $\hat{D}$,

$$\sqrt{RB(V)} = \left[\frac{1}{1000}\sum_{s=1}^{1000}\hat{V}_s(\hat{D})/\frac{1}{1000}\sum_{s=1}^{1000}(\hat{D}_s - \overline{\hat{D}})^2\right]^{1/2}. \tag{4.4}$$

The relative bias is the ratio of the average estimated variance of $\hat{D}$ to the actual variance of $\hat{D}$, as assessed from the simulation. The second entry in each cell is the actual coverage of the nominal 95% confidence interval for D. The table shows that the variance estimator is nearly unbiased, even for $D = 100$. It tends to overestimate the true variance only slightly, except in those cases where the true variance is extremely small, such as when the average class size and the sampling rate are large. The confidence interval procedure seems to work well, having coverage near the nominal level.

Table 2. $\sqrt{RB(V)}$(4.4) and coverage of nominal 95% CI for correctly specified model.

q	cv	$\overline{N}$ Small		$\overline{N}$ Large	
		$D = 100$	$D = 1000$	$D = 100$	$D = 1000$
0.05	Large	n.a.	1.01 0.96	n.a.	1.04 0.95
	Small	n.a.	1.01 0.96	n.a.	1.01 0.95
0.10	Large	1.03 0.94	1.01 0.95	1.04 0.93	1.04 0.95
	Small	1.01 0.95	1.02 0.95	1.07 0.96	1.07 0.93
0.25	Large	0.99 0.93	1.02 0.94	1.05 0.94	1.05 0.96
	Small	1.02 0.94	1.03 0.95	1.42 0.96	1.29 0.99
0.50	Large	1.02 0.95	1.06 0.96	1.06 0.95	1.07 0.97
	Small	0.96 0.93	1.08 0.95	n.a.	n.a.

4.2. Performance of $\hat{D}$ under misspecified model

The purpose of this experiment was to examine the robustness of $\hat{D}$ to model misspecification. In this simulation, it is assumed that a model of the form

(2.2)−(2.3) exists, but the exact explanatory variables are not available to the analyst. The experiment was structured like the previous one, except that only the large cv, small $\overline{N}$ population was examined. The two ($D = 100$ and 1000) populations generated were identical to those of the previous simulation, and the same 1000 Bernoulli samples ($q = 0.10$ and 0.25 for $D = 100$, $q = 0.05$, 0.10 and 0.25 for $D = 1000$) were used. The difference here was that the vector of independent variables used in estimation was not the one used to generate the population. Instead we used for estimation $X'_j = X_j + \alpha\sigma_X\varepsilon_j$, with $\varepsilon_j \sim N(0,1)$ independent of the X_j, mutually independent, and with α taking on a range of values from 0.05 to 1.0. Since X_j was a uniform random variable, $\sigma_X = \sqrt{1/12}$. This produced an independent variable that explained between 50% (for $\alpha = 1.0$) and 99% (for $\alpha = 0.05$) of the variability in the true independent variable. $\hat{D}$ was calculated for each sample. $\hat{D}_{eq}$ and $\hat{D}_{uj2a}$ are not affected by the new values of the explanatory variables.

Figures 1 and 2 display the results. In each figure, the estimated relative efficiency of $\hat{D}$ to each of the other estimators is shown. Estimated relative efficiency is calculated as the ratio of the average squared errors of the estimators. For example,

$$RE(\hat{D}, \hat{D}_{uj2a}) = \frac{\sum_{s=1}^{1000}(\hat{D}_{uj2a,s} - D)^2}{\sum_{s=1}^{1000}(\hat{D}_s - D)^2}.$$

Comparison of $\hat{D}$ with other estimators when model is misspecified and $D = 100$.

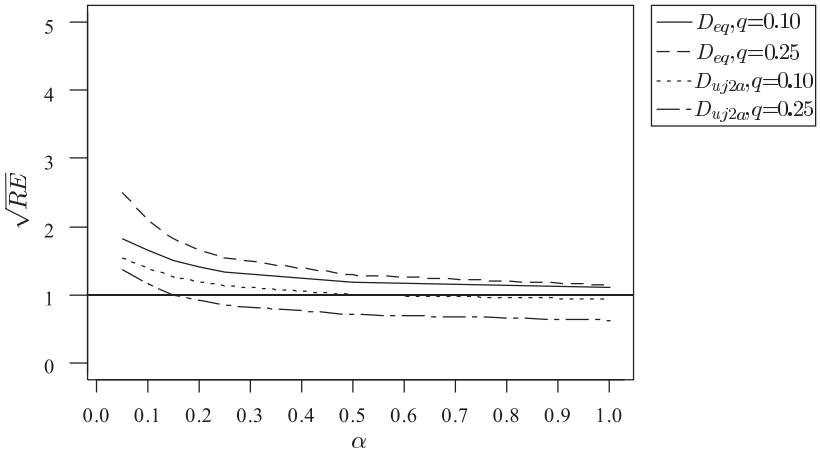

Figure 1. This figure shows $\sqrt{RE(\hat{D}, \hat{D}_{eq})}$ and $\sqrt{RE(\hat{D}, \hat{D}_{uj2a})}$ under increasing model misspecification. The comparison is shown for two sampling rates: $q = 0.10$ and $q = 0.25$, and for a simulated population for which $D = 100$.

Comparison of $\hat{D}$ with other estimators when model is misspecified and $D = 1000$.

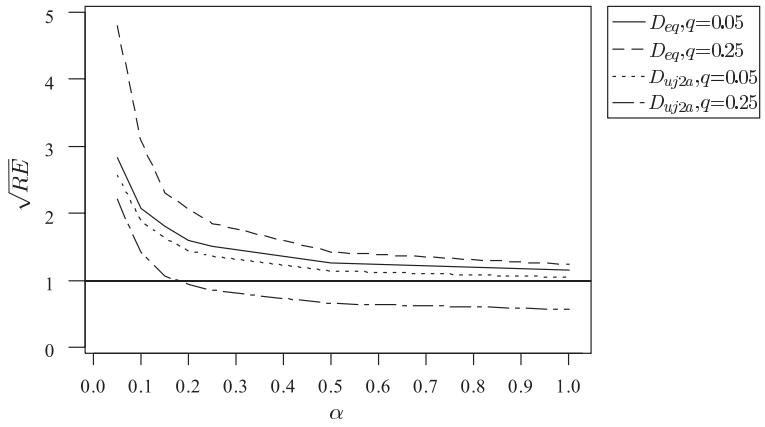

Figure 2. This figure shows $\sqrt{RE(\hat{D}, \hat{D}_{eq})}$ and $\sqrt{RE(\hat{D}, \hat{D}_{uj2a})}$ under increasing model misspecification. The comparison is shown for two sampling rates: $q = 0.05$ and $q = 0.25$, and for a simulated population for which $D = 1000$.

In Figure 1, $\sqrt{RE(\hat{D}, \hat{D}_{eq})}$ and $\sqrt{RE(\hat{D}, \hat{D}_{uj2a})}$ are shown for the two sampling rates $q = 0.10$ and $q = 0.25$ for the high cv, low $\overline{N}$ population with $D = 100$. In Figure 3, $\sqrt{RE(\hat{D}, \hat{D}_{eq})}$ and $\sqrt{RE(\hat{D}, \hat{D}_{uj2a})}$ are shown for the two sampling rates $q = 0.05$ and $q = 0.25$ for the high cv, low $\overline{N}$ population with $D = 1000$.

Comparisons of $\hat{D}$ with other estimators for Christmas Bird data.

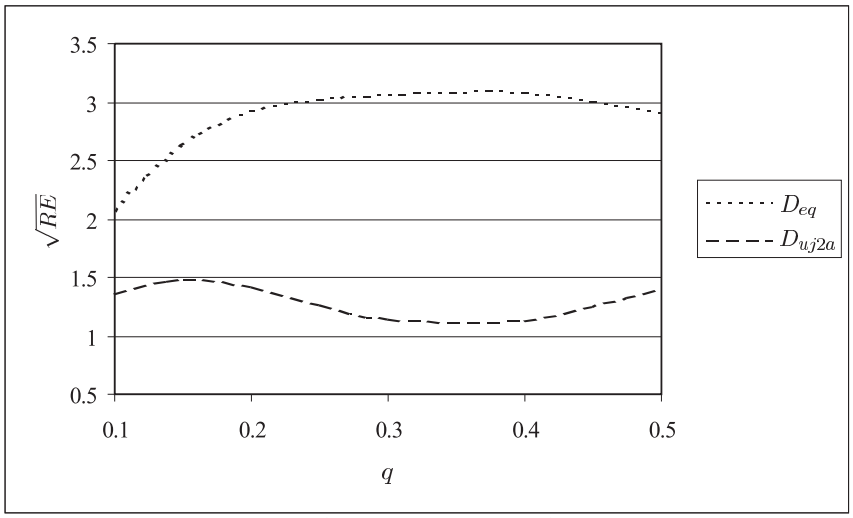

Figure 3. This figure shows $\sqrt{RE(\hat{D}, \hat{D}_{eq})}$ and $\sqrt{RE(\hat{D}, \hat{D}_{uj2a})}$ as functions of sampling rate for the Christmas Bird Count population.

The figures show that the new estimator performed better than $\hat{D}_{eq}$ even with model misspecification, and the advantage for $\hat{D}$ was larger for the larger sample size. They also show that when there was little sample information (small q), $\hat{D}$ had an advantage over $\hat{D}_{uj2a}$ as well, even with substantial model misspecification. The advantage diminished when the amount of sample information or the degree of misspecification increased. It never did completely disappear for the population having $D = 1000$ when $q = 0.05$.

4.3. Performance of $\hat{D}$ in a real population

The purpose of this experiment was to examine the performance of the new estimator in a real population, in which it is known that not all relevant independent variables are available. Table 3 is an excerpt from a table of counts of bird species observed in the 2001 Christmas Bird Count (CBC) in the Austin, Texas circle. These data can be found at www.audubon.org/bird/cbc. For purposes of this example, the birds observed were considered to be the population of interest, even though in reality the observed birds themselves are a sample of all birds present at the site on the day of the count (though hopefully one with a very large sampling rate!). We further modified this population to contain only those bird species with counts (class sizes) less than 150, since larger class sizes would be almost certain to be observed at the sampling rates considered in the simulation. Table 3 shows the five most frequent species in our population and five of the 23 least frequent (singleton) observed in 2001. The objective of this example is to estimate the number of distinct species in the area from a Bernoulli sample of the birds in the population. This population contains $N = 2916$ birds of $D = 110$ species. The average class size is large ($\overline{N} = 26.5$) and fairly variable ($cv^2 = 1.7$).

Table 3. Most and least frequent bird species in population.

Common Name of Species	Count 2001	Count 2000	Count 1999	Count 1998
Bufflehead	138	200	152	155
Carolina Wren	130	235	337	186
White-crowned Sparrow	127	444	435	275
Eastern Phoebe	116	66	86	82
Pied-billed Grebe	110	102	107	86
	...			
Black-and-white Warbler	1	0	0	0
Blue Grosbeak	1	0	0	0
Barn Owl	1	2	6	2
Brown Thrasher	1	6	1	1
Canada Goose	1	0	0	0

We attempt to improve estimation of D in this example by using auxiliary information about the size of each class. A useful predictor would be a variable that measures the rareness of each bird species in the circle. Range maps could provide this information, but I used historical data from previous Christmas Bird Counts. For each species, I recorded the number of birds observed in the Austin, Texas circle during the three previous CBC's (2000, 1999 and 1998). These counts are shown for the extreme class sizes in Table 3. This information was summarized into two predictor variables: $X_{1j} = \ln(0.1+ \text{ count in } 1998)$ and

$$X_{2j} = \begin{cases} 1 & \text{if there were fewer than 3 birds of species } j \text{ counted} \\ & \quad \text{in both 2000 and 1999} \\ 0 & \text{otherwise.} \end{cases}$$

One thousand Bernoulli samples with each of the sampling rates $q = 0.10$, 0.15, 0.20, 0.25, 0.30, 0.40 and 0.50 were drawn from the population of birds. $\hat{D}$, $\hat{D}_{uj2a}$ and $\hat{D}_{eq}$ were computed from each sample, along with their MSE over the 1000 trials. Figure 3 shows $\sqrt{RE(\hat{D}, \hat{D}_{eq})}$ and $\sqrt{RE(\hat{D}, \hat{D}_{uj2a})}$ plotted as functions of the sampling rate q. We see that the new estimator is an improvement over $\hat{D}_{eq}$ and $\hat{D}_{uj2a}$ for every sample size considered. However its advantage over the second order estimator is not monotonic over the range of sampling rates, but is greatest at both ends of the range of rates considered. The reason for this lack of monotonicity is that $\hat{D}_{uj2a}$ does not improve smoothly with increasing sample size. Table 4 shows the mean and standard deviation of $\hat{D}$, $\hat{D}_{uj2a}$ and the three components of $\hat{\boldsymbol{\beta}}$ over the 1000 simulated values for each sample size. It shows that the bias of $\hat{D}$ is consistently small, but its standard deviation is large for small sampling rates. This is likely because of the high variability in $\hat{\beta}_2$ for small sample sizes. (Few species were as rare as X_{2j} required, so small samples had little information about β_2.) By contrast, the bias of $\hat{D}_{uj2a}$ is large

Table 4. Sample characteristics for simulation of sampling from Christmas Bird counts for varying sampling rates ($D = 110,\ N = 2916$).

q	$avg(\overline{d})$	$avg(\overline{n})$	$avg(\hat{D})$ $SD(\hat{D})$	$avg(\hat{D}_{uj2a})$ $SD(\hat{D}_{uj2a})$	$avg(\hat{\beta}_0)$ $SD(\hat{\beta}_0)$	$avg(\hat{\beta}_1)$ $SD(\hat{\beta}_1)$	$avg(\hat{\beta}_2)$ $SD(\hat{\beta}_2)$
0.10	65	292	107.3 21.4	81.8 8.2	2.0 0.3	0.49 0.06	-3.3 6.0
0.20	78	594	105.8 10.2	96.5 8.0	1.9 0.2	0.51 0.04	-0.84 0.67
0.30	85	875	106.2 7.1	104.9 7.6	1.9 0.1	0.52 0.03	-0.76 0.26
0.40	91	1167	106.6 5.4	109.8 7.1	1.8 0.1	0.53 0.03	-0.72 0.19
0.50	95	1459	107.4 4.4	113.1 6.5	1.8 0.1	0.54 0.02	-0.71 0.15

until the sampling rate is high, but its standard deviation is low for all sampling rates. However, it has the peculiar characteristic that its standard deviation diminishes only slightly over the range of sampling rates considered (from 8.2 to 7.6 as sample size increases from about 292 to about 1459). We conclude that $\hat{D}_{uj2a}$ performs worse than $\hat{D}$ for small sampling rates because of its large bias, and for large sampling rates because of its large standard deviation. In the middle of the range, both are moderate and $\hat{D}_{uj2a}$ enjoys its best performance against $\hat{D}$.

References

Alho, J. M. (1990). Logistic regression in capture-recapture models, *Biometrics* **46**, 623-635.

Bunge, J. and Fitzpatrick, M. (1993). Estimating the number of species: a review. *J. Amer. Statist. Assoc.* **88**, 364-373.

Chao, A., Ma, M.-C. and Yang, M. C. K. (1993). Stopping rules and estimation for recapture debugging with unequal failure rates, *Biometrika* **80**, 193-201.

Chaudhuri, S., Motwani, R. and Narasayya, V. (1998). *Random Sampling for Histogram Construction: How Much is Enough?* In *Proceedings of the* 1998 *ACM SIGMOD International Conference on the Management of Data*, 436-447.

Deming, W. E. and Glasser, G. J. (1959). On the problem of matching lists by samples, *J. Amer. Statist. Assoc.* **54**, 403-415.

Good, I. J. (1950). *Probability and the Weighing of Evidence*. Charles Griffin, London.

Haas, P. and Stokes, L. (1998). Estimating the number of classes in a finite population, *J. Amer. Statist. Assoc.* **93**, 1475-1487.

Huggins, R. M. (1989). On the statistical analysis of capture experiments, *Biometrika* **76**, 133-140.

Lewontin, R. C. and Prout, T. (1956). Estimation of the number of different classes in a population, *Biometrics* **12**, 211-223.

Sarndal, C. Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer-Verlag, New York.

Sekar, C. C. and Deming, W. E. (1949). On a Method of Estimating Birth and Death Records and the extent of registration, *J. Amer. Statist. Assoc.* **44**, 101-115.

Sichel, H. S. (1997). Modeling species-abundance frequencies and species-individual functions with the generalized inverse Gaussian-Poisson distribution, *South African Statist. J.* **31**, 13-37.

Smith-Cayama, R. A. and Thomas, D. R. (1999). Estimating the number of distinct valid signatures in initiative petitions, *Presented at the* 1999 *Joint Statistical Meetings*.

Department of Statistical Science, Southern Methodist University, Dallas, TX, U.S.A.

E-mail: slstokes@mail.smu.edu

(Received October 2000; accepted December 2002)