Estimating the Number of Classes in a Finite Population

Peter J. HAAS and Lynne STOKES

We use an extension of the generalized jackknife approach of Gray and Schucany to obtain new nonparametric estimators for the number of classes in a finite population of known size. We also show that generalized jackknife estimators are closely related to certain Horvitz–Thompson estimators, to an estimator of Shlosser, and to estimators based on sample coverage. In particular, the generalized jackknife approach leads to a modification of Shlosser's estimator that does not suffer from the erratic behavior of the original estimator. The performance of both new and previous estimators is investigated by means of an asymptotic variance analysis and a Monte Carlo simulation study.

KEY WORDS: Census; Database; Jackknife; Number of classes; Number of species; Sample coverage.

1. INTRODUCTION

The problem of estimating the number of classes in a population has been studied for many years. A recent review article (Bunge and Fitzpatrick 1993) lists more than 125 references. In this article we consider an important special case of the general problem—estimating the number of classes in a finite population of known size. Only a handful of papers have addressed this problem and none has reached an entirely satisfactory solution, despite the fact that the first attempt at solution appeared in the statistical literature nearly 50 years ago (Mosteller 1949). The problem we consider has arisen in the literature in a variety of applications, including the following.

a. In a company-sponsored contest, many entries (say several hundred thousand) have been received. It is known that some people have entered more than once. The goal is to estimate the number of different people who have entered from a sample of entries (Mosteller 1949; Sudman 1976).
b. A sampling frame is constructed by combining a number of lists that may contain overlapping entries. It is desired to estimate, using a sample from all lists, the number of units on the combined list (Deming and Glasser 1959; Goodman 1952; Kish 1965; Sudman 1976). An important example of such a problem is an "administrative records census," currently under study by the U.S. Bureau of the Census. In such a census, several administrative files (such as AFDC or IRS records) are combined, and the total number of distinct individuals included in the combined file is determined. Exact computation of the number of distinct individuals in the combined file is extremely expensive because of the high cost of determining the number of duplicated entries. A similar problem and proposed solution was discussed in the *London Financial Times* (March 2, 1949) by C. F. Carter, who was interested in estimating the number of different investors in British industrial stocks based on samples from share registers of companies (Mosteller 1949).
c. In a relational database management system, data are organized in tables called *relations;* (e.g. Korth and Silberschatz 1991, chap. 3). In a typical relation, each row might represent a record for an individual employee in a company, and each column might correspond to a different *attribute* of the employee, such as salary, years of experience, department number, and so forth. A relational *query* specifies an output relation that is to be computed from the set of base relations stored by the system. Knowledge of the number of distinct values for each attribute in the base relations is central to determining the most efficient method for computing a specified output relation (Hellerstein and Stonebraker 1994; Selinger, Astrahan, Chamberlain, Lorie, and Price 1979). The size of the base relations in modern database systems often is so large that exact computation of the distinct-value parameters is prohibitively expensive, and thus estimation of these parameters is desired (Astrahan, Schkolnick, and Whang 1987; Flajolet and Martin 1985; Gelenbe and Gardy 1982; Hou, Ozsoyoglu, and Taneja 1988, 1989; Naughton and Seshadri 1990; Ozsoyoglu, Du, Tjahjana, Hou, and Rowland 1991; Whang, Vander-Zanden, and Taylor 1990).

In each of these applications, the size of the population (number of contest entries, total number of units over all lists, and number of rows in the base relation) is known, and this size is too large for easy computation of the number of classes.

The problem studied in this article can be described formally as follows. A population of size N consists of D mutually disjoint classes of items, labelled $C_1, C_2, \ldots, C_D$. Define N_j to be the size of class C_j, so that $N = \sum_{j=1}^{D} N_j$. A simple random sample of n items is selected (without

Peter J. Haas is Research Staff Member, IBM Research Division, Almaden Research Center, San Jose, CA 95120 (E-mail: peterh@almaden.ibm.com). Lynne Stokes is Professor, Department of Management Science and Information Systems, University of Texas, Austin, TX 78712 (E-mail: lstokes@mail.utexas.edu). Hongmin Lu made substantial contributions to the early phases of the work reported here, and an anonymous reviewer suggested the "stabilization" technique for $\hat{D}_{\mathrm{uj2}}$ used in Section 6. The Wisconsin student database was graciously provided by Bob Nolan of the University of Wisconsin-Madison Department of Information Technology (DoIT) and Jeff Naughton of the University of Wisconsin-Madison Computer Sciences Department.

Journal of the American Statistical Association
December 1998, Vol. 93, No. 444, Theory and Methods

replacement) from the population. This sample includes n_j items from class C_j. The problem we consider is estimating D using information from the sample along with knowledge of the value of N. We denote by F_i the number of classes of size i in the population, so that $D = \sum_{i=1}^{N} F_i$ and $\sum_{i=1}^{N} iF_i = N$. Similarly, we denote by f_i the number of classes represented exactly i times in the sample and by d the total number of classes represented in the sample, so that $d = \sum_{i=1}^{n} f_i$ and $\sum_{i=1}^{n} if_i = n$. Define vectors $\mathbf{N} = (N_1, N_2, \ldots, N_D)$, $\mathbf{n} = (n_1, n_2, \ldots, n_D)$, and $\mathbf{f} = (f_1, f_2, \ldots, f_n)$. Note that $\mathbf{n}$ is not observable, but $\mathbf{f}$ is. Because we sample without replacement, the random vector $\mathbf{n}$ has a multivariate hypergeometric distribution with probability mass function

$$P(\mathbf{n}|D, \mathbf{N}) = \binom{N_1}{n_1}\binom{N_2}{n_2}\cdots\binom{N_D}{n_D}\Big/\binom{N}{n}. \quad (1)$$

The probability mass function of the observable random vector $\mathbf{f}$ is simply $P(\mathbf{n}|D, \mathbf{N})$ summed over all points $\mathbf{n}$ that correspond to $\mathbf{f}$, as in

$$P(\mathbf{f}|D, \mathbf{N}) = \sum_{S} P(\mathbf{n}|D, \mathbf{N}),$$

where $S = \{\mathbf{n}\colon \#(n_j = i) = f_i \text{ for } 1 \le i \le D\}$. The probability mass function $P(\mathbf{f}|D, \mathbf{N})$ does not have a closed-form expression in general.

In Section 2 we review the estimators that have been proposed for estimating D from data generated under model (1). In Section 3 we provide several new estimators of D based on an extension of the generalized jackknife approach of Gray and Schucany (1972). We then show that generalized jackknife estimators of the number of classes in a population are closely related to certain "Horvitz–Thompson" estimators, to an estimator due to Shlosser (1981), and to estimators based on the notion of "sample coverage" (Chao and Lee 1992). In Section 4 we provide and compare approximate expressions for the asymptotic variance of several of the estimators, and in Section 5 apply our formulas to a well-known example from the literature. We provide a simulation-based empirical comparison of the various estimators in Section 6, and summarize our results and give recommendations in Section 7.

2. PREVIOUS ESTIMATORS

Bunge and Fitzpatrick (1993) mention only two non-Bayesian estimators that have been developed as estimators of D under model (1). These are the estimators of Goodman (1949) and Shlosser (1981). Goodman proved that

$$\hat{D}_{\text{Good1}} = d + \sum_{i=1}^{n} (-1)^{i+1} \frac{(N-n+i-1)!(n-i)!}{(N-n-1)!n!} f_i$$

is the unique unbiased estimator of D when $n > M \stackrel{\text{def}}{=} \max(N_1, N_2, \ldots, N_D)$. He further proved that no unbiased estimator of D exists when $n \le M$. Unfortunately, unless the sampling fraction is quite large, the variance of $\hat{D}_{\text{Good1}}$ is so great and the numerical difficulties encountered when computing $\hat{D}_{\text{Good1}}$ are so severe that the estimator is unusable. Goodman, who made note of the high variance of $\hat{D}_{\text{Good1}}$ himself, suggested the alternative estimator

$$\hat{D}_{\text{Good2}} = N - \frac{N(N-1)}{n(n-1)} f_2$$

for overcoming the variance problem. Although $\hat{D}_{\text{Good2}}$ has lower variance than $\hat{D}_{\text{Good1}}$, it can take on negative values and can have a large bias for any n if D is small. For example, consider the case in which $D = 1$ and $n > 2$, and observe that $f_2 = 0$ and $\hat{D}_{\text{Good2}} = N$.

Under the assumption that the population size N is large and the sampling fraction $q = n/N$ is nonnegligible, Shlosser (1981) derived the estimator

$$\hat{D}_{\text{Sh}} = d + f_1 \frac{\sum_{i=1}^{n}(1-q)^i f_i}{\sum_{i=1}^{n} iq(1-q)^{i-1} f_i}.$$

For the two examples considered in his article, Shlosser found that using $\hat{D}_{\text{Sh}}$ with a 10% sampling fraction resulted in an error rate below 20%. In our experiments, however, we observed root mean squared errors (RMSEs) exceeding 200%, even for well-behaved populations with relatively little variation among the class sizes; see Section 6. Considering the relationship between $\hat{D}_{\text{Sh}}$ and generalized jackknife estimators (see Sec. 3.4) provides insight into the source of this erratic behavior and suggests some possible modifications of $\hat{D}_{\text{Sh}}$ to improve performance.

In related work, Burnham and Overton (1978, 1979) proposed a family of (traditional) generalized jackknife estimators for estimating the size of a closed population when capture probabilities vary among animals. The D individuals in the population play the role of our D classes; a given individual can appear up to n times in the overall sample if captured on one or more of n possible trapping occasions. The capture probability for an individual is assumed to be constant over time, and the capture probabilities for the D individuals are modeled as D iid random samples from a fixed probability distribution. Burnham and Overton's sample design is clearly different from model (1). Under the Burnham and Overton model, for example, the quantities $f_1, f_2, \ldots, f_n$ have a joint multinomial distribution. Closely related to the work of Burnham and Overton are the ordinary jackknife estimators of the number of species in a closed region developed by Heltshe and Forrester (1983) and Smith and van Belle (1984). The sample data consist of a list of the species that appear in each of n quadrats. (The number of times that a species is represented in a quadrat is not recorded.) This setup is essentially identical to that of Burnham and Overton, with the D species playing the role of the D individuals and the n quadrats playing the role of the n trapping occasions.

3. GENERALIZED JACKKNIFE ESTIMATORS

In this section we outline an extension of the generalized jackknife approach to bias reduction and then use this approach to derive new estimators for the number of classes in a finite population. We also point out connections between

our generalized jackknife approach and several other estimation approaches in the literature.

3.1 The Generalized Jackknife Approach

Let θ be an unknown real-valued parameter. A *generalized jackknife estimator* of θ is an estimator of the form

$$G(\hat{\theta}_1, \hat{\theta}_2) = \frac{\hat{\theta}_1 - R\hat{\theta}_2}{1 - R}, \tag{2}$$

where $\hat{\theta}_1$ and $\hat{\theta}_2$ are biased estimators of θ and $R(\neq 1)$ is a real number (Gray and Schucany 1972). The idea underlying the generalized jackknife approach is to try and choose R such that $G(\hat{\theta}_1, \hat{\theta}_2)$ has lower bias than either $\hat{\theta}_1$ or $\hat{\theta}_2$. To motivate the choice of R, observe that for

$$R = \frac{E[\hat{\theta}_1] - \theta}{E[\hat{\theta}_2] - \theta}, \tag{3}$$

the estimator $G(\hat{\theta}_1, \hat{\theta}_2)$ is unbiased for θ. This optimal value of R is typically unknown, however, and can only be approximated, resulting in bias reduction but not complete bias elimination. In the following, we extend the original definition of the generalized jackknife given by Gray and Schucany (1972) by allowing R to depend on the data; that is, we allow R to be random.

Recall that d is the number of classes represented in the sample. Write d_n for d to emphasize the dependence of d on the sample size, and denote by $d_{n-1}(k)$ the number of classes represented in the sample after the kth observation has been removed. Set $d_{(n-1)} = (1/n)\sum_{k=1}^{n} d_{n-1}(k)$. We focus on generalized jackknife estimators that are obtained by taking $\hat{\theta}_1 = d_n$ and $\hat{\theta}_2 = d_{(n-1)}$ in (2); these are the usual choices for $\hat{\theta}_1$ and $\hat{\theta}_2$ in the classical first-order jackknife estimator (Miller 1974). Observe that $d_{n-1}(k) = d_n - 1$ if the class for the kth observation is represented only once in the sample; otherwise, $d_{n-1}(k) = d_n$. Thus $d_{(n-1)} = d_n - (f_1/n)$ and, by (2), $G(\hat{\theta}_1, \hat{\theta}_2) = \hat{D}$, where

$$\hat{D} = d_n + K\frac{f_1}{n} \tag{4}$$

and $K = R/(1 - R)$. It follows from (3) that the optimal choice of K is

$$K = \frac{E[d_n] - D}{E[d_{(n-1)}] - E[d_n]} = \frac{D - E[d_n]}{E[f_1]/n}. \tag{5}$$

To derive a more explicit formula for K, denote by $I[A]$ the indicator of event A and observe that

$$E[d_n] = E\left[\sum_{j=1}^{D} I[n_j > 0]\right]$$

$$= \sum_{j=1}^{D} P\{n_j > 0\} = D - \sum_{j=1}^{D} P\{n_j = 0\}.$$

Similar reasoning shows that

$$E[f_1] = \sum_{j=1}^{D} P\{n_j = 1\}, \tag{6}$$

so that

$$K = n\frac{\sum_{j=1}^{D} P\{n_j = 0\}}{\sum_{j=1}^{D} P\{n_j = 1\}}. \tag{7}$$

Following Shlosser (1981), we focus on the case in which the population size N is large and the sampling fraction $q = n/N$ is nonnegligible, and we make the approximation

$$P\{n_j = k\} \approx \binom{N_j}{k} q^k (1 - q)^{N_j - k} \tag{8}$$

for $0 \le k \le n$ and $1 \le j \le D$. That is, the probability distribution of each n_j is approximated by the probability distribution of n_j under a Bernoulli sample design in which each item is included in the sample with probability q, independently of all other items in the population. Using this approximation leads to estimators that behave almost identically to estimators derived using the exact distribution of $\mathbf{n}$ but are simpler to compute and derive (for further discussion see Haas and Stokes 1996, app. A). Substituting (8) into (7), we obtain

$$K \approx n\frac{\sum_{j=1}^{D}(1 - q)^{N_j}}{\sum_{j=1}^{D} N_j q(1 - q)^{N_j - 1}}. \tag{9}$$

The quantity K defined in (9) depends on unknown parameters $N_1, N_2, \ldots, N_D$ that are difficult to estimate. Our approach is to approximate K by a function of D and of other parameters that are easier to estimate, thereby obtaining an approximate version of (4). The estimates for these parameters, including $\hat{D}$ for D, are then substituted into the approximate version of (4), and the resulting equation is solved for $\hat{D}$.

We also consider "smoothed" jackknife estimators. The idea is to replace the quantity f_1/n in (4) by its expected value $E[f_1]/n$ in the hope that the resulting estimator of D will be more stable than the original "unsmoothed" estimator. As with the parameter K, the quantity $E[f_1]/n$ depends on the unknown parameters $N_1, N_2, \ldots, N_D$; see (6) and (8). Thus our approach to estimating $E[f_1]/n$ is the same as our approach to estimating K.

Estimators also can be based on high-order jackknifing schemes that consider the number of distinct values in the sample when two elements are removed, when three elements are removed, and so forth. Typically, using a high-order jackknifing scheme requires estimating high-order moments (skewness, kurtosis, and so forth) of the set of numbers $\{N_1, N_2, \ldots, N_D\}$. Initial experiments indicated that the reduction in estimation error due to using high-order jackknife is outweighed by the increase in error due to uncertainty in the moment estimates. Thus we do not pursue high-order jackknife schemes further.

3.2 The Estimators

Different approximations for K and $E[f_1]/n$ lead to different estimators for D. Here we develop a number of the possible estimators.

3.2.1 First-Order Estimators. The simplest estimators of D can be derived using a first-order approximation to K. Specifically, approximate each N_j in (9) by the average value $\bar{N} = (1/D)\sum_{j=1}^{D} N_j = N/D$, and substitute the resulting expression for K into (4) to obtain

$$\hat{D} = d_n + \frac{(1-q)f_1 D}{n}. \tag{10}$$

Now substitute $\hat{D}$ for D on the right side of (10) and solve for $\hat{D}$. The resulting solution, denoted by $\hat{D}_{\text{uj1}}$, is given by

$$\hat{D}_{\text{uj1}} = \left(1 - \frac{(1-q)f_1}{n}\right)^{-1} d_n. \tag{11}$$

We refer to this estimator as the "unsmoothed first-order jackknife estimator."

To derive a "smoothed first-order jackknife estimator," observe that by (6) and (8),

$$\frac{E[f_1]}{n} \approx \frac{1}{n}\sum_{j=1}^{D} N_j q(1-q)^{N_j-1}. \tag{12}$$

Approximating each N_j in (12) by $\bar{N}$, we have

$$\frac{E[f_1]}{n} \approx (1-q)^{\bar{N}-1}. \tag{13}$$

On the right side of (10), replace f_1/n by the approximate expression for $E[f_1]/n$ given in (13), yielding

$$\hat{D} = d_n + D(1-q)^{\bar{N}}.$$

Replacing D with $\hat{D}$ and $\bar{N}$ with $N/\hat{D}$ in the foregoing expression leads to the relation

$$\hat{D}(1-(1-q)^{N/\hat{D}}) = d_n.$$

We define the smoothed first-order jackknife estimator $\hat{D}_{\text{sj1}}$ as the value of $\hat{D}$ that solves the above equation. Given d_n, n, and N, $\hat{D}_{\text{sj1}}$ can be computed numerically using standard root-finding procedures. Observe that if in fact $N_1 = N_2 = \cdots = N_D = N/D$, then

$$E[d_n] \approx D(1-(1-q)^{N/D}).$$

In this case $\hat{D}_{\text{sj1}}$ can be viewed as a simple method-of-moments estimator obtained by replacing $E[d_n]$ by the estimate d_n and solving for D. If, moreover, the sampling fraction q is small enough so that the distribution of $(n_1, n_2, \ldots, n_D)$ is approximately multinomial (see Sec. 3.3), then $\hat{D}_{\text{sj1}}$ is approximately equal to the maximum likelihood estimator for D (see Good 1950). Observe that both $\hat{D}_{\text{uj1}}$ and $\hat{D}_{\text{sj1}}$ are consistent for D: $\hat{D}_{\text{uj1}} \to D$ and $\hat{D}_{\text{sj1}} \to D$ as $q \to 1$.

3.2.2 Second-Order Estimators. A second-order approximation to K can be derived as follows. Denote by γ^2 the squared coefficient of variation of the class sizes $N_1, N_2, \ldots, N_D$,

$$\gamma^2 = \frac{(1/D)\sum_{j=1}^{D}(N_j - \bar{N})^2}{\bar{N}^2}. \tag{14}$$

Suppose that γ^2 is relatively small, so that each N_j is "close" to the average value $\bar{N}$. Substitute the Taylor approximations

$$(1-q)^{N_j} \approx (1-q)^{\bar{N}} + (1-q)^{\bar{N}}\ln(1-q)(N_j - \bar{N})$$

and

$$N_j q(1-q)^{N_j-1} \approx N_j q((1-q)^{\bar{N}-1} + (1-q)^{\bar{N}-1}\ln(1-q)(N_j-\bar{N}))$$

for $1 \le j \le D$ into (9) to obtain

$$\begin{aligned} K &\approx D(1-q)\left(\frac{1}{1+\ln(1-q)\bar{N}\gamma^2}\right) \\ &\approx D(1-q)(1-\ln(1-q)\bar{N}\gamma^2). \end{aligned} \tag{15}$$

The unknown parameter γ^2 can be estimated using the following approach (cf. Chao and Lee 1992). With the usual convention that $\binom{n}{m} = 0$ for $n < m$, we find that

$$\begin{aligned} \sum_{i=1}^{N} i(i-1)E[f_i] &\approx \sum_{i=1}^{N} i(i-1)\sum_{j=1}^{D}\binom{N_j}{i} q^i(1-q)^{N_j-i} \\ &= q^2\sum_{j=1}^{D} N_j(N_j-1)\sum_{i=2}^{N_j}\binom{N_j-2}{i-2} \\ &\quad \times q^{i-2}(1-q)^{N_j-i} \\ &= q^2\sum_{j=1}^{D} N_j(N_j-1), \end{aligned}$$

so that

$$\gamma^2 \approx \frac{D}{n^2}\sum_{i=1}^{N} i(i-1)E[f_i] + \frac{D}{N} - 1.$$

Thus if D were known, then a natural method-of-moments estimator $\hat{\gamma}^2(D)$ of γ^2 would be

$$\hat{\gamma}^2(D) = \max\left(0, \frac{D}{n^2}\sum_{i=1}^{n} i(i-1)f_i + \frac{D}{N} - 1\right). \tag{16}$$

To develop a second-order estimate of D, substitute (15) into (4) to obtain

$$\hat{D} = d_n + \frac{Df_1(1-q)}{n}(1-\ln(1-q)\bar{N}\gamma^2), \tag{17}$$

from which it follows that

$$\hat{D} = d_n + \frac{Df_1(1-q)}{n} - \frac{f_1(1-q)\ln(1-q)\gamma^2}{q}.$$

Replacing D by $\hat{D}$ on the right side of this equation and solving for $\hat{D}$ yields the relation

$$\left(1 - \frac{f_1(1-q)}{n}\right)\hat{D} = d_n - \frac{f_1(1-q)\ln(1-q)\gamma^2}{q}. \tag{18}$$

An estimator of D can be obtained by substituting $\hat{\gamma}^2(\hat{D})$ for γ^2 in (18) and solving for $\hat{D}$ numerically. Alternatively, we can start with a simple initial estimator of D and then correct this estimator using (18). Following this latter approach, we use $\hat{D}_{\mathrm{uj1}}$ as our initial estimator and define

$$\hat{D}_{\mathrm{uj2}} = \left(1 - \frac{f_1(1-q)}{n}\right)^{-1} \times \left(d_n - \frac{f_1(1-q)\ln(1-q)\hat{\gamma}^2(\hat{D}_{\mathrm{uj1}})}{q}\right).$$

A smoothed second-order jackknife estimator can be obtained by replacing the expression f_1/n in (17) by the approximation to $E[f_1]/n$ given in (13), leading to

$$\hat{D} = d_n + D(1-q)^{\bar{N}}(1 - \ln(1-q)\bar{N}\gamma^2).$$

Replacing D with $\hat{D}$ and proceeding as before, we obtain the estimator

$$\hat{D}_{\mathrm{sj2}} = (1-(1-q)^{\tilde{N}})^{-1}(d_n - (1-q)^{\tilde{N}}\ln(1-q)N\hat{\gamma}^2(\hat{D}_{\mathrm{uj1}})),$$

where $\tilde{N} = N/\hat{D}_{\mathrm{uj1}}$. As with the first-order estimators $\hat{D}_{\mathrm{uj1}}$ and $\hat{D}_{\mathrm{sj1}}$, the second-order estimators $\hat{D}_{\mathrm{uj2}}$ and $\hat{D}_{\mathrm{sj2}}$ are consistent for D.

3.2.3 Horvitz–Thompson Jackknife Estimators. In this section we discuss an alternative approach to estimation of K based on a technique of Horvitz and Thompson. (See Sarndal, Swensson, and Wretman 1992 for a general discussion of Horvitz–Thompson estimators.) First, consider the general problem of estimating a parameter of the form $\theta(g) = \sum_{j=1}^{D} g(N_j)$, where g is a specified function. Observe that because $P\{n_j > 0\} > 0$ for $1 \le j \le D$, we have $\theta(g) = E[X(g)]$, where

$$X(g) = \sum_{j=1}^{D} \frac{g(N_j)I(n_j > 0)}{P\{n_j > 0\}} = \sum_{\{j:n_j>0\}} \frac{g(N_j)}{P\{n_j > 0\}}.$$

It follows from (8) that $P\{n_j > 0\} \approx 1-(1-q)^{N_j}$, and the foregoing discussion suggests that we estimate $\theta(g)$ by

$$\hat{\theta}(g) = \sum_{\{j:n_j>0\}} \frac{g(\hat{N}_j)}{1-(1-q)^{\hat{N}_j}}, \tag{19}$$

where $\hat{N}_j$ is an estimator for N_j. The key point is that we need to estimate N_j only when $n_j > 0$. To do this, observe that

$$E[n_j | n_j > 0] = \frac{E[n_j]}{P\{n_j > 0\}} \approx \frac{qN_j}{1-(1-q)^{N_j}}.$$

Replacing $E[n_j|n_j > 0]$ by n_j leads to the estimating equation

$$n_j = \frac{qN_j}{1-(1-q)^{N_j}}, \tag{20}$$

and a method-of-moments estimator $\hat{N}_j$ can be defined as the value of N_j that solves (20).

Now consider the problem of estimating K, and hence D. By (9), $K \approx \theta(f)/\theta(g)$, where $f(x) = (1-q)^x$ and $g(x) = xq(1-q)^{x-1}/n$. Thus a natural estimator of K is given by $\hat{\theta}(f)/\hat{\theta}(g)$, leading to the final estimator,

$$\hat{D}_{\mathrm{HTj}} = d_n + \frac{\hat{\theta}(f)}{\hat{\theta}(g)}\,\frac{f_1}{n}.$$

A smoothed variant of $\hat{D}_{\mathrm{HTj}}$ can be obtained by replacing f_1/n with the Horvitz–Thompson estimator of $E[f_1]/n$, namely $\hat{\theta}(g)$. The resulting estimator, denoted by $\hat{D}_{\mathrm{HTsj}}$, is given by

$$\hat{D}_{\mathrm{HTsj}} = d_n + \hat{\theta}(f).$$

Finally, a hybrid estimator can be obtained using a first-order approximation for the numerator of K and a Horvitz–Thompson estimator for the denominator. This leads to the estimator $\hat{D}_{\mathrm{hj}}$, defined as the solution $\hat{D}$ of the equation

$$\hat{D}\left(1 - \frac{f_1(1-q)^{N/\hat{D}}}{n\hat{\theta}(g)}\right) = d_n.$$

If we replace f_1/n by the Horvitz–Thompson estimator for $E[f_1]/n$ in the foregoing equation to obtain a smoothed variant of $\hat{D}_{\mathrm{hj}}$, then the resulting estimator coincides with $\hat{D}_{\mathrm{sj1}}$.

Because $D = \theta(u)$, where $u(x) \equiv 1$, it may appear that a "non-jackknife" Horvitz–Thompson estimator $\hat{D}_{\mathrm{HT}}$ can be defined by setting $\hat{D}_{\mathrm{HT}} = \hat{\theta}(u)$. It is straightforward to show, however, that $\hat{D}_{\mathrm{HT}} = \hat{D}_{\mathrm{HTsj}}$, so that $\hat{D}_{\mathrm{HT}}$ can in fact be viewed as a smoothed jackknife estimator.

Simulation experiments indicate that the behavior of the Horvitz–Thompson jackknife estimators $\hat{D}_{\mathrm{HTj}}$ and $\hat{D}_{\mathrm{HTsj}}$ is erratic (see Haas and Stokes 1996, app. D for detailed results). Overall, the poor performance of $\hat{D}_{\mathrm{HTj}}$ and $\hat{D}_{\mathrm{HTsj}}$ is caused by inaccurate estimation of $\hat{\theta}(f)$. The problem seems to be that when N_j is small, the estimator $\hat{N}_j$ is unstable and yet typically has a large effect on the value of $\hat{\theta}(f)$ through the term $(1-q)^{\hat{N}_j}/(1-(1-q)^{\hat{N}_j})$. The estimator $\hat{D}_{\mathrm{hj}}$ uses a Taylor approximation in place of $\hat{\theta}(f)$ and hence has lower bias and RMSE than the other two Horvitz–Thompson jackknife estimators. However, other estimators perform better than $\hat{D}_{\mathrm{hj}}$, and we do not consider the estimators $\hat{D}_{\mathrm{HTj}}$, $\hat{D}_{\mathrm{HTsj}}$, and $\hat{D}_{\mathrm{hj}}$ further.

3.3 Relation to Estimators Based on Sample Coverage

The generalized jackknife approach for deriving an estimator of D works for sample designs other than hypergeometric sampling. For example, the most thoroughly studied version of the number-of-classes problem is that in which the population is assumed to be infinite and $\mathbf{n}$ is assumed to have a multinomial distribution with parameter vector $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_D)$; that is,

$$P(\mathbf{n}|D, \boldsymbol{\pi}) = \binom{n}{n_1 n_2 \cdots n_D} \pi_1^{n_1}\pi_2^{n_2}\ldots\pi_D^{n_D}. \tag{21}$$

When we proceed as in Section 3.1 to derive a generalized jackknife estimator under the model in (21), the estimator turns out to be nearly identical to the "coverage-based" estimator proposed by Chao and Lee (1992). To see this, start again with (4) and select K as in (5). Because $E[d_n] - D = -\sum_{j=1}^{D}(1-\pi_j)^n$ under the model in (21), it follows that

$$K = \frac{\sum_{j=1}^{D} v_n(\pi_j)}{\sum_{j=1}^{D} \pi_j v_{n-1}(\pi_j)},$$

where $v_n(x) = (1-x)^n$. Set $\bar{\pi} = 1/D$ and use the Taylor approximations $v_n(\pi_j) \approx v_n(\bar{\pi}) + (\pi_j - \bar{\pi})v_n'(\bar{\pi})$ and $\pi_j v_{n-1}(\pi_j) \approx \pi_j(v_{n-1}(\bar{\pi}) + (\pi_j - \bar{\pi})v_{n-1}'(\bar{\pi}))$ in a manner analogous to the derivation in Section 3.2.2 to obtain

$$K \approx (D-1) + (n-1)\gamma^2, \tag{22}$$

where $\gamma^2 = -1 + D\sum_{j=1}^{D}\pi_j^2$ is the squared coefficient of variation of the numbers $\pi_1, \pi_2, \ldots, \pi_D$. Denote by $\hat{D}_{\text{mult}}$ the estimator of D under the multinomial model. Then, by (4),

$$\hat{D}_{\text{mult}} = d_n + ((D-1) + (n-1)\gamma^2)\frac{f_1}{n}. \tag{23}$$

Replace D with $\hat{D}_{\text{mult}}$ and γ^2 with an estimator $\tilde{\gamma}^2$ in (23), and solve for $\hat{D}_{\text{mult}}$ to obtain

$$\hat{D}_{\text{mult}} = \frac{d_n}{\hat{C}} + \frac{n(1-\hat{C})}{\hat{C}}\left(\frac{n-1}{n}\tilde{\gamma}^2 - \frac{1}{n}\right),$$

where $\hat{C} = 1 - (f_1/n)$. When the sample size n is large, the estimator $\hat{D}_{\text{mult}}$ is essentially the same as the estimator

$$\hat{D}_{\text{CL}} = \frac{d_n}{\hat{C}} + \frac{n(1-\hat{C})}{\hat{C}}\tilde{\gamma}^2$$

proposed by Chao and Lee (1992). The estimator $\hat{D}_{\text{CL}}$ was developed from a different point of view, using the concept of *sample coverage*. The sample coverage for an infinite population is defined as $\sum_{j=1}^{D}\pi_j I[n_j > 0]$, and the quantity $\hat{C} = 1 - (f_1/n)$ is a standard estimator of the sample coverage.

Conversely, when Chao and Lee's derivation is modified to account for hypergeometric sampling, the resulting estimator is equal to $\hat{D}_{\text{uj2}}$ (see Haas and Stokes 1996, app. B). Thus at least some estimators based on sample coverage can be viewed as generalized jackknife estimators.

3.4 Relation to Shlosser's Estimator

Observe that the estimator $\hat{D}_{\text{Sh}}$, though not developed from a jackknife perspective, can be viewed as an estimator of the form (4) with K estimated by

$$\hat{K}_{\text{Sh}} = n\frac{\sum_{i=1}^{n}(1-q)^i f_i}{\sum_{i=1}^{n} iq(1-q)^{i-1} f_i}.$$

To analyze the behavior of $\hat{D}_{\text{Sh}}$, we first rewrite the jackknife quantity K defined in (9) as follows:

$$K = n\frac{\sum_{i=1}^{N}(1-q)^i F_i}{\sum_{i=1}^{N} iq(1-q)^{i-1} F_i}. \tag{24}$$

Shlosser's justification of $\hat{D}_{\text{Sh}}$ assumes that

$$\frac{E[f_i]}{E[f_1]} \approx \frac{F_i}{F_1} \tag{25}$$

for $1 \le i \le N$. When the assumption in (25) holds and the sample size is large enough so that

$$f_i \approx E[f_i] \tag{26}$$

for $1 \le i \le N$, it is easy to see that $\hat{K}_{\text{Sh}} \approx K$, so that $\hat{D}_{\text{Sh}}$ behaves as a generalized jackknife estimator. Although the relations in (25) and (26) hold exactly for $n = N$ (implying that $\hat{D}_{\text{Sh}}$ is consistent for D), these relations can fail drastically for smaller sample sizes. For example, when $F_1 = 0$ and $F_i > 0$ for some $i > 1$, the right side of (25) is infinite, whereas the left side is finite for n sufficiently small. This observation leads one to expect that $\hat{D}_{\text{Sh}}$ will not perform well when the sample size is relatively small and $N_1, N_2, \ldots, N_D$ have similar values (with $N_j > 1$ for each j). Both the variance analysis in Section 4 and the simulation experiments described in Section 6 bear out this conjecture.

The foregoing discussion suggests that replacing $\hat{K}_{\text{Sh}}$ with

$$\hat{K}_{\text{Sh}}^{*} = \frac{K}{E[\hat{K}_{\text{Sh}}]}\hat{K}_{\text{Sh}} \tag{27}$$

in the formula for $\hat{D}_{\text{Sh}}$ might result in an improved estimator, because $\hat{K}_{\text{Sh}}^{*}$ is unbiased for K. Of course we cannot perform this replacement exactly, because K and $E[\hat{K}_{\text{Sh}}]$ are unknown, but we can approximate $\hat{K}_{\text{Sh}}^{*}$ as follows. Using the fact that

$$\begin{aligned} E[f_r] &= \sum_{j=1}^{D} P\{n_j = r\} \\ &\approx \sum_{j=1}^{D}\binom{N_j}{r} q^r(1-q)^{N_j - r} \\ &= \sum_{i=r}^{N}\binom{i}{r} q^r(1-q)^{i-r}F_i \end{aligned} \tag{28}$$

for $1 \le r \le n$, we have, to first order,

$$\begin{aligned} E[\hat{K}_{\text{Sh}}] &\approx n\frac{\sum_{i=1}^{N}(1-q)^i E[f_i]}{\sum_{i=1}^{N} iq(1-q)^{i-1}E[f_i]} \\ &= n\frac{\sum_{i=1}^{N}(1-q)^i((1+q)^i - 1)F_i}{\sum_{i=1}^{N} iq^2(1-q^2)^{i-1}F_i}. \end{aligned} \tag{29}$$

Using the first-order approximation $N_1 = N_2 = \cdots = N_D = \bar{N}$ together with (24), (27), and (29), we find that

$$\hat{K}_{\text{Sh}}^{*} \approx \left(\frac{q(1+q)^{\bar{N}-1}}{(1+q)^{\bar{N}} - 1}\right)\hat{K}_{\text{Sh}}.$$

We thus obtain a modified Shlosser estimator given by

$$\hat{D}_{\mathrm{Sh2}} = d_n + f_1 \left(\frac{q(1+q)^{\tilde{N}-1}}{(1+q)^{\tilde{N}}-1} \right) \left(\frac{\sum_{i=1}^{n}(1-q)^i f_i}{\sum_{i=1}^{n} iq(1-q)^{i-1} f_i} \right),$$

where $\tilde{N}$ is an initial estimate of $\bar{N}$ based on an initial estimate of D. We set $\tilde{N}$ equal to $N/\hat{D}_{\mathrm{uj1}}$ throughout. As with $\hat{D}_{\mathrm{Sh}}$, the estimator $\hat{D}_{\mathrm{Sh2}}$ is consistent for D.

An alternative consistent estimator of D can be obtained by directly using the expressions in (24), (27), and (29) with F_i estimated by

$$\hat{F}_i = \frac{f_1 f_i}{\sum_{j=1}^{n} jq(1-q)^{j-1} f_j} \tag{30}$$

for $1 \le i \le N$; these estimators of $F_1, F_2, \ldots, F_N$ were proposed by Shlosser (1981) in conjunction with the estimator $\hat{D}_{\mathrm{Sh}}$. Substituting the resulting estimator of K and $E[\hat{K}_{\mathrm{Sh}}]$ into (27) leads to the final estimator,

$$\hat{D}_{\mathrm{Sh3}} = d_n + f_1 \left(\frac{\sum_{i=1}^{n} iq^2(1-q^2)^{i-1} f_i}{\sum_{i=1}^{n}(1-q)^i((1+q)^i-1) f_i} \right) \times \left(\frac{\sum_{i=1}^{n}(1-q)^i f_i}{\sum_{i=1}^{n} iq(1-q)^{i-1} f_i} \right)^2.$$

As with the estimator $\hat{D}_{\mathrm{Sh}}$, Shlosser's justification of the estimators in (30) rests on the assumption in (25). Thus one might expect that, like $\hat{D}_{\mathrm{Sh}}$, the estimator $\hat{D}_{\mathrm{Sh3}}$ will be unstable when the sample size is relatively small and $N_1, N_2, \ldots, N_D$ have similar values. On the other hand, the reduction in bias of $\hat{K}^*_{\mathrm{Sh}}$ relative to $\hat{K}_{\mathrm{Sh}}$ leads one to expect that $\hat{D}_{\mathrm{Sh3}}$ will perform better than $\hat{D}_{\mathrm{Sh}}$ when γ^2 is sufficiently large. One might be tempted to avoid the assumption in (25) when estimating $F_1, F_2, \ldots, F_N$ by taking a method-of-moments approach: replace $E[f_r]$ by f_r in (28) for $1 \le r \le n$ and solve the resulting set of linear equations either exactly or approximately. As pointed out by Shlosser (1981), however, this system of equations is nearly singular, and hence extremely unstable.

4. VARIANCE AND VARIANCE ESTIMATES

Consider an estimator $\hat{D}$ that is a function of the sample only through $\mathbf{f} = (f_1, f_2, \ldots, f_M)$, where $M = \max(N_1, N_2, \ldots, N_D)$. All of the estimators introduced in Section 3 are of this type. In general, we also allow $\hat{D}$ to depend explicitly on the population size N and write $\hat{D} = \hat{D}(\mathbf{f}, N)$. Suppose that for any $N > 0$ and nonnegative M-dimensional vector $\mathbf{f} \neq \mathbf{0}$, the function $\hat{D}$ is continuously differentiable at the point $(\mathbf{f}, N)$ and

$$\hat{D}(c\mathbf{f}, cN) = c\hat{D}(\mathbf{f}, N) \tag{31}$$

for $c > 0$. Approximating the hypergeometric sample design by a Bernoulli sample design as in (8), we can obtain the following approximate expression for the asymptotic variance of $\hat{D}(\mathbf{f}, N)$ as D becomes large:

$$\mathrm{Avar}[\hat{D}(\mathbf{f}, N)] \approx \sum_{i=1}^{M} A_i^2 \mathrm{var}[f_i] + \sum_{\substack{1 \le i, i' \le M \\ i \neq i'}} A_i A_{i'} \mathrm{cov}[f_i, f_{i'}], \tag{32}$$

where A_i is the partial derivative of $\hat{D}$ with respect to f_i, evaluated at the point $(\mathbf{f}, N)$. (When computing each A_i, we replace each occurrence of n and d_n in the formula for $\hat{D}$ by $\sum_{i=1}^{M} i f_i$ and $\sum_{i=1}^{M} f_i$, respectively, before taking derivatives.) The approximation in (32) is valid when there is not too much variability in the class sizes (see Haas and Stokes 1996, app. C, for a precise formulation and proof of this result). It follows from the proof that, to a good approximation, the variance of an estimator $\hat{D}$ satisfying (31) increases linearly as D increases.

Straightforward calculations show that each of the specific estimators $\hat{D}_{\mathrm{uj1}}, \hat{D}_{\mathrm{uj2}}, \hat{D}_{\mathrm{Sh}}, \hat{D}_{\mathrm{Sh2}}$, and $\hat{D}_{\mathrm{Sh3}}$ is continuously differentiable as stated previously and also satisfies (31). Thus we can use (32) to study the asymptotic variance of these estimators. We focus on $\hat{D}_{\mathrm{uj1}}, \hat{D}_{\mathrm{uj2}}, \hat{D}_{\mathrm{Sh2}}$, and $\hat{D}_{\mathrm{Sh3}}$ because each of these estimators performs best for at least one population studied in the simulation experiments described in Section 6; we also consider $\hat{D}_{\mathrm{Sh}}$, because it is the most useful of the estimators previously proposed in the literature. Computation of the A_i coefficients for each estimator is tedious but straightforward (see Haas and Stokes 1996 for some explicit formulas).

Figures 1 and 2 compare the variances of the estimators $\hat{D}_{\mathrm{uj1}}, \hat{D}_{\mathrm{uj2}}, \hat{D}_{\mathrm{Sh}}, \hat{D}_{\mathrm{Sh2}}$, and $\hat{D}_{\mathrm{Sh3}}$ for a number of populations with equal class sizes. For these special populations, $\hat{D}_{\mathrm{uj1}}$ and $\hat{D}_{\mathrm{uj2}}$ are approximately unbiased, so that the relative variances of these estimators are appropriate measures of relative performance. It is particularly instructive to compare the variance of $\hat{D}_{\mathrm{uj1}}$ and $\hat{D}_{\mathrm{uj2}}$, because $\hat{D}_{\mathrm{uj2}}$ is obtained from $\hat{D}_{\mathrm{uj1}}$ by adjusting the latter estimator to compensate for bias induced by the assumption of equal class sizes. This adjustment is unnecessary for our special populations,

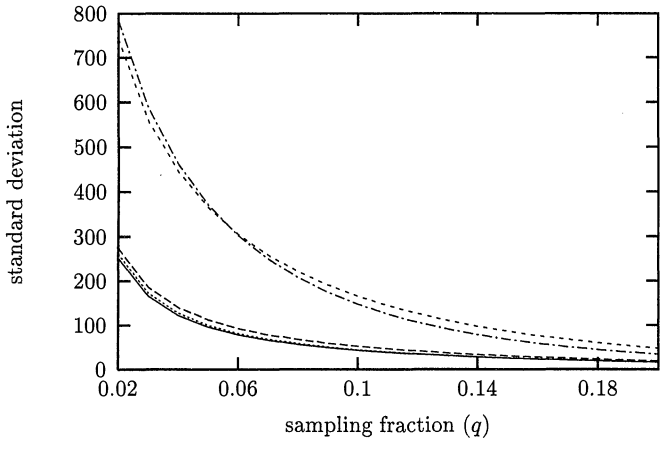

Figure 1. Standard Deviation of $\hat{D}_{uj1}$ (——), $\hat{D}_{uj2}$ (– – –), $\hat{D}_{Sh}$ (- - -), $\hat{D}_{Sh2}$ (· · ·), and $\hat{D}_{Sh3}$ (– · – · –) (D = 15,000 and $\bar{N}$ = 10).

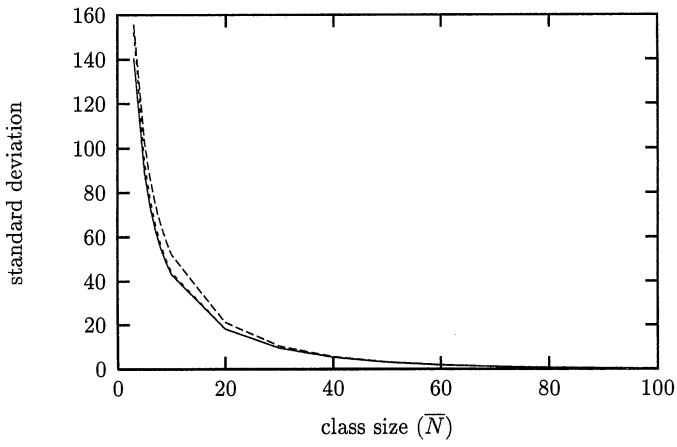

Figure 2. Standard Deviation of $\hat{D}_{uj1}$ (——), $\hat{D}_{uj2}$ (– – –), and $\hat{D}_{Sh2}$ (- - -) (D = 1,500 and q = .10).

and a comparison allows evaluation of the penalty (i.e., the increase in variance) being paid for the adjustment.

Figure 1 displays the standard deviations of $\hat{D}_{\mathrm{uj1}}, \hat{D}_{\mathrm{uj2}}$, $\hat{D}_{\mathrm{Sh}}, \hat{D}_{\mathrm{Sh2}}$, and $\hat{D}_{\mathrm{Sh3}}$ for an equal-class-size population with $N = 15{,}000$ and $D = 1{,}500$ (so that $\bar{N} = 10$) as the sampling fraction q varies. Observe that $\hat{D}_{\mathrm{uj2}}$ is only slightly less efficient than $\hat{D}_{\mathrm{uj1}}$, so that the penalty for bias adjustment is small in this case. Performance of the estimators $\hat{D}_{\mathrm{uj1}}$ and $\hat{D}_{\mathrm{Sh2}}$ is nearly indistinguishable. The most striking observation is that for this population, $\hat{D}_{\mathrm{Sh}}$ and $\hat{D}_{\mathrm{Sh3}}$ are not competitive with the other three estimators. The relative performance of $\hat{D}_{\mathrm{Sh}}$ and $\hat{D}_{\mathrm{Sh3}}$ is especially poor for small sampling fractions. On the other hand, the variance analysis indicates that modification of $\hat{D}_{\mathrm{Sh}}$ as in (27) and (29) indeed reduces the instability of the original Schlosser estimator in this case. Thus we focus on the estimators $\hat{D}_{\mathrm{uj1}}, \hat{D}_{\mathrm{uj2}}$, and $\hat{D}_{\mathrm{Sh2}}$ in the remainder of this section and in the next section. (We return to the estimator $\hat{D}_{\mathrm{Sh3}}$ in Sec. 6, where our simulation experiments indicate that $\hat{D}_{\mathrm{Sh3}}$ can exhibit smaller RMSE than the other estimators, but only at large sample sizes and for certain "ill-conditioned" populations in which γ^2 is extremely large.)

Figure 2 compares the three estimators $\hat{D}_{\mathrm{uj1}}, \hat{D}_{\mathrm{uj2}}$, and $\hat{D}_{\mathrm{Sh2}}$ for equal–class-size populations with a range of class sizes; for these calculations, the number of classes and the sampling fraction are held constant at $D = 1{,}500$ and $q = 10$. This figure illustrates the difficulty of precisely estimating D when the class size is small (but greater than 1). Again, we see that these three estimators perform similarly, with nearly equal variability when $\bar{N}$ exceeds about 40.

We checked the accuracy of the variance approximation in some example populations by comparing the values computed from (32) with results of a simulation experiment. (This experiment is discussed more completely in Sec. 6.) Simulated sampling with $q = .05, .10$, and .20 from the population examined in Figure 1 ($N = 15{,}000, D = 1{,}500$) yields variance estimates within 10% (on average) of those calculated from (32). Similar results were found in sampling from an equal-class-size population with $N = 15{,}000$ and $D = 150$.

In practice, we estimate the asymptotic variance of an estimator $\hat{D}$ by substituting estimates for $\{\mathrm{var}[f_i]\colon 1 \le i \le M\}$, and $\{\mathrm{cov}[f_i, f_{i'}]\colon 1 \le i \ne i' \le M\}$ into (32). To obtain such estimates, we approximate the true population by a population with D classes, each of size N/D. Under this approximation and the assumption in (8) of a Bernoulli sample design, the random vector $\mathbf{f}$ has a multinomial distribution with parameters D and $\mathbf{p} = (p_1, p_2, \ldots, p_n)$, where

$$p_i = \binom{N/D}{i} q^i (1-q)^{(N/D)-i}$$

for $1 \le i \le n$. It follows that $\mathrm{var}[f_i] = Dp_i(1-p_i)$ and $\mathrm{cov}[f_i, f_{i'}] = -Dp_ip_{i'}$. We can estimate each p_i by $f_i/\hat{D}$, thereby obtaining the final estimates

$$\widehat{\mathrm{var}}[f_i] = f_i\left(1 - \frac{f_i}{\hat{D}}\right)$$

and

$$\widehat{\mathrm{cov}}[f_i, f_{i'}] = -\frac{f_i f_{i'}}{\hat{D}}$$

for $1 \le i, i' \le n$. These formulas coincide with the estimators obtained using the "unconditional approach" of Chao and Lee (1992). A computer program that calculates $\hat{D}_{\mathrm{uj1}}, \hat{D}_{\mathrm{uj2}}, \hat{D}_{\mathrm{Sh2}}$, and their estimated standard errors from sample data can be obtained from the second author.

5. AN EXAMPLE

The following example illustrates how knowledge of the population size N can affect estimates of the number of classes. When the population size N is unknown, Chao and Lee (1992, sec. 3) have proposed that the estimator $\hat{D}_{\mathrm{CL}}$ defined in Section 3.3 be used to estimate the number of classes, because the formula for $\hat{D}_{\mathrm{CL}}$ does not involve the unknown parameter N. When N is known, a slight modification of the derivation of $\hat{D}_{\mathrm{CL}}$ leads to the unsmoothed second-order jackknife estimator $\hat{D}_{\mathrm{uj2}}$ (see Haas and Stokes 1996, app. B).

Our example is based on one discussed by Chao and Lee (1992), who borrowed data first described and analyzed by Holst (1981). These data arose from an application in numismatics in which 204 ancient coins were classified according to die type to estimate the number of different dies used in the minting process. Among the die types on the reverse sides of the 204 coins were 156 singletons, 19 pairs, 2 triplets, and 1 quadruplet ($f_1 = 156, f_2 = 19, f_3 = 2, f_4 = 1, d = 178$). Because the total number of coins minted is unknown in this case, model (1) is inappropriate for analyzing these data. But suppose that the same data had arisen from an application in which N was known. For example, suppose that the data were obtained by selecting a simple random sample of 204 names from a sampling frame that had been constructed by combining 5 lists of 200 names each ($N = 1{,}000$), 50 lists of 200 names each ($N = 10{,}000$), or 500 lists of 200 names each ($N = 100{,}000$). In each case our object is to estimate the number of unique individuals on the combined list, based on the sample results. We focus on the three estimators $\hat{D}_{\mathrm{uj1}}, \hat{D}_{\mathrm{uj2}}$, and $\hat{D}_{\mathrm{Sh2}}$. The estimates

Table 1. Values of $\hat{D}_{uj1}$, $\hat{D}_{uj2}$, and $\hat{D}_{Sh2}$ for Three Hypothetical Combined Lists

N	$\hat{D}_{uj1}$	$\hat{D}4_{uj2}$	$\hat{D}_{Sh2}$
1,000	455	502	455
	(47)	(60)	(51)
10,000	709	788	707
	(125)	(161)	(128)
100,000	752	835	749
	(141)	(183)	(144)

NOTE: SE in parentheses.

for the three cases are given in Table 1; the standard errors displayed in Table 1 are estimated using the procedure outlined in Section 4.

We would expect similar inferences to be made from the same data under the multinomial model and the finite population model when N is very large. Indeed, the value $\hat{D}_{\mathrm{uj2}} = 835$ agrees closely with Chao and Lee's estimate $\hat{D}_{\mathrm{CL}} = 844$ (SE 187) when $N = 100{,}000$. Moreover, when $N = 100{,}000$, we find that $\hat{\gamma}^2(\hat{D}_{\mathrm{uj1}}) \approx .13$, which is the same estimate of γ^2 given by Chao and Lee. As the population size decreases, however, both our assessment of the magnitude of D and our uncertainty about that magnitude decrease, because we are observing a larger and larger fraction of both the population and the classes.

The most extreme divergence between the estimate obtained using $\hat{D}_{\mathrm{CL}}$ and estimates obtained using $\hat{D}_{\mathrm{uj1}}, \hat{D}_{\mathrm{uj2}}$, or $\hat{D}_{\mathrm{Sh2}}$ occurs when the sample consists of all singletons ($f_1 = n$). In that case, $\hat{D}_{\mathrm{CL}} = \infty$, whereas $\hat{D}_{\mathrm{uj1}} = \hat{D}_{\mathrm{uj2}} = \hat{D}_{\mathrm{Sh2}} = N$. This result indicates that when the population size N is known, it is better to use an estimator that exploits knowledge of N than to sample with replacement and use the estimator $\hat{D}_{\mathrm{CL}}$. In some applications, sampling with replacement is not even an option. For example, the only available sampling mechanism in at least one current database system is a one-pass reservoir algorithm (as in Vitter 1985).

The empirical results in Section 6 indicate that of the three estimators displayed in Table 1, $\hat{D}_{\mathrm{uj2}}$ is the superior estimator when γ^2 is small (< 1). Thus for our example, $\hat{D}_{\mathrm{uj2}}$ would be the preferred estimator, because $\hat{\gamma}^2(\hat{D}_{\mathrm{uj1}}) \approx .13$ in all three cases. Note that $\hat{D}_{\mathrm{uj2}}$ consistently has the highest variance of the three estimators in Table 1. The bias of $\hat{D}_{\mathrm{uj2}}$ is typically lower than that of $\hat{D}_{\mathrm{uj1}}$ or $\hat{D}_{\mathrm{Sh2}}$ when γ^2 is small, however, so the overall RMSE is lower.

6. SIMULATION RESULTS

This section describes the results of a simulation study done to compare the performance of the various estimators described in Section 3. Our comparison is based on the performance of the estimators for sampling fractions of 5%, 10%, and 20% in 52 populations. (Initial experiments indicated that the performance of the various estimators is best viewed as a function of sampling fraction, rather than absolute sample size. This is in contrast to estimators of, for example, population averages.)

We consider several sets of populations. The first set comprises synthetic populations of the type considered in the literature. Populations EQ10 and EQ100 have equal class sizes of 10 and 100. In populations NGB/1, NGB/2, and NGB/4, the class sizes follow a negative binomial distribution. Specifically, the fraction $f(m)$ of classes in population NGB/k with class size equal to m is given by

$$f(m) \approx \binom{m-1}{k-1} r^k (1-r)^{m-k}$$

for $m \geq k$, where $r = .04$. Chao and Lee (1992) considered populations of this type. The populations in the second set are meant to be representative of data that could be encountered when a sampling frame for a population census is constructed by combining a number of lists which may contain overlapping entries. Population GOOD and SUDM were studied by Goodman (1949) and Sudman (1976). Population FRAME2 mimics a sampling frame that might arise in an administrative records census of the type described in Section 1. One approach to such a census is to augment the usual census address list with a small number of relatively large administrative records files, such as AFDC or food stamps, and then estimate the number of distinct individuals on the combined list from a sample. We have constructed FRAME2 so that a given individual can appear at most five times, but most individuals appear exactly once, mimicking the case in which four administrative lists are used to supplement the census address list. Population FRAME3 is similar to FRAME2, but for the FRAME3 population it is assumed that the combined list is made up of a number of small lists (perhaps obtained from neighborhood-level organizations) rather than a few large lists. The populations in the third set, denoted by Z20A, Z20B, and Z15, are used to study the behavior of the estimators when the data are extremely ill-conditioned. The class sizes in each of these populations follow a generalized Zipf distribution (see Knuth 1973, p. 398). Specifically, $N_j/N \propto j^{-\theta}$, where θ equals 1.5 or 2.0. These populations have extremely high values of γ^2. Descriptive statistics for these three sets of populations are given in Tables 2, 3, and 4. The column entitled "skew" displays the dimensionless coefficient of skewness λ, which is defined by

$$\lambda = \frac{\sum_{j=1}^{D}(N_j - \bar{N})^3/D}{(\sum_{j=1}^{D}(N_j - \bar{N})^2/D)^{3/2}}.$$

The final set comprises 40 real populations that demonstrate the type of distributions encountered when estimating the number of distinct values of an attribute in a relational database. Specifically, the populations studied correspond to various relational attributes from a database of enrollment records for students at the University of Wisconsin and a

Table 2. Characteristics of Synthetic Populations

Name	N	D	γ^2	Skew
EQ10	15,000	1,500	0	0
EQ100	15,000	150	0	0
NGB/4	82,135	874	.18	.50
NGB/2	41,197	906	.37	.81
NGB/1	20,213	930	.75	1.25

Table 3. Characteristics of "Merged List" Populations

Name	N	D	γ^2	Skew
GOOD	10,000	9,595	.04	5.64
FRAME2	33,750	19,000	.31	1.18
FRAME3	111,500	36,000	.52	1.92
SUDM	330,000	100,000	1.87	2.71

database of billing records from a large insurance company. The population size N ranges from 15,469 to 1,654,700, with D ranging from 3 to 1,547,606 and γ^2 ranging from 0 to 81.63 (see Haas and Stokes 1996, app. D, for further details). It is notable that values of γ^2 encountered in the literature (Chao and Lee 1992; Goodman 1949; Shlosser 1981; Sudman 1976) tend not to exceed the value 2, and are typically less than 1, whereas the value of γ^2 exceeds 2 for more than 50% of the real populations.

For each estimator, population, and sampling fraction, we estimated the bias and RMSE by repeatedly drawing a simple random sample from the population, evaluating the estimator, and then computing the error of each estimate. (When evaluating the estimator, we truncated each estimate below at d and above at N.) The final estimate of bias was obtained by averaging the error over all of the experimental replications, and RMSE was estimated as the square root of the averaged square error. We used 100 replications, which was sufficient to estimate the RMSE with a standard error below 5% in nearly all cases; typically the standard error was much less.

Summary results from the simulations are displayed in Tables 5 and 6. Table 5 gives the average and maximum RMSEs for each estimator of D over all populations with $0 \le \gamma^2 < 1$, with $1 \le \gamma^2 < 50$, and with $\gamma^2 \ge 50$, as well as the average and maximum RMSEs for each estimator over all populations combined. Similarly, Table 6 gives the average and maximum bias for each estimator. In these tables, the RMSE and bias are each expressed as a percentage of the true number of classes. Tables 5 and 6 also display the RMSE and bias of the estimator $\hat{\gamma}^2(\hat{D}_{\mathrm{uj1}})$ used in the second-order jackknife estimators; the RMSE and bias are expressed as a percentage of the true value γ^2 and are displayed in the column labelled $\hat{\gamma}^2$.

Comparing Tables 5 and 6 indicates that for each estimator the major component of the RMSE is almost always bias, not variance. Thus, even though the standard error can be estimated as in Section 4, this estimated standard error usually does not give an accurate picture of the error in estimation of D. Another consequence of the predominance of bias is that when γ^2 is large, the RMSE for the second-order estimator $\hat{D}_{\mathrm{uj2}}$ does not decrease monotonically as the sampling fraction increases. (In all other cases the RMSE decreases monotonically.)

Comparing $\hat{D}_{\mathrm{uj1}}$ with $\hat{D}_{\mathrm{sj1}}$ and then comparing $\hat{D}_{\mathrm{uj2}}$ with $\hat{D}_{\mathrm{sj2}}$, we see that smoothing a first-order jackknife estimator never results in a better first-order estimator. On the other hand, smoothing a second-order jackknife estimator can result in significant performance improvement when γ^2 is large.

Similarly, using higher-order Taylor expansions leads to mixed results. Second-order estimators perform better than first-order estimators when γ^2 is relatively small, but not when γ^2 is large. The difficulty is partially that the estimator $\hat{\gamma}^2(\hat{D}_{\mathrm{uj1}})$ tends to underestimate γ^2 when γ^2 is large, leading to underestimating the number of classes. Moreover, the Taylor approximations underlying $\hat{D}_{\mathrm{uj1}}, \hat{D}_{\mathrm{sj1}}, \hat{D}_{\mathrm{uj2}}$, and $\hat{D}_{\mathrm{sj2}}$ are derived under the assumption of not too much variability between class sizes; this assumption is violated when γ^2 is large. There apparently is no systematic relation between the coefficient of skewness for the class sizes and the performance of second-order jackknife estimators.

As predicted in Sections 3.4 and 4, the estimators $\hat{D}_{\mathrm{Sh}}$ and $\hat{D}_{\mathrm{Sh3}}$ behave poorly when γ^2 is relatively small, and $\hat{D}_{\mathrm{Sh3}}$ performs better than $\hat{D}_{\mathrm{Sh}}$ when γ^2 is very large. For small to medium values of γ^2, the modified estimator $\hat{D}_{\mathrm{Sh2}}$ has a smaller RMSE than $\hat{D}_{\mathrm{Sh}}$ or $\hat{D}_{\mathrm{Sh3}}$ and its performance is comparable to the generalized jackknife estimators. For extremely large values of γ^2 and also for large sample sizes, the estimator $\hat{D}_{\mathrm{Sh3}}$ has the best performance of the three Shlosser-type estimators. (For a 20% sampling fraction, $\hat{D}_{\mathrm{Sh3}}$ in fact has the lowest average RMSE of all the estimators considered.)

As indicated earlier, smoothing can improve the performance of the second-order jackknife estimator $\hat{D}_{\mathrm{uj2}}$. An alternative ad hoc technique for improving performance is to "stabilize" $\hat{D}_{\mathrm{uj2}}$ using a method suggested by Chao, Ma, and Yang (1993). Fix $c \ge 1$ and remove any class whose frequency in the sample exceeds c; that is, remove from the sample all members of classes $\{C_j\colon j \in B\}$, where $B = \{1 \le j \le D\colon n_j > c\}$. Then compute the estimator $\hat{D}_{\mathrm{uj2}}$ from the reduced sample and subsequently increment it by $|B|$ to produce the final estimate, denoted by $\hat{D}_{\mathrm{uj2a}}$. (Here $|B|$ denotes the number of elements in the set B.) When computing $\hat{D}_{\mathrm{uj2}}$ from the reduced sample, take the population size as $N - \sum_{j\in B} \hat{N}_j$, where each $\hat{N}_j$ is a method-of-moments estimator of N_j as in Section 3.2.3. If $n - \sum_{j\in B} n_j = 0$, then simply compute $\hat{D}_{\mathrm{uj2}}$ from the full sample. The idea behind this procedure is as follows. When γ^2 is large, the population consists of a few large classes and many smaller classes. By in effect removing the largest classes from the population, we obtain a reduced population for which γ^2 is smaller, so that D is easier to estimate; the contribution to D from the $|B|$ removed classes is then added back at the final step of the estimation process. (We also experimented with another stabilization technique in which the k most frequent classes are removed for some fixed k, but this technique is not as effective as the method we describe herein.) Preliminary experiments indicated that c approximately 50 yields the best performance. For larger values of c, not enough of the frequent classes are removed;

Table 4. Characteristics of "Ill-Conditioned" Populations

Name	N	D	γ^2	Skew
Z20A	50,000	247	114.38	14.60
Z15	50,000	772	166.18	23.44
Z20B	50,000	10,384	234.81	73.54

Table 5. Average RMSE for Various Estimators

Sample size	γ^2	$\hat{D}_{uj1}$	$\hat{D}_{sj1}$	$\hat{D}_{uj2}$	$\hat{D}_{sj2}$	$\hat{D}_{Sh}$	$\hat{D}_{Sh2}$	$\hat{D}_{Sh3}$	$\hat{\gamma}^2$
5%	≥0 and <1	13.48	14.20	11.84	12.27	79.17	13.23	202.16	56.65
		(43.81)	(45.14)	(39.56)	(39.67)	(428.25)	(46.59)	(3,299.10)	(96.72)
	≥1 and <50	38.14	39.17	65.34	45.25	54.30	36.67	93.92	46.51
		(70.47)	(70.48)	(186.15)	(186.15)	(218.02)	(66.82)	(1,042.73)	(91.70)
	≥50	74.11	75.92	388.77	77.78	28.13	71.23	21.45	74.72
		(85.09)	(88.49)	(564.57)	(112.13)	(47.63)	(83.71)	(38.58)	(85.55)
	All	30.95	31.91	68.61	34.44	62.33	29.86	132.06	52.78
		(85.09)	(88.49)	(564.57)	(186.15)	(428.25)	(83.71)	(3,299.10)	(96.72)
10%	≥0 and <1	11.30	12.14	9.05	9.71	33.09	11.19	22.68	49.68
		(39.80)	(42.32)	(31.73)	(31.90)	(200.79)	(44.83)	(131.15)	(90.68)
	≥1 and <50	31.41	32.59	90.96	38.74	34.96	29.16	50.17	38.34
		(61.27)	(61.28)	(267.08)	(186.15)	(107.16)	(54.03)	(357.43)	(83.12)
	≥50	63.92	65.88	682.55	115.77	15.50	58.82	11.51	64.43
		(76.47)	(81.21)	(1,133.61)	(281.98)	(28.97)	(73.14)	(21.81)	(76.89)
	All	25.79	26.89	103.38	32.94	32.71	24.18	36.10	44.93
		(76.47)	(81.21)	(1,133.61)	(281.98)	(200.79)	(73.14)	(357.43)	(90.68)
20%	≥0 and <1	8.89	9.86	5.77	6.53	12.91	8.30	9.05	40.65
		(33.01)	(37.28)	(29.82)	(27.49)	(79.16)	(30.14)	(79.16)	(81.03)
	≥1 and <50	23.44	24.81	123.00	32.79	18.14	20.88	17.91	28.65
		(46.77)	(49.73)	(369.77)	(186.15)	(49.20)	(43.38)	(74.99)	(67.42)
	≥50	50.10	52.19	1,093.07	130.30	7.73	42.58	6.32	50.51
		(62.96)	(69.06)	(2,010.61)	(381.51)	(15.12)	(56.72)	(10.62)	(63.37)
	All	19.62	20.88	150.28	29.69	15.23	17.47	13.44	35.18
		(62.96)	(69.06)	(2,010.61)	(381.51)	(79.16)	(56.72)	(79.16)	(81.03)

NOTE: Maximum RMSE values within parentheses. Both average and maximum values supplied in percentages.

for smaller c, the size of the reduced sample is too small, and the resulting inaccuracy of $\hat{D}_{\mathrm{uj2}}$ when computed from this sample offsets the benefits of the reduction in γ^2. We therefore take $c = 50$ in our experiments. As can be seen from Table 7, the RMSE for $\hat{D}_{\mathrm{uj2a}}$ is indeed much lower than that for $\hat{D}_{\mathrm{uj2}}$ when γ^2 exceeds 1. Moreover, by comparing the RMSEs of $\hat{D}_{\mathrm{sj2}}$ and $\hat{D}_{\mathrm{uj2a}}$ in Tables 5 and 7, it can be seen that stabilization is more effective than smoothing. Observe, however, that the performance of $\hat{D}_{\mathrm{uj2a}}$ is worse than that of $\hat{D}_{\mathrm{uj2}}$ when γ^2 is small. Interestingly, experiments indicate that none of the other estimators that we consider appears to benefit from stabilization, and we apply this technique only to $\hat{D}_{\mathrm{uj2}}$. Overall, the most effective estimators appear to be $\hat{D}_{\mathrm{uj2a}}$, which has the smallest average RMSE over the various populations, and $\hat{D}_{\mathrm{Sh2}}$, which has the smallest worst-case RMSE.

Table 6. Average Bias for Various Estimators

Sample size	γ^2	$\hat{D}_{uj1}$	$\hat{D}_{sj1}$	$\hat{D}_{uj2}$	$\hat{D}_{sj2}$	$\hat{D}_{Sh}$	$\hat{D}_{Sh2}$	$\hat{D}_{Sh3}$	$\hat{\gamma}^2$
5%	≥0 and <1	−12.71	−13.43	−10.76	−11.38	71.11	−10.98	90.57	−55.75
		(−43.77)	(−45.10)	(−39.51)	(−39.62)	(427.53)	(−46.59)	(958.74)	(−94.97)
	≥1 and <50	−37.95	−38.99	42.77	−16.83	39.13	−36.35	61.98	−46.19
		(−70.32)	(−70.32)	(186.15)	(186.15)	(218.01)	(−66.49)	(663.26)	(−91.70)
	≥50	−74.10	−75.91	382.88	−22.16	22.92	−71.22	3.17	−74.71
		(−85.09)	(−88.49)	(556.68)	(110.28)	(44.65)	(−83.71)	(33.44)	(−85.54)
	All	−30.54	−31.51	47.31	−15.04	50.80	−28.79	69.00	−52.24
		(−85.09)	(−88.49)	(556.68)	(186.15)	(427.53)	(−83.71)	(958.74)	(−94.97)
10%	≥0 and <1	−10.93	−11.78	−8.38	−9.31	28.12	−9.47	17.66	−48.87
		(−39.79)	(−42.31)	(−31.47)	(−31.88)	(200.49)	(−44.83)	(130.80)	(−90.59)
	≥1 and <50	−31.16	−32.34	74.62	−10.44	25.41	−28.61	35.20	−37.98
		(−61.00)	(−61.00)	(261.47)	(186.15)	(107.16)	(−53.88)	(264.38)	(−83.12)
	≥50	−63.91	−65.87	677.18	24.90	11.57	−58.78	3.10	−64.41
		(−76.47)	(−81.21)	(1,125.89)	(280.63)	(27.09)	(−73.13)	(18.47)	(−76.88)
	All	−25.51	−26.62	87.45	−7.26	25.44	−23.20	25.65	−44.41
		(−76.47)	(−81.21)	(1,125.89)	(280.63)	(200.49)	(−73.13)	(264.38)	(−90.59)
20%	≥0 and <1	−8.57	−9.55	−4.99	−6.20	9.99	−6.75	5.73	−39.71
		(−33.01)	(−37.27)	(−17.83)	(−22.67)	(45.86)	(−28.38)	(28.17)	(−81.00)
	≥1 and <50	−23.12	−24.49	112.39	−3.41	12.09	−20.13	10.02	−28.23
		(−46.54)	(−49.73)	(362.12)	(186.15)	(49.20)	(−43.38)	(49.34)	(−67.36)
	≥50	−50.09	−52.17	1,087.89	60.47	5.03	−42.53	1.72	−50.49
		(−62.96)	(−69.06)	(2,003.12)	(381.51)	(13.90)	(−56.71)	(8.23)	(−63.36)
	All	−19.32	−20.59	140.02	0.38	10.70	−16.45	7.65	−34.58
		(−62.96)	(−69.06)	(2,003.12)	(381.51)	(49.20)	(−56.71)	(49.34)	(−81.00)

NOTE: Maximum bias values within parentheses. Both average and maximum values supplied in percentages.

Table 7. Average RMSE of $\hat{D}_{uj2}$, $\hat{D}_{uj2a}$, $\hat{D}_{Sh2}$, $\hat{D}_{Sh3}$, and $\hat{D}_{hybrid}$

Sample size	γ^2	$\hat{D}_{uj2}$	$\hat{D}_{uj2a}$	$\hat{D}_{Sh2}$	$\hat{D}_{Sh3}$	$\hat{D}_{hybrid}$
5%	≥0 and <1	11.84	19.46	13.23	202.16	11.84
		(39.56)	(192.64)	(46.59)	(3,299.10)	(39.56)
	≥1 and <50	65.34	27.47	36.67	93.92	27.47
		(186.15)	(54.51)	(66.82)	(1,042.73)	(54.51)
	≥50	388.77	23.00	71.23	21.45	26.17
		(564.57)	(36.60)	(83.71)	(38.58)	(39.20)
	All	68.61	23.89	29.86	132.06	21.06
		(564.57)	(192.64)	(83.71)	(3,299.10)	(54.51)
10%	≥0 and <1	9.05	13.26	11.19	22.68	9.05
		(31.73)	(120.14)	(44.83)	(131.15)	(31.73)
	≥1 and <50	90.96	19.22	29.16	50.17	19.55
		(267.08)	(48.12)	(54.03)	(357.43)	(48.12)
	≥50	682.55	17.82	58.82	11.51	11.51
		(1,133.61)	(27.30)	(73.14)	(21.81)	(21.81)
	All	103.38	16.71	24.18	36.10	14.69
		(1,133.61)	(120.14)	(73.14)	(357.43)	(48.12)
20%	≥0 and <1	5.77	8.12	8.30	9.05	5.77
		(29.82)	(79.16)	(30.14)	(79.16)	(29.82)
	≥1 and <50	123.00	17.44	20.88	17.91	17.69
		(369.77)	(76.57)	(43.38)	(74.99)	(76.57)
	≥50	1,093.07	37.30	42.58	6.32	6.32
		(2,010.61)	(83.69)	(56.72)	(10.62)	(10.62)
	All	150.28	15.20	17.47	13.44	12.00
		(2,010.61)	(83.69)	(56.72)	(79.16)	(76.57)

NOTE: Maximum RMSE values within parentheses. Both average and maximum values supplied in percentages.

Our next observation is based on a comparison of the bias and RMSE of $\hat{D}_{\mathrm{uj1}}$ and $\hat{D}_{\mathrm{Sh2}}$ for all of the populations studied. The behavior of the two estimators is quite similar: the correlation between the bias of the estimators is .990, and the correlation between the RMSE is .993. The RMSE and bias of $\hat{D}_{\mathrm{uj1}}$ are usually slightly greater than the RMSE and bias of $\hat{D}_{\mathrm{Sh2}}$. On the other hand, using $\hat{D}_{\mathrm{Sh2}}$ requires computation of $f_1, f_2, \ldots, f_n$, whereas using $\hat{D}_{\mathrm{uj1}}$ requires computation only of f_1. Thus if computational resources are limited, then it may be desirable to use $\hat{D}_{\mathrm{uj1}}$ as a surrogate for $\hat{D}_{\mathrm{Sh2}}$; the quantity f_1 can be computed efficiently using "Bloom filter" techniques as described by Ramakrishna (1989).

The experimental results show that the relative performance of the estimators is strongly influenced by the value of γ^2. As can be seen from Table 7, the estimator $\hat{D}_{\mathrm{uj2}}$ has the smallest average RMSE when $0 \le \gamma^2 < 1$, the estimator $\hat{D}_{\mathrm{uj2a}}$ has the smallest average RMSE when $1 \le \gamma^2 < 50$, and the estimator $\hat{D}_{\mathrm{Sh3}}$ has the smallest average RMSE when $\gamma^2 \ge 50$. These results indicate that it may be desirable to allow an estimator to depend explicitly on the (estimated) value of γ^2. To illustrate this idea, we consider a simple ad hoc branching estimator, denoted by $\hat{D}_{\mathrm{hybrid}}$. The idea is to estimate γ^2 by $\hat{\gamma}^2(\hat{D}_{\mathrm{uj1}})$, fix parameters $0 < \alpha_1 < \alpha_2$, and set

$$\hat{D}_{\mathrm{hybrid}} = \begin{cases} \hat{D}_{\mathrm{uj2}} & \text{if } 0 \le \hat{\gamma}^2(\hat{D}_{\mathrm{uj1}}) < \alpha_1 \\ \hat{D}_{\mathrm{uj2a}} & \text{if } \alpha_1 \le \hat{\gamma}^2(\hat{D}_{\mathrm{uj1}}) < \alpha_2 \\ \hat{D}_{\mathrm{Sh3}} & \text{if } \hat{\gamma}^2(\hat{D}_{\mathrm{uj1}}) \ge \alpha_2. \end{cases} \quad (33)$$

Table 7 displays the estimated RMSE for $\hat{D}_{\mathrm{hybrid}}$ when $\alpha_1 = .9$ and $\alpha_2 = 30$. As can be seen, the RMSE for the combined estimator $\hat{D}_{\mathrm{hybrid}}$ almost never exceeds that for $\hat{D}_{\mathrm{uj2}}, \hat{D}_{\mathrm{uj2a}}$, or $\hat{D}_{\mathrm{Sh3}}$ separately.

7. CONCLUSIONS

Both new and previous nonparametric estimators of the number of classes in a finite population can be viewed as generalized jackknife estimators. This viewpoint has suggested ways to improve Shlosser's original estimator and has shed new light on certain Horvitz–Thompson estimators as well as estimators based on notions of "sample coverage." We have used delta-method arguments to develop estimators of the standard error of generalized jackknife estimators. As indicated by the example in Section 5, knowledge of the population size can lead to more precise estimation of the number of classes.

Of the estimators considered, the best appears to be the branching estimator $\hat{D}_{\mathrm{hybrid}}$ defined by (33), in which a modified Shlosser estimator is used when the coefficient of variation of the class sizes is estimated to be extremely large and unsmoothed second-order jackknife estimators are used otherwise. The systematic development of such branching estimators is a topic for future research. If a nonbranching estimator is desired, then we recommend the stabilized unsmoothed second-order jackknife estimator $\hat{D}_{\mathrm{uj2a}}$, followed by the modified Shlosser estimator $\hat{D}_{\mathrm{Sh2}}$. If computing resources are scarce, then $\hat{D}_{\mathrm{uj1}}$ is a reasonable estimator.

The various estimators of D discussed in this article embody different approaches for dealing with the difficulties caused by variation in the class sizes $N_1, N_2, \ldots, N_D$. Such variation is reflected by large values of γ^2. First-order estimates simply approximate each N_j by $\bar{N}$. It is well known

in the literature that such an approach tends to yield downwardly biased estimates (cf. Bunge and Fitzpatrick 1993). More sophisticated approaches considered here include

- Taylor corrections to the first-order approximation, as in the estimators $\hat{D}_{\mathrm{uj2}}$ and $\hat{D}_{\mathrm{sj2}}$
- the stabilization technique of Section 6, in which the population is in effect modified so that the variation in class sizes is reduced
- the Horvitz–Thompson approach, in which the first-order assumption is avoided by estimating explicitly each N_j such that $n_j > 0$
- Shlosser's approach, which replaces the first-order assumption with the assumption in (25) and in its purest form results in the estimators $\hat{D}_{\mathrm{Sh}}$ and $\hat{D}_{\mathrm{Sh3}}$.

The poor performance of the Horvitz–Thompson estimators indicates that approaches based on direct estimation of the N_j's are unlikely to be successful. The second-order Taylor correction is effective mainly for small values of γ^2, and both the stabilization technique and Shlosser's approach are effective mainly for large values of γ^2. Thus, until a better solution is found, the best estimators will result from a judicious combination of the various approaches considered here.

[Received May 1996. Revised March 1998.]

REFERENCES

Astrahan, M., Schkolnick, M., and Whang, K. (1987), "Approximating the Number of Unique Values of an Attribute Without Sorting," *Information Systems*, 12, 11–15.

Bunge, J., and Fitzpatrick, M. (1993), "Estimating the Number of Species: A Review," *Journal of the American Statistical Association*, 88, 364–373.

Burnham, K. P., and Overton, W. S. (1978), "Estimation of the Size of a Closed Population When Capture Probabilities Vary Among Animals," *Biometrika*, 65, 625–633.

——— (1979), "Robust Estimation of Population Size When Capture Probabilities Vary Among Animals," *Ecology*, 60, 927–936.

Chao, A., and Lee, S. (1992), "Estimating the Number of Classes via Sample Coverage," *Journal of the American Statistical Association*, 87, 210–217.

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.

Deming, W. E., and Glasser, G. J. (1959), "On the Problem of Matching Lists by Samples," *Journal of the American Statistical Association*, 54, 403–415.

Flajolet, P., and Martin, G. N. (1985), "Probabilistic Counting Algorithms for Data Base Applications," *Journal of Computer and System Sciences*, 31, 182–209.

Gelenbe, E., and Gardy, D. (1982), "On the Sizes of Projections: I," *Information Processing Letters*, 14, 18–21.

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

Goodman, L. A. (1949), "On the Estimation of the Number of Classes in a Population," *Annals of Mathematical Statistics*, 20, 572–579.

——— (1952), "On the Analysis of Samples From k Lists," *Annals of Mathematical Statistics*, 23, 632–634.

Gray, H. L., and Schucany, W. R. (1972), *The Generalized Jackknife Statistic*, New York: Marcel Dekker.

Haas, P. J., and Stokes, L. (1996), "Estimating the Number of Classes in a Finite Population," IBM Research Report RJ 10025, IBM Almaden Research Center, San Jose, CA, Revised March 1998.

Hellerstein, J. M., and Stonebraker, M. (1994), "Predicate Migration: Optimizing Queries With Expensive Predicates," in *Proceedings of the 1994 ACM SIGMOD International Conference on Management of Data*, pp. 267–276.

Heltshe, J. F., and Forrester, N. E. (1983), "Estimating Species Richness Using the Jackknife Procedure," *Biometrics*, 39, 1–11.

Holst, L. (1981), "Some Asymptotic Results for Incomplete Multinomial or Poisson Samples," *Scandinavian Journal of Statistics*, 8, 243–246.

Hou, W., Ozsoyoglu, G., and Taneja, B. (1988), "Statistical Estimators for Relational Algebra Expressions," in *Proceedings of the Seventh ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pp. 276–287.

——— (1989), "Processing Aggregate Relational Queries With Hard Time Constraints," in *Proceedings of the 1989 ACM SIGMOD International Conference on Management of Data*, pp. 68–77.

Kish, L. (1965), *Survey Sampling*, New York: Wiley.

Knuth, D. E. (1973), *The Art of Computer Programming, Vol. 3: Sorting and Searching*, Reading, MA: Addison-Wesley.

Korth, H. F., and Silberschatz, A. (1991), *Database System Concepts* (2nd ed.), New York: McGraw-Hill.

Miller, R. G. (1974), "The Jackknife—A Review," *Biometrika*, 61, 1–17.

Mosteller, F. (1949), "Questions and Answers," *American Statistician*, 3, 12–13.

Naughton, J. F., and Seshadri, S. (1990), "On Estimating the Size of Projections," in *Proceedings of the Third International Conference on Database Theory*, pp. 499–513.

Ozsoyoglu, G., Du, K., Tjahjana, A., Hou, W., and Rowland, D. Y. (1991), "On Estimating COUNT, SUM, and AVERAGE Relational Algebra Queries," in *Database and Expert Systems Applications, Proceedings of the International Conference in Berlin, Germany, 1991 (DEXA 91)*, pp. 406–412.

Ramakrishna, M. V. (1989), "Practical Performance of Bloom Filters and Parallel Free-Text Searching," *Communications of the ACM*, 32, 1237–1239.

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

Selinger, P. G., Astrahan, D. D., Chamberlain, R. A., Lorie, R. A., and Price, T. G. (1979), "Access Path Selection in a Relational Database Management System," in *Proceedings of the 1979 ACM SIGMOD International Conference on Management of Data*, pp. 23–34.

Shlosser, A. (1981), "On Estimation of the Size of the Dictionary of a Long Text on the Basis of a Sample," *Engineering Cybernetics*, 19, 97–102.

Smith, E. P., and van Belle, G. (1984), "Nonparametric Estimation of Species Richness," *Biometrics*, 40, 119–129.

Sudman, S. (1976), *Applied Sampling*, New York: Academic Press.

Vitter, J. S. (1985), "Random Sampling With a Reservoir," *ACM Transactions on Mathematical Software*, 27, 703–718.

Whang, K., Vander-Zanden, B. T., and Taylor, H. M. (1990), "A Linear-Time Probabilistic Counting Algorithm for Database Applications," *ACM Transactions on Database Systems*, 15, 208–229.