Journal of Official Statistics, Vol. 21, No. 3, 2005, pp. 489–504

Estimating the Number of Distinct Valid Signatures in Initiative Petitions

Ruben A. Smith[1] *and David R. Thomas*[2]

In some states in the U.S.A., if citizens are dissatisfied with certain laws or feel that new laws are needed, they can petition to place proposed legislation on the ballot. For the petition to be certified for the ballot, its sponsor must circulate the complete text of the proposal among voters and obtain signatures of those in favor. Petitions will contain both invalid and valid signatures. Valid signatures from registered voters can appear more than once. To qualify a petition as a ballot measure, the total number of distinct valid signatures collected must exceed a required number. We are considering the case when a simple random sample of signatures is drawn from the entire petition, and all signatures in the sample are verified. The problem is to estimate the total number of distinct valid signatures based on the sample information and the knowledge of the total number of signatures collected in the petition. We consider several linear estimators and one nonlinear estimator. Expressions for the variance of the linear estimators are provided. The performance of the estimators is evaluated using data from several Washington State petitions that have been completely verified.

Key words: Estimation; sampling signatures; replicated signatures.

1. Introduction

Some state constitutions in the U.S.A. give initiative and referendum power to the people. If citizens from these states are dissatisfied with certain laws or feel that new laws are needed, they can petition to propose legislation, either to the legislature or to the ballot. The sponsor of the petition must circulate the complete text of the proposed legislation among voters and collect signatures of those in favor.

After signatures are collected they are filed as a petition with the state office in charge, usually the Secretary of State. The office in charge determines, by some procedure established by state law, if the petition is certified or not. A petition is certified by state law if the number of distinct valid signatures in the petition is equal to or exceeds the minimum required.

In this article, we are considering the case when a petition of known size contains both invalid and valid signatures. Valid signatures from registered voters can appear more than once. It is assumed that a simple random sample of signatures is drawn from the entire petition and all signatures in the sample are verified. Our interest is to estimate the number of distinct valid signatures in the petition based on the sample information and the

[1] Department of Statistics, University of Los Andes, Venezuela and Department of Statistics, Oregon State University, USA. Email: rsmith@science.oregonstate.edu
[2] Department of Statistics, Oregon State University, USA. Email: thomasd541@msn.com

knowledge of the petition size. Many states use this approach, including California, Illinois, Oregon, and Washington (Hauser 1985).

When no invalid signatures are present, the estimation problem reduces to one known as estimation of the number of classes in a finite population. A class here is equivalent to a unique valid signature. Goodman (1949) showed that the linear unbiased estimator for the total number of classes in a finite population is unique under the assumption that the sample size is no smaller than the maximum number of elements in any class. Deming and Glasser (1959) proposed an estimator for the number of classes on matching sampling frames. Chao and Lee (1992) proposed a nonparametric estimator for the unknown number of classes based on sample coverage. Bunge and Fitzpatrick (1993) provided a review of applications and techniques proposed to estimate the number of classes in finite and infinite populations. Haas and Stokes (1998) proposed nonlinear estimators, based on the generalized jackknife technique, for the number of classes in a finite population with small variation in the population size classes. Recently, Stokes (2003) proposed an estimator for the number of distinct species in a finite population that incorporates auxiliary information correlated with the class size.

Following Goodman's approach, we consider a linear unbiased estimator for the number of distinct valid signatures in the petition. Several other linear estimators and one nonlinear estimator are also considered. In Section 2 we introduce terminology and notation pertinent to our problem. The estimators are described in Section 3. Expressions for the variance of the linear estimators are also provided. In Section 4 we compare the performance of all estimators, and in Section 5 we give a summary.

2. Terminology and Notation

After petition signatures are collected, the state elections office reviews each sheet and removes all the signature pages obtained that do not satisfy state regulations. This procedure leads to a subset of the total number of signature pages originally collected, which will be subject to a verification procedure. This subset of signatures is called the petition here.

Signatures in the petition can be classified as valid (from registered voters) or invalid signatures, for example: illegible writing and signatures different from the ones contained in the registration records. Let N denote the size of the petition, and U and V the unknown number of invalid and distinct valid signatures in the petition, respectively. Further, let D denote the total number of duplicates (replicates) of valid signatures in the petition. Note that "duplicate" is used here to describe all signatures by an elector after his or her first signature. Therefore, $N = U + V + D$.

In this article we are interested in estimating the unknown number of distinct valid signatures in the petition, V, which can also be expressed as:

$$V = N - U - D \tag{1}$$

Since N is known, an estimator for V can be obtained by determining estimators for U and D. As an unbiased estimator for U under simple random sampling design is given by $\widehat{U} = \frac{N}{n} u$, our problem reduces to the estimation of D.

Let N_j be the number of times the *j*th distinct valid signature appears in the petition, $j = 1, \ldots, V$. Therefore, the *j*th distinct valid signature has $N_j - 1$ duplicated signatures in the petition, $j = 1, \ldots, V$, so D can be expressed as

$$D = \sum_{j=1}^{V} (N_j - 1)$$

Usually, one would expect N_j to be small, such as 1 or 2 for most registered voters (electors) signing a petition.

Let F_j be the number of electors with i valid signatures in the petition, $i = 1, \ldots, N - U$. Observe that by definition $0 \le F_i \le V$, and satisfies the equation

$$F_i = \sum_{j=1}^{V} I(N_j = i) \tag{2}$$

where $I(\cdot)$ denotes the indicator function. Based on Equation (2), we obtain expressions for N and V

$$N = U + \sum_{i=1}^{N-U} iF_i \tag{3}$$

$$V = \sum_{i=1}^{N-U} F_i \tag{4}$$

From Equations (3) and (4) we can rewrite D as

$$D = \sum_{i=2}^{N-U} (i-1)F_i$$

Assume a sample of n signatures is drawn at random without replacement from the petition. Let u be the observed number of invalid signatures in the sample and f_i be the number of electors in the sample with i valid signatures. Then n can be written as $n = u + \sum_{i=1}^{n-u} if_i$

3. Theoretical Background

3.1. Estimators for D

First, the form of the unbiased estimator, $\hat{D}_{\text{unbias}}$, for D is determined. Let $k = \max(N_1, \ldots, N_V)$. Suppose a sample of $n\,(n \ge k)$ signatures is drawn without replacement from a petition of size N. Let P_{ij} denote the hypergeometric probability,

$$P_{ij} = \frac{\binom{j}{i}\binom{N-j}{n-i}}{\binom{N}{n}}$$

that i signatures from an elector who signed j times in the petition of size N will be observed in a random sample of size n,

$$c_2 = 1, \text{ and } c_j = (j-1) - \sum_{i=2}^{j-1} c_i \frac{P_{ij}}{P_{ii}} \quad \text{for } j = 3, 4, \ldots, n$$

Then, under the assumption that $n \geq k$, an unbiased estimator of D is given by

$$\hat{D}_{\text{unbias}} = \sum_{i=2}^{n} \frac{c_i}{P_{ii}} f_i \tag{5}$$

The proof of this result is given in Lemma 1 of the Appendix. It can be shown that $\hat{D}_{\text{unbias}}$ is equivalent to the unique unbiased estimator for D derived by Goodman (1949). Goodman proved that the unbiased estimator of D exists only when $n \geq k$. In large state initiative petitions k is expected to be very small relative to n, for example see Table 1 where $k \leq 12$.

Observe that the expansion factors, c_i/P_{ii}, for f_i, can take positive or negative values. These expansion factors can be very large in absolute value, because the selection probabilities P_{ii} can be very small depending on the petition and sample sizes. As a result the estimator $\hat{V}_{\text{unbias}}$ obtained by using $\hat{D}_{\text{unbias}}$ in Equation (1) can be unreasonable. This characteristic of the unbiased estimator was first discussed by Goodman (1949). He also pointed out that in many cases the estimator $\hat{D}_{\text{unbias}}$ rendered a large variance. Hou and Ozsoyoglu (1991) and Hou, Ozsoyoglu, and Taneja (1988) found that unless the sampling fraction is quite large $\hat{D}_{\text{unbias}}$ can result in unreasonable estimates. To avoid this difficulty, we consider alternative linear estimators, which ignore the valid signatures appearing more than two or three times in the sample:

$$\hat{D}_2 = \frac{N(N-1)}{n(n-1)} f_2 \tag{6}$$

$$\hat{D}_3 = \frac{N(N-1)}{n(n-1)} f_2 - \frac{N(N-1)(N-3n+4)}{n(n-1)(n-2)} f_3 \tag{7}$$

Goodman (1949) proposed $\hat{D}_2$ for estimating the number of duplicates of classes in a finite population. The next estimator considered is used by the Washington Elections Division Office[3]:

$$\hat{D}_{2+} = \frac{N(N-1)}{n(n-1)} f_{2+} \quad \text{where } f_{2+} = \sum_{i=2}^{n} f_i \tag{8}$$

Notice that f_{2+} is the number of electors in the sample with valid signatures appearing two or more times.

Note that if at most pairs of valid signatures occur in the petition ($F_j = 0 \quad \text{for } j \geq 3$) then the estimators (6–8) are equal to the unbiased estimator, $\hat{D}_{\text{unbias}}$. Similarly, $\hat{D}_3 = \hat{D}_{\text{unbias}}$ when at most triplicate valid signatures occur in the petition ($F_j = 0 \quad \text{for } j \geq 4$), but otherwise, the estimators (6–8) are biased.

[3] Pamela Floyd, Elections Division, Voter Registration Services, Office of Secretary of State, telephone interview, February 9, 1999.

To possibly reduce the bias of a linear estimator for D, we consider using information from other fully verified petitions to approximate a bias adjustment factor (BAF). The adjusted estimators for D are defined for any linear estimator $\hat{D}$ as

$$\hat{D}_{\text{adj}} = \hat{D} \times \text{BAF}$$

with bias adjustment factor

$$\text{BAF} = \frac{D}{\text{E}(\hat{D})}$$

where $D = \sum_{j=2}^{k} (j-1)F_j$, F_j denotes the number of electors with j valid signatures in the petition, and k is the maximum number of times any valid signatures appear in the petition, $k = \max\{j : F_j > 0\}$.

Let $\hat{D}$ be the biased estimators defined in Equations (6) and (8). Then the BAF for $\hat{D}$ can be written as

$$B_{n,N,k,r}^{\hat{D}} = \begin{cases} \dfrac{1 + \sum_{j=3}^{k} (j-1)r_j}{1 + \sum_{j=3}^{n} P_{2j} r_j / P_{22}} & \text{for } \hat{D} = \hat{D}_2 \\[2ex] \dfrac{1 + \sum_{j=3}^{k} (j-1)r_j}{1 + \sum_{j=3}^{n} r_j \sum_{i=2}^{j} P_{ij} / P_{22}} & \text{for } \hat{D} = \hat{D}_{2+} \end{cases}$$

which is a function of N, n, k and $\boldsymbol{r}$, where $\boldsymbol{r} = (r_3, r_4, \ldots, r_k)$ with $r_j = F_j/F_2$ for $j = 3, \ldots, k$. The BAF depends on the unknown values k and r_j for $j = 3, \ldots, k$, which are unknown. However, when prior information is available from previous fully verified petitions, this information can be used to approximate the corresponding unknown values in the BAF for the sampled petition, and an approximation for the BAF can be obtained, resulting in possible bias reduction. In some states, including Washington, some petitions are fully verified.

By approximating the hypergeometric probabilities P_{ij} by the corresponding binomial probabilities, $P_{ij} \approx \begin{pmatrix} j \\ i \end{pmatrix} q^i (1-q)^{j-i}$, where $q = n/N$ is the sampling fraction, the BAF can be simplified as

$$B_{q,k,\boldsymbol{r}}^{\hat{D}} = \begin{cases} \dfrac{1 + \sum_{i=3}^{k} (i-1)r_i}{1 + \dfrac{1}{2(1-q)^2} \displaystyle\sum_{i=3}^{k} i(i-1)(1-q)^i r_i} & \text{for } \hat{D} = \hat{D}_2 \\[3ex] \dfrac{1 + \sum_{i=3}^{k} (i-1)r_i}{1 + \dfrac{1}{q^2} \displaystyle\sum_{i=3}^{k} (1 - (1 + (i-1)q)(1-q)^{i-1}) r_i} & \text{for } \hat{D} = \hat{D}_{2+} \end{cases}$$

Then the adjusted estimator for D is approximated as

$$\hat{D}_{\text{adj}} = \hat{D} \times \text{B}_{q,k,r}^{\hat{D}} \tag{9}$$

The binomial approximation gives values of the BAF, which are very similar to those obtained using the exact hypergeometric sampling distribution of signatures when N and n

are large. The binomial sampling approximation was also used by Goodman (1949) and Haas and Stokes (1998) to simplify calculations in their work. For comparing the performance of the estimators in Section 4, we use for each of the elements of $\boldsymbol{r} = (r_3, r_4, \ldots, r_k)$ the average of the corresponding r-values over three fully verified petitions.

3.2. Estimators for V

Estimators for V can be obtained by substituting in Equation (1) any of the estimators for D presented in Equations (5–9)

$$\hat{V} = N - \hat{U} - \hat{D} \quad \text{with } \hat{D} = B\sum_{i=2}^{t} A_i f_i \tag{10}$$

for constants B, t, and A_i, with $B = \hat{B}^{\hat{D}}_{q,k,\boldsymbol{r}}$ for the adjusted estimators and $B = 1$ for the unadjusted estimators. The constant t is defined as $t = 2$ for $\hat{D}_2$ and $\hat{D}_{2\text{adj}}$, $t = 3$ for $\hat{D}_3$ and $t = n$ for $\hat{D}_{2+}$, $\hat{D}_{2+\text{adj}}$ and $\hat{D}_{\text{unbias}}$. The coefficients $A_i \neq 0$ corresponding to Equations (5–8) are

$$A_2 = \frac{c_2}{P_{22}} = \frac{N(N-1)}{n(n-1)} \text{ for all estimators}$$

$$A_i = A_2 \text{ for } i = 3, \ldots, n \text{ for } \hat{D}_{2+} \text{ and } \hat{D}_{2+\text{adj}}$$

$$A_3 = \frac{c_3}{P_{33}} = \frac{N(N-1)(N-3n+4)}{n(n-1)(n-2)} \text{ for } \hat{D}_3 \text{ and } \hat{D}_{\text{unbias}}$$

$$A_i = \frac{c_i}{P_{ii}} \text{ for } i = 4, \ldots, n \text{ for } \hat{D}_{\text{unbias}}$$

Haas and Stokes (1998) considered generalized jackknife estimators for the number of classes in a finite population. They recommended a second-order jackknife estimator, $\hat{V}_{uj2}$, for the number of classes in a finite population, V, when the squared coefficient of variation (γ^2) of the class sizes, $N_1, N_2, \ldots, N_V$

$$\gamma^2 = \frac{(1/V)\sum_{j=1}^{V}(N_j - \bar{N})^2}{\bar{N}^2} \quad \text{where } \bar{N} = (1/V)\sum_{j=1}^{V} N_j$$

is relatively small, $\gamma^2 \leq 1$. With this estimator Haas and Stokes attempted to reduce the bias of a first order estimator, $\hat{V}_{uj1}$, for V. In petitions, the squared coefficient of variation for the number of replication of valid signatures $(N_1, N_2, \ldots, N_V)$, γ^2, is expected to be small. Then, their estimator $\hat{V}_{uj2}$ will be directly applicable to the estimation of the number of distinct valid signatures in initiative petitions if no invalid signatures are present, $U = 0$. In initiative petitions we are interested in the number of unique signatures, V, in the sub-population of valid signatures. The size of this subpopulation is $N - U$, and the corresponding sample size is $n - u$. A direct modification of the Haas and Stokes second-order jackknife estimator, that accommodates the additional class of invalid signatures for

$u < n$ is given by replacing n by $n - u$ and N by $N - \hat{U}$. The resulting estimator for V is

$$\hat{V}_{uj2m} = \hat{V}_{uj1m} \times \left(1 - \frac{f_1(1-q)\ln(1-q)\hat{\gamma}^2\left(\hat{V}_{uj1m}\right)}{q\sum_{i=1}^{n-u} f_i}\right) \text{ where } \hat{V}_{uj1m} = \left(1 - \frac{(1-q)f_1}{n-u}\right)^{-1} \sum_{i=1}^{n-u} f_i$$

is a direct modification of the Haas and Stokes first-order estimator and

$$\hat{\gamma}^2\left(\hat{V}_{uj1m}\right) = \max\left(0, \frac{\hat{V}_{uj1m}}{(n-u)^2}\sum_{i=1}^{n-u} i(i-1)f_i + \frac{\hat{V}_{uj1m}}{N - \hat{U}} - 1\right)$$

is an estimator squared coefficient of variation (γ^2) among $N_1, N_2, \ldots, N_V$.

3.3. *Expectation and variance of $\hat{V}$*

The expected value and variance for any estimator of the general form given in Equation (10) is obtained as

$$E(\hat{V}) = N - U - B\sum_{i=2}^{t} A_i \sum_{j=1}^{n} P_{ij}F_j \tag{11}$$

$$\mathrm{Var}(\hat{V}) = \mathrm{Var}(\hat{U}) + \mathrm{Var}(\hat{D}) + 2\mathrm{Cov}(\hat{U}, \hat{D}) \tag{12}$$

where

$$\mathrm{Var}(\hat{U}) = \frac{N^2}{n}\left(\frac{N-n}{N-1}\right)\frac{U}{N}\left(1 - \frac{U}{N}\right)$$

$$\mathrm{Var}(\hat{D}) = B^2\sum_{i=2}^{t}\sum_{k=2}^{t} A_i A_k \sum_{j=i}^{n}\sum_{l=k}^{n} v_{ijkl}$$

$$\mathrm{Cov}(\hat{U}, \hat{D}) = -\frac{BU}{n}\sum_{i=2}^{t} A_i \sum_{j=1}^{n}\left(\frac{iN - jn}{N - j}\right) P_{ij}F_j$$

$$v_{ijkl} = \begin{cases} (1 + (P_{ij.ij} - P_{ij})F_j - P_{ij.ij})P_{ij}F_j & \text{for } i = k,\ j = l \\ ((P_{kj.ij} - P_{kj})F_j - P_{kj.ij})P_{ij}F_j & \text{for } i \neq k,\ j = l \\ (P_{kl.ij} - P_{kl})P_{ij}F_jF_l & \text{for } j \neq l \end{cases}$$

$$P_{ij} = \frac{\binom{j}{i}\binom{N-j}{n-i}}{\binom{N}{n}} \quad \text{and} \quad P_{kl.ij} = \frac{\binom{l}{k}\binom{N-j-l}{n-i-k}}{\binom{N-j}{n-i}}$$

The expression for the expected value of the linear estimator $\hat{V}$ follows from Equation (10), the unbiasedness of $\hat{U}$, and Equation (A.2) of the Appendix. The expression for $\mathrm{Var}(\hat{U})$ is well-known, and the expressions for $\mathrm{Var}(\hat{D})$ and $\mathrm{Cov}(\hat{U}, \hat{D})$, are derived in the Appendix.

4. Performance of the Estimators

In this section, the estimators for the number of distinct valid signatures, V, are compared with regard to their bias and root mean squared error (RMSE) for four fully verified Washington State petitions, denoted as A, B, C, and D. In Washington, if the random sample indicates that V attains the required number then the measure is certified. Otherwise, complete verification of the petition is required.

Table 1 describes the four petitions with regard to: petition size (N), numbers of invalid signatures (U), duplicates of valid signatures (D), distinct valid signatures (V), the number of electors with i valid signatures in the petition (F_i), and the squared coefficient of variation, γ^2, for the number of times (N_j) distinct valid signatures appear in the petition. Also included is the year that each petition was submitted for verification. The petition sizes range from 162,324 to 231,723, the proportion of invalid signatures from 12.0 to 20.4 percent, the duplication rates from 2.0 to 5.6 percent, and the proportion of distinct valid signatures from 76.4 to 85.4 percent. The petitions C and D with the largest percentage of pairs (F_2) also have the largest percentage of triplicates (F_3) and quadruples (F_4). Only two petitions have electors who signed more than four times, petition B has one elector who signed twelve times and petition D has two electors who signed six times, and three electors who signed five times. For all four petitions, the proportion of electors with triplicates or higher, is small ($<0.24\%$). As expected, all four petitions have small values of γ^2.

Table 2 displays the expected frequency for replications of distinct valid signatures in the sample for each sampling fraction and petition. For sampling fractions 3%, 5%, and 10%, and all four petitions, the expected number of distinct valid signatures that appear more than twice in a random sample is less than one. When the sampling fraction is increased to 20%, the expected number of triplicate valid signatures exceeds one only for petitions B, C, and D, and the expected number of quadruples or higher is less than 0.22.

To calculate the bias adjustment factors, $B^{\hat{D}}_{q,k,\boldsymbol{r}}$, we need to specify k and $r_i = F_i/F_2$ for $i = 3, \ldots, k$, where $q = n/N$. When sampling is used the values of k and r_i, $i = 3, \ldots, k$, are unknown. Here, for each petition, $i = \mathrm{A, B, C,}$ and D, information from only the other three petitions is used to specify values for the unknown k and $r_3, \ldots, r_k$. For each

Table 1. Description of the petitions A, B, C, and D

	A (1984)	B (1995)	C (1989)	D (1996)
N	162,324	231,723	173,561	228,148
U (%)	19,437 (12.0)	47,383 (20.4)	31,325 (18.0)	34,542 (15.1)
D (%)	4,256 (2.6)	4,546 (2.0)	9,738 (5.6)	11,584 (5.1)
V (%)	138,631 (85.4)	179,794 (77.6)	132,448 (76.3)	182,022 (79.8)
F_1 (%)	134,489 (82.9)	175,363 (75.7)	123,205 (71.0)	170,988 (74.9)
F_2 (%)	4,031 (2.5)	4,331 (1.9)	8,878 (5.1)	10,518 (4.6)
F_3 (%)	108 (0.07)	93 (0.04)	385 (0.22)	489 (0.21)
F_4	3	6	30	22
F_5		0		3
F_6		0		2
F_{12}		1		
γ^2	0.0296	0.0252	0.0652	0.0584

Table 2. Expected frequency for replications of valid signatures, $E(f_i)$[1]

Sampling fraction	i	A	B	C	D
3%	2	3.93	4.22	9.15	10.91
	3	0.0032	0.0077	0.0135	0.0173
	≥ 4	<0.0001	0.0003	<0.0001	<0.0001
5%	2	10.89	11.67	25.34	30.20
	3	0.0149	0.0318	0.0624	0.07924
	≥ 4	<0.0001	0.0023	0.0002	<0.0001
10%	2	43.37	46.34	100.63	119.87
	3	0.1188	0.1998	0.4929	0.6216
	≥ 4	0.0003	0.0262	0.0030	0.0061
20%	2	172.02	183.37	396.69	472.15
	3	0.9407	1.1338	3.8479	4.7925
	≥ 4	0.0050	0.2149	0.0480	0.0893

[1] $E(f_i) = \sum_{j=1}^{n} P_{ij}F_j$

petition, the specified value, $\tilde{k}$, was determined as the maximum of the observed k-values from the other three petitions. Similarly, the specified vector, $\tilde{\boldsymbol{r}}$, is calculated as the average of the known entries for the other three petitions. Table 3 gives the true and specified values for k and $\boldsymbol{r}$ for each petition. Observe that for each petition, the specified values for r_i, ($\tilde{r}_i, i = 3, \ldots, k$) are similar to the true ones. Furthermore, $\tilde{r}_i$ are very close to zero, for $i \geq 4$, suggesting that the corresponding class frequency F_i is relatively small.

Table 4 gives values for the bias adjustment factors, $B^{\hat{D}}_{q,k,\boldsymbol{r}}$, using $k = 3$ and $k = 12$ for each petition, estimator, and sampling fraction (q): 3%, 5%, 10%, and 20%. From Table 4, we can see that the values of the BAF corresponding to $\tilde{\boldsymbol{r}} = \tilde{r}_3$ and $\tilde{\boldsymbol{r}} = (\tilde{r}_3, \ldots, \tilde{r}_{12})$ are similar in all cases. Therefore, we consider only the bias adjustment factor based on at most triplicate valid signatures, $\tilde{\boldsymbol{r}} = \tilde{r}_3$, hereafter.

For each linear estimator, we use Equations (11) and (12) to compute the bias and root mean squared error (RMSE)

$$\mathrm{Bias}(\hat{V}) = \mathrm{E}(\hat{V}) - V \ \text{ and } \ \mathrm{RMSE} = \sqrt{\mathrm{Var}(\hat{V}) - \{\mathrm{Bias}(\hat{V})\}^2}$$

Table 3. True values of k and $r_3, \ldots, r_k$, and specified values $\tilde{k}$ and $\tilde{r}_3, \ldots, \tilde{r}_k$

	A		B		C		D	
	True	Spec	True	Spec	True	Spec	True	Spec
k	4	12	12	6	4	12	6	12
r_3	0.0268	0.0371	0.0215	0.0389	0.0434	0.0316	0.0465	0.0306
r_4	0.0007	0.0023	0.0014	0.0021	0.0034	0.0014	0.0465	0.0018
r_5	0	0.0001	0	0.0001	0	0.0001	0.0003	0
r_6	0	0.0001	0	0.0001	0	0.0001	0.0002	0
r_{12}	0	0.0001	0.0002	0	0	0.0001	0	0.0001

Note: The entries of $\boldsymbol{r} = (r_3, \ldots, r_{12})$ and $\tilde{\boldsymbol{r}} = (\tilde{r}_3, \ldots, \tilde{r}_{12})$ not displayed are equal to zero.

Table 4. Specified values of the bias adjusted factor, $\boldsymbol{B}^{\hat{D}}_{q,k,r}$, for each petition, adjusted estimator and sampling fraction (q): 3%, 5%, 10%, and 20%

		A		B		C		D	
q	Estimator	$k = 3$	$k = 12$	$k = 3$	$k = 12$	$k = 3$	$k = 12$	$k = 3$	$k = 12$
3%	$\hat{D}_{2\text{adj}}$	0.970	0.960	0.968	0.962	0.974	0.967	0.974	0.966
	$\hat{D}_{2+\text{adj}}$	0.969	0.958	0.967	0.961	0.974	0.966	0.973	0.965
	$\hat{D}_{d\text{adj}}$	0.968	0.957	0.966	0.960	0.973	0.964	0.972	0.963
5%	$\hat{D}_{2\text{adj}}$	0.971	0.970	0.963	0.965	0.976	0.970	0.975	0.969
	$\hat{D}_{2+\text{adj}}$	0.970	0.961	0.969	0.963	0.975	0.967	0.974	0.967
	$\hat{D}_{d\text{adj}}$	0.968	0.958	0.967	0.961	0.973	0.965	0.973	0.964
10%	$\hat{D}_{2\text{adj}}$	0.976	0.971	0.975	0.971	0.980	0.976	0.980	0.976
	$\hat{D}_{2+\text{adj}}$	0.973	0.966	0.972	0.967	0.977	0.972	0.977	0.971
	$\hat{D}_{d\text{adj}}$	0.970	0.961	0.969	0.963	0.975	0.967	0.974	0.966
20%	$\hat{D}_{2\text{adj}}$	0.986	0.985	0.986	0.984	0.989	0.988	0.988	0.987
	$\hat{D}_{2+\text{adj}}$	0.980	0.976	0.979	0.976	0.983	0.980	0.982	0.979
	$\hat{D}_{d\text{adj}}$	0.973	0.966	0.972	0.967	0.977	0.972	0.977	0.971

Table 5. Bias (RMSE) of estimators for V expressed as the ratio to the true number of distinct valid signatures in the petition and multiplied by 1,000

Sampling fraction	Estimator	Petitions			
		A	B	C	D
3%	$\hat{V}_{\text{unbias}}$	0 (24.878)	0 (6,384,553)	0 (56.621)	0 (271.840)
	$\hat{V}_3$	0.019 (21.636)	0.667 (21.540)	0.200 (39.375)	0.255 (31.901)
	$\hat{V}_2$	−0.766 (16.644)	−0.769 (13.960)	−3.243 (26.130)	−2.940 (20.822)
	$\hat{V}_{2+}$	−0.792 (16.651)	−0.818 (13.972)	−3.357 (26.162)	−3.046 (20.851)
	$\hat{V}_{\text{2adj}}$	0.193 (16.171)	0.060 (13.578)	0.060 (13.578)	−1.238 (20.155)
	$\hat{V}_{\text{2+adj}}$	0.198 (16.163)	0.039 (13.576)	−1.268 (25.315)	−1.286 (20.157)
	$\hat{V}_{uj2m}$	−25.881 (37.665)	−22.244 (31.838)	−56.326 (68.162)	−51.090 (59.733)
5%	$\hat{V}_{\text{unbias}}$	0 (12.546)	0 (262,326)	0 (24.380)	0 (58.260)
	$\hat{V}_3$	0.018 (11.867)	0.524 (11.094)	0.185 (20.318)	0.230 (16.364)
	$\hat{V}_2$	−0.714 (10.261)	−0.679 (8.815)	−3.017 (16.021)	−2.730 (12.806)
	$\hat{V}_{2+}$	−0.758 (10.269)	−0.755 (8.829)	−3.206 (16.074)	−2.905 (12.858)
	$\hat{V}_{\text{2adj}}$	0.181 (9.990)	0.094 (8.600)	−1.131 (15.426)	−1.142 (12.294)
	$\hat{V}_{\text{2+adj}}$	0.191 (9.982)	0.065 (8.596)	−1.207 (15.428)	1.222 (12.300)
	$\hat{V}_{uj2m}$	−25.273 (30.225)	−21.466 (25.543)	−54.686 (59.255)	−49.622 (53.015)
10%	$\hat{V}_{\text{unbias}}$	0 (5.737)	0 (2,960)	0 (9.300)	0 (9.483)
	$\hat{V}_3$	0.014 (5.677)	0.287 (5.163)	0.149 (8.893)	2178 (7.109)
	$\hat{V}_2$	−0.586 (5.430)	−0.491 (4.873)	−2.455 (8.422)	−2.213 (6.779)
	$\hat{V}_{2+}$	−0.672 (5.445)	−0.617 (4.891)	−2.830 (8.551)	−2.558 (6.910)
	$\hat{V}_{\text{2adj}}$	0.153 (5.306)	0.144 (4.782)	−0.907 (7.972)	0.910 (6.368)
	$\hat{V}_{\text{2+adj}}$	0.172 (5.298)	0.110 (4.775)	−1.057 (7.985)	−1.065 (6.389)
	$\hat{V}_{uj2m}$	−23.065 (24.470)	−19.221 (20.455)	−49.610 (50.845)	−44.443 (45.400)
20%	$\hat{V}_{\text{unbias}}$	0 (2.925)	0 (22.966)	0 (4.252)	0 (3.421)
	$\hat{V}_3$	0.010 (2.922)	0.102 (2.769)	0.100 (4.230)	0.109 (3.368)
	$\hat{V}_2$	−0.330 (2.916)	−0.213 (2.756)	−1.353 (4.363)	−1.207 (3.514)
	$\hat{V}_{2+}$	−0.500 (2.943)	−0.401 (2.778)	−2.088 (4.651)	−1.878 (3.803)
	$\hat{V}_{\text{2adj}}$	0.093 (2.875)	0.150 (2.736)	−0.473 (4.141)	−0.468 (3.307)
	$\hat{V}_{\text{2+adj}}$	0.133 (2.868)	0.144 (2.729)	−0.763 (4.175)	−0.763 (3.355)
	$\hat{V}_{uj2m}$	−17.884 (18.303)	−14.652 (15.050)	−38.728 (39.090)	−34.569 (34.840)

For the nonlinear estimator, $\hat{V}_{uj2m}$, we estimate the bias and RMSE from 10,000 independent simulated random samples, drawn without replacement from each petition. When evaluating the nonlinear estimator, following Haas and Stokes (1998), we truncated each estimate, below at $\sum_{i=1}^{n-u} f_i$ and above $N - \hat{U}$.

In Table 5, the bias and RMSE are given for the nine estimators of V for Petitions A–D and sampling fractions: 3%, 5%, 10%, and 20%. In this table, the bias and RMSE are each expressed as the ratio to the true number of distinct valid signatures in the petition and multiplied by 1,000 ($\mathrm{Bias}/V \times 1,000$ and $\mathrm{RSME}/V \times 1,000$). For the adjusted estimator, Equation (9), $k = 3$ $(\tilde{\boldsymbol{r}} = \tilde{r}_3)$ is used for the bias adjusted factor, $B^{\hat{D}}_{q,k,\boldsymbol{r}}$.

In Table 5 the estimator $\hat{V}_3$ tends to have a relatively small positive bias (≤ 0.667) in all cases. The biases of $\hat{V}_2$ and $\hat{V}_{2+}$ are negative in all cases, corresponding to positive biases in the estimators for the number of duplicates of valid signatures $\hat{D}_2$ and $\hat{D}_{2+}$. Note that the difference between the biases of these estimators tends to increase as the sampling fraction increases. This is expected since the number of triplicate and quadruple valid signatures increases with sample size (Table 2). The two adjusted estimators show a small reduction in the absolute bias when compared with their nonadjusted counterparts. The nonlinear estimator, $\hat{V}_{uj2m}$, tends to have a relatively large negative bias ranging from -56.326 to -14.652. A possible reason for this result may originate in the development of the jackknife estimators, where all the N_j are estimated by $\bar{N}$. It appears that the variability induced by this approximation may affect the jackknife estimator performance. This point was recently addressed by Stokes (2003), who discusses the effect of populations with small average class sizes in estimating the number of classes in a finite population, V, and suggests a tendency of the jackknife estimators to underestimate V when the class sizes vary.

From Table 5, it can be seen that the RMSE decreases at a faster rate than $1/\sqrt{n}$ for all estimators and petitions. This results from the corresponding property of the estimators for D in Equation (10). The estimator $\hat{V}_3$ has always smaller RMSE than $\hat{V}_{\mathrm{unbias}}$. The estimator $\hat{V}_2$ has smaller RMSE than $\hat{V}_3$, except for the 20% sampling fraction for petitions C and D. The estimators $\hat{V}_2$ and $\hat{V}_{2+}$ tend to have similar RMSE's for the sampling fractions of 3%, 5%, and 10% over all four petitions. This is as expected from the form of the estimators and the very small expected number of triplicate or higher replications of distinct valid signatures (Table 2). For the 20% sampling fraction, the RMSE for $\hat{V}_{2+}$ is slightly larger than the RMSE's for $\hat{V}_2$ for petitions C and D, and similar for petitions A and B. The adjusted estimators $\hat{V}_{2\mathrm{adj}}$ and $\hat{V}_{2+\mathrm{adj}}$ show a slight reduction in the RMSE compared to their nonadjusted counterparts. These two adjusted estimators have similar RMSE's in all cases. The RMSE for the nonlinear estimator $\hat{V}_{uj2m}$ is relatively large in all cases.

5. Summary

In this article we have compared several estimators for the number of distinct valid signatures in a petition. Explicit forms for the bias and RMSE were provided for the linear estimators. Simulated random samples were used to estimate the bias and RMSE of the nonlinear estimator, $\hat{V}_{uj2m}$, adapted from Haas and Stokes (1998).

Small sampling fractions less than or equal to 10% are typically used for sampling state petitions. For these sample sizes it was difficult to improve much on the Goodman-type estimator $\hat{V}_2$, which is unbiased when valid signatures are duplicated at most once.

This results from the very small probability of observing higher duplicate replication from typical petitions. When duplication data is available from similar fully verified petitions, it might be possible to reduce the bias of the (biased) linear estimators.

Appendix

Calculation of $E(\hat{D})$, $\mathrm{Var}(\hat{D})$, *and* $\mathrm{Cov}(\hat{U}, \hat{D})$

Consider a random sample of n signatures drawn without replacement from a petition of size N. Let $\delta_{j\alpha}$ denote the number of valid signatures in the sample from the αth elector who signed j valid signatures in the petition, for $\alpha = 1, 2, \ldots, F_j$ and $j = 1, \ldots, N - U$. Note that $\delta_{j\alpha}$ has the hypergeometric (N, n, j, i) with $P_{ij} = P(\delta_{j\alpha} = i)$ given by

$$P_{ij} = \frac{\binom{j}{i}\binom{N-j}{n-i}}{\binom{N}{n}} \quad \text{for } i = 0, 1, \ldots, j$$

Similarly, the conditional distribution of $\delta_{j\beta}$, given $\delta_{j\alpha} = i$ is hypergeometric $(N - j, n - i, l, k)$ with $P_{kl\cdot ij} = P(\delta_{j\beta} = k | \delta_{j\alpha} = i)$ given by

$$P_{kl\cdot ij} = \frac{\binom{l}{k}\binom{N-j-l}{n-i-k}}{\binom{N-j}{n-i}} \quad \text{for } k = 0, 1, \ldots, l$$

For the number of electors in the sample with i valid signatures, f_i, write

$$f_i = \sum_{j=i}^{n} f_{ij} \quad \text{with } f_{ij} = \sum_{\alpha=1}^{F_j} I(\delta_{j\alpha} = i)$$

where f_{ij} is the number of electors with i signatures in the sample and j signatures in the petition $(i \leq j)$ and $I(\cdot)$ is the indicator function. Note that f_{ij} is not observable, but f_i is. Then,

$$\mathrm{E}(f_{ij}) = P_{ij}F_j \quad \text{and} \quad E(f_i) = \sum_{j=i}^{n} P_{ij}F_j \tag{A.1}$$

Thus, from the general form of the linear estimator $\hat{D} = B\sum_{i=2}^{t} A_i f_i$ we have

$$E(\hat{D}) = B\sum_{i=2}^{t} A_i \sum_{j=i}^{n} P_{ij}F_j \tag{A.2}$$

for constants B, t, and A_i, which are given in Equation (10).

LEMMA 1. Let $k = \max(N_1, \ldots, N_V)$. Suppose a sample of n $(n \geq k)$ signatures is drawn without replacement from a petition of size N. Define

$$c_2 = 1 \quad \text{and} \quad c_j = (j-1) - \sum_{i=2}^{j-1} c_i \frac{P_{ij}}{P_{ii}} \quad \text{for } j = 3, 4, \ldots, n$$

Then, an unbiased estimator of D is given by

$$\hat{D}_{\text{unbias}} = \sum_{i=2}^{n} \frac{c_i}{P_{ii}} f_i \tag{A.3}$$

Proof. The unbiasedness property for $\hat{D}_{\text{unbias}}$ follows from substitution of the expectation for f_i in Equation (A.3).

$$E(\hat{D}_{\text{unbias}}) = \sum_{i=2}^{n} \frac{c_i}{P_{ii}} E(f_i) = \sum_{i=2}^{n} \frac{c_i}{P_{ii}} \sum_{j=i}^{n} P_{ij} F_j = F_2 + \sum_{j=3}^{n} F_j \left(\sum_{i=2}^{j-1} c_i \frac{P_{ij}}{P_{ii}} + c_j \right)$$

$$= \sum_{j=2}^{n} F_j (j-1) \text{ since } c_j = (j-1) - \sum_{i=2}^{j-1} c_i \frac{P_{ij}}{P_{ii}} \quad j = 3, 4, \ldots, n$$

$$= \sum_{j=2}^{N-U} F_j (j-1) \text{ since } F_j = 0 \quad \text{for } j = k+1, k+2, \ldots, N-U \text{ and } n \geq k$$

$$= D$$

The next result is used for the calculation of $\text{Var}(\hat{D})$ and $\text{Cov}(\hat{U}, \hat{D})$.

Lemma 2. The $\text{Cov}(f_{ij}, f_{kl}) = v_{ijkl}$ $i \leq j$, $k \leq l$, where

$$v_{ijkl} = \begin{cases} (1 + (P_{ij\cdot ij} - P_{ij})F_j - P_{ij\cdot ij})P_{ij}F_j & \text{for } i = k, j = l \\ ((P_{kj\cdot ij} - P_{kj})F_j - P_{kj\cdot ij})P_{ij}F_j & \text{for } i \neq k, j = l \\ (P_{kl\cdot ij} - P_{kl})P_{ij}F_jF_l & \text{for } j \neq l \end{cases}$$

Proof. Substitute (A.1) in

$$\text{Cov}(f_{ij}, f_{kl}) = E(f_{ij}, f_{kl}) - E(f_{ij})E(f_{kl}) = \sum_{\alpha=1}^{F_j} \sum_{\beta=1}^{F_l} P(\delta_{j\alpha} = k, \delta_{j\beta} = k) - (P_{ij}F_j)(P_{kl}F_l)$$

1. Case where $i = k, j = l$

$$\text{Cov}(f_{ij}, f_{kl}) = \sum_{\alpha=1}^{F_l} P(\delta_{j\alpha} = i) + \sum_{\alpha=1}^{F_j} \sum_{\substack{\beta=1 \\ \beta \neq \alpha}}^{F_l} P(\delta_{j\alpha} = i, \delta_{j\beta} = i) - (P_{ij}F_j)^2$$

$$= P_{ij}F_j + P_{ij}P_{ij\cdot ij}F_j(F_j - 1) - (P_{ij}F_j)^2 = (1 + (P_{ij\cdot ij} - P_{ij})F_j - P_{ij\cdot ij})P_{ij}F_j$$

2. Case where $i \neq k, j = l$

$$\text{Cov}(f_{ij}, f_{kj}) = 0 + \sum_{\alpha=1}^{F_j} \sum_{\substack{\beta=1 \\ \beta \neq \alpha}}^{F_j} P(\delta_{j\alpha} = i, \delta_{j\beta} = k) - (P_{ij}F_j)(P_{kj}F_j)$$

$$= P_{ij}P_{kj\cdot ij}F_j(F_j - 1) - P_{ij}P_{kj}F_j^2 = ((P_{kj\cdot ij} - P_{kj})F_j - P_{kj\cdot ij})P_{ij}F_j$$

3. Case where $j \neq l$

$$\text{Cov}(f_{ij}, f_{kl}) = \sum_{\alpha=1}^{F_j} \sum_{\beta=1}^{F_l} P(\delta_{j\alpha} = k, \delta_{l\beta} = k) - (P_{ij}F_j)(P_{kl}F_l) = (P_{kl\cdot ij} - P_{kl})P_{ij}F_jF_l$$

The variance for the general form of the linear estimator $\hat{D} = B\sum_{i=2}^{t} A_i f_i$,

$$\mathrm{Var}(\hat{D}) = B^2 \sum_{i=2}^{t} \sum_{k=2}^{t} A_i A_k \sum_{j=i}^{n} \sum_{l=k}^{n} v_{ijkl}$$

then follows from the covariance of the sums $f_i = \sum_{j=i}^{n} f_{ij}$ and $f_k = \sum_{l=k}^{n} f_{kl}$ and Lemma 2.

Lemma 3. Under the assumption for $F_j = 0, \ j > n$

$$\mathrm{Cov}(u, f_{ij}) = -\left(\frac{iN - jn}{N - j}\right)\left(\frac{U}{N}\right) P_{ij} F_j$$

Proof. For fixed i and j write

$$u = n - \sum_{k=1}^{n} k f_k = n - \sum_{k=1}^{n} k \sum_{l=k}^{n} f_{kl} = n - \sum_{\substack{l=1 \\ l \neq j}}^{n} \sum_{k=1}^{l} k f_{kl} - \sum_{k=1}^{j} k f_{kj}$$

$$= n - \sum_{\substack{l=1 \\ l \neq j}}^{n} \sum_{k=1}^{l} k f_{kl} - i f_{ij} - \sum_{\substack{k=1 \\ k \neq i}}^{i} k f_{kj}$$

Then,

$$\mathrm{Cov}(u, f_{ij}) = -\sum_{\substack{l=1 \\ l \neq j}}^{n} \sum_{k=1}^{l} k \mathrm{Cov}(f_{kl}, f_{ij}) - i \mathrm{Var}(f_{ij}) - \sum_{\substack{k=1 \\ k \neq i}}^{j} k \mathrm{Cov}(f_{kl}, f_{ij})$$

From Lemma 2,

$$\mathrm{Cov}(u, f_{ij}) = \sum_{\substack{l=1 \\ l \neq j}}^{n} \sum_{k=1}^{l} k(P_{kl.ij} - P_{kl}) P_{ij} F_j F_l - i(1 + (P_{ij.ij} - P_{ij}) F_j - P_{ij.ij}) P_{ij} F_j$$

$$- \sum_{\substack{k=1 \\ k \neq i}}^{j} k((P_{kj.ij} - P_{kj}) F_j - P_{kj.ij}) P_{ij} F_j$$

$$= -\left(\sum_{\substack{l=1 \\ l \neq j}}^{n} \sum_{k=0}^{l} k(P_{kl.ij} - P_{kl}) F_l + i(1 + (P_{ij.ij} - P_{ij}) F_j - P_{ij.ij}) \right.$$

$$\left. + \sum_{\substack{k=0 \\ k \neq i}}^{j} k((P_{kj \cdot ij} - P_{kj}) F_j - P_{kj \cdot ij}) \right) P_{ij} F_j$$

$$= -\left(i - \sum_{k=0}^{j} k P_{kj.ij} + \sum_{l=1}^{n} \left(\sum_{k=0}^{j} k(P_{kl.ij} - P_{kl}) \right) F_l \right) P_{ij} F_j$$

Using the expectation of hypergeometric distributions then gives the reduction

$$\mathrm{Cov}(u,f_{ij}) = -\left(i - \frac{j(n-i)}{N-j} + \sum_{l=1}^{n}\left(\frac{l(n-i)}{N-j} - \frac{ln}{N}\right)F_l\right)P_{ij}F_j$$

$$= -\left(\frac{iN-jn}{N-j}\right)\left(1 - \frac{1}{N} + \sum_{l=1}^{n} lF_l\right)P_{ij}F_j = -\left(\frac{iN-jn}{N-j}\right)\left(\frac{U}{N}\right)P_{ij}F_j$$

From Lemma 3, the covariance of $\hat{U} = \frac{N}{n}u$ and $\hat{D} = B\sum_{i=2}^{t} A_i f_i = B\sum_{i=2}^{t}\sum_{j=i}^{n} A_i f_{ij}$ is then

$$\mathrm{Cov}(\hat{U},\hat{D}) = -\frac{BU}{n}\sum_{i=2}^{t} A_i \sum_{j=i}^{n}\left(\frac{iN-jn}{N-j}\right)P_{ij}F_j$$

6. References

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

Chao, A. and Lee, S.M. (1992). Estimating the Number of Classes via Sample Coverage. Journal of the American Statistical Association, 47, 210–217.

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.

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

Haas, P.J. and Stokes, L. (1998). Estimating the Number of Classes in a Finite Population. Journal of the American Statistical Association, 93, 1475–1487.

Hauser, J. (1985). Validating Initiative and Referendum Petition Signatures. Research Monograph. Legislative Research. S420 State Capitol, Salem, Oregon.

Hou, W., Ozsoyoglu, G., and Taneja, B.K. (1988). Statistical Estimators for Relational Algebra Expressions. Proceedings of the ACM Symposium on Principles of Database Systems, 276–287.

Hou, W. and Ozsoyoglu, G. (1991). Statistical Estimators for Aggregate Relational Algebra Queries. ACM Transactions on Database Systems, 16, 600–654.

Stokes, L. (2003). Using Auxiliary Information for Improving Estimation in the Number of Species Problem. Statistica Sinica, 13, 655–671.

Received May 2003
Revised September 2004