ESTIMATING THE NUMBER OF DISTINCT VALID SIGNATURES IN INITIATIVE PETITIONS

Rubén A. Smith-Cayama, Department of Statistics, University of Los Andes, Venezuela (smith@stat.orst.edu)
David R. Thomas, Department of Statistics, Oregon State University

Key works: biased estimator; number of classes; duplication.

Abstract:

In some states, if citizens are dissatisfied with certain laws or feel that new laws are needed, they can petition to place proposed legislation on the ballot. To be certified for the ballot, the sponsor of the petition 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 non-linear 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.

1. Introduction

Some state constitutions 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 paper, 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 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 valid signature. Bunge and Fitzpatrick (1993) provided a review of applications and techniques proposed to estimate the number of classes in finite and infinite populations. 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. Recently, Haas and Stokes (1998) proposed non-linear estimators based on the generalized jackknife technique.

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 non-linear 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 M the unknown number of invalid and distinct valid signatures in the petition, respectively.

Let N_j be the number of times the jth distinct valid signature appears in the petition, $j = 1, \dots, M$. Therefore, the jth distinct valid signature has $N_j - 1$ duplicated signatures in the petition, $j = 1, \dots, M$. We denote by D the total number of duplicates (replicates) of valid signatures in the petition, which can be expressed as,

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

Note that 'duplicate' is used here to describe all signatures by an elector after his or her first signature. Also, F_i is the number of electors with i valid signatures in the petition, $i = 1, \ldots, N$. Observe that $0 \leq F_i \leq M$, so that

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

where $I(\cdot)$ denotes the indicator function. Based on Equation (1), we obtain expressions for N and M

$$N = U + \sum_{i=1}^{N} iF_i \tag{2}$$

$$M = \sum_{i=1}^{N} F_i \tag{3}$$

From Equations (2) and (3) we can rewrite D as

$$D = \sum_{i=2}^{N} (i - 1)F_i. \tag{4}$$

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 $n = u + \sum\limits_{i=1}^{n} if_i$.

3. Theoretical Background

From Equations (2), (3), and (4) we have

$$M = N - U - D \tag{5}$$

Since N is known, an estimator for M can be obtained by determining estimators for U and D. Since 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.

3.1. Estimators for D

First, the form of the unbiased estimator, $\widehat{D}_{\text{unbias}}$, for D is determined. Let $k = \max(N_1, \ldots, N_M)$. Suppose a sample of n $(n \geq k)$ signatures is drawn without replacement from a petition of size N. Define $P_{ij} = \frac{\binom{j}{i}\binom{N-j}{n-i}}{\binom{N}{n}}$ and,

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

Then, an unbiased estimator of D is given by,

$$\widehat{D}_{\text{unbias}} = \sum_{i=2}^{n} \frac{c_i}{P_{ii}} f_i. \tag{6}$$

The proof of this result is given in Smith-Cayama (1999). Observe that the expansion factors, $\frac{c_i}{P_{ii}}$, for f_i, can take positive or negative values. These expansion factors can be very large in absolute value, depending on the petition and sample sizes. As a result the estimator $\widehat{M}_{\text{unbias}}$, obtained by using $\widehat{D}_{\text{unbias}}$ in Equation (5), can be unreasonable. To avoid this difficulty, we consider alternative linear estimators, which ignore the valid signatures appearing more than two or three times in the sample,

$$\widehat{D}_2 = \frac{N(N-1)}{n(n-1)} f_2, \tag{7}$$

$$\widehat{D}_3 = \frac{N(N-1)}{n(n-1)} f_2 - \frac{N(N-1)(N-3n+4)}{n(n-1)(n-2)} f_3 \tag{8}$$

Goodman (1949) proposed $\widehat{D}_2$ for estimating the number of duplicates of classes in a finite population. The next estimator considered is used by the Washington Elections Division Office[1],

$$\widehat{D}_{2+} = \frac{N(N-1)}{n(n-1)} f_{2+} \text{ where } f_{2+} = \sum_{i=2}^{n} f_i. \tag{9}$$

Notice that f_{2+} is the number of electors in the sample with valid signatures appearing two or more times. A more intuitive estimator is one that replaces f_{2+} by the total number of duplicates in the sample

$$\widehat{D}_d = \frac{N(N-1)}{n(n-1)} d \quad \text{where} \quad d = \sum_{i=2}^{n} (i-1) f_i. \tag{10}$$

Note that if at most pairs of valid signatures occur in the petition ($F_j = 0$ for $j \geq 3$) then the estimators (7-10) are equal to the unbiased estimator, $\widehat{D}_{\text{unbias}}$. Similarly, $\widehat{D}_3 = \widehat{D}_{\text{unbias}}$ when at most triplicate valid signatures occur in the petition ($F_j = 0$ for $j \geq 4$).

When prior information is available, it may be possible to reduce the bias of the estimators by incorporating a bias adjustment factor (BAF), denoted as $B^{\widehat{D}}_{q,k,r}$, which is a function of q, k, and r, where

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

$q = n/N$ is the sampling fraction, $\boldsymbol{r} = (r_3, r_4, \dots, r_k)$ with $r_i = F_i/F_2$ for $i = 3, \dots, k$, and k is the maximum number of times any valid signature appears in the petition, $k = \max\{j : F_j > 0\}$. The BAF for $\widehat{D}$ is defined as,

$$B^{\widehat{D}}_{q,k,\boldsymbol{r}} = \frac{D_k}{\mathrm{E}(\widehat{D}|q,k)}$$

where $D_k = F_2 + 2F_3 + \cdots + (k-1)F_k$, and $\mathrm{E}(\widehat{D}|q,k)$ denotes the expectation of $\widehat{D}$ given q, and k. The BAF is approximated using binomial sampling (Smith-Cayama, 1999),

$$B^{\widehat{D}}_{q,k,\boldsymbol{r}} = \begin{cases} \dfrac{1+\sum\limits_{i=3}^{k}(i-1)r_i}{1+\frac{1}{2(1-q)^2}\sum\limits_{i=3}^{k} i(i-1)(1-q)^i r_i} & \text{for } \widehat{D} = \widehat{D}_2 \\ \dfrac{1+\sum\limits_{i=3}^{k}(i-1)r_i}{1+\frac{1}{q^2}\sum\limits_{i=3}^{k}(i+(1+(i-1)q)(1-q)^{i-1})r_i} & \text{for } \widehat{D} = \widehat{D}_{2+} \\ \dfrac{1+\sum\limits_{i=3}^{k}(i-1)r_i}{1+\frac{1}{q^2}\sum\limits_{i=3}^{k}(iq-1+(1-q)^i)r_i} & \text{for } \widehat{D} = \widehat{D}_d. \end{cases}$$

Then, the adjusted estimator of D is

$$\widehat{D}_{\text{adj}} = \widehat{D}\, B^{\widehat{D}}_{q,k,\boldsymbol{r}} \tag{11}$$

where $\widehat{D}$ is any of the biased estimators defined in Equations (7), (9), and (10). The binomial approximation give values of $\mathrm{E}(\widehat{D}|q,k)$ which are very similar to those obtained using the exact distribution, when N and n are large. The binomial sampling approximation was also used by Goodman (1949), and Haas and Stokes (1998). Observe that the population values k and $\boldsymbol{r}$ are unknown and need to be specified using prior information. In some states, including Washington, duplication data from previous fully verified petitions might be used.

3.2. Estimators for M

Estimators for M can be obtained by substituting in Equation (5) any of the estimators for D presented in Equations (6-11)

$$\widehat{M} = N - \widehat{U} - \widehat{D} \quad \text{with} \quad \widehat{D} = B\sum_{i=2}^{t} A_i f_i \tag{12}$$

for constants B, t, and A_i, with

$$B = \begin{cases} 1 & \text{for } \widehat{D} = \widehat{D}_{\text{unbias}}, \widehat{D}_3, \widehat{D}_2, \widehat{D}_{2+}, \widehat{D}_d \\ B^{\widehat{D}}_{q,k,\boldsymbol{r}} & \text{for the adjusted estimators.} \end{cases}$$

In petitions, the coefficient of variation for the number of times valid signatures appear in the petition is expected to be small. The square of this coefficient of variation is

$$\gamma^2 = \frac{(1/M)\sum\limits_{j=1}^{M}(N_j - \overline{N})^2}{\overline{N}^2} \quad \text{where } \overline{N} = \frac{1}{M}\sum_{j=1}^{M} N_j = \frac{N-U}{M}$$

A second-order jackknife estimator, $\widehat{M}_{\text{uj2}}$, was recommended by Haas and Stokes (1998) for applications where γ^2 is relatively small. The following estimator is a modification of the Haas and Stokes second-order jackknife estimator to accommodate the additional class of invalid signatures

$$\widehat{M}_{\text{uj2m}} = \left(1 - \frac{f_1(1-q^*)}{n^*}\right)^{-1}\left(\sum_{i=1}^{n^*} f_i - \frac{f_1(1-q^*)\ln(1-q^*)\widehat{\gamma}^2\left(\widehat{M}_{\text{uj1}}\right)}{q^*}\right)$$

where $n^* = n - u$ is the reduced sample size obtained after removing all invalid signatures, and $N^* = N - \widehat{U}$ is an unbiased estimator of the number of valid signatures in the petition and

$$\begin{aligned} q^* &= n^*/N^* \\ \widehat{M}_{\text{uj1}} &= \left(1 - \tfrac{(1-q^*)f_1}{n^*}\right)\sum_{i=1}^{n^*} f_i \\ \widehat{\gamma}^2(M) &= \max\left(0, \tfrac{M}{n^{*2}}\sum_{i=1}^{n^*} i(i-1)f_i + \tfrac{M}{N^*} - 1\right). \end{aligned}$$

3.3. Expectation and Variance of $\widehat{M}$

The expected value and variance for any estimator of the general form given in Equation (12) is obtained as (Smith-Cayama, 1999)

$$\mathrm{E}(\widehat{M}) = N - U - B\sum_{i=2}^{t} A_i \sum_{j=i}^{n} P_{ij}F_j \tag{13}$$

$$\mathrm{Var}(\widehat{M}) = \mathrm{Var}(\widehat{U}) + \mathrm{Var}(\widehat{D}) + 2\mathrm{Cov}(\widehat{U}, \widehat{D}) \tag{14}$$

where

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

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

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

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

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

4. Performance of the Estimators

In this section, the estimators for the number of distinct valid signatures, M, 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 M attains the required number then the measure is certified. Otherwise, complete verification of the petition is required.

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

	A (1984)		B (1995)		C (1989)		D (1996)	
N	162,324		231,723		173,858		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)
M (%)	138,631	(85.4)	179,794	(77.6)	132,795	(76.4)	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 1 describes the four petitions with regard to: petition size (N), numbers of invalid signatures (U), duplicates of valid signatures (D), distinct valid signatures (M), 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 numbersof 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.

Table 2 Expected frequency for replications of valid signatures, $\mathrm{E}(f_i)^1$.

Sampling Fraction	i	A	B	C	D
	2	3.93	4.22	9.15	10.91
3%	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	0.0792
	≥ 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] $\mathrm{E}(f_i) = \sum_{j=i}^{n} P_{ij}F_j$.

To calculate the bias adjustment factors, $B^{\widehat{D}}_{q,k,\boldsymbol{r}}$, 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, we apply a jackknife approach where for each petition, $i = \mathrm{A, B, C, D}$, information from only the remaining three petitions is used to specify values for the unknown k and $r_3, \ldots, r_k$. For each petition, the specified value, $\widetilde{k}$, was determined as the maximum of the observed k-values from the other three petitions. Similarly, the specified vector, $\widetilde{\boldsymbol{r}}$, 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.

Table 4 gives values for the bias adjustment factors, $B^{\widehat{D}}_{q,k,\widetilde{\boldsymbol{r}}}$, using $k = 3$ and 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 $\widetilde{\boldsymbol{r}} = \widetilde{r}_3$ and $\widetilde{\boldsymbol{r}} = (\widetilde{r}_3, \ldots, \widetilde{r}_{12})$ are similar in all cases. Therefore, we consider only bias adjustment based on triplicate valid signatures, $\widetilde{\boldsymbol{r}} = \widetilde{r}_3$, hereafter.

For each linear estimator, we use Equations (13) and (14) to compute the bias and root mean squared error (RMSE)

$$\mathrm{Bias}(\widehat{M}) = \mathrm{E}(\widehat{M}) - M \quad \text{and} \quad \mathrm{RMSE} = \sqrt{\mathrm{Var}(\widehat{M}) + \{\mathrm{Bias}(\widehat{M})\}^2}.$$

Table 3 True values of k and $r_3, \ldots, r_k$, and specified values $\widetilde{k}$ and $\widetilde{r_3}, \ldots, \widetilde{r_k}$.

	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 $\widetilde{\boldsymbol{r}} = (\widetilde{r_3}, \ldots, \widetilde{r_{12}})$ not displayed are equal to zero.

For the nonlinear estimator, $\widehat{M}_{\text{uj2m}}$, we estimate the bias and RMSE from 10,000 independent simulated random samples, drawn without replacement from each petition.

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

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

In Tables 5 and 6, the bias and RMSE are given for the nine estimators of M for Petitions A-D and sampling fractions: 3%, 5%, 10%, and 20%. In these tables, the bias and RMSE are expressed as a percentage of the true number of distinct valid signatures. For the adjusted estimator, Equation (11), $k = 3$ ($\boldsymbol{r} = \widetilde{r_3}$) is used for the bias adjusted factor, $B^{\widehat{D}}_{q,k,\boldsymbol{r}}$.

In Table 5, the estimator $\widehat{M}_3$ tends to have a relatively small positive bias ($< 0.07\%$) in all cases. The biases of $\widehat{M}_2$, $\widehat{M}_{2+}$, and $\widehat{M}_d$ are negative in all cases, corresponding to positive biases in the estimators for the number of duplicates of valid signatures $\widehat{D}_2$, $\widehat{D}_{2+}$, and $\widehat{D}_d$. Note that the difference between the bias of these estimators tend 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 three adjusted estimators show a small reduction in the absolute bias when compared with their non-adjusted counterparts. The non-linear estimator, $\widehat{M}_{\text{uj2m}}$, tends to have a relatively large negative bias ranging from -4.52% to -1.19%.

Table 5 BIAS (%) of estimators for M.

$\frac{n}{N}$	Estimator	A	B	C	D
3%	$\widehat{M}_{\text{unbias}}$	0	0	0	0
	$\widehat{M}_3$	0.00	0.07	0.02	0.03
	$\widehat{M}_2$	-0.08	-0.08	-0.32	-0.29
	$\widehat{M}_{2+}$	-0.08	-0.08	-0.34	-0.30
	$\widehat{M}_d$	-0.08	-0.09	-0.35	-0.32
	$\widehat{M}_{2\text{adj}}$	0.02	0.01	-0.12	-0.12
	$\widehat{M}_{2+\text{adj}}$	0.02	0.00	-0.13	-0.13
	$\widehat{M}_{d\text{adj}}$	0.02	0.00	-0.13	-0.13
	$\widehat{M}_{\text{uj2m}}$	-1.83	-1.54	-4.51	-4.14
5%	$\widehat{M}_{\text{unbias}}$	0	0	0	0
	$\widehat{M}_3$	0.00	0.05	0.02	0.02
	$\widehat{M}_2$	-0.07	-0.07	-0.30	-0.27
	$\widehat{M}_{2+}$	-0.08	-0.08	-0.32	-0.29
	$\widehat{M}_d$	-0.08	-0.08	-0.34	-0.31
	$\widehat{M}_{2\text{adj}}$	0.02	0.01	-0.11	-0.11
	$\widehat{M}_{2+\text{adj}}$	0.02	0.01	-0.12	-0.12
	$\widehat{M}_{d\text{adj}}$	0.02	0.00	-0.13	-0.13
	$\widehat{M}_{\text{uj2m}}$	-1.88	-1.57	-4.52	-4.17
10%	$\widehat{M}_{\text{unbias}}$	0	0	0	0
	$\widehat{M}_3$	0.00	0.03	0.02	0.02
	$\widehat{M}_2$	-0.06	-0.05	-0.24	-0.22
	$\widehat{M}_{2+}$	-0.07	-0.06	-0.28	-0.26
	$\widehat{M}_d$	-0.08	-0.08	-0.32	-0.29
	$\widehat{M}_{2\text{adj}}$	0.02	0.01	-0.09	-0.09
	$\widehat{M}_{2+\text{adj}}$	0.02	0.01	-0.11	-0.11
	$\widehat{M}_{d\text{adj}}$	0.02	0.01	-0.12	-0.12
	$\widehat{M}_{\text{uj2m}}$	-1.69	-1.42	-4.21	--3.88
20%	$\widehat{M}_{\text{unbias}}$	0	0	0	0
	$\widehat{M}_3$	0.00	0.01	0.01	0.01
	$\widehat{M}_2$	-0.03	-0.02	-0.13	-0.12
	$\widehat{M}_{2+}$	-0.05	-0.04	-0.21	-0.19
	$\widehat{M}_d$	-0.07	-0.06	-0.28	-0.26
	$\widehat{M}_{2\text{adj}}$	0.01	0.02	-0.05	-0.05
	$\widehat{M}_{2+\text{adj}}$	0.01	0.01	-0.08	-0.08
	$\widehat{M}_{d\text{adj}}$	0.01	0.01	-0.11	-0.11
	$\widehat{M}_{\text{uj2m}}$	-1.43	-1.19	-3.63	-3.33

From Table 6, it can be seen that the RMSE decreases at a faster rate than $1/\sqrt{n}$ for all estimators and petitions. This results from corresponding property of the estimators for D in Equation (12). The estimator $\widehat{M}_3$ has smaller RMSE than $\widehat{M}_{\text{unbias}}$, except for the 20%

Table 6 RMSE (%) of estimators for M.

$\frac{n}{N}$	Estimator	A	B	C	D
3%	$\widehat{M}_{\text{unbias}}$	2.49	638,455	5.65	27.20
	$\widehat{M}_3$	2.16	2.15	3.93	3.19
	$\widehat{M}_2$	1.66	1.40	2.61	2.08
	$\widehat{M}_{2+}$	1.66	1.40	2.61	2.09
	$\widehat{M}_d$	1.67	1.40	2.62	2.09
	$\widehat{M}_{\text{2adj}}$	1.62	1.36	2.53	2.02
	$\widehat{M}_{\text{2+adj}}$	1.62	1.36	2.53	2.02
	$\widehat{M}_{\text{dadj}}$	1.62	1.36	2.53	2.02
	$\widehat{M}_{\text{uj2m}}$	2.84	2.39	5.47	4.87
5%	$\widehat{M}_{\text{unbias}}$	1.26	26,247	2.43	5.83
	$\widehat{M}_3$	1.19	1.11	2.03	1.64
	$\widehat{M}_2$	1.03	0.88	1.60	1.28
	$\widehat{M}_{2+}$	1.03	0.88	1.60	1.29
	$\widehat{M}_d$	1.03	0.89	1.61	1.29
	$\widehat{M}_{\text{2adj}}$	1.00	0.86	1.54	1.23
	$\widehat{M}_{\text{2+adj}}$	1.00	0.86	1.54	1.23
	$\widehat{M}_{\text{dadj}}$	1.00	0.86	1.54	1.23
	$\widehat{M}_{\text{uj2m}}$	2.42	2.03	5.06)	4.57
10%	$\widehat{M}_{\text{unbias}}$	0.57	296	0.93	0.95
	$\widehat{M}_3$	0.57	0.52	0.89	0.71
	$\widehat{M}_2$	0.54	0.49	0.84	0.68
	$\widehat{M}_{2+}$	0.54	0.49	0.85	0.69
	$\widehat{M}_d$	0.55	0.49	0.87	0.71
	$\widehat{M}_{\text{2adj}}$	0.53	0.48	0.80	0.64
	$\widehat{M}_{\text{2+adj}}$	0.53	0.48	0.80	0.64
	$\widehat{M}_{\text{dadj}}$	0.53	0.48	0.80	0.64
	$\widehat{M}_{\text{uj2m}}$	2.02	1.71	4.56	4.14
20%	$\widehat{M}_{\text{unbias}}$	0.29	2.30	0.42	0.34
	$\widehat{M}_3$	0.29	0.28	0.42	0.34
	$\widehat{M}_2$	0.29	0.28	0.44	0.35
	$\widehat{M}_{2+}$	0.29	0.28	0.46	0.38
	$\widehat{M}_d$	0.30	0.28	0.51	0.42
	$\widehat{M}_{\text{2adj}}$	0.29	0.27	0.41	0.33
	$\widehat{M}_{\text{2+adj}}$	0.29	0.27	0.42	0.34
	$\widehat{M}_{\text{dadj}}$	0.29	0.27	0.43	0.35
	$\widehat{M}_{\text{uj2m}}$	1.70	1.43	3.92	3.56

sampling fraction for petition A where the RMSE's are equal. The estimator $\widehat{M}_2$ has smaller RMSE than $\widehat{M}_3$, except for the 20% sampling fraction for petitions C and D. The estimators $\widehat{M}_2$, $\widehat{M}_{2+}$, and $\widehat{M}_d$ tend to have similar RMSE's for the sample 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 2.2). For the 20% sampling fraction, the RMSE for $\widehat{M}_d$ is slightly larger than the RMSE's for $\widehat{M}_2$ and $\widehat{M}_{2+}$ for petitions B and C, and similar for petitions A and B. The adjusted estimators $\widehat{M}_{\text{2adj}}$, $\widehat{M}_{\text{2+adj}}$, and $\widehat{M}_{\text{dadj}}$ show a slight reduction in the RMSE compared to their non-adjusted counterparts. These three adjusted estimators have similar RMSE's in all cases. The RMSE for the non-linear estimator $\widehat{M}_{\text{uj2m}}$ is relatively large in all cases.

5. Summary

In this paper we 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 non-linear estimator, $\widehat{M}_{\text{uj2m}}$, adapted from Haas and Stokes (1998).

Small sampling fractions less 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 $\widehat{M}_2$, 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 duplicate replication data is available from similar fully-verified petitions, it is possible to reduce the bias of the (biased) linear estimators.

References

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

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.

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

Smith-Cayama, R. A. (1999), Statistical Estimation for Initiative Petitions and Performance of the Decision Rule for Oregon State Petitions, unpublished Ph.D. dissertation, Oregon State University.