Journal of Official Statistics, Vol. 21, No. 1, 2005, pp. 21–38

The Accuracy of Estimators of Number of Signatories to a Petition Based on a Sample

Duncan I. Hedderley[1] *and Stephen J. Haslett*[2]

Petitions are a relatively widespread international phenomenon. In some countries, including New Zealand, and in several U.S. states, they have legal status and there is legislation which obligates the legislature to react to petitions which have widespread popular support. Normal practice at present is to check a sample of the signatures and from that estimate the number of eligible electors who have signed a petition, making allowance for signatories who are not eligible and multiple signatures from eligible electors.

The problem is related to a number of others, e.g., number of species in an ecosystem, but was found via a simulation study to be sufficiently different that a universally best estimator does not exist. The simulation drew samples from artificial petitions with known distributions of multiple signatures to assess the performance of several estimators described in the literature. The effect of sampling fraction on bias, distribution, variability and estimated variance of the estimators was also investigated. Bias adjustment factors previously proposed in the literature were investigated and found not to be particularly useful.

Ineligible and duplicate signatures often occur in the same petition. Extending the simulation to include ineligible signatures showed that estimating their number added to the variability of the overall estimate of number of eligible signatories. Although the estimated number of multiple signatures and the estimated number of ineligible signatures are correlated, the simulations suggest the correlation is small and can generally be ignored.

Key words: Biased estimators; number of classes; population size; referendum; duplicate signatures; multiple signatures.

1. Background

A number of countries, including New Zealand, and U.S. states including Washington, Oregon, and California have legislation which obliges the legislature to react to petitions which have widespread popular support. The New Zealand Citizens' Initiated Referenda Act, 1993, states that if a petition presented to the Clerk of the House of Representatives has been signed by at least 10 percent of registered voters, then the House of Representatives is required to hold an indicative referendum on the petition.

[1] New Zealand Institute of Crop and Food Research Ltd, Private Bag 11-600, Palmerston North, New Zealand. Email: hedderleyd@crop.cri.nz

[2] Statistics Research and Consulting Centre, Massey University, Private Bag 11-222, Palmerston North, New Zealand. Email: s.j.haslett@massey.ac.nz

Acknowledgements: This research was carried out as part of the first author's MSc thesis, and was funded by Massey University's staff study provision and the Clerk of the New Zealand House of Representatives through Statistics New Zealand. We would like to thank Mike Doherty at Statistics New Zealand for initially raising the problem, sharing his thoughts and experience, and explaining the practical constraints the process operates under. We would also like to thank JOS reviewers whose comments led to clarification to a number of issues in this article.

Clearly, it is important to establish reliably the number of electors who have signed a petition. The task of checking the number of signatories is substantial: the petition is bound to be large (approximately 250,000 electors' signatures are needed to trigger a referendum), and checking for multiple signatures requires some effort. Because of this, normal practice at present is to take a sample (between 8 and 10 percent) of the signatures and check them for eligibility and multiple signatures, and to estimate the number of eligible signatories based on this. In New Zealand, this last task falls to the Government Statistician.

For example, a recent petition on tougher sentencing for violent criminals had 252,336 signatures. A sample of 28,704 (11%) was taken; of these, 4,454 could not be confirmed as registered voters, 23,842 were those of registered voters who appeared once in the sample, 402 were those of registered voters who appeared twice in the sample (i.e., 201 people), and six were those of registered voters who appeared three times in the sample (i.e., two people). This was the first petition in recent years where the sample contained triple signatures.

The problem of estimating the number of individuals in a population based on a sample arises in a wide range of contexts from archaeology to database management and ecology, and a wide variety of solutions have been suggested. Bunge and Fitzpatrick (1993) review the problem and various proposed solutions.

More recently, Haas and Stokes (HS) (1998) used simulation to test the performance of a number of estimators on a range of problems; Brutlag and Richardson (2002) looked at estimators used in the database field, and cite an interesting result on the distributions for which specific estimators will perform best; and Smith-Cayama and Thomas (SCT) (1999) specifically consider the problem of estimating the number of eligible signatories to a petition.

This article starts with a review of the estimators, describing them using a common notation for clarity; it then reports the results of two simulation studies based on samples from three synthetically generated petitions. The first investigates the performance of the estimators, the method for estimating variance proposed by HS, and the bias adjustments proposed by SCT. The second simulation study investigates how estimating the number of ineligible signatures (i.e., those which cannot be confirmed as belonging to registered electors) in the petition affects the overall variability of the estimate of the number of signatories.

2. The Problem, Formally

Following HS notation, the problem can be stated as:

We have a population of size N, whose members can each be classified as falling into one (and only one) of D classes. In this application, each person constitutes a single class. These classes are labelled C_j ($1 \leq j \leq D$), and the jth, class has N_j members (i.e., N_j signatures) in the population. Because the classes are disjoint,

$$\sum_{j=1}^{D} N_j = N$$

A simple random sample of size n is drawn without replacement from the population. This sample contains n_j members of C_j. The problem is to estimate the value of D, given the $\{n_j\}$ and knowledge of N.

In the petition problem, the sizes of the individual classes are not important; we are more concerned with how many classes of a given size there are (the "frequency of frequencies"). The number of classes of size i in the population will be written F_i; this means that $\Sigma_{i=1}^{N} F_i = D$, and $\Sigma_{i=1}^{N} iF_i = N$.

Similarly, the number of classes appearing exactly i times in the sample is written f_i, and the total number of classes in the sample is written d. This means that $\Sigma_{i=1}^{n} f_i = d$ and $\Sigma_{i=1}^{n} if_i = n$.

Because we have sampled without replacement, the probability of the sample consisting of a particular vector $(n_1, n_2,. . ., n_D)$ is multivariate hypergeometric:

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

Obviously, $(n_1,\ n_2,. . ., n_D)$ is unobservable; we know the values of the $n_j \geq 1$, but knowing how many $n_j = 0$ would be equivalent to knowing D. All we can observe is the vector $(f_1, f_2,. . ., f_n)$. The probability mass function of $(f_1, f_2,. . ., f_n)$ is the sum of the $p((n_1,\ n_2,. . ., n_D)|D,\ (N_1,\ N_2,. . ., N_D))$ over all combinations of $(n_1,\ n_2,. . ., n_D)$ which correspond to $(f_1, f_2,. . ., f_n)$; in other words, those combinations which have exactly $(D\text{-}d)$ of the n_i's equal to 0, f_1 of the n_i's equal to 1, f_2 of the n_i's equal to 2, etc.

2.1. Goodman's estimator

Most of the estimators discussed by SCT are variants of Goodman's (1949) estimator, which they state as:

$$D_{Goodman} = N - \sum_{i=2}^{n} \frac{c_i}{p_{ii}} f_i$$

where

$$p_{ij} = \frac{\binom{j}{i}\binom{N-j}{n-i}}{\binom{N}{n}} = p\begin{pmatrix} \textit{a sample of } n \textit{ from } N \textit{ will contain } i \textit{ members} \\ \textit{of a class with a total of } j \textit{ members} \end{pmatrix}$$

and

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

If the sample is larger than the largest class in the population (i.e., $n >$ max $\{N_1,\ N_2,. . ., N_D\}$), this is the *only* unbiased estimator of D. If $n <$ max $\{N_1, N_2,. . ., N_D\}$, no unbiased estimator exists. Unfortunately its variance can be very large. Because the weight given to f_i is approximately $\left(\frac{N-n}{n}\right)^i$ (see Kish 1965), the size of the estimate can be very heavily influenced by the high-i classes (i.e., multiple signatures) in the sample. When the sampling fraction is small, the chances of a class appearing several times in the sample are low, so the number of classes appearing several times can be subject to proportionally

quite large sampling variation, and the large weights mean this variability can have a considerable influence on the estimate. Goodman was aware of this problem and suggested several alternative estimators which were not as susceptible to it but not (generally) unbiased. The simplest

$$D_{Goodman2} = N - \frac{N(N-1)}{n(n-1)} f_2$$

is D_{Goodman} with the very large weights removed, just leaving the term for f_2. SCT also suggest a D_{Goodman3}

$$D_{Goodman3} = N - \frac{N(N-1)}{n(n-1)} f_2 + \frac{N(N-1)(N-3n+4)}{n(n-1)(n-2)} f_3$$

the first three terms of D_{Goodman}.

HS observe that estimators like these will not work well when the number of classes in the population, D, is small and so the sample contains comparatively few single and duplicate observations, but more higher multiples. However, one would expect large public petitions to have large D; samples from recent CIR petitions in New Zealand have consisted overwhelmingly of single and duplicate signatures, with only a handful of triplicate signatures in one petition.

SCT mention two other variations on Goodman's estimator. One, used by the state of Washington Elections Division Office, is simply D_{Goodman2} with f_2 replaced by the total number of people who have multiple signatures in the sample

$$D_{Goodman2+} = N - \frac{N(N-1)}{n(n-1)} \sum_{i=2}^{n} f_i$$

The second is an extension of this, replacing the number of people who have signed multiple times, $\Sigma_{i=2}^{n} f_i$, with the number of duplicate signatures, $\Sigma_{i=2}^{n}(i-1)f_i$ (by "duplicate" they mean any signatures beyond that person's first; so a person who has signed twice has one valid signature and one duplicate; a person who has signed 17 times has one valid signature and 16 duplicates)

$$D_{GoodmanDup} = N - \frac{N(N-1)}{n(n-1)} \sum_{i=2}^{n} (i-1)\ f_i$$

One point to note is that if the sample contains only single and double signatures (which has been the case with most recent CIR petitions), then Goodman's estimator and all its variants are equivalent.

2.2. *Haas and Stokes' estimators*

One way to improve biased estimators is to apply bias-reduction techniques to them. HS use two jackknife approaches to bias reduction (the generalized jackknife, and Horvitz-Thompson jackknife estimators) to develop a range of estimators.

To test the various estimators, HS created a number of data sets and ran simulation studies, drawing samples of between 5% and 20% of the observations. They group the data sets by γ^2, the coefficient of variation of the class sizes $N_1, N_2, \ldots N_D$. Judging by the fully

enumerated Washington state petitions cited in SCT, $\gamma^2 \ll 1$. The estimator which had the lowest RMSE under those conditions was

$$D_{uj2} = \left(1 - \frac{f_1(1-q)}{n}\right)^{-1} \left(d - \frac{f_1(1-q)\ln(1-q)\gamma^2(D_{uj1})}{q}\right)$$

where $q = \dfrac{n}{N}$, the sampling fraction,

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

and

$$D_{uj1} = d\left(1 - \frac{(1-q)\ f_i}{n}\right)^{-1}, \text{ an initial estimate of } D$$

2.3. Variance of the estimates

HS also present a way of estimating the asymptotic variance of an estimator, which is a function of the frequency of frequencies and the population size, and which is continuously differentiable. Using the delta method, the general form is:

$$\textit{Asymptotic Var } [\hat{D}(f,N)] \approx \sum_{i=1}^{M} A_i^2 var[f_i] + \sum_{i=1}^{M}\sum_{\substack{j=1 \\ j\neq i}}^{M} A_i A_j cov[f_i, f_j]$$

where $\hat{D}$ is the estimator

N is the size of the population
A_i is the partial derivative of $\hat{\mathrm{D}}$ with respect to f_i
and $M = \max(N_1, N_2, \ldots, N_D)$, i.e., the size of the largest class in the population.

They derive approximate values for $var(f_i)$ and $cov(f_i, f_j)$ by assuming that all classes are of equal size (N/D). In this case the frequency of frequencies is approximately multinomial, and so:

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

and

$$\hat{cov}[f_i, f_j] = -\frac{f_i f_j}{\hat{D}}$$

2.4. The structure of the petition

The performance of an estimator, in particular its sampling variability, is likely to be influenced by the distribution of the number of times each individual appears in the population, i.e., what proportion of the people who have signed the petition have signed only once, what proportion have signed the petition twice, three times, etc. This may

explain why the estimator that HS found performed best (over a wide range of data sets of different sizes with different distributions) performed so poorly in SCT's study (which was based on data from fully enumerated petitions from Washington state), and why initial tests of several estimators used in the database field (Jeffries's, Chao and Lee's, and Shlosser's; see Brutlag and Richardson 2002) found that they were either badly biased or far more variable than those proposed by SCT.

Unfortunately, it is not possible to conclude much about the proportions of single, duplicate, triplicate, etc signatures in petitions solely on the basis of an 8% – 10% sample; there are simply too many unknowns. The sample contains f_i signatures which appear i times in the sample (in New Zealand petitions, signatures typically appear only once or twice in the sample; the "tougher sentencing" petition was unusual in having signatures which appear three times in the sample); the petition contains F_j signatures which appear j times. Obviously, $\max\{j|F_j > 0\} \geq \max\{i|f_i > 0\}$, but we do not know how much larger. Presumably, if one assumes a suitable tightly-defined structure for $\{F_j\}$ and a value for $\max\{j|F_j > 0\}$, the $\{F_j\}$ most likely to generate the actual sample could be determined; but such assumptions are difficult if not impossible to check in practice.

None of the New Zealand Citizens' Initiated Referenda petitions have been completely checked for the number of single, duplicate, triplicate etc signatures, so there is no firm information on the structure of New Zealand petitions. However, SCT present data on four petitions from Washington which were completely checked, because it was not possible to conclusively decide that they were above or below the threshold on the basis of a sample. In all four, the distribution of the number of signings is roughly geometric. This is in contrast to many other applications, such as estimating the number of species, where the distributions tend to have much longer tails.

3. First Simulation Study

To check the performance of the estimators described by SCT, and HS's D_{uj2}, computer simulation was used to generate samples from six petitions with a known number of signatories. The petitions either had 100,000 or 250,000 signatories, and one of three "structures" (proportions of duplicate, triplicate etc signatures). These were:

- One with a genuine geometric distribution, with an r of 0.95 (i.e., 95% of the signatories had signed once, 95% of the signatories who had signed more than once had signed twice, 95% of the signatories who had signed more than twice had signed three times, and so on; in the four petitions cited by SCT, r ranged from 93% to 98%). This meant individuals signed up to five times. With 100,000 signatories the petition had 105,264 signatures; with 250,000 signatories, the petition had 263,157 signatures.
- One with fewer multiple signatures than the geometric, 95% of the signatories had signed once, the remaining 5% had signed twice. With 100,000 signatories this meant the petition had 105,000 signatures; with 250,000 signatories the petition had 262,500 signatures.
- One with more multiple signatures than the geometric: 95% of the signatories had signed once and the remaining 5% were equally likely to have signed twice, three times, four times or five times. With 100,000 signatures this meant the petition had 112,500 signatures; with 250,000 signatories, the petition had 281,250 signatures.

The computer was used to repeatedly draw random samples; on the basis of each sample, the number of signatories was estimated using Goodman's estimator, D_2, D_3, D_{2+}, D_d, and D_{uj2}. The program (written in SAS) drew 500 samples at sampling fractions of 5%, 10% and 20% (for the 100,000-signature petition) and 5%, 8% and 10% (for the 250,000-signature petition). These levels were chosen so that two matched for the different petition sizes, while the third ensured that it was possible to make comparison between samples of approximately the same size. The code is available from the authors on request.

3.1. Results

3.1.1. Point estimates

Table 1 shows summaries of the samples drawn. They indicate that, given a set of samples from one of these petitions, it should be reasonably easy to decide which petition they are from, largely on the proportion of samples with triplicate signatures. Unfortunately, in practice, only one sample is taken and checked; in which case, even deciding between three alternatives is difficult, let alone the much wider range of possible structures.

Table 2a and 2b summarises the estimates from the petitions. Clearly, D_{uj2} performs much worse than the other estimators, having substantial bias and also greater sampling variability. Despite the warnings about the variability of Goodman's estimator, it produced a very similar set of estimates to the modified Goodman's estimators (95,000 to 108,000 and 241,000 to 266,000 from the 5% sampling fractions; 94,000 to 103,000 and 245,000 to 258,000 from the 10% samples) for the geometric petition. The estimates were identical for the singles-and-doubles petition. The 5% sample from the uniform petition did produce the occasional absurd estimate (the lowest was $-166{,}200$ for the 100,000 signatory petition and $-26{,}500$ for the 250,000 signatory petition; the highest 139,000 and 2,756,000 respectively); the ranges with the 8%, 10%, and 20% samples were large but not as extreme (79,000 to 164,000 and 210,000 to 354,000 for the 10% samples). For comparison, D_2, D_{2+} and D_{dup} gave much narrower

Table 1. Number of samples containing duplicate, triplicate, etc signatures

Petition	Geometric			Singles-and-doubles			Uniform		
100,000 Signatories									
Sampling fraction	5%	10%	20%	5%	10%	20%	5%	10%	20%
Number of samples with duplicates	500	500	500	500	500	500	500	500	500
Triplicates	16	115	453	0	0	0	452	500	500
Quadruples	0	1	13	0	0	0	19	250	500
Quintuples	0	0	0	0	0	0	0	3	179
250,000 Signatories									
Sampling fraction	5%	8%	10%	5%	8%	10%	5%	8%	10%
Number of samples with Duplicates	500	500	500	500	500	500	500	500	500
Triplicates	47	148	274	0	0	0	497	500	500
Quadruples	0	0	2	0	0	0	57	252	405
Quintuples	0	0	0	0	0	0	3	2	19

*Table 2a. Summary of estimates for first simulation study: bias**

	Goodman	*D*2	*D*3	*D*2 +	*Ddup*	*Duj*2
100,000 Signatories						
Geometric petition						
5% Sample	−10.8	207.0	−10.8	219.8	232.6	4,203.2
10% Sample	34.0	197.3	20.8	222.7	248.3	3,836.8
20% Sample	5.6	117.7	−1.6	177.9	238.8	2,970.5
Singles-and-doubles petition						
5% Sample	67.1	**	**	**	**	3,600.8
10% Sample	−7.2	**	**	**	**	3,367.8
20% Sample	14.5	**	**	**	**	2,697.8
Uniform petition						
5% Sample	386.2	9,656.5	−4,835.7	10,524.7	11,408.8	18,575.5
10% Sample	218.4	7,349.2	−3,812.3	9,010.3	10,739.1	15,669.6
20% Sample	3.9	2,953.5	−2,351.8	5,870.2	9,062.0	9,751.0
250,000 Signatories						
Geometric petition						
5% Sample	−107.3	559.4	−107.3	598.6	637.8	10,661.9
8% Sample	−76.3	484.9	−76.3	544.0	603.1	9,974.4
10% Sample	−112.4	469.6	−86.3	549.4	629.6	9,569.9
Singles-and-doubles petition						
5% Sample	182.5	**	**	**	**	9,634.1
8% Sample	−32.5	**	**	**	**	8,987.7
10% Sample	−30.4	**	**	**	**	8,415.7
Uniform petition						
5% Sample	−11,128.7	24,272.1	−11,916.2	26,450.8	28,682.3	46,692.4
8% Sample	1,730.6	20,490.6	−10,387.4	23,856.4	27,338.4	42,113.8
10% Sample	7.5	18,274.8	−9,039.8	22,355.3	26,618.9	39,049.2

*Bias = True number of signatories − mean estimate

***D*2, *D*3 *D*2 + and *Ddup* are equivalent to Goodman for singles-and-doubles petition

ranges of estimates, but were heavily biased; the D_3 estimates showed less bias than the D_2 estimates, and a narrower range than the unmodified Goodman estimates (93,000 to 114,000 and 245,000 to 276,000 with the 10% sample).

As for the other estimators, they perform much better with the geometric and singles-and-doubles petitions than for the uniform one; and for those petitions, increasing the sampling fraction has a substantial effect on the RMSE. Since the bias is not consistently reduced, the improvement must reflect considerable reduction in the sampling variability, in most cases much more than would be expected simply on the basis of the increase in the sample size (see Figure 1); presumably, as the sample size increases, the composition of the sample becomes more stable.

The distribution of the estimates is roughly normal in most cases; however, Goodman's estimator with the uniform petition produces a distribution which tends to have wider tails than would be expected of a normal distribution. Even with this distribution, the 95% confidence interval calculated assuming the distribution of the estimates is normal and the empirical 95% confidence interval give results which are close (in the worst case, with the 5%

Table 2b. Summary of estimates for first simulation study: RMSE

	Goodman	*D*2	*D*3	*D*2 +	*Ddup*	*Duj*2
100,000 Signatories						
Geometric petition						
5% Sample	1,853.3	1,455.7	1,853.3	1,460.9	1,469.5	4,784.2
10% Sample	852.9	769.6	812.6	778.7	791.6	4,003.8
20% Sample	371.5	377.5	367.7	401.8	437.1	3,015.3
Singles-and-doubles petition						
5% Sample	1,347.2	**	**	**	**	4,201.2
10% Sample	671.4	**	**	**	**	3,526.6
20% Sample	348.7	**	**	**	**	2,744.9
Uniform petition						
5% Sample	27,974.8	10,063.3	10,838.3	10,913.2	11,808.8	18,806.6
10% Sample	7,575.9	7,473.2	4,884.4	9,117.0	10,848.8	15,741.4
20% Sample	1,288.3	3,003.1	2,477.4	5,897.8	9,090.5	9,775.0
250,000 Signatories						
Geometric petition						
5% Sample	3,268.5	2,363.9	3,268.5	2,365.1	2,373.5	11,266.3
8% Sample	1,732.8	1,501.5	1,732.8	1,523.1	1,553.2	10,217.5
10% Sample	1,365.5	1,237.0	1,286.5	1,274.6	1,322.1	9,730.4
Singles-and-doubles petition						
5% Sample	2,224.7	**	**	**	**	10,274.4
8% Sample	1,399.9	**	**	**	**	9,255.5
10% Sample	1,073.6	**	**	**	**	8,580.5
Uniform petition						
5% Sample	195,563.4	24,712.5	20,188.1	26,864.8	29,104.6	46,938.9
8% Sample	21,818.8	20,685.8	12,682.7	24,033.6	27,521.1	42,224.9
10% Sample	16,440.0	18,393.7	10,372.1	22,459.2	26,729.7	39,117.7

**$D2$, $D3$ $D2+$ and *Ddup* are equivalent to Goodman for singles-and-doubles petition

sample, the normal theory interval is − 121,937.3 to 644,195.3, while the empirical interval, based on percentiles, is − 121,547.4 to 643,804.8).

3.1.2. Estimates of variability

Similarly, one can compare HS's variance estimates with the actual variability of the estimators (Tables 3a and 3b). Given the poor performance of D_{uj2}, estimated standard errors were not calculated for it.

Figure 2 shows the mean of the estimated standard errors and the actual variability for Goodman's estimator, D_2 and D_3 on each petition based on sampling fractions of between 5% and 20% (for the 100,000-signatory petition) and 5% to 15% (for the 250,000-signatory petition). D_{2+} and D_{dup} performed similarly to D_2, and so were excluded for clarity.

For the uniform petition, the estimates appear reasonable; for Goodman's estimator nominal 95% confidence intervals based on them have a coverage of 89% to 95%. The bias of D_2, D_{2+} and D_{dup} meant coverage ranged from 0% to 7%. For D_3, coverage ranges from 30% to 94%.

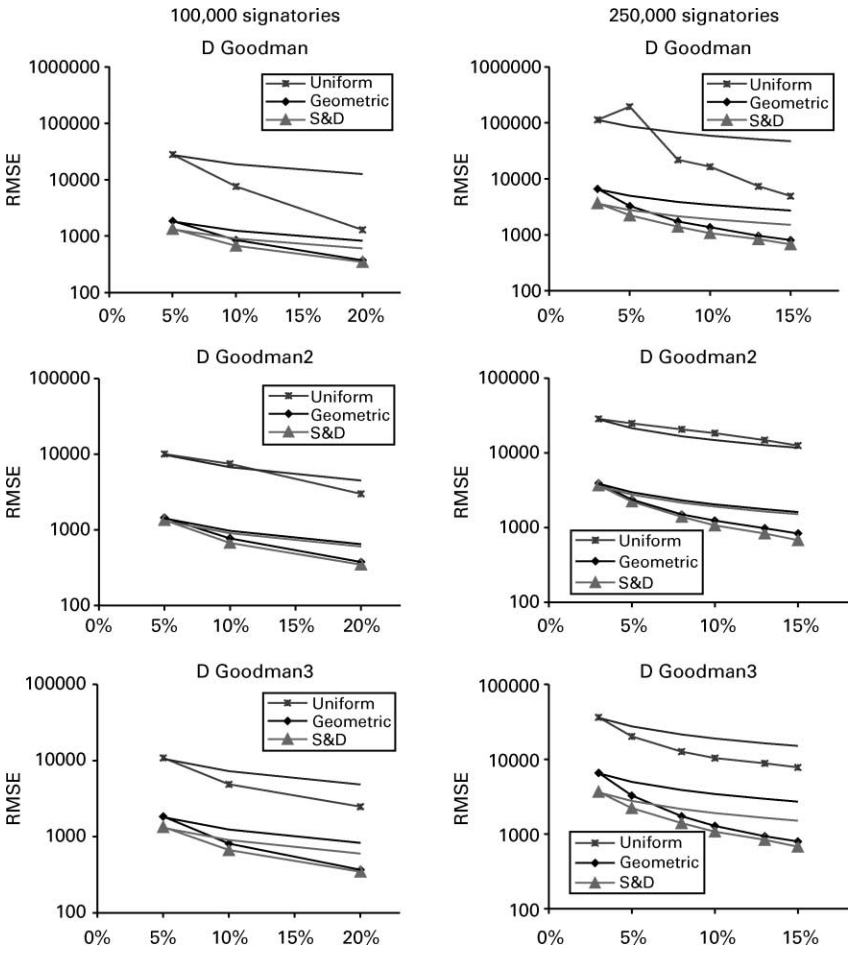

Fig. 1. Plots of RMSE against sampling fraction (with reference lines indicating $1/\sqrt{n}$)

For the geometric and singles-and-doubles petitions, the estimated standard errors are slightly too high, the overestimation being more marked with the 250,000-signatory petitions and with higher sampling fractions. Nominal 95% confidence intervals for these petitions using the estimated standard errors have 99% + coverage.

3.2. Bias adjustment strategies

SCT propose the use of Bias Adjustment Factors (BAFs) to improve the biased estimators such as D_2, D_3, D_{2+} and D_{dup}. These factors are calculated assuming a distribution for the data (SCT use the data from the four fully-enumerated Washington petitions to produce estimates for each BAF) and then applied to the estimate. The result, assuming that the distribution used to derive the BAF is similar to the distribution of the actual signatures in the petition, should be to make the mean adjusted estimate equal to the true number of eligible signatories to the petition.

As can be seen from comparing the RMSE and SD around own mean values in Table 3 (which differ only by the bias squared), bias is not a particularly large component of the error

Table 3a. Summary of estimates for first simulation study: SD about own mean

	Goodman	*D*2	*D*3	*D*2 +	*Ddup*
100,000 Signatories					
Geometric petition					
5% Sample	1,853.3	1,440.9	1,853.3	1,444.2	1,451.0
10% Sample	852.3	743.9	812.3	746.2	751.7
20% Sample	371.4	358.7	367.7	360.3	366.1
Singles-and-doubles petition					
5% Sample	1,345.6	**	**	**	
10% Sample	671.4	**	**	**	
20% Sample	348.4	**	**	**	
Uniform petition					
5% Sample	27,972.1	2,832.2	9,699.7	2,886.1	3,047.4
10% Sample	7,572.8	1,355.6	3,053.5	1,390.5	1,538.8
20% Sample	1,288.3	543.5	779.0	570.2	719.6
250,000 Signatories					
Geometric petition					
5% Sample	3,270.0	2,299.0	3,270.0	2,290.4	2,288.5
8% Sample	1,732.8	1,422.4	1,732.8	1,424.0	1,432.8
10% Sample	1,362.3	1,145.5	1,284.8	1,151.3	1,163.7
Singles-and-doubles petition					
5% Sample	2,219.4	**	**	**	**
8% Sample	1,399.5	**	**	**	**
10% Sample	1,073.2	**	**	**	**
Uniform Petition					
5% Sample	195,442.0	4,649.2	16,312.4	4,703.1	4,944.7
8% Sample	21,771.8	2,837.9	7,284.2	2,915.5	3,168.3
10% Sample	16,456.5	2,090.0	5,090.7	2,159.7	2,434.5

**$D2$, $D3$ $D2+$ and *Ddup* are equivalent to Goodman for singles-and-doubles petition

for the geometric and singles-and-doubles petitions; thus reducing the bias does not necessarily reduce the variability of the estimate much. Bias is a large component in the RMSE of the estimators with the uniform petition, although even if it could be completely eliminated, the variability of the estimators about their own mean is larger than for the geometric and singles-and-doubles petitions. To eliminate the bias, it would be necessary to have a good guess-timate of the true distribution of the numbers of signatures in the petition. Often this is not available, and any assumption could be difficult to justify in a legal context.

The results of estimating BAFs based on the four fully-enumerated Washington petitions and then applying them to the results from the simulation study are shown in Table 4. Although the bias adjustment reduces the sampling variation, the reduction is not large. For the geometric and uniform distributions, bias is also reduced, producing a 9%–11% reduction in the RMSE for the geometric petition, and a 7%–8% reduction in the RMSE for the uniform petition. The bias-adjusted results for the singles-and-doubles petition are slightly more biased than the original results (because the bias adjustment factors were calculated from

Table 3b. Summary of estimates for first simulation study: Mean estimated SE

	Goodman	*D*2	*D*3	*D*2 +	*Ddup*
100,000 Signatories					
Geometric petition					
5% Sample	2,098	1,932	2,098	1,932	1,932
10% Sample	1,193	1,128	1,182	1,128	1,128
20% Sample	680	671	679	671	671
Singles-and-doubles					
5% Sample	1,891	**	**	**	**
10% Sample	1,122	**	**	**	**
20% Sample	678	**	**	**	**
Uniform petition					
5% Sample	14,584	3,029	9,926	3,028	3,027
10% Sample	5,683	1,478	3,252	1,476	1,473
20% Sample	1,271	727	984	721	714
250,000 Signatories					
Geometric petition					
5% Sample	3,500	3,069	3,500	3,071	3,073
8% Sample	2,303	2,124	2,303	2,127	2,128
10% Sample	1,926	1,796	1,907	1,796	1,796
Singles-and-doubles petition					
5% Sample	3,012	**	**	**	**
8% Sample	2,092	**	**	**	**
10% Sample	1,774	**	**	**	**
Uniform petition					
5% Sample	44,192	4,822	16,203	4,895	4,968
8% Sample	15,580	2,979	7,519	3,041	3,104
10% Sample	11,609	2,375	5,128	2,376	2,375

**$D2$, $D3$ $D2+$ and *Ddup* are equivalent to Goodman for singles-and-doubles petition

petitions which contained triplicate etc signatures), although the increase in RMSE is only about 2%.

4. Second Simulation Study

One distinctive aspect of applying these estimation procedures to petitions is that there are in fact two estimation problems: under the New Zealand Citizens' Initiated Referenda Act, only people eligible to vote (i.e., on the electoral roll) qualify, so the number of ineligible signatures in the petition must be estimated, as well as the number of multiple signatures from eligible individuals. The first problem is simple, so most attention has focused on the second, especially as solutions to it can be applied to a wide variety of fields. However, as SCT observe, the two estimates are not independent: if 100x% of the signatures in the sample are ineligible, a point estimate of the number of unique eligible signatures can be obtained by applying one of the existing estimators with a population size of $N(1-x)$; however, the standard error of the estimate will not simply be the standard error of the number of unique signatures (nor even the square root of the sum of their squared standard errors).

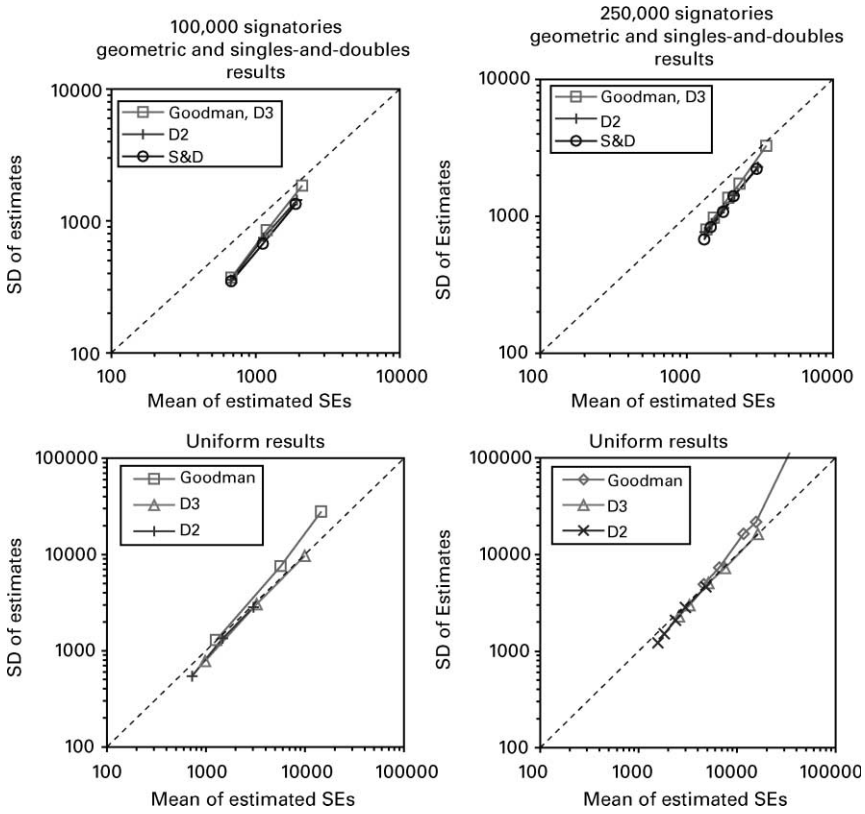

Fig. 2. Comparing estimated standard errors with sampling variability

The estimated number of unique eligible signatures is:

$$N - \hat{U} - \hat{D}$$

where N = Number of signatures in petition,

$\hat{U}$ = Estimated number of ineligible signatures,
$\hat{D}$ = Estimated number of multiple signatures (beyond the first),

which clearly has a variance of

$$Var(\hat{U}) + Var(\hat{D}) + 2Cov(\hat{U}, \hat{D})$$

If the sampling can be treated as approximately Bernoulli, then

$$Var(\hat{D}) = \frac{u(n-u)}{q^2 n}$$

where u is the number of ineligible signatures in the sample.

Since we already have expressions for the variance of the number of signatories ($N - \hat{D}$) derived in the situation when N is a constant, and there are no ineligibles, these can be used to calculate Var($\hat{D}$).

Table 4. Summary of bias-adjusted estimates for 8% sample from 250,000 signatories petitions

RMSE	*D*2	Bias Adjusted *D*2	*D*2+	Bias Adjusted *D*2+	*Ddup*	Bias Adjusted *Ddup*
Geometric	1,501.5	1,385.0	1,523.1	1,381.1	1,553.2	1,384.1
Singles-and-doubles	1,399.9	1,412.8	1,399.9	1,421.8	1,399.9	1,432.4
Uniform	20,685.8	19,240.3	24,033.6	22,272.2	27,521.1	25,413.1
Bias*						
Geometric	484.9	103.0	544.0	105.6	603.1	107.7
Singles-and-doubles	− 32.5	− 381.6	− 32.5	− 431.5	− 32.5	− 481.3
Uniform	20,490.6	19,041.9	23,856.4	22,093.0	27,338.4	25,229.3
SD about own mean						
Geometric	1,422.4	1,382.6	1,424.0	1,375.5	1,432.8	1,381.2
Singles-and-doubles	1,399.5	1,361.7	1,399.5	1,356.1	1,399.5	1,350.5
Uniform	2,837.9	2,758.5	2,915.5	2,822.2	3,168.3	3,054.2

*Bias = 250,000 – Mean (estimate)

SCT present a formula for the covariance term when one is using a linear estimator. However, to get a feel for the relative importance of the three terms, we modified the program used in the first simulation study to generate a random number of ineligible signatures before randomly selecting the rest of the sample from the set of eligible signatures.

4.1. Results

The results of the simulations are shown in Figure 3. The correlation is significant at $p = 0.05$ for $n = 500$ if $|r| > 0.0895$. Clearly, none of the correlations are particularly strong, but as the proportion of invalid signatures or the sampling fraction increases, the strength of the correlation also tends to increase; and with the uniform petition, the correlation is stronger for D_2 than for Goodman or D_3.

Figure 4 shows mean calibrated standard error estimates (derived from Figure 2), estimates of standard error based on $\sqrt{(Var(\hat{U}) + Var(\hat{D}))}$, and estimates based on $\sqrt{(Var(\hat{U}) + Var(\hat{D}) + 2\mathrm{Cov}(\hat{U}, \hat{D}))}$ for the 250,000-signatory petitions. Clearly, for the geometric and singles-and-doubles petitions, as the sampling fraction increases and the proportion of invalid signatures increases, including the estimated standard error of the number of invalid signatures has an appreciable effect on the overall standard error of the estimate. Making allowance for the covariance has no appreciable effect; in fact, the difference between $\sqrt{(Var(\hat{U}) + Var(\hat{D}))}$ and $\sqrt{(Var(\hat{U}) + Var(\hat{D}) + 2Cov(\hat{U}, \hat{D}))}$ was less than one signature in all cases.

The variability of Goodman's estimator and D_3 with the uniform petition was so large that even the difference between the mean calibrated standard error estimates and $\sqrt{(Var(\hat{U}) + Var(\hat{D}))}$ is only minor.

5. Conclusions

The simulations have provided useful quantitative information on the performance of the various estimators, allowing an assessment of various issues such as the degree to which D_2, D_{2+}, D_3 and D_{dup} are biased, how much more variable the results of Goodman's estimator are than alternatives like D_2, and the importance of allowing for the variability of the estimate of the number of ineligible signatures.

On the geometric petition, Goodman's estimator was $\sim 20\%$ more variable than D_2 at an 8% sampling fraction, and $\sim 8\%$ more variable at a 10% sampling fraction. The bugbear of absurdly high or low estimates only arose with the 5% sample from the uniform petition.

Given that no one estimator clearly performs better than the others, it may be best to calculate confidence intervals for the number of signatures using several estimators (Goodman's, D_3, and D_2 or one of its variants). If all three agree as to whether the petition is sufficiently large or not, use that conclusion. If there is no consensus, then simulations based on a number of scenarios about the actual distribution of signatures may be informative; or it may be necessary to increase the sample size; or even completely enumerate the entire petition.

The biased variants of Goodman's estimator (D_2, D_{2+}, D_{dup} and D_3) were only slightly biased on the geometric and singles-and-doubles petitions. This meant that adjusting for their estimated bias produced only minor improvements in their performance, while adding a number of extra assumptions about the distribution of the number of signatures in

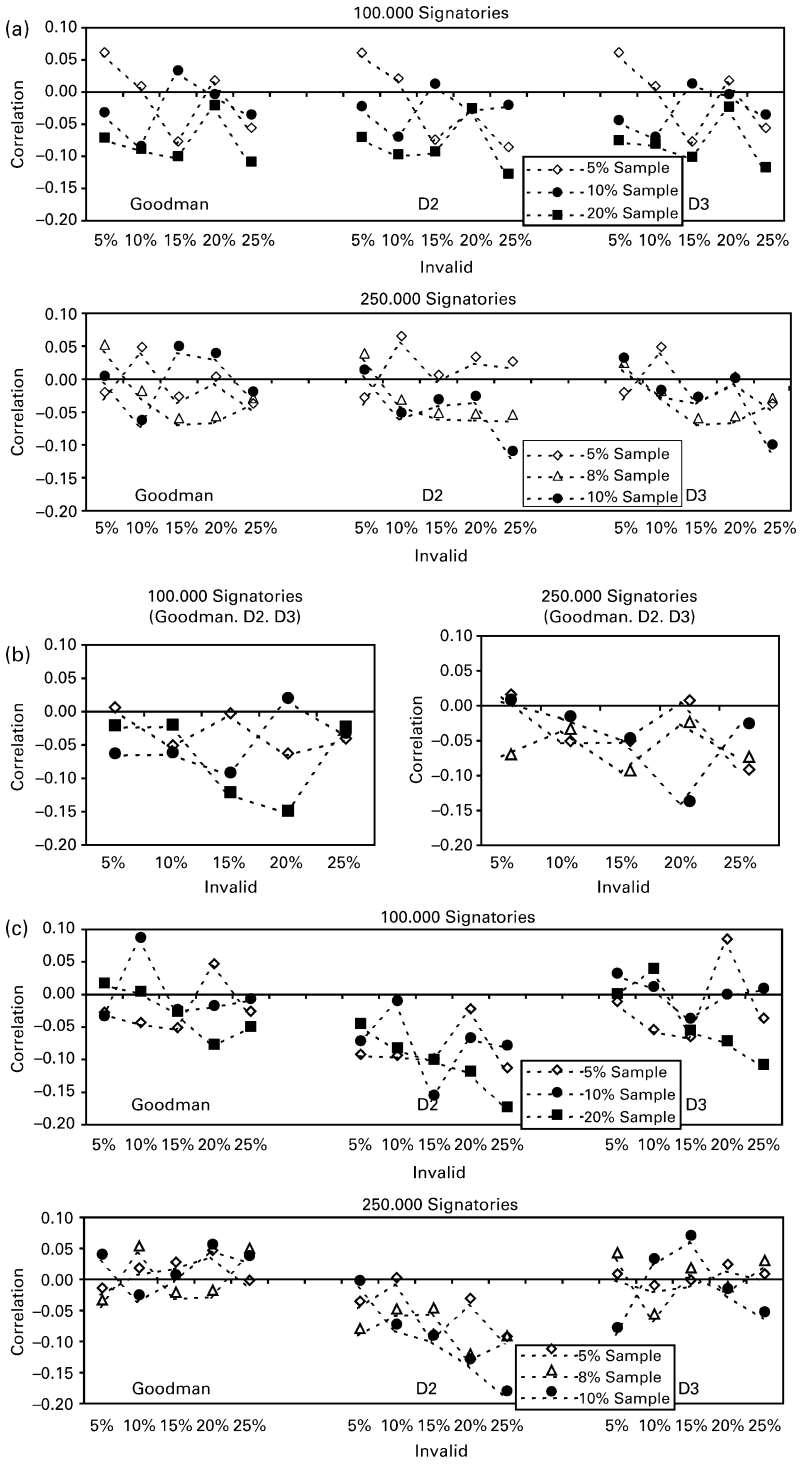

Fig. 3. (a) Correlation between Estimates of U and D – Geometric Petition; (b) Correlation between Estimates of U and D – Singles and Doubles Petition; (c) Correlation between Estimates of U and D – Uniform Petition

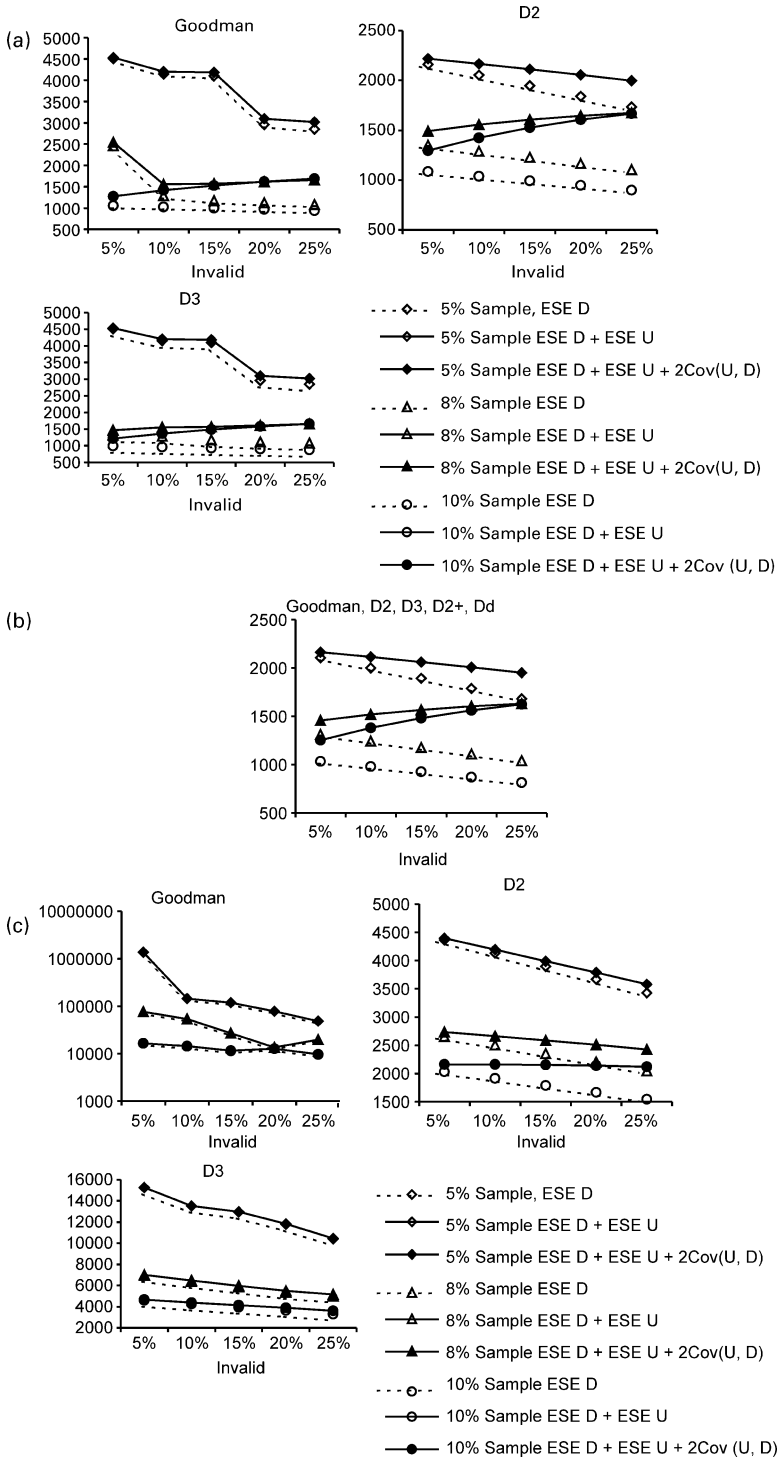

Fig. 4. (a) Estimated Standard Errors – Geometric Petition (250,000 Signatories); (b) Estimated Standard Errors – Singles and Doubles Petition; (c) Estimated Standard Errors – Uniform Petition (250,000 Signatories)

the petition. On the uniform petition the bias was much more marked. Bias adjustment actors based on petition with a different structure (the fully enumerated Washington petitions, which are roughly geometric) only produced minor improvements.

Haas and Stokes's variance estimates appear to have a rather complex relationship with the actual variability of estimators, the exact nature of which depends on the size of the petition and the distribution of the number of signatures. This is presumably because of HS's assumption that all classes are the same size (i.e., everyone has signed the petition the same number of times), which seems unlikely but makes the problem tractable. Any other assumption about the distribution of the number of signatures complicates the problem considerably, and still leaves open the question of whether it is appropriate for a particular sample. By making some assumptions about the likely structure of a petition, it might be possible to run simulations and obtain a graph similar to Figure 2, which could be used to improve the estimated standard errors produced by HS's formula. A conservative alternative might be to use HS's variance estimates, since if they err, they are likely to err on the high side.

Including the variability of the estimate of ineligible signatures in the estimate of the variability of the overall result appears wise. Although the estimate of the number of ineligible signatures and the estimate of the number of eligible but duplicated signatures in the petition are correlated, the correlation appears to be small and does not alter the estimate of the variability appreciably.

6. References

Brutlag, J.D. and Richardson, T.S. (2002). A Block Sampling Approach to Distinct Value Estimation. Journal of Computational and Graphical Statistics, 11, 389–404.

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

Goodman, L. (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.

Kish, L. (1965). Survey Sampling. John Wiley.

Smith-Cayama, R.A. and Thomas, D.R. (1999) Estimating the Number of Distinct Valid Signatures in Initiative Petitions. Proceedings of the American Statistical Association, Section on Survey Research Methods, 38–243.

Received March 2003
Revised September 2004