# The Performance of Techniques for Estimating the Number of Eligible Signatories to a Large Petition on the Basis of a Sample of Signatures

A thesis presented in partial fulfilment of the requirements for the degree of

Master of Science

in

Statistics

at Massey University, Palmerston North, New Zealand

Duncan Hedderley

2002

## Abstract

The New Zealand Citizens' Initiated Referenda Act, 1993, states that if a petition signed by at least 10 percent of eligible electors is presented to the House of Representatives, then parliament is required to hold an indicative referendum on the petition. 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.

We review a number of techniques used for similar problems such as estimating the size of a population through capture-recapture studies, or estimating the number of duplicate entries in a mailing list. One suitable estimator was developed by Goodman (1949). A number of variants on it are reported by Smith-Cayama & Thomas (1999).

An estimator proposed by Esty (1985) was found to give unreasonable estimates, and so a modification was developed. In order to test the performance of the modified estimator, simulations, drawing repeated samples from artificial petitions with known distributions of multiple signatures, were performed.

The simulation results allowed us to investigate bias in the estimators and the accuracy of the variance estimates proposed by Hass & Stokes (1998). The effect of sampling fraction on bias, variability and estimated variance of the estimators was also investigated.

The simulation program was modified to include ineligible signatures. Results of these simulations showed that estimating the number of ineligible signatures added to the variability of the overall estimate of number of eligible signatories. Although Smith-Cayama & Thomas (1999) mention that the estimated number of multiple eligible signatures and the estimated number of ineligible signatures are correlated, the simulations suggest the correlation is small and makes little difference to the final estimate of variability.

## Acknowledgements

I would like to acknowledge the advice, support and occasional harrying of my supervisor, Associate Professor Stephen Haslett. Wearing his Director of the Statistics Research and Consulting Centre hat, Steve also deserves thanks for letting me work on the problem when consulting work was thin on the ground.

I would also like to thank Mike Doherty at Statistics New Zealand for initially raising the problem; freely sharing his initial thoughts, his gleanings from the literature, and experience with previous petitions; and explaining the practical constraints that the process operates under.

Many thanks also to Wendy Browne (Institute of Information Sciences and Technology, Massey University) and Kathy Hamilton (Office of the Pro Vice Chancellor, College of Sciences, Massey University) for cheerfully and capably shepherding me through the narrow paths and steep slopes of university procedures.

And finally, thanks to David Fletcher for permission to use his 'The Politician' cartoon (p72), which first appeared in The Dominion on 2 April 2002.

# Contents

# Chapter 1

## Literature Review

A number of countries, including New Zealand, and US 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 eligible electors, then the House of Representatives is required to hold an indicative referendum on the petition.

Against this background, it is important to establish reliably the number of eligible 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 whether a signature is eligible (ie the person is on the electoral roll, and discarding multiple signatures, so that if a person has signed the petition several times, they are only counted once) requires some effort. Because of this, normal practice at the present is to take a sample (between 8 and 10 percent) of the signatures and check them for eligibility and multiple signatures.

The task of estimating the number of people who have signed a petition on the basis of a sample can be seen as a special case of a wider class of problem: estimating the number of *types* of observation in a population (where the observations are partitioned into classes) from a sample. Other examples include estimating the number of species in a biological population; estimating the number of types of coin in circulation from archaeological finds; or estimating the size of an author's vocabulary on the basis of their published work. Bunge & Fitzpatrick (1993) give a review of this type of problem, and the various ways people have attempted to solve it.

Not all of these approaches appear to be relevant to the petition problem. For instance, estimating 'coverage', the proportion of the population in the classes which appear in the sample, may provide useful information in ecology or numismatics where some classes will have many members and some only a few members; but with a CIR petition it is expected that there are many classes (signatories), most of whom only appear

once in the population (have signed the petition only once). Similarly, models which assume that one is sampling from an infinite population, or a large population where the sampling procedure is unlikely to substantially reduce the numbers in (and so probability of selecting) any given class are unlikely to be a good approximation to the CIR petition problem. Bunge & Fitzpatrick identify several approaches which may be relevant, based on assuming hypergeometric sampling from a finite population.

**Hypergeometric Sampling Models**

The basic hypergeometric distribution is the equivalent of the binomial distribution for sampling a finite population without replacement. Under the binomial distribution, observations can fall into one of two classes, and each observation has a probability p of being in the first class. Under the hypergeometric distribution, observations are drawn without replacement from a finite population of size N consisting of two classes of observation; each observation is equally likely to be drawn and initially there are A observations in the first class; so when the first observation of the sample is drawn, it has a probability A/N of being from the first class; however, once n observations have been drawn, of which a are from the first class, the probability that the next observation will be from the first class is (A-a)/(N-n).

Just as the binomial distribution can be extended to cover more than two classes, producing the multinomial distribution, the hypergeometric can be generalised to cover a finite population consisting of C classes. In this case, the $i^{th}$ class initially has $N_i$ members, and the probability that a sample of size n contains $n_i$ members of the $i^{th}$ class is

$$\binom{N}{n}^{-1} \times \prod_{i=1}^{C} \binom{N_i}{n_i} \qquad \text{if } n_i \le N_i \text{ for all i, where } n = \sum_{i=1}^{C} n_i$$

Goodman (1949) develops a hypergeometric model, and from that an unbiased estimator of the number of classes; however, the estimator is very variable because it

involves the observed numbers of singles, pairs, triples, quadruples etc. As Kish (1965) notes, if a sample of fraction f is taken from a population, then each class with just one member has a probability f of being in the sample; each class with 2 members has a probability $f^2$ of both members appearing in the sample; and so on. Thus to estimate the number of classes with one member in the population, one could take the number of classes with one member in the sample and multiply by 1/f; to estimate the number of classes with two members in the population one would need to multiply the number of classes with two members in the sample by $1/f^2$ . For classes with more than two members, the chances of them all appearing in the sample are even lower, and their weight in the estimate of the population correspondingly higher. Thus, observing a class with two or more members in the sample is a rare event with high weight, which contributes considerably to the variability of the estimate of the number of classes in the population. Goodman presents a number of alternative estimators, which while not unbiased are less variable, the simplest of which is simply the first two terms (ie for single observations and duplicate observations) from the full estimator.

Shlosser (1981) develops an estimator of the number of classes with k members in a population, and from that the total number of classes in the population, on the basis of binomial sampling and asymptotic behaviour. He notes that it is biased, and that it will perform better when the sampling fraction is closer to unity, and the number of classes with k>1 members is small compared to the number with one member.

Hill (1968) presents a Bayesian model for the problem of estimating the number of classes in a population. Some of the aspects of the underlying model are a bit strange: for instance, individual observations are ranked, as well as being assigned to classes, and much of the development concerns inference about the ranks; and the model does not explicitly take account of the size of the classes, just the overall size of the population and the overall number of classes. Hill (1979) presents formulae for the mean and variance of the posterior distribution, assuming the prior distribution of the number of classes is uniform on the range from 1 to the size of the population.

From a numismatic perspective Esty (1985) develops a model based on a negative binomial distribution of the class sizes, and binomial sampling. The assumption that the distribution of the class sizes is known simplifies development considerably. However, much seems to rest on the choice of a shape parameter for the initial negative binomial. The figures Esty quotes as likely for the number of coins produced by an individual die in the ancient world[1] are clearly not appropriate for the numbers of times an individual signs a petition. Another section of Esty's paper reports the results of a simulation study on whether it is possible to estimate the value of the shape parameter from a sample. The results are most disappointing, and eventually Esty recommends using rule of thumb values (k=1 for a pure geometric distribution; k=2 for many numismatic problems)

One oddity of Esty's model is that by using the negative binomial, it includes classes of size zero in its total population (in the petition case, people who didn't sign the petition even once). In Chapter 3 of this thesis, a version of the estimator which assumes that the *additional* signatures follow a negative binomial distribution has been developed; with a shape parameter (k) of 1 and figures typical of recent petitions (a sample of 12500 signatures of which 50-60 turn out to be duplicates) gives an estimate which is at least of the right order of magnitude.

Chao & Lee (1992) build on various papers from the ecological and numismatic literature (including Esty, 1985) which have discussed coverage estimators. The two estimators they develop are 'non-parametric' in that they allow different classes different probabilities of being drawn in the sample, but unlike Esty make no assumptions about the distribution of those probabilities. However, they do assume that the sampling is multinomial rather than hypergeometric; this implies either a population very much larger than the sample, so that the probability a specific observation is a specific class remains essentially the same as the sample is drawn, or sampling with replacement. If we apply either Chao & Lee estimator to the typical results from recent petitions (a sample of 12500 signatures of which 50 are duplicates), the estimated coverage is very low (0.8%) and the estimate of the number of unique signatures (about 1.5 million) is about 5 times the total number of signatures on a typical petition (between 250,000 and 300,000).

---

[1] 'It appears that 10000 coins per die is quite possible'

Based on these initial investigations, it appears that the difference between a population with finite and (effectively-) infinite class sizes is substantial.

Shuster (1974) presents a decision rule for determining whether a petition has sufficient signatures based on a sample from it. The method uses stochastic minimisation to simplify the range of possible problems to one which simply involves single and duplicate signatures, and then develops a Poisson approximation to an earlier result from Raj (1961). One interesting aspect of the paper is that it is specifically phrased in terms of hypothesis testing, with the null hypothesis being that the petition does not have sufficient signatures. It is not clear from the Citizens' Initiated Referenda Act which way a hypothesis might be phrased: is the onus on the petition organiser to show that there are sufficient signatories, or on the Clerk of the House to show that there are not? One approach Statistics New Zealand have considered (but not yet attempted) is reversing Shuster's approach, to produce an upper estimate of the number of signatories consistent with the data from the sample.

Bunge & Handley (1991) look at the problem of estimating the number of duplicate entries in a database. Their approach is to draw a small sample of records and then check the rest of the database (all the rest of the database) to find how often they occur. In simulations, a sample of 100 records from a database of approximately 20 million records produced estimates with coefficients of variation between 0.018 and 0.118 (The larger the average size of classes, the higher the CV). Although the approach appears promising, it is probably more practical for data held electronically, since that is considerably easier to search completely than paper records like the sheets on which a petition has been submitted.

## Capture-Recapture Models

One situation where one might wish to estimate the number of classes in a population is when there is no complete list of the population, just a set of incomplete lists, some of which may contain some of the same individuals. Capture-recapture studies fall into this class of problem; similar approaches have been used to estimate the number of diabetics in a region from a number of registers, and to estimate the size of the World-Wide Web (cited in Fienberg *et al* 1999).

The simplest of these models just look at the number of times the same individual turns up, which is comparable to the petition problem. However, one of the issues in this field is that some individuals might be more likely to appear than others (for instance, they may be easier to catch); similarly, some lists or samplings may be more comprehensive than others. Existing approaches have been based on analysing contingency tables of the number of individuals in list/ sample A which also appear in list/sample B using log-linear models. Fienberg *et al* (1999) summarise these before presenting a Bayesian approach which appears to perform better, although at the cost of increased computation.

In trying to apply these to the petition situation, the question which arises is, "what are the lists?" For the simple models, which individual appears on which list is not important; all that matters is how many times they appear on the composite list. In that case there is no practical difference between the way we would estimate the size of a petition from a sample in which an individual appears no more than $m$ times, and the way we would estimate the size of a population compiled from $l$ ($\geq m$) lists. However, one might argue that people who have signed several times are more likely to appear in a list/sample, and so models which take account of the 'catchability' of individuals might be more appropriate. But to fit these models we need more information about the lists, such as how many individuals are common between any two lists. To answer that question, we need a better concept of what the 'lists' are in this situation. Individual sheets of the petition might serve, on the assumption that someone is unlikely to sign their name twice on the same sheet of paper; however, a typical sample of 12,000 signatures from a petition might be spread over 1000 or more petition sheets; recording

how the names on the sheets relate to those on other sheets would complicate the data collection considerably, to say nothing of the demands of analysing a (sparse) $2^{1000}-1$ contingency table.

**Recent Papers**

Two more recent papers make some attempt to compare and contrast techniques, rather than just continuing the proliferation.

Haas & Stokes (1998) dismiss Goodman's estimator as too variable, and Goodman's proposed biased estimator based on the numbers of singletons and doubles in the sample, because in some situations you may never have singles and doubles, only higher multiplicities in the sample (This sort of situation may be conceivable, but does not seem to be the case with petitions). They discard Hill's estimator as Bayesian (and thus, presumably, subjective and suspect). They develop two modifications of Shlosser's estimator, as well as a number of estimators based on the Generalised Jackknife approach. They then test these against a variety of (created) data sets, covering a range of conditions (skewness of the distribution of multiples, sampling fractions). Their conclusion is that for data which is not seriously skewed, a second-order generalised jackknife estimator gives the best performance; they also suggest other refinements (such as a 'stabilisation' technique, which basically post-stratifies the sample into low and high multiplicity classes) which do not seem appropriate for the results typical of petitions.

Haas & Stokes also present a delta-method approach to estimating the variance of the estimator.

In a paper specifically concerned with the petition problem, Smith-Cayama & Thomas (1999) review the literature, and the estimators used by a number of US states which have legislation similar to the Citizens' Initiated Referenda Act. They then develop formulae to estimate the variance of a variety of linear estimators derived from Goodman's original suggestions. Since these estimators are biased, they also develop formulae for bias; however, to apply these, one needs to have some prior information

about the likely distribution of the numbers of multiple signatures in the petition as a whole. Fortunately, in Oregon State, when a petition is neither clearly large enough, nor clearly too small, the whole petition is checked; so Smith-Cayama & Thomas had access to the complete distribution of multiple signatures in four petitions from the 1980s and 90s which were checked completely.

**Table 1.1 Completely Enumerated Oregon State Petitions**

**(from Smith-Cayama & Thomas, 1999)**

| | Petition A (1984) | Petition B (1995) | Petition C (1989) | Petition D (1996) |
|---|---|---|---|---|
| Number of Signatures | 162324 | 231723 | 173858 | 228148 |
| Number Invalid | 19437 (12.0%) | 47383 (20.4%) | 31325 (18.0%) | 34542 (15.1%) |
| Number Duplicated | 4256 (2.6%) | 4546 (2.0%) | 9738 (5.6%) | 11584 (5.1%) |
| Number of Unique, Valid Signatures | 138631 (85.4%) | 179794 (77.6%) | 132795 (76.4%) | 182022 (79.8%) |
| Number Signing | | | | |
| ... Once | 134489 | 175363 | 123205 | 170988 |
| ... Twice | 4031 | 4331 | 8878 | 10518 |
| ... Three Times | 108 | 93 | 385 | 489 |
| ... Four Times | 3 | 6 | 30 | 22 |
| ... Five Times | | | | 3 |
| ... Six Times | | | | 2 |
| ... Twelve Times | | 1 | | |
| Coefficient of Variation Squared | 0.0296 | 0.0252 | 0.0652 | 0.0584 |

With these, they are able to estimate the RMSE for the various estimation techniques; they also compare these against Haas & Stokes' recommended method, a second-order generalised jackknife estimator. Their conclusion is that for this application 'it was difficult to improve much on the Goodman-type estimator' based on the first two terms (singletons and doubles) of Goodman's full estimator. They make the point that often in

a sample from a petition one will only have singleton and double signatures; in that case, this estimator is equivalent to the full estimator, and so is unbiased.

**Meanwhile, in New Zealand...**

None of the recent petition submitted under the Citizens' Initiated Referenda Act have been completely counted; however, as an example, the results of the second submission of Norm Withers' petition on tougher sentencing for violent criminals had 252,336 signatures. A sample of 28,704 (11.4%) was taken; of these 4,454 were invalid, 23,842 were valid single signatures, 201 were valid pairs of signatures, and 2 were valid triples. This was the first petition in recent times to have triple signatures in the sample.

# Chapter 2
# The Models

The task of assessing the number of eligible people who have signed a petition actually consists of two problems:

- estimating the number of signatures from *eligible* people (ie people on the electoral roll), and then
- estimating the number of eligible *individuals* who have signed the petition from the number of eligible signatures.

The first step is comparatively simple: since we are not interested in estimating the number of eligible *individuals* who have signed the petition, but just in the number if eligible *signatures* at this stage, we can simply use the proportion of eligible signatures in a sample as an estimate of the proportion of eligible signatures in the petition. So, if we draw a sample of n signatures from a petition with a total of N signatures, and it contains u ineligible signatures, then our estimate of the number of ineligible signatures in the petition would be $\frac{N}{n}u$ . This estimate is unbiased, and its variance can easily be estimated.

The rest of this chapter reviews a number of the models proposed for solving the second step: estimating the number of classes in a finite population.

**The Problem, Formally**

Almost every author on this topic seems to have used their own notation. We will follow Haas & Stokes (1998) system, which states the problem as:

We have a population of size N, whose members can each be classified as falling into one (and only one) of D classes. These classes are labelled $C_j$ $(1 \leq j \leq D)$, and the $j^{th}$ class has $N_j$ members in the population. Because the classes are disjoint (ie members of the

population belong to just one class, and every member of the population belongs to a class) $\sum_{j=1}^{N} N_j = N$

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

In this 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 $N_i$; this means that $\sum_{i=1}^{N} N_i = D$, and $\sum_{i=1}^{N} iN_i = N$

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

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

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

Obviously, $(n_1, n_2, \ldots n_D)$ 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 $(f_1, f_2, \ldots f_n)$. The probability mass function of $(f_1, f_2, \ldots f_n)$ is the sum of the $p( (n_1, n_2, \ldots n_D) \mid D, (N_1, N_2, \ldots N_D) )$ over all combinations of $(n_1, n_2, \ldots n_D)$ which correspond to $(f_1, f_2, \ldots f_n)$; in other words, which have exactly (D-d) $n_i$ equal to 0, $f_1$ $n_i$ equal to 1, $f_2$ $n_i$ equal to 2, etc.

**Goodman's Estimator**

Working from the basis of hypergeometric sampling, Goodman (1949) showed that:

$$D_{Goodman} = d + \sum_{i=1}^{n} (-1)^{i+1} \frac{(N-n+i-1)!\,(n-i)!}{(N-n-1)!\,n!} f_i$$

is the *only* unbiased estimator of D, so long as $n > \max(N_1, N_2, \ldots N_D)$. If $n < \max(N_1, N_2, \ldots N_D)$, no unbiased estimator exists.

Smith-Cayama & Thomas (1999) presented an alternative form

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

$$\text{where } p_{ij} = \frac{\binom{j}{i}\binom{N-j}{n-i}}{\binom{N}{n}} = p\left(\begin{array}{c}\text{a sample of n from N will contain i members} \\ \text{of a class with a total of j members}\end{array}\right)$$

and

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

It is comparatively easy to demonstrate that the two formulations are equivalent when the sample only contains single and double signatures; proving equivalence in general is complicated by the fact that the formula for $c_j$ refers to preceding $c_i$ s. However, Smith-

Cayama & Thomas (1999) show that their formulation is an unbiased estimate of D, and since Goodman (1949) showed that $D_{Goodman}$ is the only unbiased estimator of D, this implies that Smith-Cayama & Thomas' formulation is equivalent to Goodman's.

Although this estimator is unbiased, its variance can be very high (The estimate need not even be positive!). This is because the proportion of duplicates in the sample tends to be considerably lower than in the population. The under-representation is even more marked as the multiplicity increases (so even if $N_i = N_j$ , $i<j$ , $f_i$ will tend to be larger than $f_j$ ). To correct for this, the $f_i$ s are given different weights in the estimator; the weight given to $f_i$ is approximately $\left(\frac{N-n}{n}\right)^i$ , which means that the size of the estimate can be very heavily influenced by the number of rare, high-i classes there are in the sample[1], especially if the sampling fraction , $\frac{n}{N}$ , is small. Because the classes are rare, this means that their numbers can be subject to proportionally quite large sampling variation. The high weight means that this can feed through to have a considerable

---

[1] If a sample of n is taken from a population of N, then each class with just one member has a probability

$$\left(\frac{n}{N}\right)$$

of being in the sample; each class with 2 members has a probability

$$\left(\frac{n}{N}\right)\left(\frac{n-1}{N-1}\right)$$

of both members appearing in the sample; and so on. Thus to estimate the number of classes with one member in the population, one could take the number of classes with one member in the sample and multiply by

$$\left(\frac{N}{n}\right)$$

To estimate the number of classes with two members in the population one would need to multiply the number of classes with two members in the sample by

$$\left(\frac{N}{n}\right)\left(\frac{N-1}{n-1}\right)$$

and so on. See Kish (1965)

influence on the estimate, increasing its sampling variability. Goodman was aware of this, and suggested several alternative estimators which were less variable, but biased. The simplest

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

was simply $D_{Goodman}$ with the very high weights removed, just leaving the terms for $f_1$ and $f_2$. Smith-Cayama & Thomas (1999) 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 $D_{Goodman}$.

Haas & Stokes (1998) observe that these estimators will not work well when the number of classes in the population, D, is small and so the sample contains comparatively few singles and duplicate observations. However, samples from recent CIR petitions have consisted overwhelmingly of single and duplicate signatures, with only a handful of triplicate signatures in one petition.

Smith-Cayama & Thomas 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 a refinement of this, replacing the number of people who have signed multiple times, $\sum_{i=2}^{n} f_i$, with the number of duplicate signatures, $\sum_{i=2}^{n} (i-1) f_i$ (By 'duplicate' they mean any signatures beyond that person's first; so a person who has

signed twice has 1 valid signature and 1 duplicate; a person who has signed 17 times has 1 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. Similarly, if the sample only contains single, double and triple signature, then $D_{Goodman}$ and $D_{Goodman3}$ are equivalent. In this case, though, the other variants of Goodman's estimator are biased (Since they are not equivalent to the full Goodman's estimator, and it is the only unbiased estimator)

Smith-Cayama & Thomas suggest improving the biased estimators by calculating bias-correction factors for them. To do this you need to have an idea of the actual distribution of the signatures, so that you can calculate the ratio of the actual number of signatures to the estimated number of signatures. Smith-Cayama & Thomas use the results from the four fully counted Oregon petitions to estimate the bias of each estimator, and then test the results on the same petitions, jack-knife fashion (ie use the results from petitions B, C and D to calculate bias-correction factors, and then apply them to a sample from petition A, etc) Their results suggest that the bias-correction factors tend to reduce bias somewhat, but do not substantially reduce the overall Root MSE of the estimators.

**Shlosser's Estimator**

Based on asymptotic results and binomial sampling, Shlosser proposed

$$D_{Shlosser} = d + f_1 \frac{\sum_{i=1}^{n} (1-q)^i f_i}{\sum_{i=1}^{n} iq(1-q)^{i-1} f_i} \quad \text{where } q = \frac{n}{N}, \text{ the sampling fraction}$$

as an estimate of D.

Shlosser's estimator is, of course, biased; Shlosser comments that the bias will be less when the sampling fraction is larger and ratios $f_i/f_1$ are smaller. However, he gives two examples where even with a sampling fraction of 10% the errors are comparatively small. In early tests we found Shlosser's estimator very heavily biased; Haas & Stokes (1999) make the same observation.

### Haas & Stokes' Estimators

The estimators so far have either offered unbiasedness, but at the cost of potentially high variance, or less variable but biased estimates. One way to improve these latter estimates is to apply bias-reduction techniques to them. Haas & Stokes use two jackknife approaches to bias reduction (the Generalized Jackknife, and Horvitz-Thompson Jackknife Estimators) to develop a range of estimators. They also discuss Shlosser's estimator and why it appears to perform poorly with low sample sizes, and suggest two variations on it.

To test the various estimators, Haas & Stokes 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 $\gamma^2$, the coefficient of variation of the class sizes $N_1, N_2, \ldots N_D$. Judging by the Oregon petitions cited in Smith-Cayama & Thomas, $\gamma^2 << 1$. The estimator which performed best[2] 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 $q = \frac{n}{N}$, the sampling fraction

[2] Had the lowest RMSE

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

and

$$D_{uj1} = \frac{d}{\left(1 - \frac{(1-q)f_1}{n}\right)}, \text{ an initial estimate of D}$$

This was closely followed by a smoothed version of the same estimator, which has $E[f_1]/n$ in place of $f_1/n$ . The next best estimators were a variant of Shlosser's estimator and $D_{uj1}$ .

**Variance of the Estimates**

Haas & Stokes (1999) present a way of estimating the asymptotic variance of an estimator, based on the delta method. The general form of Haas & Stokes' variance estimator is:

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

where

$\hat{D}$ is the estimator

N is the size of the population

$A_i$ is the partial derivative of $\hat{D}$ with respect to $f_i$

They derive approximate values for $\text{var}(f_i)$ and $\text{cov}(f_i, f_j)$ 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

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

For most applications, this is sufficient; but as Smith-Cayama & Thomas (1999) point out, in the case of petitions, we are estimating the number of eligible *signatures* in the petition before estimating the number of eligible *individuals* who have signed the petition. The variability of this estimate needs to be included in our overall estimate of the variance; as does the covariance between the two estimates.

If there are u ineligible signatures in a sample of n signatures from N, then our estimate of the number of ineligible signatures in the petition would be $\frac{N}{n}u$ , and assuming that u follows a hypergeometric distribution (since we are sampling without replacement), the variance of the estimate is:

$$\frac{N^2}{n^2} n\left(\frac{u}{n}\right)\left(1-\frac{u}{n}\right)\left(\frac{N-n}{N-1}\right) = \frac{N^2}{n}\left(\frac{u}{n}\right)\left(1-\frac{u}{n}\right)\left(\frac{N-n}{N-1}\right)$$

The covariance is more complicated. The sample size is clearly the sum of the total number of eligible *signatures* plus the total number of ineligible signatures:

$$n = u + \sum_{k=1}^{n} k \sum_{l=k}^{n} f_{kl} \quad \text{so } u = n - \sum_{k=1}^{n} k \sum_{l=k}^{n} f_{kl}$$

where $f_{kl}$ is the number of people who have signed the petition l times and appear in the sample k times.

So for $f_{ij}$ with a given i and j

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

And

$$\text{cov}\left(u, f_{ij}\right) = -\sum_{k=1}^{l}\sum_{\substack{l=1 \\ l\neq j}}^{n} \text{kcov}\left(f_{kl}, f_{ij}\right) - \text{ivar}\left(f_{ij}\right) - \sum_{\substack{k=1 \\ k\neq i}}^{j} \text{kcov}\left(f_{kj}, f_{ij}\right)$$

Since $f_i = \sum_{j=i}^{n} f_{ij}$, we can calculate cov(u, $f_i$).

Goodman's estimator and its variants can be written as linear combinations of the $\{f_i\}$, which means that calculating cov(u, $D_{Goodmanx}$) should be feasible, although long-winded (Smith-Cayama & Thomas, 1999). For Schlosser's estimator and the Haas & Stokes estimators, which are non-linear combinations of the $\{f_i\}$, the task will be harder

# Chapter 3
# Improving Esty's Estimator

Esty (1985) presents an estimator based on the assumption that the size of each class (eg number of signatures from a person) follows a Negative Binomial distribution, with parameters k and r

$$p(\text{class size} = y) = \frac{\Gamma(k+y)}{\Gamma(k)\, y!}\, r^k\, (1-r)^y \qquad \text{where } k > 0,\ 0 < r < 1 \text{ and } y = 0,1,2,\ldots$$

and then the number of elements of that class in the sample follow a Binomial distribution with parameters y and q. In developing the estimator, the parameters r and q get absorbed into n (the sample size) and d (the number of distinct classes in the sample), but k remains explicitly in the model. Esty's simulations suggest that estimating k from the observed data is extremely unreliable even if the sample is perfectly random and the distribution of the y's is exactly Negative Binomial. Ultimately he suggests using an educated guess at k; for numismatic problems, he suggests k=2. We also tried k=1, which gives a Geometric distribution.

Applying Esty's estimator to typical petition data (a sample of 12500 signatures, with 50 duplicates) produced estimates which were far too high ($3.1 \times 10^6$ for k=1; $2.3 \times 10^6$ for k=2). However, while discussing this, Prof Haslett pointed out that if the class size, y, follows a Negative Binomial distribution, this will create some classes with sizes of 0; in other words, using the Negative Binomial distribution for class size means that among the total number of classes being estimated there will be classes of size 0 (ie those who didn't sign the petition at all). As estimates of the total electorate, the figures above are not unreasonable.

This prompted us to investigate a new estimator which assumed that the number of *additional* signatures a person made followed a negative binomial distribution, ie

p(person has signed y times) = Negative Binomial $_{(k,r)}$ (y - 1)

where Negative Binomial$_{(k,r)}$(y) denotes the probability of a value y under a Negative Binomial distribution with parameters k and r

Following Esty, it is fairly easy to show that the distribution of the number of signatures from an individual in the sample, x, is

$$(1-q)\text{ x Negative Binomial}_{\left(k,\frac{1}{(r+q-rq)}\right)}(x) + q\text{ x Negative Binomial}_{\left(k,\frac{1}{(r+q-rq)}\right)}(x-1)$$

where q is the proportion of the total population (petition) drawn in the sample.

Following Esty

$$E(x) = q + k \times \left(\frac{q - rq}{r}\right) = m'$$

$$p(X = 0) = (1 - q) \times \frac{r^k}{(r + q - rq)^k} = (1 - q) \times \left(1 + \frac{(m' - q)}{k}\right)^{-k}$$

and if there are D classes in the population, then there are d classes in the sample. Under the assumption of binomial sampling, the size of the sample itself is a random variable, n, and

$$E(n) = Dm'$$
$$E(d) = Dp(X > 0) = D \times (1 - p(X = 0))$$

Thus

$$E(n)/E(d) = \frac{E(X)}{(1 - p(X = 0))} = \frac{m'}{1 - (1 - q)\left(1 + \frac{(m - q)}{K}\right)^{-k}}$$

$$= \frac{(k + m' - q)^k m'}{(k + m' - q)^k - k^k (1 - q)}$$

If we substitute $n/d$ for $E(n)/E(d)$ then for a given k and q, we can find m', and since

$$N = E(n)/m'$$

we can then get an estimate of N by substituting n for E(n).

So for **k=1** (a geometric distribution for the number of additional signatures)

$$m' = n/d + q - 1$$

Given a sample of 12500 signatures (1/20 of the petition), of which 50 are duplicates (so d=12450)

$m' = 0.054$

and $\hat{N} = n/m' = 12500/0.054 = 231412.6$

which is a lot more credible than the original Esty estimator.

For **k=2**, the maths is messier; m' is the solution of

$$m^3 + m^2\left(4 - 2q - {}^{n}/_{d}\right) + m\left(4 - 4q + q^2 - 4{}^{n}/_{d} - 2q{}^{n}/_{d}\right) - q^2{}^{n}/_{d} = 0$$

Given a sample of 12500 signatures (1/20 of the petition), of which 50 are duplicates (so d=12450)

$m' = 0.112$ (Excel Solver solution)

and $\hat{N} = {}^{n}/_{m'} = {}^{12500}/_{0.112} = 111763.2$

At least this is the right order of magnitude, unlike the original estimates. That it is not a very likely total number of signatures on the petition probably reflects more on the credibility of the assumed distribution; a Negative Binomial with a shape parameter of 2 implies that multiple signatures are considerably more common than with a shape parameter of 1.

**Variance Estimates**

Haas & Stokes describe a delta-function approach to estimating the variance of an estimator.

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

where

$\hat{D}$ is the estimator

N is the size of the population

$A_i$ is the partial derivative of $\hat{D}$ with respect to $f_i$

Applying this to the modified Esty estimator with k=1 (The simplest case, and also the one that seems closest to our actual situation)

$$\hat{D} = \hat{N} = {}^{n}\!/_{m'} = \frac{n}{\left({}^{n}\!/_{d} + q - 1\right)} = \frac{nd}{(n + dq - d)}$$

where

n is fixed

$\left(\text{it can be written} \sum_i i f_i \text{ , but in this case we are analysing the sample } \textit{conditional} \text{ on n}\right)$

$d = \sum_i f_i$

we get

$$A_i = \left(\frac{n}{n + dq - d}\right)^2 = \left(\frac{\hat{D}}{d}\right)^2$$

This still leaves the question of estimating $\text{var}(f_i)$ and $\text{cov}(f_i , f_j)$. The approximation that the number of members of a class in the sample follows a binomial distribution, given the total size of that class is known, is widely used in the literature; however, to go from that to a distribution of number of members of a class in the sample requires that we either need to know the total size of the class (which we do not) or make an assumption about the distribution of the total sizes of classes. Since we do not know the former, any attempt to solve the problem has to be based on the latter assumption. Obviously, the details of this assumption will influence the estimated variance. Since there is no way of checking its validity (other than enumerating the whole petition), a useful assumption is one which gives reasonable results even for petitions which have different distributions of class sizes. Chapter 4 explores this issue using a series of simulation studies, using several distributions of class sizes.

Haas & Stokes (1998) present approximations, based on the assumption that all classes are of equal size (N/D). From this, and the approximation that the frequency of frequencies is multinomial, they derive:

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

and

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

So the asymptotic variance of the modified Esty estimator is

$$\left(\frac{\hat{D}}{d}\right)^2\left(\sum_{i=1}^{m} f_i\left(1-\frac{f_i}{\hat{D}}\right) - \sum_{i=1}^{m}\sum_{\substack{j=1 \\ j\neq i}}^{m}\frac{f_i f_j}{\hat{D}}\right)$$

So for a typical sample of 12500 signatures (1/20 of the petition) of which 50 are duplicates (so d=12450) and assuming k=1, the modified estimate of the number of signatories is 231412.6. The asymptotic variance estimate is 4069936.11, giving a standard error of the estimate of 2017.41

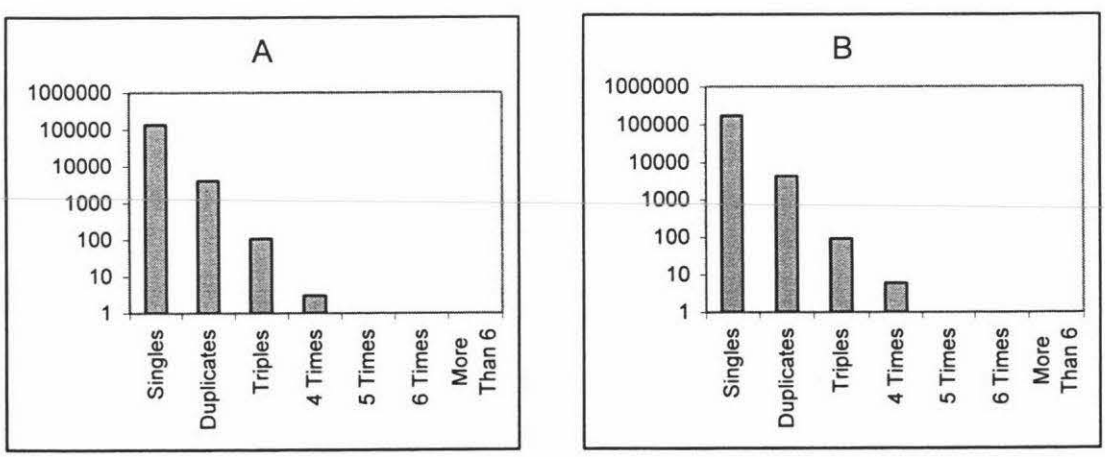
**Is the Negative Binomial Distribution Appropriate for Petitions?**

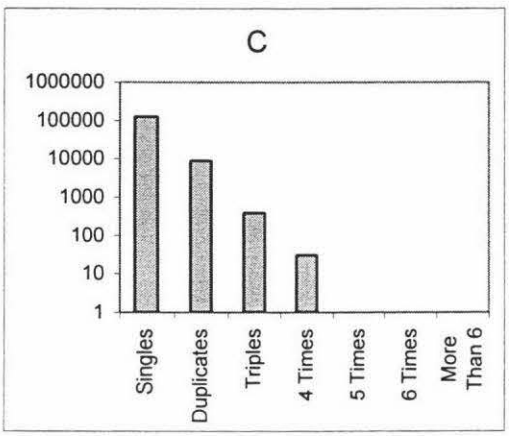
Determining the proportion of duplicate, triplicate, etc signatures in a petition from a (small) sample is complicated by the fact that an individual who has signed the petition twice is far more likely to have *one* of their signatures appear in the sample than both of them; similarly, an individual who has signed three times is more likely to have their signature appear once or even twice in the sample than for all three signatures to occur (This is discussed in more detail at the start of Chapter 4). It is possible to imagine a number of different structures of petitions (in terms of the proportion of duplicate,

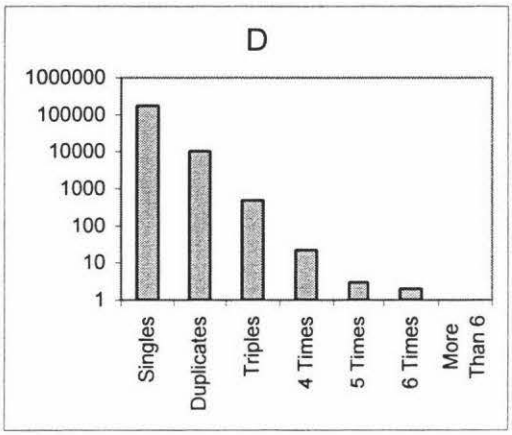
triplicate, etc signatures, and what the highest number of signatures from an individual is) which would typically generate any sample. To complicate matters further, the logistics of taking and checking a sample mean that only one sample is taken; so we have no guarantee that the sample we have *is* a typical sample from that petition. This means that the range of possible petition structures is even wider. There is not even any theory to put constraints on the structure (eg that $n_i > n_j$ if $i< j$), nor even any good reason to believe that all petitions are subject to the same constraints, even if those constraints are unknown.

Completely checking a petition (or, ideally, several petitions) for duplicate, triplicate, etc signatures, would be useful, but involve considerable effort and expense. This has not been done with any petitions in New Zealand; however, in the US state of Oregon, which has similar legislation on citizens' petitions, when a petition is in the grey area between being clearly large enough to pass or being clearly too small, the number of signatures on the entire petition is checked. Smith-Cayama & Thomas (1999) give the distribution of the number of signatures on four recent petitions. All four roughly follow a Geometric distribution, which is equivalent to a Negative Binomial distribution with k=1 (Fig 3.1 a-d. The histograms have logarithmic vertical scales; on these, a geometric distribution would appear as a constant downward slope, the gradient of which is determined by the ratio between successive categories)

**Fig 3.1 Numbers Signing Oregon Petitions (Table 1.1) Once, Twice, etc**






The Oregon results are encouraging, but they are not a strong argument that an estimator based on the assumption that the number of signatures is Geometric (or any other Negative Binomial distribution) will give useful estimates of the total number of signatories in New Zealand. It would be more reassuring to see that the estimator performs reasonably even when the underlying distribution of the number of signatures is not Negative Binomial.

**A Simulation Study**

As a rapid check on the performance of the modified Esty's estimator, three distributions were used to generate samples from a petition which had 250000 unique signatories. The sampling fraction was 8%. The three distributions were:

- a genuine Geometric distribution, with an r of 0.95 (ie 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). With 250,000 signatories, this meant the petition had 263,157 signatures on it, with individuals having signed up to 5 times. All samples had

duplicate signatures in them; 148 of the 500 samples had one, two or three triplicate signatures in them.

- a distribution where 95% of the signatories had signed once, and the remaining 5% had signed twice. With 250,000 signatories, this meant the petition had 262,500 signatures on it. All samples had duplicate samples in them; there were, naturally, no triplicates or higher. This is referred to as the Singles And Doubles or S & D petition.

- a Uniform distribution, where 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 250,000 signatories, this meant the petition had a total of 281,250 signatures on it. All of the 500 samples had one or more triplicate signatures in them; 252 of the 500 samples had between one and four quadruple signatures in them; 2 samples had a quintuple signature in them.

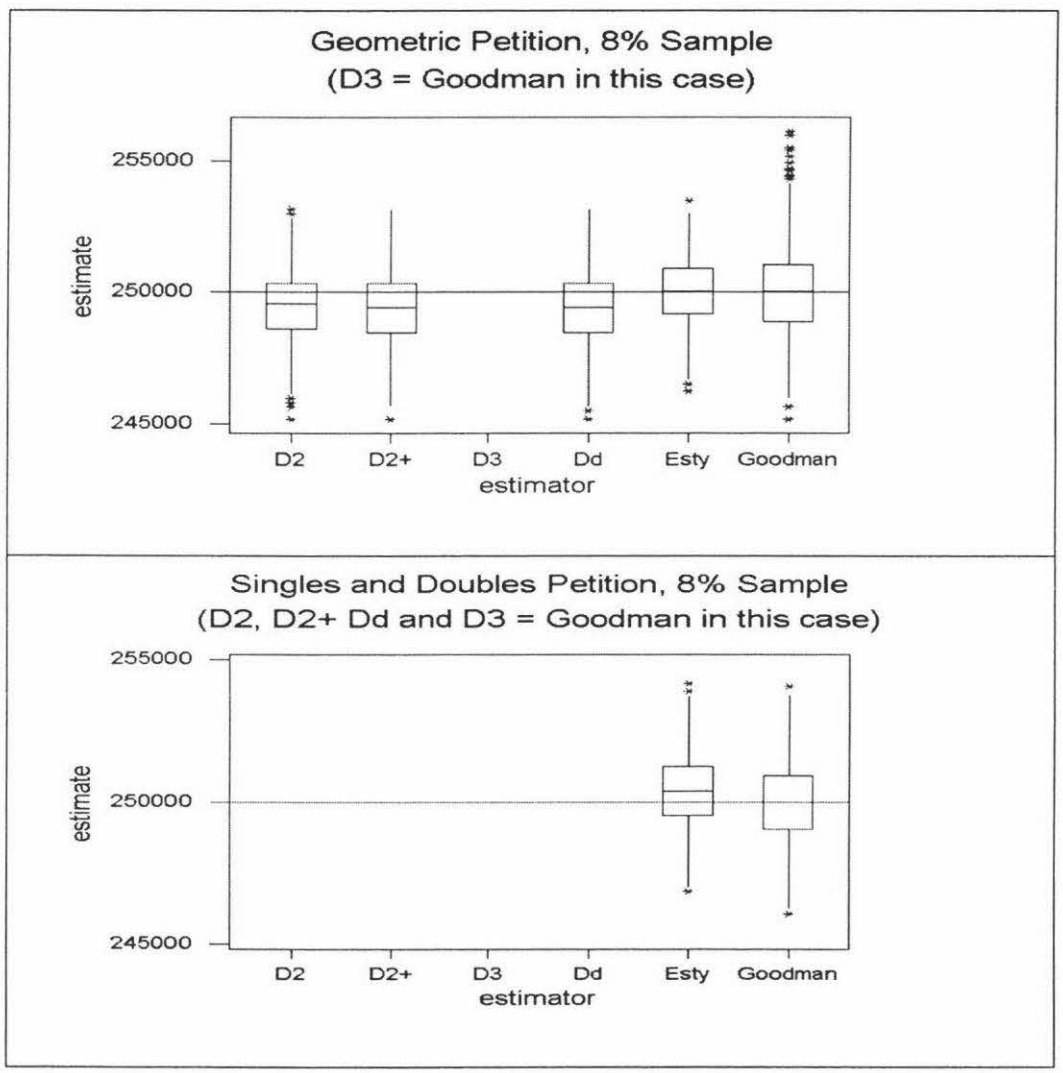
For comparison, estimates were also calculated using Goodman's estimator, $D_2$ (the variant of Goodman's estimator just using the number of duplicates), $D_3$ (a variant of Goodman's estimator just using the terms for duplicates and triplicates), $D_{2+}$ (a variant of $D_2$ mentioned by Smith-Cayama & Thomas) and $D_d$ (another variant of $D_2$ mentioned by Smith-Cayama & Thomas).

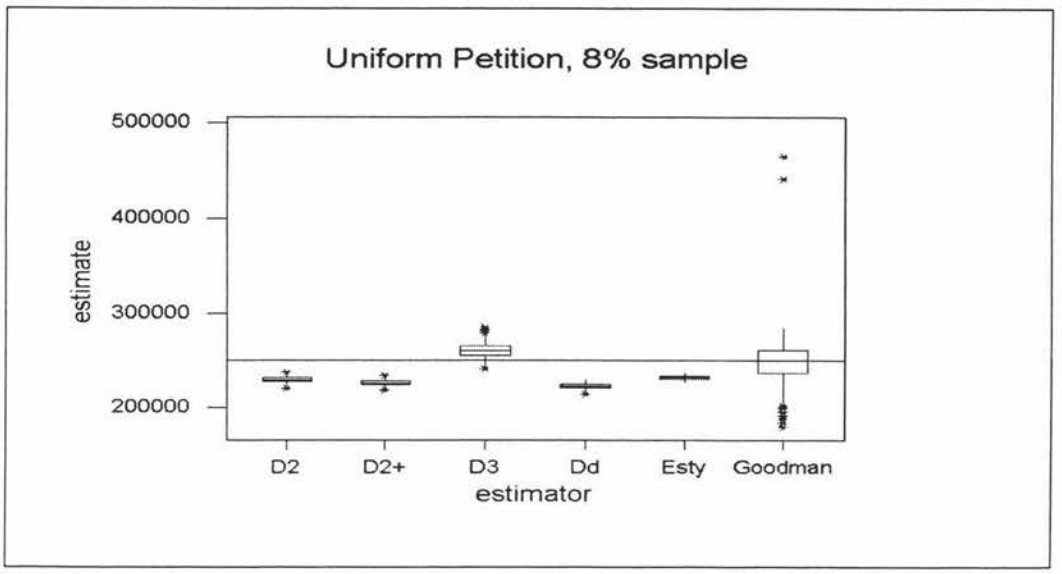
All of the estimators performed well on the Geometric distribution and the distribution with just singles and duplicates (The two distributions are very similar – under the Geometric distribution 95% of signatories signed once, 4.75% signed twice, and only 0.25% signed more than twice; whereas under the singles-and-duplicates distribution, 95% of signatories signed once and 5% signed twice).

With the Uniform distribution, the modified Esty's estimator gave very biased results, consistently underestimating the number of the signatories. Goodman's estimator gave very variable results, and tended to over-estimate the number of signatories. $D_2$, $D_{2+}$ and $D_d$ all performed similarly, producing estimates in a narrow range, but underestimating the number of signatories slightly more than the modified Esty estimator.

$D_3$ performed similarly to Goodman's estimator, tending to overestimate the number of signatories. It too was more variable than either modified Esty's estimator or $D_2$ *et al*, but it avoided the extremes of the full Goodman's estimator.

**Fig 3.2 Box-and-Whisker Plots of Estimated Number of Signatories**






**The Distribution of the Estimators**

One issue which arose during the writing of this chapter was whether the estimated standard error of the estimate can be used to obtain reasonable confidence intervals for the number of signatories. If the estimates from different samples drawn from the same population (petition) are Normally distributed, an estimate of the standard error (and the estimated number of signatories) is all that is required to calculate a confidence interval. Other distributions may require additional information.

The Chebyshev Inequality (DeGroot,1986) puts an absolute upper limit on the probability that an observation is more than a given distance from the expected value of the estimator, regardless of the distribution of the observations. However, the generality of this result means that the interval produced may be much wider than is needed for specific distributions. The Chebyshev Inequality states that

For any random variable X for which Var(X) exists, for any $t > 0$

$$p(|X - E[X]| \geq t) \leq \frac{Var(X)}{t^2}$$

For instance, a 95% confidence interval is formally $\{t \mid p(|X - E[X]| > t) = 0.05\}$. Apply Chebyshev's Inequality we find

$$0.05 \leq \frac{\text{Var}(X)}{t^2} \text{ or } t \leq \sqrt{20}\ \text{SD}(X) \simeq 4.47\ \text{SD}(X)$$

So regardless of the distribution, at least 95% of observations will fall within 4.47 Standard Deviations of the expected value. If the data were (say) Normally distributed, only 0.0008% of observations would lie outside that interval, and a considerably narrower interval would serve to describe 95% of the observations.

The simulation study results offer an opportunity to check the sampling distribution of the estimators. Normal probability plots show that most of the distributions are close to Normal, with perhaps some small deviations in the tails. A notable exception is $D_{Goodman}$ with the Uniform petition; the bulk of the data appears close to Normally distributed, but a few serious overestimates (441,307 and 464,394) stick out. However, comparing the actual range which contains 95% of the observations (ie the $2.5^{th}$ percentile to the $97.5^{th}$ percentile) with what would be projected using the observed means and SDs and assuming the data are Normally distributed, the match is generally good. A similar test was performed with the $25^{th}$ and $75^{th}$ percentiles, and again the match was generally good. The only discrepancy on the order of thousands occurs with the quartiles of the estimates from $D_{Goodman}$, which Normal theory estimates as being more widely spread than they actually are.

**Table 3.1 Comparing Normal Theory and Empirical Results from Simulations**

| Geometric Petition | | | 95% CI | | Lower Quartile | Upper Quartile |
|---|---|---|---|---|---|---|
| Mod Esty | | | | | | |
| Mean | 250029 | Normal Theory | 247474.5 | 252583.5 | 249149.9 | 250908.1 |
| StDev | 1303.34 | Percentiles | 247477.0 | 252580.9 | 249180.0 | 250887.0 |
| Goodman | | | | | | |
| Mean | 250076 | Normal Theory | 246679.7 | 253472.3 | 248907.2 | 251244.8 |
| StDev | 1732.82 | Percentiles | 246683.4 | 253469.2 | 248878.0 | 251027.0 |
| D2 | | | | | | |
| Mean | 249515 | Normal Theory | 246727.1 | 252302.9 | 248555.6 | 250474.4 |
| StDev | 1422.41 | Percentiles | 246730.0 | 252300.1 | 248624.0 | 250343.0 |

D2+ and DDup excluded because they behave similarly to D2
D3 excluded because it is equivalent to Goodman with these samples

| Singles and Doubles | | | 95% CI | | Lower Quartile | Upper Quartile |
|---|---|---|---|---|---|---|
| Mod Esty | | | | | | |
| Mean | 250440 | Normal Theory | 247921.3 | 252958.7 | 249573.2 | 251306.8 |
| StDev | 1285.05 | Percentiles | 247924.4 | 252956.6 | 249537.0 | 251256.0 |
| Goodman | | | | | | |
| Mean | 250033 | Normal Theory | 247287.3 | 252778.7 | 249088.1 | 250977.9 |
| StDev | 1400.89 | Percentiles | 247289.6 | 252775.5 | 249051.0 | 250926.0 |

D2, D3, D2+ and DDup excluded because they are equivalent to Goodman with these samples

| Uniform Petition | | | 95% CI | | Lower Quartile | Upper Quartile |
|---|---|---|---|---|---|---|
| Mod Esty | | | | | | |
| Mean | 232004 | Normal Theory | 227637.0 | 236371.0 | 230501.2 | 233506.8 |
| StDev | 2228.04 | Percentiles | 227641.2 | 236366.2 | 230480.0 | 233531.0 |
| Goodman | | | | | | |
| Mean | 248269 | Normal Theory | 205596.3 | 290941.7 | 233583.9 | 262954.1 |
| StDev | 21771.8 | Percentiles | 205640.1 | 290898.7 | 237176.0 | 260810.0 |
| D2 | | | | | | |
| Mean | 229509 | Normal Theory | 223946.7 | 235071.3 | 227594.8 | 231423.2 |
| StDev | 2837.92 | Percentiles | 223952.7 | 235066.0 | 227659.0 | 231409.0 |
| D3 | | | | | | |
| Mean | 260387 | Normal Theory | 246110.0 | 274664.0 | 255473.8 | 265300.2 |
| StDev | 7284.19 | Percentiles | 246124.9 | 274649.8 | 255432.0 | 265002.0 |

D2+ and DDup were excluded because they behave similarly to D2

# Chapter 4

## Simulation Studies

The simulation program developed to investigate the behaviour of the modified Esty's estimator and compare it to other estimators (Ch 3 and Appendix 2) can be used for other comparisons: for instance, to check how changing the sampling fraction affects the estimates (and thus whether increasing it slightly from the ~8% presently used in New Zealand would substantially improve the accuracy of the conclusions; or conversely, whether using a lower sampling fraction would reduce the workload without substantially affecting the accuracy of the result); or to check how well the variance estimates (Appendix 1 and Ch 3) reflect the actual variability of the results. But first we will look at the issue of bias in $D_2$ , $D_{2+}$, $D_{Dup}$ , and $D_{Mod\ Esty}$.

### Why are $D_{Goodman\ 2}$, $D_{Goodman\ 2+}$ and $D_{Dup}$ Biased?

It has already been noted that since $D_{Goodman}$ is the sole unbiased estimator of the number classes in the population, $D_{Goodman\ 2}$, $D_{Goodman\ 2+}$ and $D_{Dup}$ must be biased estimators. In the simulation study the bias is comparatively small with the Geometric petition; with the Singles-and-Doubles petition, there are only single and double signatures in the sample, so $D_{Goodman\ 2}$, $D_{Goodman\ 2+}$ and $D_{Dup}$ are equivalent to $D_{Goodman}$, and so are not biased; with the Uniform petition the size of the petition is underestimated by between 20000 and 30000.

$D_{Goodman\ 2}$, $D_{Goodman\ 2+}$ and $D_{Dup}$ are simplifications of $D_{Goodman}$, and the main reason for the bias is what the simplification process has done to $D_{Goodman}$. If there are individuals who have signed the petition more than twice, then some of their signatures will appear twice in the sample. In fact, with small sampling fractions, an individual who has signed the petition more than twice is more likely to have their signature appear exactly twice in the sample than an individual who has signed twice . At first sight, this appears fair, because the individual who has signed more than twice is also causing a big discrepancy between the number of signatures in the petition (N) and the number of

signatories (D), and so the higher probability of their signature appearing twice means that a bigger correction will be made to $\hat{D}$. Unfortunately, only in very specific circumstances do the two considerations cancel each other out and with small sampling fractions the adjustments are often larger than are needed (see below). In Goodman's full estimator, the later terms include an adjustment for any over-correction in the earlier terms, which is why the form of the later terms becomes more and more complex. However, $D_{Goodman2}$, $D_{Goodman2+}$ and $D_{Dup}$ do not include these later terms, and so at low sampling fractions the over-correction remains unadjusted

$D_{Goodman2+}$ and $D_{Dup}$ exacerbate the problem by treating signatures which appear more than twice in the sample as double signatures, and making an additional correction for them.

The problem is more extreme with the Uniform petition than with the Geometric petition because the former has many more signatures appearing three or more times in the petition, and so the over-correction is larger.

Formally, we can consider the effect that having d people signing the petition j times has on each of the estimators. If they are the only people who have signed multiple times, then number of signatories (D) is the number of signatures (N) minus d(j-1).

If a sample of n (=fN) signatures is taken, then we would expect

$d\frac{j!}{2!(j-2)!}f^2(1-f)^{j-2}$ of the people who have signed the petition j times to have their signatures appear exactly twice in the sample. Thus we would expect $\hat{D}_2$ to be

$$N-\frac{N(N-1)}{n(n-1)}d\frac{j!}{2(j-2)!}\left(\frac{n}{N}\right)^2\left(1-\frac{n}{N}\right)^{j-2}$$

$$\approx N-\frac{dj(j-1)}{2}(1-f)^{j-2}$$

So if $\frac{j}{2}(1-f)^{j-2} > 1$ $\hat{D}_2$ will tend to underestimate the true number of signatories;

conversely, if $\frac{j}{2}(1-f)^{j-2} < 1$ it will tend to overestimate the true number of signatories.

| For j= | $\hat{D}_2$ will tend to underestimate the true number of signatories if... |
|---|---|
| 3 | f<0.33 |
| 4 | f<0.29 |
| 5 | f<0.26 |
| 6 | f<0.24 |

To do a similar calculation for $\hat{D}_{2+}$ we need to calculate the expected number of people who have signed the petition j times whose signatures appear two or more times in the sample. This is simply d minus the expected number whose signatures either do not appear in the sample, or only appear once in the sample, ie $d-d(1-f)^j - djf(1-f)^{j-1}$. Thus the expected value of $\hat{D}_{2+}$ would be

$$N - \frac{N(N-1)}{n(n-1)} d\left(1-(1-f)^j - jf(1-f)^{j-1}\right)$$

$$\approx N - \frac{d}{f^2} + \frac{d(1-f)^j}{f^2} + \frac{dj(1-f)^{j-1}}{f}$$

So if $\frac{d}{f^2} - \frac{d(1-f)^j}{f^2} - \frac{dj(1-f)^{j-1}}{f} > d(j-1)$, $\hat{D}_{2+}$ will underestimate the number of signatories; conversely, if $\frac{d}{f^2} - \frac{d(1-f)^j}{f^2} - \frac{dj(1-f)^{j-1}}{f} < d(j-1)$, $\hat{D}_{2+}$ will overestimate the number of signatories

| For j= | $\hat{D}_{2+}$ will tend to underestimate the true number of signatories if... |
|---|---|
| 3 | f<0.50 |
| 4 | f<0.45 |
| 5 | f<0.41 |
| 6 | f<0.39 |

For $\hat{D}_{Dup}$ we need to calculate the expected number of signatures of people who have signed the petition j times which appear multiple times in the sample (ie for individuals whose signatures appear twice in the sample, one is a duplicate; for individuals whose signatures appear three times, two are duplicates, etc). The number of signatures which appear two or more times in the sample (from the people who have signed j times) is the expected number of signatures from the group minus the expected number of signatures which appear only once; ie $djf - djf(1-f)^{j-1}$ . The expected number of individuals whose signatures appear more than once in the sample is $d-d(1-f)^{j} - djf(1-f)^{j-1}$ . So the expected number of duplicate signatures is

$$djf - djf(1-f)^{j-1} - (d-d(1-f)^{j} - djf(1-f)^{j-1})$$

Thus the expected value of $\hat{D}_{Dup}$ would be

$$N - \frac{N(N-1)}{n(n-1)} d\left(jf - 1 + (1-f)^{j}\right)$$

$$\approx N - \frac{dj}{f} + \frac{d}{f^2} - \frac{d(1-f)^{j}}{f^2}$$

So if $\frac{djf}{f^2} - \frac{d}{f^2} + \frac{d(1-f)^{j}}{f^2} > d(j-1)$, $\hat{D}_{Dup}$ will underestimate the number of signatories;

for values of j between 3 and 6, $\frac{djf}{f^2} - \frac{d}{f^2} + \frac{d(1-f)^{j}}{f^2} > d(j-1)$ for all f<1.

**Why is $D_{Mod\ Esty}$ Biased?**

$D_{Mod\ Esty}$ was derived assuming that the distribution of the number of times individuals had signed the petition followed an exponential decay. This is a strong assumption; having made it, one can estimate D on the basis of just N, q and the number of duplicate signatures in the sample (n-d) without needing to know the distribution of the multiple signatures (ie how many appear twice, three times, etc in the sample). Whatever the number of duplicate signatures in the sample, one can work back to an estimate of D*on the assumption that the frequency of frequencies decays exponentially.* If that assumption is wrong, then the estimate of D will be wrong also. For instance, in the Uniform petition, there are a higher proportion of people who have signed the petition three or more times compared to those who have signed twice, than there would be if there were the same number of duplicate signatures (ie N-D) in the petition, but the distribution of the number of times individuals had signed the petition followed an exponential decay. As noted above, people who have signed the petition three or more times are disproportionately more likely to have their signatures appear multiple times in the sample; thus the sample contains more multiple signatures than there would be if there were the same number of duplicate signatures (ie N-D) in the petition, but the distribution of the number of times individuals had signed the petition followed an exponential decay; in other words, the sample looks like a sample from a petition where the distribution of the number of times individuals had signed the petition followed an exponential decay, but with more duplicate signatures than there actually are. Thus $D_{Mod\ Esty}$ over-corrects the number of duplicate signatures.

In a petition with a lower proportion of people who have signed three or more times compared to those who have signed twice than would be predicted by an exponential decay, $D_{Mod\ Esty}$ would under-correct, producing estimates which are too high. This occurs with the Singles-and-Doubles petition, but only on a small scale, as the difference between the Singles-and-Doubles distribution and an exponential decay is comparatively small.

**Bias Adjustment Factors**

Clearly, bias is potentially a major problem with $D_{Goodman2}$, $D_{Goodman2+}$, $D_{Dup}$ and $D_{Mod\ Esty}$. Smith-Cayama & Thomas (1999) propose a Bias Adjustment Factor

$$B^{\hat{D}}_{q,k,r} = \frac{D_k}{E(\hat{D} \mid q,k)}$$

q is the sampling fraction;

k is the maximum number of times a signature appears in the petition[1];

$r = (r_3, r_4, ..., r_k)$ and $r_i = N_i/N_2$

In Smith-Cayama's notation, $D_k = N_2 + 2N_3 + 3N_4 + ... + (k-1)N_k = N - D$ is the number of duplicate signatures in the petition [2]

The catch is that in order to calculate $E(\hat{D}|q,k)$ you need to know the distribution of the number of times individuals have signed the petition, or at least make some assumptions about it.

We do not have a complete count of any New Zealand petitions; however, the four petitions from Oregon quoted by Smith-Cayama & Thomas are of a similar total size to New Zealand petitions, and the expected numbers of duplicates, triplicates etc in a

---

[1] Note, this is different from the k used as the shape parameter in the Negative Binomial distribution in Chapter 3

[2] So our existing estimators $D = N - D_k$ .

$D_{k\ adj} = B \times D_k$ should be a better estimator of the number of duplicate signatures; so $N - D_{k\ adj}$ should be a better estimator of the number of individuals signing.

Since $D_{k\ adj} = B \times D_k = B \times (N - D)$, this means that

$N - B \times (N - D) = (1 - B)N + BD$

should be a better estimator of the number of individuals signing in our notation

sample from them are similar to the actual results obtained in New Zealand, making them reasonable candidates as sources of information for calculating B.

Smith-Cayama & Thomas present formulae for calculating B for $D_{Goodman2}$, $D_{Goodman2+}$ and $D_{Dup}$[3]. For $D_{Mod\ Esty}$, a first order approximation has been derived (See Appendix 3), although the results below are based on simulation as the approximation did not appear to work well for one of the petitions.

Calculating B on the basis of the four Oregon petitions gives the following values:

**Table 4.1 Bias Adjustment Factors**

| q | Estimator | Min | Max | Geometric Mean |
|---|---|---|---|---|
| 5% | D Goodman2 | 0.959 | 0.977 | 0.968 |
| | D Goodman2+ | 0.956 | 0.976 | 0.965 |
| | D Dup | 0.954 | 0.975 | 0.963 |
| | D Mod Esty | 0.931 | 1.008 | 0.982 |
| 8% | D Goodman2 | 0.963 | 0.980 | 0.972 |
| | D Goodman2+ | 0.959 | 0.978 | 0.968 |
| | D Dup | 0.955 | 0.975 | 0.964 |
| | D Mod Esty | 0.925 | 1.012 | 0.980 |
| 10% | D Goodman2 | 0.966 | 0.981 | 0.974 |
| | D Goodman2+ | 0.961 | 0.979 | 0.970 |
| | D Dup | 0.956 | 0.976 | 0.965 |
| | D Mod Esty | 0.926 | 1.003 | 0.977 |

If we then apply these to the estimators, we find:

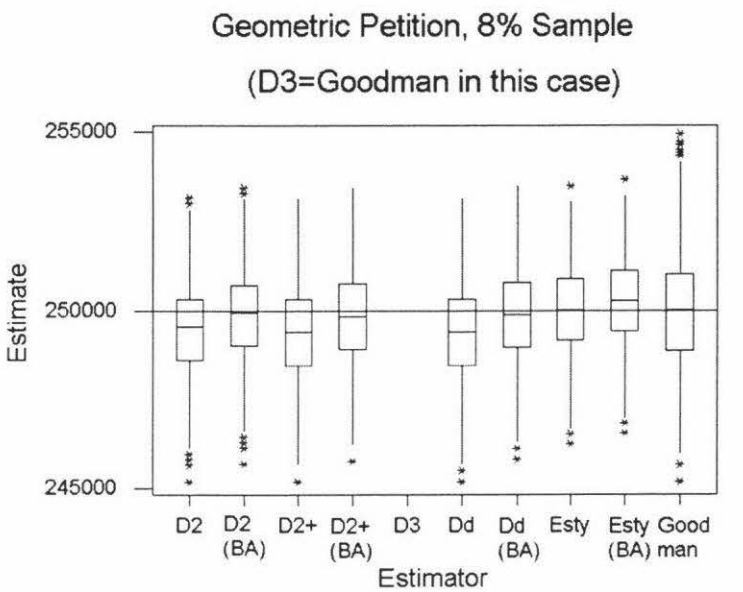
---

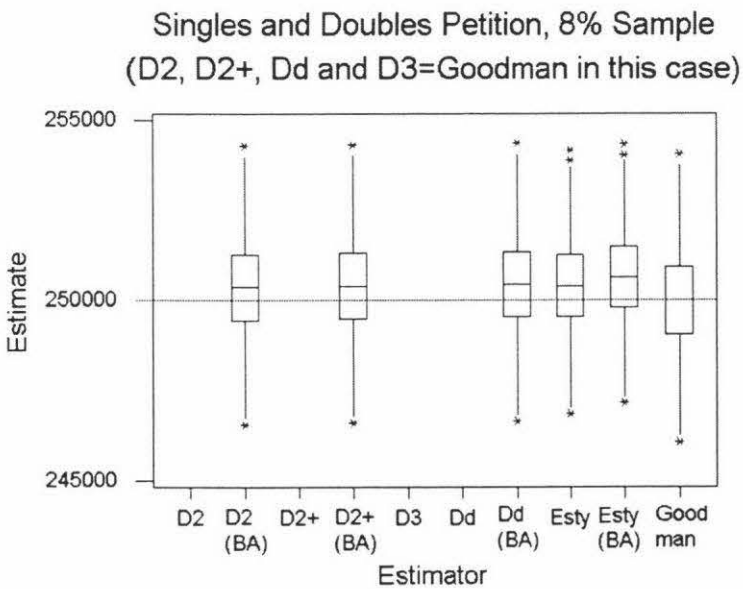
[3] There appears to be a typographical error in the formula for B for $D_{Goodman2+}$. It should read

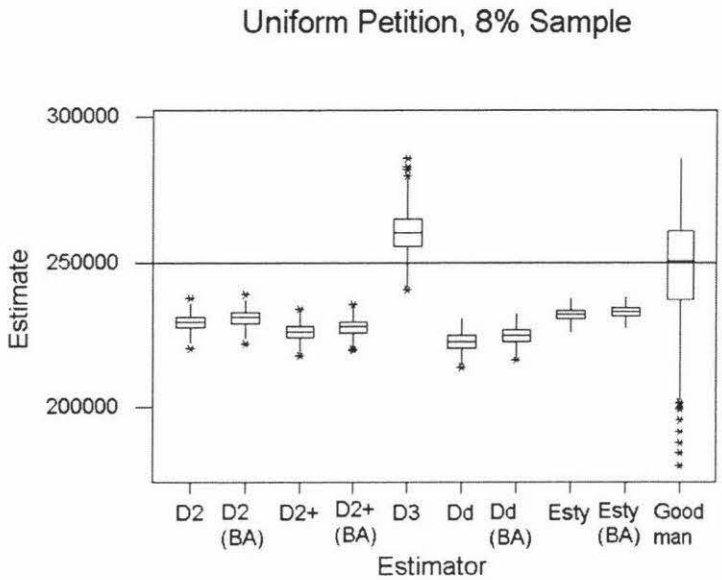
$$B^{\hat{D}_{Goodman2+}}_{q,k,r} = \frac{1+\sum_{i=3}^{k}(i-1)r_i}{1+\frac{1}{q^2}\sum_{i=3}^{k}\left[1-\{1+(i-1)q\}(1-q)^{i-1}\right]r_i}$$

The figures for B in Table 4 are consistent with this formula rather than the printed formula.

**Fig 4.1 Box-and-Whisker Plots of Estimated Number of Signatories with and without Bias Adjustment (BA)**






With the Geometric petition, Bias Adjustment virtually eliminates the (small) bias in $D_{Goodman2}$, $D_{Goodman2+}$ and $D_{Dup}$ (which is hardly surprising, given that the Geometric petition was designed to be similar to the four Oregon petitions used to derive the Bias Adjustment Factors). With the Singles-and-Doubles petition, where $D_{Goodman2}$, $D_{Goodman2+}$ and $D_{Dup}$ are equivalent to the unbiased estimator, the bias adjustment tends to *increase* their bias. For $D_{Mod\ Esty}$, the Bias Adjustment appears, paradoxically, to increase the bias with these two petitions. With the Uniform petition, all the Bias Adjusted figures are an improvement over the unadjusted estimates, but not by a great deal.

**Bias – Conclusions**

We have seen that $D_{Goodman2}$ and $D_{Goodman2+}$ tend to underestimate the number of unique signatures in petitions at low (<0.33) sampling fractions, while $D_{Dup}$ always tends to underestimate the number of unique signatures, and the performance of $D_{Mod\ Esty}$ depends on the distribution of the number of individuals signing multiple times; if there are more multiple signatories than would be expected from a Geometric distribution, it tends to underestimate the number of unique signatures; while if there are fewer, it tends to overestimate.

The Bias Adjustment Factors make a small improvement to the performance of $D_{Goodman2}$ , $D_{Goodman2+}$ and $D_{Dup}$ . This is sufficient to substantially reduce their (small) bias on petitions similar to the ones the BAFs were calculated from; and although the improvement is much less marked with Uniform petition, it is encouraging that they do provide an improvement even when the distribution of signatures in the actual petition is different from the petitions they were calculated from.

The results from the Singles-and-Doubles petition, however, are a warning against blindly applying the BAFs; if a sample only contains single and double signatures, $D_{Goodman2}$ , $D_{Goodman2+}$ and $D_{Dup}$ are equivalent to the unbiased estimator, and will not benefit from adjusting.

For $D_{Mod\ Esty}$ the BAFs do not appear to improve performance.

**Variance Estimates**

Based on the 8% samples used in Ch 3, the following variance estimates were calculated. The Root Mean Squared Error of the estimators was also calculated.

The estimates of the standard errors are actually slightly higher than the RMSEs for all the estimators for the Geometric and Singles-and-Doubles samples.

For the Uniform sample, the estimated SEs $D_{Goodman}$ and $D_{Goodman3}$ underestimate the actual RMSEs, although for $D_{Goodman3}$, the estimate is close to the estimator's variability about its own mean. For $D_{Goodman\ 2}$, $D_{Goodman\ 2+}$ and $D_{Dup}$ the estimated SEs are barely larger than those for the Geometric and Singles-and-Doubles petitions; not only do they not come within an order of magnitude of the RMSE (which is high largely because of bias), but they underestimate even the variation about their own means. For $D_{Mod\ Esty}$ the estimated SEs are actually smaller than for the Geometric and Singles-and-Doubles petition; they give no indication of the large bias, and underestimate even the variation about the mean estimate.

**Table 4.2 Root Mean Squared Errors and Estimated Standard Errors for 8% Samples**

| | Geometric | | | | | |
|---|---|---|---|---|---|---|
| | Mod Esty | Goodman | D 2 | D 3 | D 2+ | D Dup |
| RMSE | 1302.4 | 1732.8 | 1501.5 | 1732.8 | 1523.2 | 1553.2 |
| Standard Error Estimates | | | | | | |
| Min | 1627.8 | 1762.1 | 1761.8 | 1762.1 | 1762.1 | 1762.1 |
| L Quartile | 1647.3 | 1768.7 | 1768.2 | 1768.7 | 1768.3 | 1768.3 |
| Median | 1652.9 | 1771.8 | 1770.1 | 1771.8 | 1770.4 | 1770.4 |
| U Quartile | 1658.5 | 2332.1 | 1772.2 | 2332.1 | 1772.5 | 1772.5 |
| Max | 1675.7 | 3175.1 | 1779.8 | 3175.1 | 1779.8 | 1779.8 |

| | Singles and Doubles | | | | | |
|---|---|---|---|---|---|---|
| | Mod Esty | Goodman | D 2 | D 3 | D 2+ | D Dup |
| RMSE | 1357.2 | 1399.9 | 1399.9 | 1399.9 | 1399.9 | 1399.9 |
| Standard Error Estimates | | | | | | |
| Min | 1633.9 | 1756.3 | 1756.3 | 1756.3 | 1756.3 | 1756.3 |
| L Quartile | 1651.6 | 1763.3 | 1763.3 | 1763.3 | 1763.3 | 1763.3 |
| Median | 1657.3 | 1765.4 | 1765.4 | 1765.4 | 1765.4 | 1765.4 |
| U Quartile | 1663.0 | 1767.5 | 1767.5 | 1767.5 | 1767.5 | 1767.5 |
| Max | 1682.3 | 1774.1 | 1774.1 | 1774.1 | 1774.1 | 1774.1 |

| | Uniform | | | | | |
|---|---|---|---|---|---|---|
| | Mod Esty | Goodman | D 2 | D 3 | D 2+ | D Dup |
| RMSE | 18133.5 | 21818.8 | 20685.8 | 12682.7 | 24033.6 | 27521.1 |
| SD about own mean | 2228.0 | 21771.8 | 2837.9 | 7284.2 | 2915.5 | 3168.3 |
| Standard Error Estimates | | | | | | |
| Min | 1444.9 | 5392.9 | 1889.2 | 4707.4 | 1898.9 | 1907.4 |
| L Quartile | 1474.1 | 7227.2 | 1903.3 | 6730.0 | 1911.6 | 1920.4 |
| Median | 1484.1 | 18241.1 | 1907.1 | 7226.7 | 1915.8 | 1924.6 |
| U Quartile | 1493.3 | 19108.7 | 1911.5 | 7692.7 | 1920.1 | 1929.5 |
| Max | 1520.8 | 202084.6 | 1927.2 | 9198.2 | 1933.6 | 1942.7 |

One would expect the Bias Adjusted estimates to have lower variance than the original estimates, because $D_{Adjusted} = (1 - B)N + BD$; however, since $Var(D_{Adjusted}) = B^2 Var(D)$, the difference is not large.

**Table 4.3 Root Mean Squared Errors and Estimated Standard Errors for 8% Samples with Bias Adjustments**

| | Geometric | | | |
|---|---|---|---|---|
| | Mod Esty (BA) | D2 (BA) | D2+ (BA) | DDup (BA) |
| RMSE | 1351.6 | 1385.0 | 1381.1 | 1384.1 |
| Standard Error Estimates | | | | |
| Min | 1569.2 | 1712.5 | 1705.7 | 1698.6 |
| L Quartile | 1588.0 | 1718.7 | 1711.8 | 1704.7 |
| Median | 1593.4 | 1720.5 | 1713.7 | 1706.7 |
| U Quartile | 1598.8 | 1722.6 | 1715.8 | 1708.7 |
| Max | 1615.4 | 1730.0 | 1722.9 | 1715.8 |

| | Singles and Doubles | | | |
|---|---|---|---|---|
| | Mod Esty (BA) | D2 (BA) | D2+ (BA) | DDup (BA) |
| RMSE | 1515.4 | 1412.8 | 1421.8 | 1432.4 |
| Standard Error Estimates | | | | |
| Min | 1575.1 | 1707.1 | 1700.1 | 1693.1 |
| L Quartile | 1592.2 | 1713.9 | 1706.9 | 1699.8 |
| Median | 1597.6 | 1716.0 | 1708.9 | 1701.8 |
| U Quartile | 1603.1 | 1718.0 | 1710.9 | 1703.9 |
| Max | 1621.7 | 1724.4 | 1717.3 | 1710.2 |

| | Uniform | | | |
|---|---|---|---|---|
| | Mod Esty (BA) | D2 (BA) | D2+ (BA) | DDup (BA) |
| RMSE | 16364.9 | 19240.3 | 22272.2 | 25413.1 |
| SD about own mean | 2185.26 | 2757.04 | 2822.20 | 3055.48 |
| Standard Error Estimates | | | | |
| Min | 1417.4 | 1836.3 | 1838.1 | 1838.8 |
| L Quartile | 1446.1 | 1850.0 | 1850.4 | 1851.2 |
| Median | 1455.9 | 1853.7 | 1854.5 | 1855.3 |
| U Quartile | 1464.9 | 1858.0 | 1858.6 | 1860.0 |
| Max | 1491.9 | 1873.2 | 1871.7 | 1873.7 |

**Sampling Fractions**

The simulations were repeated with the same petitions, but with either a 5% or a 10% sampling fraction. The overall pattern of results was similar to the 8% sampling fraction: all estimators performed well, with little bias on the Geometric and Singles-and-Doubles petitions. On the Uniform petition, $D_{Mod\ Esty}$, $D_{Goodman\ 2}$, $D_{Goodman\ 2+}$ and $D_{Dup}$ all showed considerable downwards bias; $D_{Goodman\ 3}$ showed a smaller, upwards bias; and $D_{Goodman}$ showed no bias but greater variability than the other estimators.

Interestingly, although signatures which appear three or more times in the sample are seen to be a major source of variability in Goodman's estimator (which is why the various modifications of $D_{Goodman}$ exclude them, or treat them like signatures which appear twice), the estimates from the 10% samples are less
variable than those from the 5% sample, even though the 10% samples contain more triple signatures[4] .

Comparing the results, one might expect the figures from the 10% sample to be approximately 71% of the figures from the 5% sample (since the sample size has doubled, the variance should have halved, and the standard errors dropped by $1/\sqrt{2}$ ; if one incorporates a finite population correction, the figure should be approximately 69% ). Many of the RMSEs have in fact halved: for instance, the RMSEs for all the estimators on the Singles-and-Doubles petition, and the RMSEs for $D_{Mod\ Esty}$, $D_{Goodman2}$, $D_{Goodman2+}$ and $D_{Dup}$ on the Geometric petition. The RMSEs for $D_{Goodman}$ and $D_{Goodman3}$ on the Geometric petition have more than halved. With the Uniform petition, where $D_{Mod\ Esty}$, $D_{Goodman2}$, $D_{Goodman2+}$ and $D_{Dup}$ show serious bias, increasing the sample size doses not reduce the RMSE as markedly: the figures from the 10% sample are between 74% and 94% of the figures from the 5% sample. Presumably this is because increasing the sample size does not reduce the bias component of RMSE as much as it reduces the

---

[4] With the Geometric petition, 9.4% of the 5% samples included triple signatures, whereas 54.8% of the 10% samples did – and 2 of the 500 contained quadruple signatures.
With the Uniform petition, 99.4% of the 5% samples contained triple signatures and 11.4% contained quadruple signatures, whereas 100% of the 10% samples contained triple signatures, and 81.0% contained quadruple signatures

**Table 4.4 Root Mean Squared Errors and Estimated Standard Errors for 5% Samples**

| | Geometric | | | | | |
|---|---|---|---|---|---|---|
| | Mod Esty | Goodman | D 2 | D 3 | D 2+ | D Dup |
| RMSE | 2071.3 | 3268.5 | 2363.9 | 3268.5 | 2365.1 | 2373.5 |
| Standard Error Estimates | | | | | | |
| Min | 2065.7 | 2261.9 | 2261.9 | 2261.9 | 2261.9 | 2261.9 |
| L Quartile | 2113.6 | 2278.0 | 2276.7 | 2278.0 | 2276.7 | 2276.8 |
| Median | 2125.8 | 2283.4 | 2282.0 | 2283.4 | 2282.0 | 2282.0 |
| U Quartile | 2138.3 | 2290.1 | 2287.4 | 2290.1 | 2287.4 | 2287.4 |
| Max | 2173.1 | 9970.6 | 2308.7 | 9970.6 | 2308.7 | 2308.7 |

| | Singles and Doubles | | | | | |
|---|---|---|---|---|---|---|
| | Mod Esty | Goodman | D 2 | D 3 | D 2+ | D Dup |
| RMSE | 2064.3 | 2224.7 | 2224.7 | 2224.7 | 2224.7 | 2224.7 |
| Standard Error Estimates | | | | | | |
| Min | 2080.5 | 2253.6 | 2253.6 | 2253.6 | 2253.6 | 2253.6 |
| L Quartile | 2119.9 | 2271.2 | 2271.2 | 2271.2 | 2271.2 | 2271.2 |
| Median | 2129.2 | 2276.5 | 2276.5 | 2276.5 | 2276.5 | 2276.5 |
| U Quartile | 2141.7 | 2280.6 | 2280.6 | 2280.6 | 2280.6 | 2280.6 |
| Max | 2183.3 | 2297.9 | 2297.9 | 2297.9 | 2297.9 | 2297.9 |

| | Uniform | | | | | |
|---|---|---|---|---|---|---|
| | Mod Esty | Goodman | D 2 | D 3 | D 2+ | D Dup |
| RMSE | 18860.7 | 195563.4 | 24712.5 | 20188.1 | 26864.8 | 29104.6 |
| SD about own mean | 3419.0 | 195442.0 | 4649.2 | 16312.4 | 4703.1 | 4944.7 |
| Standard Error Estimates | | | | | | |
| Min | 1816.6 | 2459.7 | 2445.3 | 2459.7 | 2456.6 | 2459.7 |
| L Quartile | 1883.5 | 13946.6 | 2478.6 | 13945.7 | 2485.9 | 2493.0 |
| Median | 1901.6 | 16987.1 | 2487.8 | 15544.8 | 2495.8 | 2504.2 |
| U Quartile | 1922.2 | 19564.7 | 2498.4 | 18322.7 | 2505.9 | 2514.0 |
| Max | 1984.4 | 2481865.9 | 2540.8 | 24868.2 | 2543.6 | 2553.3 |

**Table 4.5 Root Mean Squared Errors and Estimated Standard Errors for 10% Samples**

| | Geometric | | | | | |
|---|---|---|---|---|---|---|
| | Mod Esty | Goodman | D 2 | D 3 | D 2+ | D Dup |
| RMSE | 1059.0 | 1365.5 | 1237.0 | 1286.5 | 1274.6 | 1322.1 |
| Standard Error Estimates | | | | | | |
| Min | 1439.8 | 1558.4 | 1557.9 | 1558.4 | 1558.0 | 1558.0 |
| L Quartile | 1457.6 | 1563.0 | 1561.4 | 1563.0 | 1561.4 | 1561.4 |
| Median | 1462.4 | 1722.6 | 1562.7 | 1722.6 | 1562.8 | 1562.8 |
| U Quartile | 1466.2 | 1726.6 | 1564.0 | 1726.6 | 1564.0 | 1564.0 |
| Max | 1476.4 | 6826.0 | 1570.1 | 2140.3 | 1570.1 | 1570.1 |

| | Singles and Doubles | | | | | |
|---|---|---|---|---|---|---|
| | Mod Esty | Goodman | D 2 | D 3 | D 2+ | D Dup |
| RMSE | 1127.6 | 1073.6 | 1073.6 | 1073.6 | 1073.6 | 1073.6 |
| Standard Error Estimates | | | | | | |
| Min | 1446.1 | 1554.1 | 1554.1 | 1554.1 | 1554.1 | 1554.1 |
| L Quartile | 1463.6 | 1557.6 | 1557.6 | 1557.6 | 1557.6 | 1557.6 |
| Median | 1466.9 | 1559.0 | 1559.0 | 1559.0 | 1559.0 | 1559.0 |
| U Quartile | 1471.2 | 1560.0 | 1560.0 | 1560.0 | 1560.0 | 1560.0 |
| Max | 1482.1 | 1565.8 | 1565.8 | 1565.8 | 1565.8 | 1565.8 |

| | Uniform | | | | | |
|---|---|---|---|---|---|---|
| | Mod Esty | Goodman | D 2 | D 3 | D 2+ | D Dup |
| RMSE | 17649.6 | 16440.0 | 18393.7 | 10372.1 | 22459.2 | 26729.7 |
| SD about own mean | 1734.5 | 16456.5 | 2090.0 | 5090.7 | 2159.7 | 2434.5 |
| Standard Error Estimates | | | | | | |
| Min | 1287.0 | 4014.8 | 1661.9 | 3809.9 | 1663.2 | 1663.9 |
| L Quartile | 1307.8 | 7997.7 | 1669.6 | 4572.0 | 1670.5 | 1671.5 |
| Median | 1314.9 | 10215.7 | 1671.7 | 4852.9 | 1672.9 | 1673.9 |
| U Quartile | 1321.5 | 12234.5 | 1674.2 | 5068.5 | 1675.4 | 1676.4 |
| Max | 1346.3 | 84965.3 | 1682.1 | 5891.6 | 1682.9 | 1683.6 |

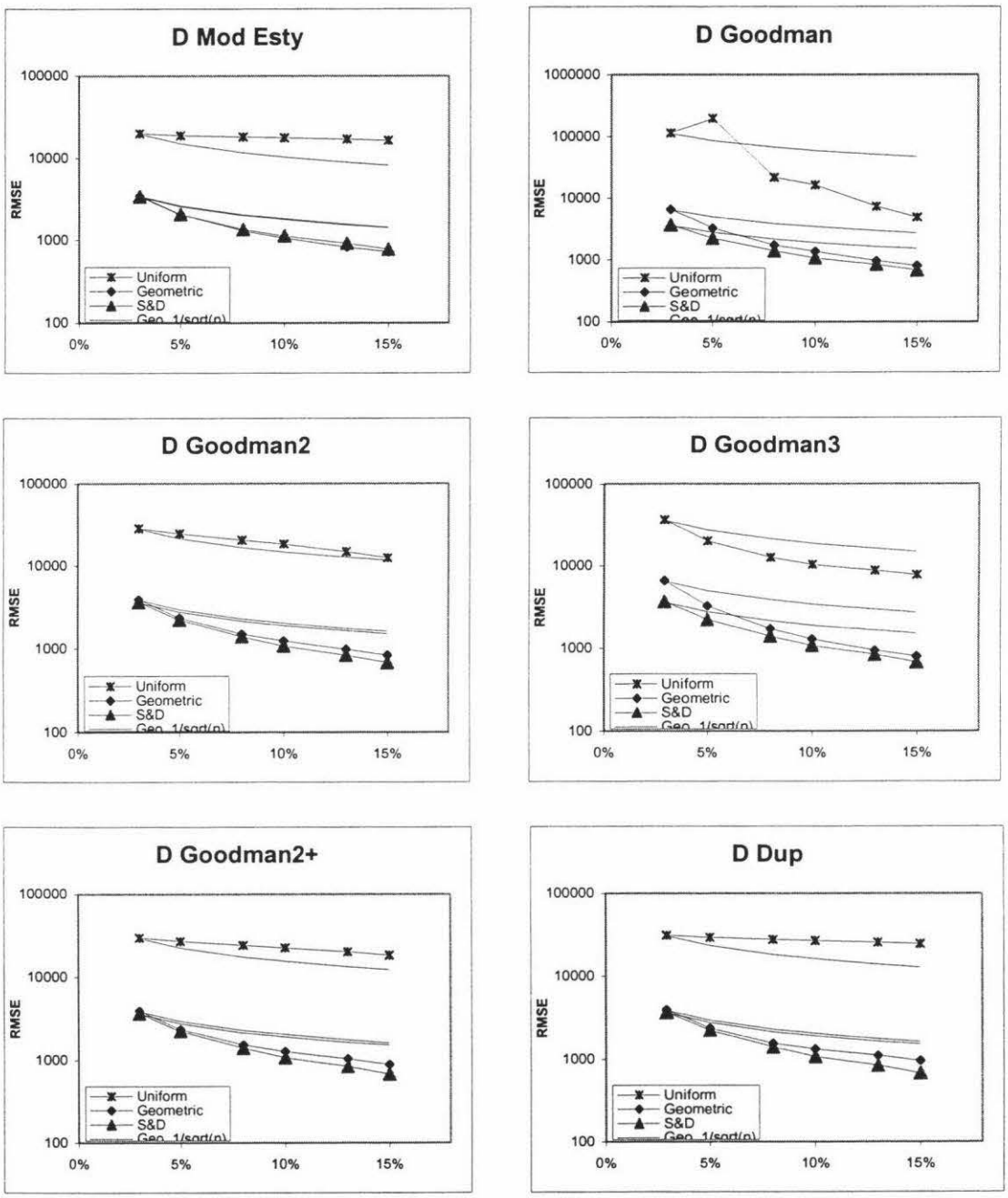
random variation component. For $D_{Goodman3}$, doubling the sample size halves the RMSE for the Uniform petition; for $D_{Goodman}$, the RMSE from the 10% sample is only 8% of the RMSE from the 5% sample.

Simulations were also run with sampling fractions of 3%, 13% and 15%. Plotting the RMSEs against sample size (with traces to indicate the reduction which could be expected due to increase in sample size and the finite population correction alone, Fig 4.2) confirms the observations above; all the estimators show considerable improvement (faster than $1/\sqrt{n}$, even allowing for the finite population) as sample size increases for the Geometric and Singles-and-Doubles petition. With the Uniform petition, $D_{Goodman}$ and $D_{Goodman3}$ improve faster than the sample size alone would suggest; $D_{Goodman2}$ improves at about the rate of $1/\sqrt{n}$; while $D_{Mod\ Esty}$, $D_{Goodman2+}$ and $D_{Dup}$ only improve very slowly because of the role of bias in their RMSEs. However, the graphs suggest that all the estimators improve their accuracy gradually as the sample size increases; at least in the range 5% to 15% there is no evidence that there is a threshold sampling fraction above which accuracy markedly improves.

An estimate of the Standard Error for each estimator can be calculated from each simulation, using Haas & Stokes' delta method approximation. For all three petitions, the median and the maximum Standard Error Estimates for $D_{Goodman2}$, $D_{Goodman2+}$, $D_{Dup}$ and $D_{Mod\ Esty}$ with a 10% sampling fraction are between 66% and 68% of the estimates from a 5% sample – roughly the amount predicted by the increase in sample size.

With the Singles-and-Doubles petition, the estimates for $D_{Goodman}$ and $D_{Goodman3}$ also drop to 68% of the value from the 5% sample. With the Geometric sample, the median estimated Standard Error only drops to 75% of the value from the 5% sample, though the maximum estimate drops to 68% ($D_{Goodman}$) or 21% ($D_{Goodman3}$). With the Uniform petition, the drop is more marked: the median estimated Standard Error is 60% ($D_{Goodman}$) or 31% ($D_{Goodman3}$); the maximum estimated Standard Error is 3% ($D_{Goodman}$) or 24% ($D_{Goodman3}$).

**Fig 4.2 Plots of RMSE against Sampling Fraction**
**(with reference lines indicating $1/\sqrt{n}$)**




The 'blip' in the graph for DGoodman with the Uniform petition at a 5% sampling fraction is genuine, but is the product a few very high estimates inflating the mean squared error.

Comparing the estimated Standard Errors with the RMSEs, $D_{Goodman2}$, $D_{Goodman2+}$, $D_{Dup}$ and $D_{Mod\ Esty}$ at 5% sampling on the Geometric and Singles-and-Doubles petition have similar RMSEs and estimated Standard Errors. As the sampling fraction rises, the RMSE drops faster than the estimated Standard Errors, so with a 8% or 10% sampling fraction the estimated Standard Error is larger than the RMSE.

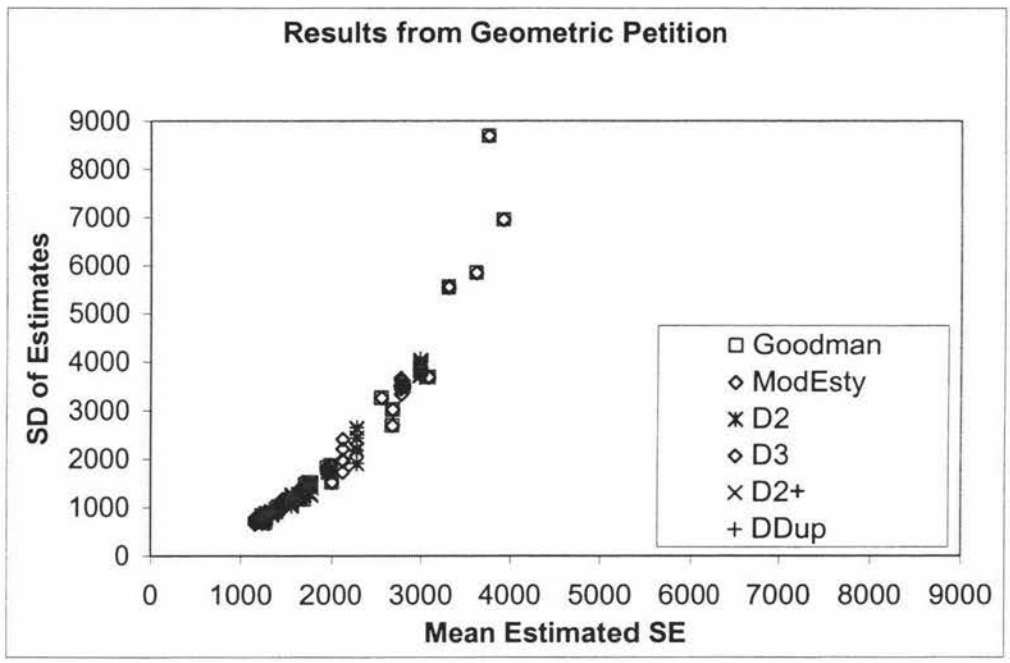
$D_{Goodman}$ and $D_{Goodman3}$ show an even sharper drop in RMSE; at 5% sampling fraction, the estimated Standard Errors are smaller than the RMSEs (for the Geometric and Singles-and-Doubles petitions), but by 10% sampling fraction, the estimated Standard Errors exceed the RMSEs.

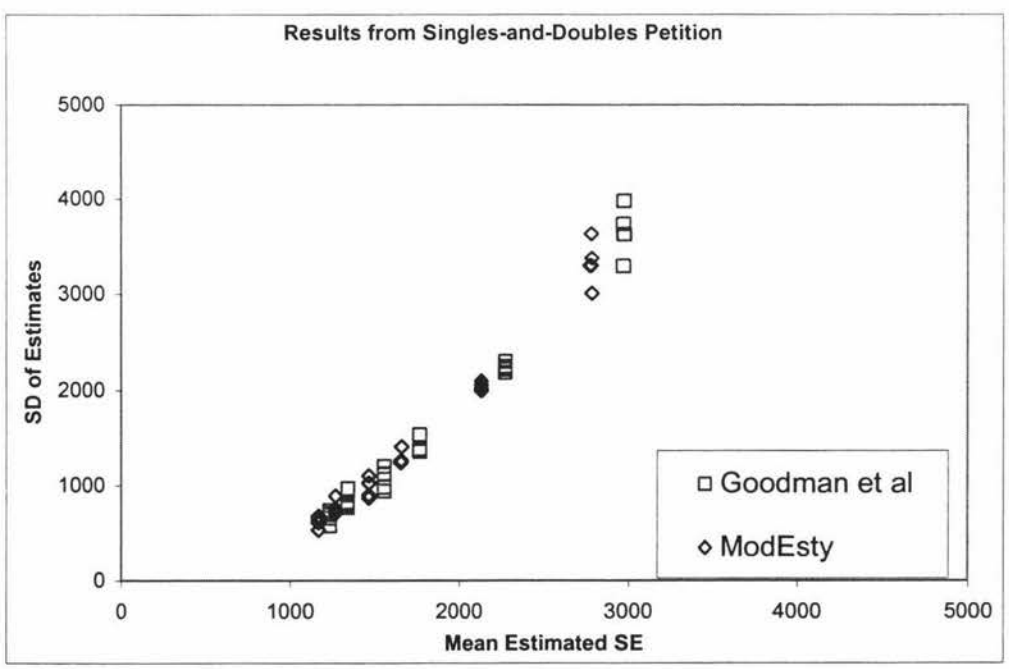
For the Uniform petition, $D_{Goodman2}$, $D_{Goodman2+}$, $D_{Dup}$ and $D_{Mod\ Esty}$ have large RMSEs (mostly due a large bias component), but much smaller estimated Standard Errors. $D_{Goodman}$ and $D_{Goodman3}$ have much larger estimated Standard Errors, but still they are smaller than the RMSEs.

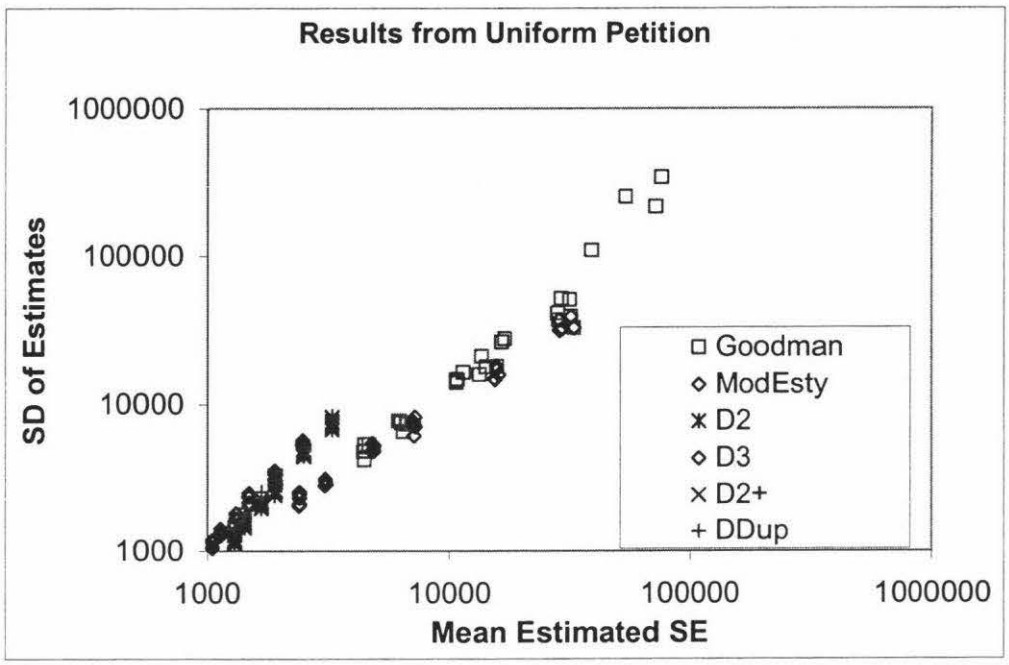
If bias means that the Estimated Standard Errors produced by Haas & Stokes' delta method approximations are not consistently good indicators of the Root Mean Squared Error, then perhaps they reflect the variation of the estimators around their own mean. Using the simulations used to construct Fig 4.2, estimated standard errors were calculated for each estimate from each sample; the mean estimated standard error was then compared to the standard deviation of the actual estimates. The 500 simulated samples generated from each petition at each sampling fraction (3%, 5%, 8%, 10%, 13% and 15%) were split into 5 blocks of 100 each, and the Mean ESE and Standard Deviation of the actual estimates were determined for each block (See Appendix 4 for data)

Fig 4.3 shows the mean Estimated Standard Error and the standard deviation of the actual estimates for each estimator on each block of each petition. Plotted on a conventional scale (as the Geometric and Singles-and-Doubles results are), the graphs show the curve and increasing variance typical of a power relationship. The range of values on the Uniform chart makes plotting the data against log scales necessary; but it also underlines the linearity of the relationship between Log(Sampling Variability) and Log(Mean ESE).

**Fig 4.3 Sampling Variability of Estimators vs Mean Estimated Standard Error**






It is possible to calculate 'calibration equations' to relate the Mean Estimated Standard Errors to the observed sampling variability (Or, more likely Log(Mean ESE) to Log(Sampling Variability) ). Since the variability of the estimated standard errors can vary considerably, particularly at low sampling fractions or with the Uniform petition, the observations were weighted by 1/Variance(ESE) (Or equivalently for the Log-Log regression by Mean$^2$/Variance, which is a first-order delta method approximation to the variance of Log(MeanESE) ).

Models for predicting Log(Sampling Variability) from Log(Mean ESE) were fitted for each estimator for each petition. Initially quadratic terms were included in the models, but they were found not to be significant. It transpired that a single model fitted the Geometric petition data from all the estimators. Another model fitted both the $D_{Mod\ Esty}$ and $D_{Goodman}$ data from the Singles-and-Doubles petition. For the data from the Uniform petition, a single model fits both $D_{Goodman}$ and $D_3$. The models for $D_{Mod\ Esty}$ and $D_{Dup}$ have similar slopes, but different intercepts. The models for $D_2$ and $D_{2+}$ also have similar slopes but different intercepts. (Table 4.6)

**Table 4.6 Calibration Equations Relating Log(MeanESE) to Log(SD Estimates)**

Regression weighted by $MeanESE^2/Variance(ESE)$

SEs of coefficients in brackets

Geometric petition

| | | | | | |
|---|---|---|---|---|---|
| Common Curve | Log(SD Estimates) = | -5.58 | + 1.72 Log(MeanESE) | | |
| for all Estimators | | (0.25) | (0.03) | | |
| | | | | $R^2 = 0.955$ | (Mod Esty) |
| | | | | $R^2 = 0.869$ | (Goodman) |
| | | | | $R^2 = 0.921$ | (D2) |
| | | | | $R^2 = 0.761$ | (D3) |
| | | | | $R^2 = 0.933$ | (D2+) |
| | | | | $R^2 = 0.939$ | (DDup) |

S & D Petition

| | | | | | |
|---|---|---|---|---|---|
| Common Curve | Log(SD Estimates) = | -7.61 | + 1.98 Log(MeanESE) | | |
| for all Estimators | | (0.61) | (0.09) | | |
| | | | | $R^2 = 0.973$ | (Mod Esty) |
| | | | | $R^2 = 0.900$ | (Goodman *et al*) |

Uniform Petition

| | | | | | |
|---|---|---|---|---|---|
| Goodman and D3 | Log(SD Estimates) = | -0.63 | + 1.07 Log(MeanESE) | | |
| | | (0.21) | (0.03) | | |
| | | | | $R^2 = 0.948$ | (Goodman) |
| | | | | $R^2 = 0.965$ | (D3) |
| Mod Esty | Log(SD Estimates) = | -5.89 | + 1.86 Log(MeanESE) | | |
| | | (0.42) | (0.06) | $R^2 = 0.973$ | |
| D2 | Log(SD Estimates) = | -7.94 | + 2.10 Log(MeanESE) | | |
| | | (0.68) | (0.09) | $R^2 = 0.946$ | |
| D2+ | Log(SD Estimates) = | -7.48 | + 2.04 Log(MeanESE) | | |
| | | (0.58) | (0.08) | $R^2 = 0.959$ | |
| DDup | Log(SD Estimates) = | -5.61 | + 1.81 Log(MeanESE) | | |
| | | (0.56) | (0.08) | $R^2 = 0.951$ | |

Regressions of the form Log(Y) = a + bLog(X) correspond to power models of the form $Y = Exp(a) \times X^b$ ; so for instance, the common model for the Geometric petition data is equivalent to SD Estimates = 0.0038 ($MeanESE^{1.72}$ ).

**The Distribution of the Estimators**

Applying a similar approach to that used in Ch 3 to the 5% and 10% samples, the only results where the Normal plot looked problematic were $D_{Goodman}$ for the 5% sample from the Geometric and Uniform petitions, and the 10% sample from the Uniform petition.

**Table 4.7 Comparing Normal Theory and Empirical Results from Simulations (Selected Estimators, Petitions and Sampling Fractions)**

| 5% sample Geometric | | | 95% CI | | Lower Quartile | Upper Quartile |
|---|---|---|---|---|---|---|
| Goodman | | | | | | |
| Mean | 250107 | Normal Theory | 243697.7 | 256516.3 | 247901.4 | 252312.6 |
| StDev | 3270.0 | Percentiles | 243704.6 | 256510.1 | 247954.0 | 251555.0 |
| 5% sample Uniform | | | 95% CI | | Lower Quartile | Upper Quartile |
| Goodman | | | | | | |
| Mean | 261129 | Normal Theory | -121937.3 | 644195.3 | 129303.4 | 392954.6 |
| StDev | 195442.0 | Percentiles | -121547.4 | 643804.8 | 246148.0 | 271257.0 |
| 10% sample Uniform | | | | | | |
| Goodman | | | | | | |
| Mean | 249992 | Normal Theory | 217737.3 | 282246.7 | 238892.1 | 261091.9 |
| StDev | 16456.5 | Percentiles | 217770.6 | 282214.3 | 241573.0 | 255751.0 |

Clearly, the skewness does not seriously affect the sample from the Geometric petition; interestingly, the samples from the Uniform petition have quite reasonable estimates of their 95% confidence intervals, but the quartiles are estimated to be further apart than they are in actuality. This suggests that for estimators other than Goodman's, the sampling distribution of their estimates is Normal, and that (assuming one can obtain a reasonable estimate of the variance), confidence intervals can be calculated using Normal theory. For petitions where the bulk of the multiple signatures are duplicates (eg the enumerated petitions reported by Smith-Cayama & Thomas, the Geometric and

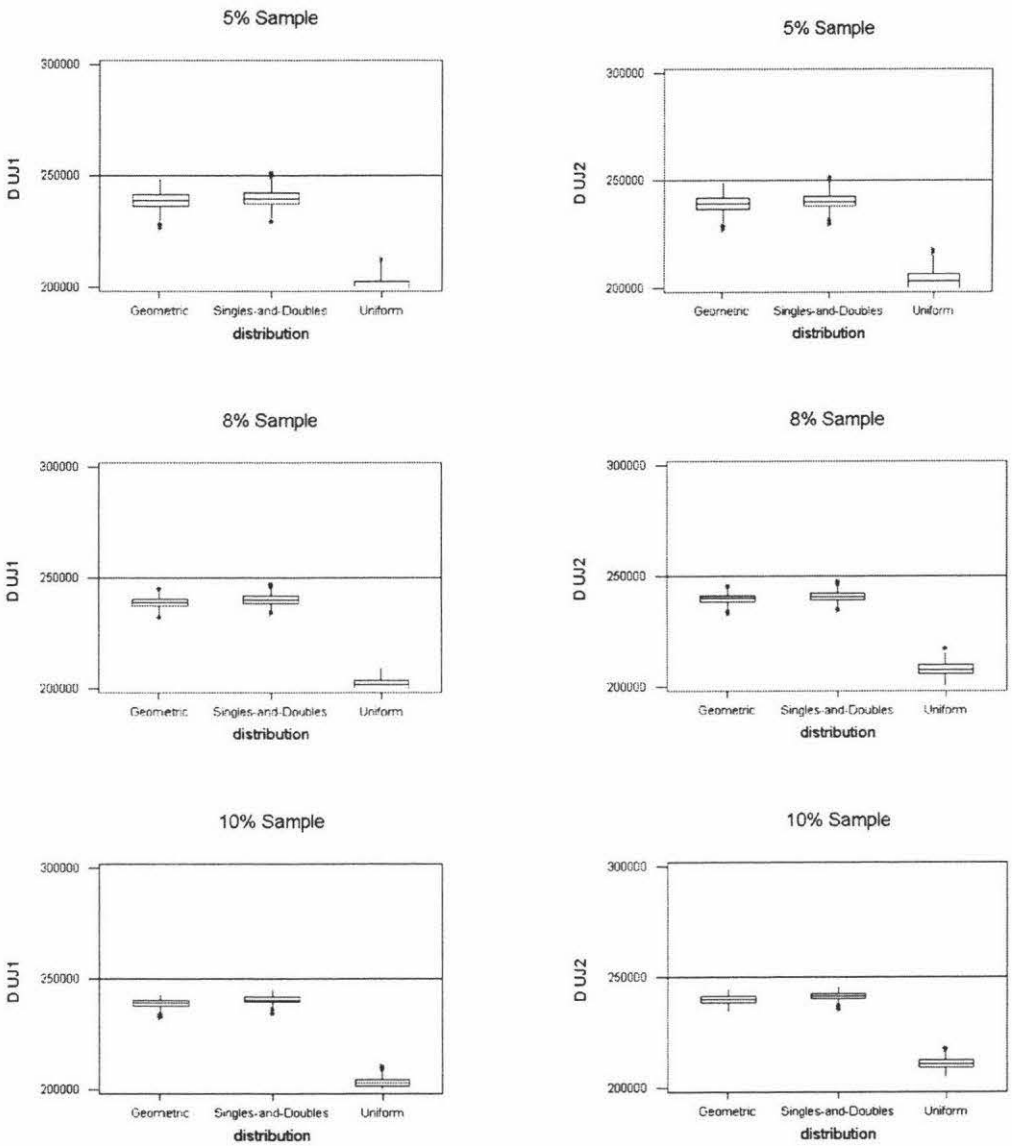
Singles-and-Doubles petition in the simulation studies), the sampling distribution of Goodman's estimator also appears to be roughly Normal. With the Uniform petition, the sampling distribution has more widely-spread tails. Even so, the Normal theory estimates for the 95% confidence interval match the observed results fairly well; and there is a suggestion that for smaller coverage (ie x% confidence intervals, where $x < 95$) the Normal theory results will tend to be conservative (ie wider than is necessary to enclose x% of the data).

**Haas & Stokes' Jackknife Estimators**

Smith-Cayama & Thomas (1999) also tested the jack-knife estimator recommended by Haas & Stokes (1998) but found it performed poorly, with a large negative bias (ie underestimating the number of individuals who have signed the petition)

Applying $D_{uj2}$ to the simulated samples with 5%, 8% and 10% sampling fractions, we found the same behaviour as Smith-Cayama & Thomas; a pronounced negative bias with all the petitions (even the Geometric and Singles-and-Doubles petitions, where the other estimators performed reasonably well). As sample size increases, the estimates tend to cluster tighter around their means; unfortunately, the bias does not decrease, meaning that the RMSE drops only gradually. The first-order unsmoothed estimator ($D_{uj1}$, which needs to be calculated in order to calculate $D_{uj2}$ ) tends to be somewhat more biased than the second-order estimator ($D_{uj2}$ )

**Fig 4.4 Box-and-Whisker Plots of Estimated Number of Signatories Using Jack-Knife Estimators**




## Conclusions

This chapter has covered a wide range of topics arising from the use of simulation studies. We have investigated the conditions under which $D_{Goodman2}$, $D_{Goodman2+}$, $D_{Dup}$ and $D_{Mod\ Esty}$ tend to underestimate the number of unique signatures in the petition; for

$D_{Goodman2}$, $D_{Goodman2+}$ this is a function of the sampling fraction; for $D_{Mod\ Esty}$ it is a function of the distribution of the number of times signatories have signed the petition.

Nonetheless, with petitions which have an approximately Geometric distribution of numbers of signatures (such as the four Oregon petitions Smith-Cayama & Thomas, 1999, present), biases are small. This means that Bias Adjustment Factors are close to 1, and adjusting the estimators for bias has a comparatively small effect. It does almost eradicate bias in the Geometric petitions; but in the Uniform petitions they only have a minor impact on the large bias. With the Singles-and-Doubles petition, where $D_{Goodman2}$, $D_{Goodman2+}$ and $D_{Dup}$ are effectively unbiased, applying the bias adjustment makes them somewhat more biased.

Haas & Stokes' variance estimates appear to have a rather complex relationship with the actual variability of estimators, the exact nature of which depends on the distribution of the number of signatures in the petition. This is presumably because of their assumption that all classes are the same size (ie 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.

Looking at the RMSEs at different sampling fractions, all estimators show great improvements (better than would be expected simply as a result of the increase in sample size) as the sampling fraction increases for the Geometric and Singles-and-Doubles petitions. Only $D_{Goodman}$ and $D_{Goodman3}$ also show this improvement with the Uniform petition; the RMSEs of the other estimators being strongly affected by bias.

Finally, we investigated the performance of the estimator recommended by Haas & Stokes (1998). Although it performed well on the wide range of problems Haas & Stokes considered, for petitions data it appears to give heavily biased results, even on problems where the other estimators are not.

# Chapter 5

## The Problem of Ineligible Signatures

One distinctive aspect of applying these estimation procedures to petitions is that, in fact, there are two estimation problems; under the New Zealand Citizens' Initiated Referenda Act, only people eligible to vote (ie on the electoral roll) qualify; thus from the sample one must estimate the number of ineligible signatures in the petition, and also the number of duplicate signatures from eligible individuals. The first problem is simple, so most attention has focused on the second, especially as solutions to that can be applied to a wide variety of fields. However, as Smith-Cayama & Thomas observe, the two estimates are not independent[1]; if 100x% of the signatures in the sample are ineligible, it is not appropriate to simply estimate the number of unique eligible signatures by applying one of the existing estimators with a population size of N(1-x). The point estimate will be correct, but the correlation will mean that the standard error of the estimate will not simply be the standard error of the number of unique signatures (Nor even $\square$ ($SE^2_{\text{number of unique signatures}} + SE^2_{\text{number of ineligible signatures}}$) )

Smith-Cayama & Thomas' notation, which focuses on estimating the number of ineligible signatures, and the number of eligible but duplicated[2] signatures makes estimating the correlation between the two estimates easier than the approach most other papers have taken, of estimating the number of unique classes. In this notation, the number of unique eligible signatures is:

---

[1] Clearly, the more ineligible signatures there are in a sample of a given size, the fewer eligible signatures there must be, which may lead to the conclusion that there are fewer unique eligible signatures in the petition.
Similarly, for a petition of a given size, if there are more ineligible signatures in it, it may be that the number of unique eligible signatures is smaller.

[2] 'Duplicated' here means the number of additional signatures beyond one; it does not just refer to the second signature from someone who has signed twice. Someone who has signed five times would generate four 'duplicated' signatures

$N - \hat{U} - \hat{D}$ where N = Number of Signatures in Petition,

$\hat{U}$ = Estimated Number of Ineligible Signatures,

$\hat{D}$ = Estimated Number of Duplicated Signatures

which clearly has a variance of

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

If the sampling can be treated as approximately binomial, then the variance of the estimated number of ineligible signatures in the petition is $\frac{u(n-u)}{q^2 n}$, where u is the number of ineligible signatures in the sample.

We already have expressions for the variance of the estimated number of duplicated signatures (or equivalently for the variance of the estimated number of unique signatures, since these calculations were done assuming all signatures are eligible, and so the total number of eligible signatures is the sum of the number of unique signatures and the number of duplicated signatures).

Smith-Cayama & Thomas present a formula for the covariance term when one is using a linear estimator. Deriving the covariance term for the Modified Esty's estimator is more complex, because it is non-linear (See Appendix 4 for an algebraic approach. The first-order approximation derived depends on $Cov(f_{i|g}, f_{m|n}\, f_{r|t})$, which I have not been able to find an expression for).

**Simulation Study**

As it was proving difficult to obtain an analytical solution to the correlation between the estimated number of ineligible signatures and the estimated number of duplicate signatures according to the modified $D_{Esty}$ , a simulation study was conducted.

The basic code was similar to that used in Chapters 3 and 4, but in addition to varying the structure of the petition (ie the proportion of individuals who have signed it once, twice, three-times, etc) and the sampling fraction, we also varied the proportion of signatures in the petition which were invalid[3]. The proportions used were 5%, 10%, 15%, 20% and 25%, reflecting the range observed in the fully enumerated Oregon petitions.

Five hundred samples were drawn from each combination of sampling fraction, petition structure and proportion invalid. Estimates of the number of invalid signatures and the number of duplicated signatures (based on the modified $D_{Esty}$) were calculated for each sample; and the five hundred pairs of estimates were correlated. The correlations proved to be comparatively small, and five hundred simulations were used so that if a confidence interval was calculated using Fisher's z transformation[4], it was sufficiently small to be helpful.

---

[3] This was done by determining the sample size, as before, and then randomly determining the number of invalid signatures in the sample using a Binomial (sample size, proportion of invalid signatures in petition) distribution. Having done that, the remaining signatures in the sample were drawn, as usual, from the valid signatures in the petition.

[4] $z = (1/2) \log_e ( (1+r)/(1-r) )$ . If the two variables being correlated are Multivariate Normally distributed and have a true correlation of □, then z is approximately Normally distributed with mean $\tanh^{-1}$ □ + □/2(n-1) and variance 1/(n-3) (Lindley & Scott, 1984)

**Table 5.1 Correlation Coefficients plus 95% Confidence Interval based on Fisher's z-transformation**

| 5% Sample | Geometric | Singles & Doubles | Uniform |
|---|---|---|---|
| 5% Invalid | -0.0289<br>(-0.1181, 0.0607) | 0.0151<br>(-0.0745, 0.1044) | -0.0422<br>(-0.1312, 0.0474) |
| 10% Invalid | 0.0640<br>(-0.0256, 0.1526) | -0.0511<br>(-0.1399, 0.0386) | 0.0010<br>(-0.0884, 0.0905) |
| 15% Invalid | 0.0106<br>(-0.0789, 0.1000) | -0.0513<br>(-0.1401, 0.0384) | **-0.0989**<br>(-0.1868, -0.0095) |
| 20% Invalid | 0.0355<br>(-0.0541, 0.1246) | 0.0073<br>(-0.0822, 0.0967) | -0.0503<br>(-0.1392, 0.0393) |
| 25% Invalid | 0.0350<br>(-0.0547, 0.1240) | **-0.0918**<br>(-0.1798, -0.0024) | **-0.1068**<br>(-0.1945, -0.0175) |

| 8% Sample | Geometric | Singles & Doubles | Uniform |
|---|---|---|---|
| 5% Invalid | 0.0384<br>(-0.0512, 0.1275) | -0.0700<br>(-0.1585, 0.0196) | -0.0863<br>(-0.1744, 0.0032) |
| 10% Invalid | -0.0327<br>(-0.1219, 0.0569) | -0.0337<br>(-0.1228, 0.0559) | -0.0445<br>(-0.1334, 0.0452) |
| 15% Invalid | -0.0506<br>(-0.1395, 0.0390) | **-0.0930**<br>(-0.1810, -0.0036) | -0.0557<br>(-0.1444, 0.0340) |
| 20% Invalid | -0.0518<br>(-0.1406, 0.0379) | -0.0234<br>(-0.1126, 0.0662) | **-0.1294**<br>(-0.2163, -0.0404) |
| 25% Invalid | -0.0570<br>(-0.1457, 0.0327) | -0.0742<br>(-0.1626, 0.0154) | **-0.1106**<br>(-0.1981, -0.0213) |

| 10% Sample | Geometric | Singles & Doubles | Uniform |
|---|---|---|---|
| 5% Invalid | 0.0109<br>(-0.0787, 0.1003) | 0.0081<br>(-0.0815, 0.0975) | 0.0124<br>(-0.0772, 0.1017) |
| 10% Invalid | -0.0553<br>(-0.1441, 0.0343) | -0.0159<br>(-0.1052, 0.0737) | -0.0895<br>(-0.1775, 0.0000) |
| 15% Invalid | -0.0321<br>(-0.1212, 0.0575) | -0.0472<br>(-0.1361, 0.0425) | **-0.1162**<br>(-0.2036, -0.0270) |
| 20% Invalid | -0.0306<br>(-0.1197, 0.0591) | **-0.1377**<br>(-0.2244, -0.0488) | **-0.1427**<br>(-0.2293, -0.0540) |
| 25% Invalid | **-0.1104**<br>(-0.1979, -0.0211) | -0.0261<br>(-0.1153, 0.0635) | **-0.1827**<br>(-0.2678, -0.0948) |

In an attempt to derive some overall patterns from this data, an Analysis of Variance was performed on the correlation coefficients, using the Sampling Fraction, the Proportion of Invalid Signatures, and the Structure of the Petition as factors.

**Table 5.2 Analysis of Variance in Correlation Coefficients from Table 5.1**

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Sampling Fraction | | | | | |
| Linear Effect | 1 | 0.0130 | 0.0130 | 6.45 | 0.0218 |
| NonLinear Effect | 1 | 0.0010 | 0.0010 | 0.50 | 0.4891 |
| Proportion Invalid | | | | | |
| Linear Effect | 1 | 0.0224 | 0.0224 | 11.07 | 0.0043 |
| NonLinear Effect | 3 | 0.0017 | 0.0006 | 0.28 | 0.8403 |
| Structure | | | | | |
| Uniform vs Others | 1 | 0.0258 | 0.0258 | 12.79 | 0.0025 |
| Geo vs S&D | 1 | 0.0062 | 0.0062 | 3.05 | 0.0999 |
| Samp Fract * Struct | | | | | |
| Lincar*Uni/Other | 1 | 0.0000 | 0.0000 | 0.01 | 0.9194 |
| Linear*Geo/S&D | 1 | 0.0041 | 0.0041 | 2.02 | 0.1749 |
| NonLin*Uni/Other | 1 | 0.0008 | 0.0008 | 0.38 | 0.5444 |
| NonLin*Geo/S&D | 1 | 0.0000 | 0.0000 | 0.02 | 0.8900 |
| Prop Inval * Struct | | | | | |
| Linear*Uni/Other | 1 | 0.0037 | 0.0037 | 1.85 | 0.1924 |
| Linear*Geo/S&D | 1 | 0.0000 | 0.0000 | 0.00 | 0.9657 |
| NonLin*Uni/Other | 3 | 0.0007 | 0.0002 | 0.12 | 0.9467 |
| NonLin*Geo/S&D | 3 | 0.0004 | 0.0001 | 0.07 | 0.9735 |
| Samp Fract * Prop Inval | | | | | |
| Linear * Linear | 1 | 0.0056 | 0.0056 | 2.78 | 0.1146 |
| Remainder | 7 | 0.0098 | 0.0014 | 0.70 | 0.6750 |
| Error | 16 | 0.0323 | 0.0020 | | |
| Corrected Total | 44 | 0.1277 | | | |

Judging by the F values, the principal effects are the main effects, in particular the structure of the petition (largely driven by differences between the extreme case of the Uniform petition and the more lifelike Geometric and Singles and Doubles petitions; the correlations tend to be more negative with the Uniform petition), and the Proportion of Invalid Signatures (the higher the proportion of signatures which are invalid, the more negative the correlation). The Sampling Fraction also makes some difference (the higher it is, the more negative the correlation), but not as markedly. Interactions and non-linear effects only play a comparatively minor role.

(Although the main aim of this analysis was simply to summarise the effects in the data, inspection of the residuals suggests that they are close to Normally distributed, and have homogeneous variances, so the significance tests have some validity – assuming that using the three-way interaction between the factors as an 'error' term is appropriate)

The samples generated by the simulation study can be used to calculate Estimated Standard Errors (ESEs) and correlations between $\hat{U}$ and $\hat{D}$ for other estimators as well (The figures quoted for $\hat{D}$ are Calibrated Standard Error Estimates, ie the results of applying the calibration equations derived in Chapter 4 to the Haas & Stokes Estimated Standard Errors). Summaries of these figures are given in the following tables and graphs.

**Table 5.3a Mean Estimated Standard Errors for $\hat{U}$ and $\hat{D}$, plus correlation between $\hat{U}$ and $\hat{D}$**

| Geometric Petition | | $D_{Mod\ Esty}$ | | $D_{Goodman}$ | | $D_2$ | | $D_{2+}$ | | $D_{Dup}$ | | $D_3$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean ESE $\hat{U}$ | Mean CSEE $\hat{D}$ | Corr $(\hat{U},\hat{D})$ | Mean CSEE $\hat{D}$ | Corr $(\hat{U},\hat{D})$ | Mean CSEE $\hat{D}$ | Corr $(\hat{U},\hat{D})$ | Mean CSEE $\hat{D}$ | Corr $(\hat{U},\hat{D})$ | Mean CSEE $\hat{D}$ | Corr $(\hat{U},\hat{D})$ | Mean CSEE $\hat{D}$ | Corr $(\hat{U},\hat{D})$ |
| 5% sample | | | | | | | | | | | | | |
| 5% invalid | 499.9 | 1915.1 | -0.0289 | 2783.7 | -0.0200 | 2158.2 | -0.0277 | 2158.2 | -0.0278 | 2158.2 | -0.0278 | 2783.7 | -0.0200 |
| 10% invalid | 687.9 | 1836.9 | 0.0640 | 2633.6 | 0.0485 | 2060.6 | 0.0650 | 2060.6 | 0.0652 | 2060.6 | 0.0653 | 2633.6 | 0.0485 |
| 15% invalid | 819.4 | 1755.4 | 0.0106 | 2547.6 | -0.0264 | 1962.2 | 0.0064 | 1962.2 | 0.0089 | 1962.3 | 0.0114 | 2547.6 | -0.0264 |
| 20% invalid | 917.6 | 1673.1 | 0.0355 | 2196.2 | 0.0036 | 1863.7 | 0.0336 | 1863.7 | 0.0351 | 1863.7 | 0.0366 | 2196.2 | 0.0036 |
| 25% invalid | 993.3 | 1589.0 | 0.0350 | 2092.6 | -0.0374 | 1764.0 | 0.0265 | 1764.0 | 0.0309 | 1764.0 | 0.0351 | 2092.6 | -0.0374 |
| 8% sample | | | | | | | | | | | | | |
| 5% invalid | 632.3 | 1243.7 | 0.0384 | 1692.7 | 0.0522 | 1396.0 | 0.0385 | 1396.0 | 0.0391 | 1396.0 | 0.0393 | 1598.3 | 0.0245 |
| 10% invalid | 870.0 | 1192.7 | -0.0327 | 1563.8 | -0.0176 | 1335.1 | -0.0301 | 1335.1 | -0.0311 | 1335.1 | -0.0319 | 1563.8 | -0.0176 |
| 15% invalid | 1036.3 | 1140.3 | -0.0506 | 1436.1 | -0.0597 | 1273.2 | -0.0527 | 1273.2 | -0.0511 | 1273.2 | -0.0494 | 1436.1 | -0.0597 |
| 20% invalid | 1160.6 | 1086.3 | -0.0518 | 1360.5 | -0.0562 | 1211.2 | -0.0536 | 1211.2 | -0.0524 | 1211.3 | -0.0510 | 1360.5 | -0.0562 |
| 25% invalid | 1256.6 | 1032.4 | -0.0570 | 1301.7 | -0.0287 | 1147.8 | -0.0522 | 1147.8 | -0.0542 | 1147.8 | -0.0561 | 1301.7 | -0.0287 |
| 10% sample | | | | | | | | | | | | | |
| 5% invalid | 706.5 | 1007.4 | 0.0109 | 1256.0 | 0.0047 | 1127.9 | 0.0169 | 1127.9 | 0.0142 | 1127.9 | 0.0115 | 1242.8 | 0.0322 |
| 10% invalid | 973.5 | 966.0 | -0.0553 | 1202.5 | -0.0624 | 1079.5 | -0.0475 | 1079.5 | -0.0511 | 1079.5 | -0.0543 | 1188.8 | -0.0168 |
| 15% invalid | 1158.3 | 923.2 | -0.0321 | 1147.1 | 0.0501 | 1030.9 | -0.0306 | 1031.0 | -0.0308 | 1031.0 | -0.0313 | 1134.0 | -0.0273 |
| 20% invalid | 1297.7 | 879.8 | -0.0306 | 1078.3 | 0.0393 | 981.5 | -0.0221 | 981.5 | -0.0257 | 981.5 | -0.0294 | 1065.1 | 0.0017 |
| 25% invalid | 1405.1 | 836.4 | -0.1104 | 1000.8 | -0.0189 | 931.0 | -0.1093 | 931.0 | -0.1094 | 931.0 | -0.1092 | 987.9 | -0.0997 |

ESE = Estimated Standard Error CSEE = Calibrated Standard Error Estimate

**Table 5.3b Mean Estimated Standard Errors for $\hat{U}$ and $\hat{D}$, plus correlation between $\hat{U}$ and $\hat{D}$**

| Singles and Doubles Petition | | $D_{Mod\ Esty}$ | | $D_{Goodman}$ , $D_2$ , $D_{2+}$ , $D_{Dup}$ , $D_3$ | |
|---|---|---|---|---|---|
| | Mean ESE $\hat{U}$ | Mean CSEE $\hat{D}$ | Corr $(\hat{U},\hat{D})$ | Mean CSEE $\hat{D}$ | Corr $(\hat{U},\hat{D})$ |
| 5% sample | | | | | |
| 5% invalid | 499.4 | 1843.1 | 0.0151 | 2085.8 | 0.0164 |
| 10% invalid | 686.9 | 1756.5 | -0.0511 | 1973.2 | -0.0507 |
| 15% invalid | 818.0 | 1668.6 | -0.0513 | 1860.5 | -0.0500 |
| 20% invalid | 916.0 | 1578.4 | 0.0073 | 1749.8 | 0.0080 |
| 25% invalid | 992.2 | 1485.1 | -0.0918 | 1638.5 | -0.0913 |
| 8% sample | | | | | |
| 5% invalid | 631.4 | 1121.7 | -0.0700 | 1262.1 | -0.0695 |
| 10% invalid | 869.5 | 1067.8 | -0.0337 | 1194.4 | -0.0331 |
| 15% invalid | 1034.3 | 1014.6 | -0.0930 | 1127.0 | -0.0925 |
| 20% invalid | 1158.9 | 958.7 | -0.0234 | 1060.1 | -0.0230 |
| 25% invalid | 1255.1 | 903.8 | -0.0742 | 992.3 | -0.0730 |
| 10% sample | | | | | |
| 5% invalid | 705.8 | 880.0 | 0.0081 | 986.7 | 0.0090 |
| 10% invalid | 971.2 | 837.8 | -0.0159 | 934.3 | -0.0149 |
| 15% invalid | 1156.9 | 795.1 | -0.0472 | 881.5 | -0.0462 |
| 20% invalid | 1296.3 | 752.6 | -0.1377 | 828.7 | -0.1368 |
| 25% invalid | 1403.1 | 709.1 | -0.0261 | 776.4 | -0.0251 |

**Table 5.3c Mean Estimated Standard Errors for $\hat{U}$ and $\hat{D}$, plus correlation between $\hat{U}$ and $\hat{D}$**

| Uniform Petition | | $D_{Mod\ Esty}$ | | $D_{Goodman}$ | | $D_2$ | | $D_{2+}$ | | $D_{Dup}$ | | $D_3$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean ESE $\hat{U}$ | Mean CSEE $\hat{D}$ | Corr $(\hat{U},\hat{D})$ | Mean CSEE $\hat{D}$ | Corr $(\hat{U},\hat{D})$ | Mean CSEE $\hat{D}$ | Corr $(\hat{U},\hat{D})$ | Mean CSEE $\hat{D}$ | Corr $(\hat{U},\hat{D})$ | Mean CSEE $\hat{D}$ | Corr $(\hat{U},\hat{D})$ | Mean CSEE $\hat{D}$ | Corr $(\hat{U},\hat{D})$ |
| 5% sample | | | | | | | | | | | | | |
| 5% invalid | 516.7 | 3370.4 | -0.0422 | 28894.6 | -0.0138 | 4540.3 | -0.0353 | 4508.4 | -0.0377 | 4876.3 | -0.0378 | 15028.9 | 0.0088 |
| 10% invalid | 711.7 | 3251.1 | 0.0010 | 23531.3 | 0.0184 | 4263.6 | 0.0029 | 4241.1 | 0.0042 | 4618.7 | 0.0049 | 14000.4 | -0.0092 |
| 15% invalid | 846.6 | 3144.3 | -0.0989 | 20343.6 | 0.0278 | 3983.6 | -0.0880 | 3970.1 | -0.0915 | 4355.7 | -0.0926 | 12655.7 | -0.0010 |
| 20% invalid | 948.2 | 3010.2 | -0.0503 | 16258.9 | 0.0465 | 3720.0 | -0.0305 | 3714.5 | -0.0374 | 4105.8 | -0.0437 | 11483.3 | 0.0241 |
| 25% invalid | 1027.3 | 2889.5 | -0.1068 | 12689.9 | -0.0016 | 3446.6 | -0.0918 | 3448.9 | -0.0964 | 3844.1 | -0.0979 | 10064.9 | 0.0089 |
| 8% sample | | | | | | | | | | | | | |
| 5% invalid | 653.7 | 2117.6 | -0.0863 | 16737.2 | -0.0332 | 2602.8 | -0.0671 | 2627.2 | -0.0793 | 3021.4 | -0.0813 | 6661.1 | 0.0429 |
| 10% invalid | 899.1 | 2048.3 | -0.0445 | 13306.2 | 0.0536 | 2447.7 | -0.0572 | 2474.7 | -0.0477 | 2865.0 | -0.0380 | 6103.9 | -0.0558 |
| 15% invalid | 1071.1 | 1975.2 | -0.0557 | 11218.8 | -0.0207 | 2292.1 | -0.0393 | 2321.6 | -0.0458 | 2706.9 | -0.0477 | 5593.4 | 0.0190 |
| 20% invalid | 1199.0 | 1897.2 | -0.1294 | 8373.3 | -0.0174 | 2142.0 | -0.1133 | 2173.6 | -0.1203 | 2552.9 | -0.1210 | 5089.4 | -0.0117 |
| 25% invalid | 1299.1 | 1797.9 | -0.1106 | 7835.5 | 0.0502 | 1989.7 | -0.0755 | 2023.2 | -0.0910 | 2395.4 | -0.1024 | 4737.7 | 0.0306 |
| 10% sample | | | | | | | | | | | | | |
| 5% invalid | 730.2 | 1692.8 | 0.0124 | 9352.3 | 0.0406 | 1974.9 | -0.0240 | 2009.9 | -0.0017 | 2383.0 | 0.0182 | 4371.3 | -0.0776 |
| 10% invalid | 1006.2 | 1635.9 | -0.0895 | 8232.2 | -0.0249 | 1858.3 | -0.0549 | 1894.2 | -0.0724 | 2260.6 | -0.0814 | 4038.7 | 0.0334 |
| 15% invalid | 1198.5 | 1575.1 | -0.1162 | 7422.3 | 0.0074 | 1742.2 | -0.0633 | 1778.9 | -0.0899 | 2137.9 | -0.1074 | 3752.4 | 0.0708 |
| 20% invalid | 1341.9 | 1513.7 | -0.1427 | 6653.3 | 0.0567 | 1628.1 | -0.1149 | 1665.4 | -0.1281 | 2016.2 | -0.1345 | 3456.8 | -0.0150 |
| 25% invalid | 1451.7 | 1447.8 | -0.1827 | 5866.4 | 0.0382 | 1516.7 | -0.1754 | 1554.5 | -0.1798 | 1896.4 | -0.1739 | 3135.7 | -0.0523 |

**Fig 5.1 Estimated Standard Errors for Samples from Geometric Petition**

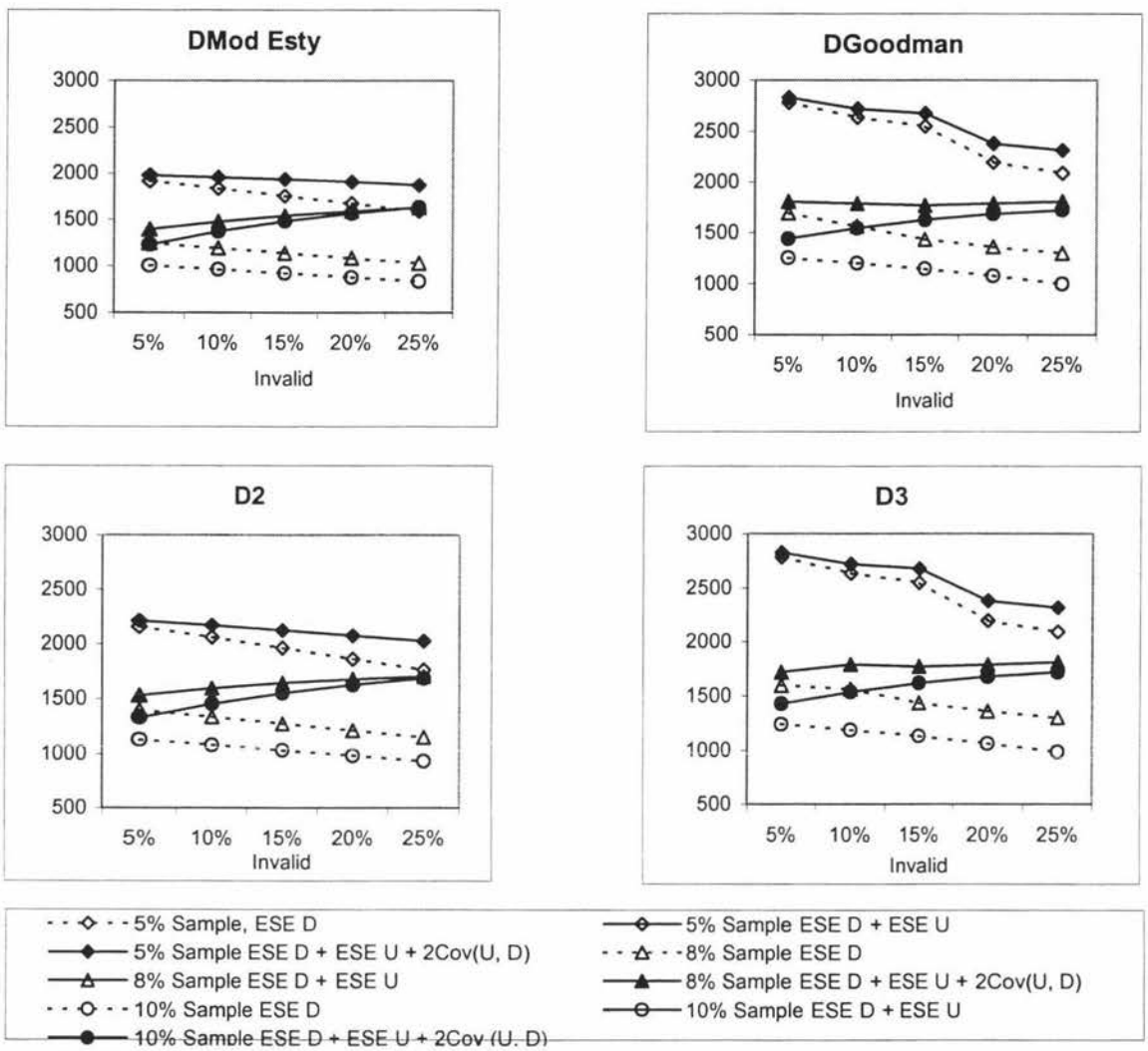
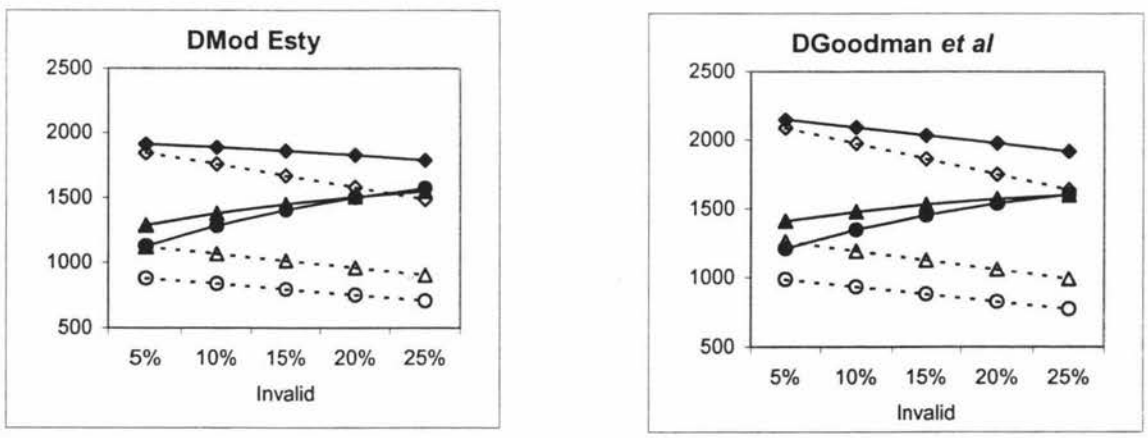
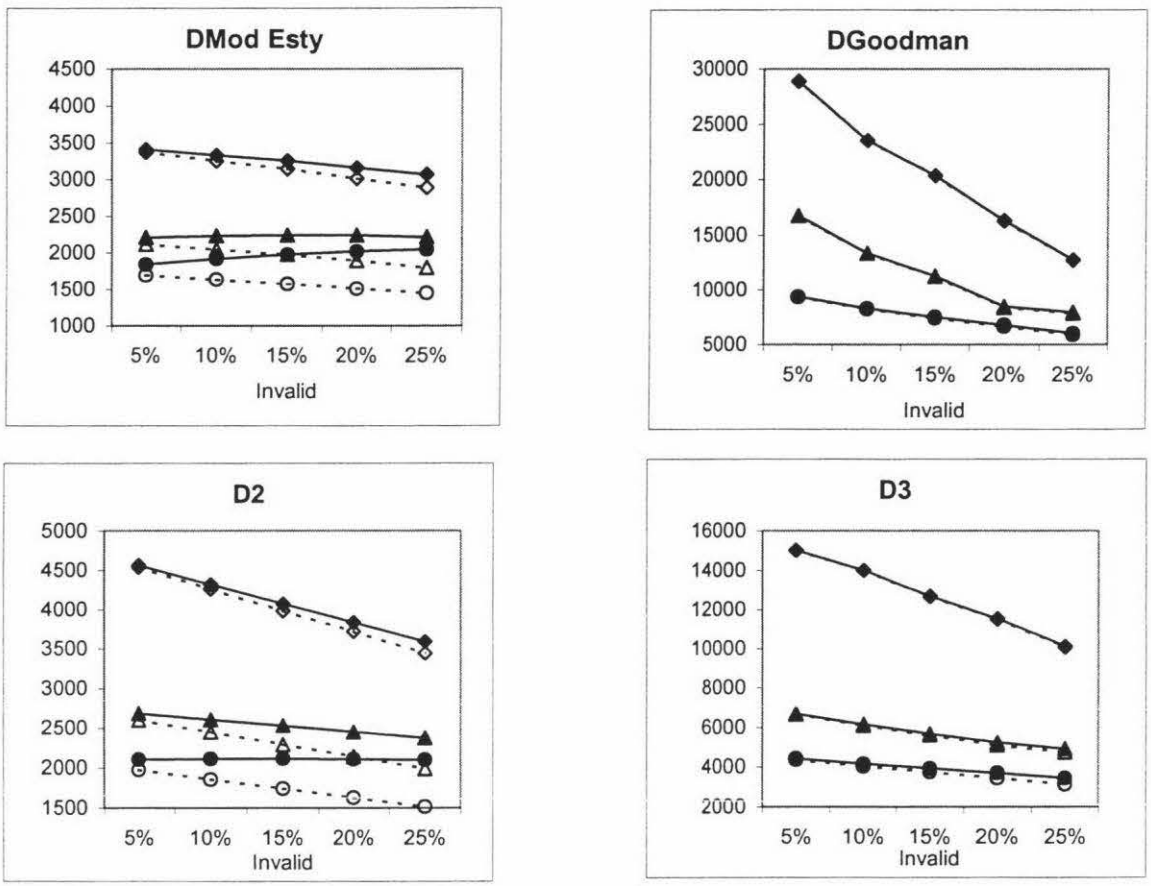



**Fig 5.2 Estimated Standard Errors for Samples from Singles and Doubles Petition**




$D_2$, $D_3$, $D_{2+}$ and $D_{Dup}$ are all equal to $D_{Goodman}$ with the Singles and Doubles petition.

**Fig 5.3 Estimated Standard Errors for Samples from Uniform Petition**




Graphs for $D_{2+}$ and $D_{Dup}$ are similar to D2

The graphs illustrate that, as the proportion of invalid signatures rises, the estimated standard error of D alone drops (as the effective size of the petition is dropping); but for the Geometric and Singles-and-Doubles petitions the estimated standard error for D+U is appreciably higher than the CSEE of D alone.

Making allowance for the covariance has no appreciable effect; in fact, the difference between $\sqrt{(\text{Var}(\hat{U}) + \text{Var}(\hat{D}))}$ and $\sqrt{(\text{Var}(\hat{U}) + \text{Var}(\hat{D}) + 2\,\text{Cov}(\hat{U}, \hat{D}))}$ was less than 1 signature in all cases, meaning that the points cannot be distinguished on the graphs.

With the Uniform petition, the story is more complex. For $D_{Goodman}$ and $D_3$, the sampling variability in $\hat{D}$ dwarfs the variability in $\hat{U}$, meaning that adjusting the CSEE for the variability in $\hat{U}$ (whether or not one also makes allowance for the covariance between the two estimates) will not substantially alter the ESE. For $D_{Mod\ Esty}$, $D_2$, $D_{2+}$ and $D_{Dup}$ adjusting the CSEE for the variability in $\hat{U}$ does appreciably increase the estimated standard error. However, it should be remembered (from Ch 4) that these estimators have considerable bias, which showed little sign of dropping over the range of sampling fractions tested.

In either case it appears that, given the typical parameters of the petitions problem, there is little to be gained by adjusting the variance estimate for the covariance between $\hat{D}$ and $\hat{U}$.

# Chapter 6
# Conclusions

We have seen in Chapter 1 that the problem of estimating the number of unique individuals in a population, based on a sample arises in a wide range of contexts. A wide variety of solutions have been suggested. Often these solutions have been developed in a particular context, and either explicitly or implicitly make use of some features of that context, which may or may not apply in other contexts; for instance, the capture-recapture models assume that a comparatively small number of samples are taken, and each individual appears at most once in each sample. It is hard to see how this applies to the problem of calculating the number of duplicate signatures in a petition; one might argue that it is unlikely an individual would sign the same sheet of a petition twice, and so one could treat the petition sheets as separate samples. But since in checking a petition a large number of petition sheets are checked, one would then be left trying to calculate estimates from a very high-dimensional, very sparse contingency table. The effect of different assumptions about the sampling process, and the likely distribution of the numbers of times individuals occur in the population may well explain why the estimators which Haas & Stokes (1998) found gave the best performance over a range of data sets performed so poorly in the simulations performed by Smith-Cayama & Thomas (1999) and us on typical petition data.

In Chapter 2 we covered a number of the more promising models and estimators. A key point is that because the sample is only a (small) fraction of the population, the proportion of duplicates in the sample is (often) considerably lower than the proportion in the population. The effect is even more marked for triplicates and higher order multiples. In order to get an unbiased estimate of the number of unique individuals in the population, one needs to give high weights to the number of higher-order multiples in the sample. Because the higher-order multiples are rare, their numbers have (proportionally) quite large sampling variation. Combined with the high weight, this means the unbiased estimate can have considerable sampling variability. Reducing the weights (even to zero) reduces the sampling variation but produces biased estimates. Approximate variance

estimates (introduced by Haas & Stokes, 1998), and bias adjustment factors (introduced by Smith-Cayama & Thomas, 1999) were also described.

The estimator proposed by Esty (1985) produced estimates of the number of unique signatures which almost an order of magnitude higher than the *total* number of signatures when applied to data typical of the samples from recent New Zealand petitions. The cause of the problem was the way the underlying model was formulated, and in Chapter 3 a modified version of the estimator was developed, which gave much more credible estimates. The problem of determining the sampling variability of the estimator when the proportion of single, duplicate, triplicate etc signatures in the petition is not known was raised. Four US petitions were shown to have roughly Geometric distributions. At the end of the chapter, results of some simulation studies were presented, which compared the modified estimator to the most promising estimators identified by Smith-Cayama & Thomas (1999). The modified Esty's estimator performed similarly to $D_2$ (which has been used for recent petitions in New Zealand) and $D_{2+}$ (which Smith-Cayama & Thomas report as being used for petitions in Washington state).

Chapter 4 presented more results derived from the simulation studies. The causes of the bias observed in several of the estimators was discussed, and the circumstances when those estimators would tend to underestimate or overestimate the number of unique signatures were derived. The performance of the Bias Adjustment Factors proposed by Smith-Cayama & Thomas (1999) was investigated. Variance estimates were calculated for each estimator under each of the sets of conditions using Haas & Stokes' (1998) delta-method approach. These were compared with the actual Root Mean Squared Errors, and the intrinsic variation of the estimates. The variance estimates and the intrinsic variation appear to be related, but the relationship is not linearly, and varies depending on the distribution of the number of signatures in the petition. The chapter finished with a brief investigation of the performance of the two estimators which Haas & Stokes (1998) found gave the best average performance across a wide range of data sets. Although they

may have been good all-rounders, on data typical of the samples from recent petitions (in New Zealand and the USA) they performed very poorly.

One aspect of the petition problem which does not arise in the more general literature on estimating the number of unique individuals in a population is that there are two estimation problems; what proportion of signatures in the petition are from people eligible to sign the petition[1]; and then, how many eligible individuals are represented by the eligible signatures? The first estimate is simple, but some allowance for its variability ought to be built into the estimate of the precision of the estimated number of eligible signatories. Smith-Cayama and Thomas also point out that the two estimates (for the number of ineligible signatures and the number of duplicate signatures from eligible signatories) are correlated, and provide a general formula for this correlation if an estimator is a linear function of the number of multiple signatures. Unfortunately, the modified Esty's estimator is a non-linear function of the number of multiple signatures, and deriving a formula for the correlation proved difficult. However, using a simulation study, we obtained estimates of its value under a range of conditions. These indicated that although the variability of the estimated number of ineligible signatures is not negligible, the correlation between it and the estimate of the number of eligible signatories is small, and so
the variance of the total is approximately the sum of the variances of the two estimates. Using the samples generated by the simulation study, we found that this pattern also applied to the Goodman's estimator and its variants, at least with the petitions similar to those reported by Smith-Cayama and Thomas.

**Recommendations**

In the simulation studies, the Geometric and Single-and-Doubles petitions produced samples similar to those obtained from recent New Zealand petitions; ie consisting mostly of unique signatures, with a few signatures appearing twice. The

[1] In New Zealand, registered electors

samples from the Uniform petition were quite different, regularly containing triple and quadruple signatures. This suggests that the results from the Geometric and Singles-and-Doubles petitions are most likely to be relevant, with the results form the Uniform petitions indicating how the estimators perform in very unusual situations. It would have been nice to investigate a less extreme 'unusual situation' but it was not clear how that might best be defined.

On the Geometric petition, $D_{Mod\ Esty}$, $D_2$, $D_{2+}$ and $D_{Dup}$, despite being biased, had the lowest Root Mean Square errors, because they were not influenced by the appearance of the occasional triple signature. Bias adjustment reduced the RMSE for $D_2$, $D_{2+}$ and $D_{Dup}$ by between 8 and 11%. However, it was impressive how similar the RMSEs of all the estimators were: with an 8% sampling fraction, $D_{Mod\ Esty}$ had an RMSE of 1302 (around a true value of 250,000); the bias-adjusted $D_2$, $D_{2+}$ and $D_{Dup}$ had RMSEs of between 1380 and 1385; the unadjusted $D_2$, $D_{2+}$ and $D_{Dup}$ had RMSEs of between 1500 and 1555; while the 'highly variable' $D_{Goodman}$ had an RMSE of 1733. With a 10% sampling fraction, the distinctions were smaller. With a 5% sample the differences were more marked; $D_{Mod\ Esty}$ had an RMSE of 2071, $D_2$ *et al* had RMSEs between 2360 and 2370, and $D_{Goodman}$ had an RMSE of 3269.

On the Singles-and-Doubles petition, all the variants of $D_{Goodman}$ ($D_2$, $D_3$, $D_{2+}$, $D_{Dup}$) are equivalent to $D_{Goodman}$. RMSEs for $D_{Mod\ Esty}$ and the $D_{Goodman}$ variants were similar; with a 5% sampling fraction between 2064 and 2225; with a 8% sampling fraction between 1357 and 1400; and with a 10% sampling fraction between 1074 and 1128. Because the $D_{Goodman}$ variants were all equivalent to $D_{Goodman}$ itself, they were all unbiased. Applying bias adjustment factors to them actually increased the bias (and their RMSEs), although only slightly (RMSE increased by 1-2%). This illustrates a useful point – blindly applying the bias adjustment factors does not guarantee an improvement in accuracy; one needs to consider whether, given the results from a particular sample, the result from one of these 'biased' variants of $D_{Goodman}$ is equivalent to the unbiased result from $D_{Goodman}$ itself, in which case it does not need adjustment.

On the basis of these results, from petitions where the bulk of the multiple signatures are from people who have signed twice, it appears there is little to chose

between using $D_{Mod\ Esty}$, $D_2$, $D_{2+}$ or $D_{Dup}$ , with bias adjustment for the $D_{Goodman}$ variants if the sample contains triple, quadruple, etc signatures.

The catch is that if the multiple signatures include a substantial proportion of people who have signed more than twice, the simulation results suggest that $D_{Mod\ Esty}$, $D_2$, $D_{2+}$ and $D_{Dup}$ are heavily biased. Alarmingly, there is no indication that this might be a problem; the estimated standard errors are similar to those for the Geometric and Singles-and-Doubles petitions. Estimators which specifically account for the triple, quadruple, etc signatures, like $D_3$ and $D_{Goodman}$ , are less biased; and although they are somewhat more variable, the simulations suggest that with the structures of petitions reported by Smith-Cayama and Thomas, the difference is not especially large. For petitions, the bugbear of estimates which are negative, or which are absurdly large does not appear a major problem; the minimum and maximum estimates from the simulations reported in chapter 4 (n=500 for each sampling fraction-petition combination) are shown below

**Table 6.1 Minimum and Maximum $D_{Goodman}$ and $D_3$ Estimates from 500 Samples**

| | DGoodman | | D3 | |
|---|---|---|---|---|
| 5% sample | min | max | min | max |
| Geometric | 241552.6 | 266361.0 | 241552.6 | 266361.0 |
| Singles & Doubles | 243298.6 | 256499.6 | 243298.6 | 256499.6 |
| Uniform | -26511.3 | 2756378.6 | 222844.6 | 314872.4 |
| 8% sample | | | | |
| Geometric | 245186.5 | 256125.5 | 245186.5 | 253156.0 |
| Singles & Doubles | 246082.1 | 254050.4 | 246082.1 | 254050.4 |
| Uniform | 179983.3 | 464394.2 | 240705.9 | 285705.5 |
| 10% sample | | | | |
| Geometric | 245455.5 | 257580.5 | 245455.5 | 253656.4 |
| Singles & Doubles | 246099.4 | 252799.7 | 246099.4 | 252799.7 |
| Uniform | 210241.4 | 354393.6 | 245650.5 | 276153.7 |

The biggest problems occur with the 5% sample from the Uniform petition; of the 500 estimates under this condition, the 10 smallest and the 10 largest were...

**Table 6.2 10 Smallest and 10 Largest $D_{Goodman}$ Estimates based on 5% Sample Drawn From Uniform Petition**

| Smallest | Largest |
|---|---|
| -26511.3 | 302468.0 |
| -14105.5 | 303269.4 |
| -5304.23 | 303668.1 |
| 103969.5 | 304471.0 |
| 106768.6 | 305268.4 |
| 111171.9 | 306468.5 |
| 111569.2 | 314872.4 |
| 111972.0 | 2587887.0 |
| 113169.4 | 2747976.0 |
| 113570.9 | 2756379.0 |

The simulation studies (Chapter 4) indicate that for all of the estimators increasing the sample size with the Geometric and Singles-and-Doubles petitions reduces the RMSE by more than the usual factor of $1/\sqrt{n}$ ($D_{Goodman}$ and $D_3$ show similar behaviour with the Uniform petition) This suggests that there is not some ideal sample size above which there is little point in increasing further; it also suggests that reducing the sample size will have even more than the usual deleterious effects on accuracy; for instance, reducing the sampling fraction from 8% to 5% would reduce the size of the task by 37%, but increase the RMSE by between 50% and 90%.

The fact that the only method of estimating the sampling variation of an estimator on the basis of a sample (Haas & Stokes' 1998 approach) produces estimates which have a non-linear relationship with the actual variability presents a challenge, especially as the relationship also appears to depend on the distribution of the number of signatures in the petition, which is unknown in practice. Presumably the problem stems from their assumption that all classes are the same size (ie everyone has signed the petition the same number of times). Although this seems unlikely, it 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. Since the four fully-enumerated petitions cited by Smit-Cayama & Thomas (1999) have distributions which are roughly Geometric; and since most samples from recent New Zealand petitions have been similar to those obtained in the simulation study from the Geometric petition (ie mostly single and duplicate signatures, with maybe

a few triplicates), a practical strategy might be to use the calibration equations for the Geometric petition to obtain calibrated standard error estimates.

**In Short**

- On the basis of completely checked petition in the US, and from comparing the results of simulation studies with results from recent NZ petitions suggest that most people only sign petitions once, and those who do sign more than once do not sign many times.

- With petitions like these, where the sample contains mostly single and duplicate signatures, there is little to choose between Goodman's estimator, its various variants, and the modified Esty's estimator described here. All show little or no bias and similar degrees of accuracy.

- For petitions where multiple signing is more of an issue and the sample contains signatures which occur three or more times, $D_{\text{Mod Esty}}$, $D_2$ and its variants give biased results. $D_{\text{Goodman}}$ is unbiased, while $D_3$ shows less bias than $D_{\text{Mod Esty}}$ and $D_2$, while at the same time being somewhat less variable than $D_{\text{Goodman}}$.

- Simulation studies suggest that increasing the size of sample taken will improve the accuracy of the estimates roughly in proportion to 1/sample size.

- Haas & Stokes variance estimates need calibration to produce good estimates of standard error. For petitions where the sample contains mostly single and duplicate signatures, the Geometric petition calibrations may be appropriate.

- The number of ineligible signatures in the petition also needs to be estimated from the sample. Allowance for the variability of this estimate should be included with the sampling variability of the estimator of the number of duplicate signatures. Although the two figures are correlated, the correlation is small.

# Appendix1

## Variance Estimators for Various Estimators

In Chapter 3 we present a formula for the approximate variance of the modified Esty's estimator, based on Haas & Stokes' (1999) delta method. Below are similar formulae for the other estimators mentioned in Chapter 2.

### General Form

The general form of Haas & Stokes' variance estimator is:

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

where

$\hat{D}$ is the estimator

N is the size of the population

$A_i$ is the partial derivative of $\hat{D}$ with respect to $f_i$

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

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

and

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

### Goodman's

Goodman's estimator is in effect

$$D_{\text{Goodman}} = \sum_{i=1}^{n} f_i + \sum_{i=1}^{n} (-1)^{i+1} \frac{(N-n+i-1)!(n-i)!}{(N-n-1)!n!} f_i$$

$$\text{so } A_i = \frac{\delta D}{\delta f_i} = 1 + (-1)^{i+1}\frac{(N-n+i-1)!(n-i)!}{(N-n-1)!n!}$$

This means that for $\mathbf{D_{Goodman2}}$, **or** $\mathbf{D_{Goodman}}$ **if the sample only contains singles and duplicate samples**:

$$\text{Var} = \left(1+\frac{(N-n)!\ (n-1)!}{(N-n-1)!\ n!}\right) f_1\left(1-\frac{f_1}{\hat{D}}\right) + \left(1-\frac{(N-n+1)!\ (n-2)!}{(N-n-1)!\ n!}\right) f_2\left(1-\frac{f_2}{\hat{D}}\right)$$

$$-2\left(1+\frac{(N-n)!\ (n-1)!}{(N-n-1)!\ n!}\right)\left(1-\frac{(N-n+1)!\ (n-2)!}{(N-n-1)!\ n!}\right)\frac{f_1 f_2}{\hat{D}}$$

$$=\left(\frac{N}{n}\right)^2 f_1\left(1-\frac{f_1}{\hat{D}}\right) + \left(\frac{n(n-1)-(N-n+1)(N-n)}{n(n-1)}\right)^2 f_2\left(1-\frac{f_2}{\hat{D}}\right)$$

$$-2\left(\frac{N}{n}\right)\left(\frac{n(n-1)-(N-n+1)(N-n)}{n(n-1)}\right)\frac{f_1 f_2}{\hat{D}}$$

Similarly, for $\mathbf{D_{Goodman3}}$

$$\text{Var} = \left(\frac{N}{n}\right)^2 f_1\left(1-\frac{f_1}{\hat{D}}\right) + \left(\frac{n(n-1)-(N-n+1)(N-n)}{n(n-1)}\right)^2 f_2\left(1-\frac{f_2}{\hat{D}}\right)$$

$$+\left(\frac{n(n-1)(n-2)+(N-n+2)(N-n+1)(N-n)}{n(n-1)(n-2)}\right)^2 f_3\left(1-\frac{f_3}{\hat{D}}\right)$$

$$-2\left(\frac{N}{n}\right)\left(\frac{n(n-1)-(N-n+1)(N-n)}{n(n-1)}\right)\frac{f_1 f_2}{\hat{D}}$$

$$-2\left(\frac{N}{n}\right)\left(\frac{n(n-1)(n-2)+(N-n+2)(N-n+1)(N-n)}{n(n-1)(n-2)}\right)\frac{f_1 f_3}{\hat{D}}$$

$$-2\left(\frac{n(n-1)-(N-n+1)(N-n)}{n(n-1)}\right)\left(\frac{n(n-1)(n-2)+(N-n+2)(N-n+1)(N-n)}{n(n-1)(n-2)}\right)\frac{f_2 f_3}{\hat{D}}$$

**D** $_{\text{Goodman2+}}$ is similar to D $_{\text{Goodman2}}$ , meaning that:

$$\text{Var} = \left(\frac{N}{n}\right)^2 f_1\left(1-\frac{f_1}{\hat{D}}\right) + \sum_{i=2}^{n}\left(\frac{n(n-1) - (N-n+1)(N-n)}{n(n-1)}\right)^2 f_i\left(1-\frac{f_i}{\hat{D}}\right)$$

$$- \sum_{i=2}^{n} 2\left(\frac{N}{n}\right)\left(\frac{n(n-1) - (N-n+1)(N-n)}{n(n-1)}\right)\frac{f_1 f_i}{\hat{D}}$$

$$- \sum_{i=2}^{n}\sum_{j=i+1}^{n} 2\left(\frac{n(n-1) - (N-n+1)(N-n)}{n(n-1)}\right)\left(\frac{n(n-1) - (N-n+1)(N-n)}{n(n-1)}\right)\frac{f_i f_j}{\hat{D}}$$

$$= \left(\frac{N}{n}\right)^2 f_1\left(1-\frac{f_1}{\hat{D}}\right) + \left(\frac{n(n-1) - (N-n+1)(N-n)}{n(n-1)}\right)^2 \sum_{i=2}^{n} f_i\left(1-\frac{\sum_{i=2}^{n} f_i}{\hat{D}}\right)$$

$$- 2\left(\frac{N}{n}\right)\left(\frac{n(n-1) - (N-n+1)(N-n)}{n(n-1)}\right)\frac{f_1\sum_{i=2}^{n} f_i}{\hat{D}}$$

Similarly, **D** $_{\text{GoodmanDup}}$ is

$$\text{Var} = \left(\frac{N}{n}\right)^2 f_1\left(1-\frac{f_1}{\hat{D}}\right) + \left(\frac{n(n-1) - (N-n+1)(N-n)}{n(n-1)}\right)^2 \sum_{i=2}^{n}(i-1)f_i\left(1-\frac{\sum_{i=2}^{n}(i-1)f_i}{\hat{D}}\right)$$

$$- 2\left(\frac{N}{n}\right)\left(\frac{n(n-1) - (N-n+1)(N-n)}{n(n-1)}\right)\frac{f_1\sum_{i=2}^{n}(i-1)f_i}{\hat{D}}$$

**Shlosser's Estimator**

This can be written

$$D_{Shlosser} = \sum_{i=1}^{n} f_i + f_1 \frac{\sum_{i=1}^{n} (1-q)^i f_i}{\sum_{i=1}^{n} iq(1-q)^{i-1} f_i} \quad \text{where the sampling fraction } q = \frac{n}{N}$$

so

$$A_i = \frac{dD}{df_i} = 1 + \left( \frac{f_1 (1-q)^i}{\left( \sum_{j=1}^{n} jq(1-q)^{j-1} f_j \right)} - \frac{f_1 \left( \sum_{j=1}^{n} (1-q)^j f_j \right) iq(1-q)^{i-1}}{\left( \sum_{j=1}^{n} jq(1-q)^{j-1} f_j \right)^2} \right) \text{ for } i > 1$$

or

$$A_i = \frac{dD}{df_i} = 1 + \left( \frac{2(1-q)}{\left( \sum_{j=1}^{n} jq(1-q)^{j-1} f_j \right)} - \frac{f_1 q \left( \sum_{j=1}^{n} (1-q)^j f_j \right)}{\left( \sum_{j=1}^{n} jq(1-q)^{j-1} f_j \right)^2} \right) \text{ for } i = 1$$

**Haas & Stokes' Estimators**

Haas & Stokes' second-order unsmoothed jackknife estimator $D_{uj2}$ can be written:

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

where $q = \frac{n}{N}$, 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

$$D_{uj1} = \frac{n\sum_{i=1}^{n} f_i}{\left(n - (1-q)f_1\right)}, \text{ an initial estimate of D}$$

This means that

$$\text{for i>1} \quad A_i = \frac{\delta D}{\delta f_i} = \left(\frac{n}{n\text{-}f_1(1-q)}\right)\left(1 - \frac{f_1(1\text{-}q)\ \ln(1\text{-}q)}{q}\ \frac{\delta\gamma^2}{\delta f_i}\right)$$

$$\text{if } \gamma^2(D_{uj1}) > 0, \text{ then } \frac{\delta\gamma^2}{\delta f_i} = \frac{i(i\text{-}1)\sum_{j=1}^{n} f_j}{n\left(n - (1\text{-}q)f_1\right)} + \frac{\sum_{j=1}^{n} j(j\text{-}1)f_j}{n\left(n - (1\text{-}q)f_1\right)} + \frac{n}{n\left(n - (1\text{-}q)f_1\right)N}$$

so

$$A_i = \frac{1}{q\left(n\text{-}f_1(1-q)\right)^2}\left(qn\left(n - (1\text{-}q)f_1\right) - f_1(1\text{-}q)\ln(1\text{-}q)\left(i(i\text{-}1)\sum_{j=1}^{n} f_j + \sum_{j=1}^{n} j(j\text{-}1)f_j + \frac{n}{N}\right)\right)$$

$$= \frac{1}{\left(n\text{-}f_1(1-q)\right)^2}\left(n^2 - f_1(1\text{-}q)\left(n + \ln(1\text{-}q) + \frac{\ln(1\text{-}q)}{q}\left(i(i\text{-}1)\sum_{j=1}^{n} f_j + \sum_{j=1}^{n} j(j\text{-}1)f_j\right)\right)\right)$$

$$= \frac{n^2 - f_1(1\text{-}q)\left(n + \ln(1\text{-}q)\right)}{\left(n\text{-}f_1(1-q)\right)^2} + \frac{\ln(1\text{-}q)\sum_{j=1}^{n} j(j\text{-}1)f_j}{q\left(n\text{-}f_1(1-q)\right)^2} - \frac{f_1(1\text{-}q)\ln(1\text{-}q)\sum_{j=1}^{n} f_j}{q\left(n\text{-}f_1(1-q)\right)^2}\left(i^2\text{-}i\right)$$

$$\text{for i = 1} \quad A_1 = \frac{\delta D}{\delta f_1} = \left(\frac{n}{n\text{-}f_1(1-q)}\right)\left(1 - \frac{f_1(1\text{-}q)\ \ln(1\text{-}q)}{q}\ \frac{\delta\gamma^2}{\delta f_1} - \frac{(1\text{-}q)\ln(1-q))}{q}\gamma^2\right)$$

$$\text{if } \gamma^2(D_{uj1}) > 0, \text{ then } \frac{\delta\gamma^2}{\delta f_i} = \frac{i(i\text{-}1)\sum_{j=1}^{n} f_j}{n\left(n - (1\text{-}q)f_1\right)} + \frac{\sum_{j=1}^{n} j(j\text{-}1)f_j}{n\left(n - (1\text{-}q)f_1\right)} + \frac{n}{n\left(n - (1\text{-}q)f_1\right)N}$$

so

$$A_1 = \left(\frac{n}{n\text{-}f_1(1-q)}\right) - \left(\frac{n}{n\text{-}f_1(1-q)}\right)\left(\frac{(1\text{-}q)\ln(1\text{-}q)}{q}\right) f_1 \left(\frac{\sum_{j=1}^{n} j(j\text{-}1)f_j + q}{n\left(n - (1\text{-}q)f_1\right)}\right)$$

$$+ \left(\frac{n}{n\text{-}f_1(1-q)}\right)\left(\frac{(1\text{-}q)\ln(1\text{-}q)}{q}\right)\left(\frac{\sum_{i=1}^{n} f_i \sum_{j=1}^{n} j(j\text{-}1)f_j}{n\left(n - (1\text{-}q)f_1\right)} + \frac{n\sum_{j=1}^{n} f_j}{N\left(n - (1\text{-}q)f_1\right)} - 1\right)$$

$$= \frac{n}{\left(n\text{-}f_1(1-q)\right)} - \frac{1}{\left(n\text{-}f_1(1-q)\right)^2}\left(\frac{f_1(1\text{-}q)\ln(1\text{-}q)}{q}\right)\left(\sum_{j=1}^{n} j(j\text{-}1)f_j + q\right)$$

$$+ \frac{1}{\left(n\text{-}f_1(1-q)\right)^2}\left(\frac{(1\text{-}q)\ln(1\text{-}q)}{q}\right)\left(\sum_{i=1}^{n} f_i \sum_{j=1}^{n} j(j\text{-}1)f_j + nq\sum_{j=1}^{n} f_j - n\left(n\text{-}f_1(1-q)\right)\right)$$

# Appendix 2

# Computer Programs

This section lists examples of the SAS code used to simulate sampling from petitions with patterns of repeated signatures.

All the simulations have been based on petitions signed by 250,000 individuals. The different patterns of replication used meant that the total number of signatures on the petitions varies slightly; for instance, a petition where 5% of signatories have signed twice has a total of 262,500 signatures, whereas one where 5% of signatories have signed more than once and are equally likely to have signed two, three, four or five time has a total of 281,250. While there is an argument that it would be better to compare petitions with the same number of total signatures, this would mean that the number of individuals signing would differ depending on the pattern of replication; we decided that it would be easier to compare the accuracy of the estimates for the different patterns of replication if the correct answer was the same for each (A third option would have been to vary the proportion of individuals signing more than once, to keep the total number of signatures and the total numbers of individuals signing constant. However, because one of the factors being considered for the simulation studies at an early stage was the proportion of individuals signing more than once, we did not consider using it to control other variables)

The first programs I wrote to do this were slow:

1. Create 250,000 unique ID numbers (to represent the individuals signing the petition),
2. Replicate the unique IDs to fit the desired pattern of repeated signatures (for instance if 4.5% of the individuals had signed twice, 4.5% of the ID numbers were duplicated; if 0.475% had signed three times, another 0.475% of the ID numbers were copied twice).
3. Sort the list into random order by:
   a. Assigning each entry (signature) a random number from a Uniform (0,1) distribution, then
   b. Sorting the list of IDs by those random numbers.
4. If a sample of n (= q x 250000) is required, use the first n entries in the sorted list.

This guaranteed hypergeometric sampling, but used considerable computing time and capacity (approximately 2 minutes per sample on a 166 Mhz PC with 32 Mb RAM) in randomising a large file.

Professor Haslett pointed out that the procedure did not actually need to sort the whole file; a lot of effort was going into sorting the 90% - 95% of the file which would not appear in the sample. The first attempt to take advantage of this used the mechanics of the sorting procedure. Since the random numbers used to sort the list of ID numbers were from a Uniform distribution between 0 and 1, the sample (the first (100q) % of the list) would have random numbers between 0 and (approximately) q. Thus one could safely discard any entry with a sufficiently large random number (say, above (100q+1)/100).

1. Create 250,000 unique ID numbers (to represent the individuals signing the petition),
2. Replicate the unique IDs to fit the desired pattern of repeated signatures (for instance if 4.5% of the individuals had signed twice, 4.5% of the ID numbers were duplicated; if 0.475% had signed three times, another 0.475% of the ID numbers were copied twice).
3. Sort the list into random order by:
   a. Assigning each entry (signature) a random number from a Uniform (0,1) distribution,
   b. Deleting any with random numbers > (100q+1)/100, then
   c. Sorting the list of IDs by those random numbers.
4. If a sample of n (= q x 250000) is required, the first n entries in the sorted list are used.

This reduced the size of the data set which had to be sorted and the time and resources taken to 20-30 seconds per sample.

Several other approaches were tried. One was based on the idea that if the sample consisted of n signatures, then no more than n individuals could appear in the sample. Thus we need only generate n individuals; randomly determine how many times they had signed the petition (based on the proportion of individuals signing once, twice, three times, etc); replicate the ID numbers the necessary number of times; and then draw a sample of n from that smaller data set. This was faster, but produced far too many

multiple signatures; the process was effectively doing hypergeometric sampling with the same sample size (n) and proportions of individuals signing once, twice, etc; but the total pool it was sampling from was much smaller, which raised the sampling fraction, and so the probability that multiple signatures would appear in the sample.

A second approach tried accumulating the sample, using the binomial approximation to hypergeometric sampling. An individual was picked; using the proportion of signatures in the petition from people who had signed once, twice, etc, it was randomly determined whether this person had signed the petition once, twice, etc. If they had signed the petition more than once, a conditional Binomial distribution (Binomial(q, no. times signed petition|number of signatures in sample≥1) was used to determine how many times that individual's signature appeared in the sample. The process then moved on to the next individual in the sample, until n signatures had been generated. The procedure was fast, but also produced slightly more multiple signatures than would be expected (Eg in an 8% sample from a petition where 5% of signatories have signed twice, under Binomial sampling one would expect 80 duplicate signatures, with a SD of 8.92. Five hundred simulations gave an average of 83 with a SD of 9.90)

So we eventually used the procedure which generated the whole petition, allocated random numbers to each signature, and then just sampled from those with random number below (100q+1)/100. An example of the code is given below:

```
/*************************************************/
/* Code for simulation procedure in SAS          */
/* Uses SAS/Stat and Macro Language              */
/* (Note: variables in Macro Language can only have*/
/* integer values;  hence &PMULT and &PSAMPLE being*/
/* expressed as percentages, and the decay rate of */
/* the Exponential distribution as a fraction    */
/* rather than a decimal)                        */
/*************************************************/

/*********************************************************/
/* core macro to do draw samples from a specific petition */
/*********************************************************/

%macro cirloop;

/* draws 500 samples */
%let sim=1;
%do %until(&sim>500);

data cir3;
/* whole petition is in CIR2 */
set cir2;
/* ORDER = random numbers which will be used to sort petition */
order=ranuni(0);
/* remove any signature which is unlikely to appear in sample */
if order>((&sampperc+1)/100) then delete;
run;

/* sort the remaining signatures into random order */
proc sort data=cir3 out=cir3;
by order;
run;

/* select just the first n signatures from the random order */
data cir4;
set cir3(obs=&nsample);
run;

/* how often does each signature/ID number appear in sample */
/* saved to data set SUML                                   */
proc freq data=cir4;
table i/norow nocol nocum nopercent noprint out=sum1;
run;

/* calculate frequencies of frequencies - how many signatures  */
/* appear once, twice, etc in the sample (as recorded in SUML) */
/* saved to a daya set SUM2                                    */
proc freq data=sum1(rename=count=c);
table c/ out=sum2 noprint;
run ;
```

```
/* transpose SUM2 so that instead of rows for C1, C2, etc, */
/* have columns for them. Saved to a data set SUM3          */
proc transpose data=sum2 out=sum3 prefix=c;
var count;
id c;
run;

/* if this is the first simulated sample, create a data set */
/* RESULTS from SUM3                                        */
%if &sim=1 %then %do;
data results;
set sum3;
sim=&sim;
run;
%end;

/* if this is not the first simulated sample, add SUM3 as an */
/* extra row to RESULTS                                      */
%if &sim>1 %then %do;
data results;
set results sum3(in=new);
if new then sim=&sim;
run;
%end;

/* increment the counter by 1 */
%let sim=%eval(&sim+1);
%end;
/* and return to the start of the %DO loop */

%mend;

/*************************/
/* end of sampling macro */
/*************************/
```

```
/*********************************************/
/* Now generate petition with an Exponential    */
/* distribution of signatures  eg 95% singles,  */
/* 95% of the remaining 5% doubles 95% of the   */
/* remaining 0.25% triples, etc                 */
/*********************************************/

/* numerator and denominator of decay rate parameter */
/* eg 95%=19/20                                      */
%let rnum=19;
%let rdenom=20;

/* percentage of multiple signatures in petition*/
%let pmult=5;

/* sampling fraction as a percentage*/
%let sampperc=8;

/* calculate number of individuals who have signed once */
/* (PETSINGL), twice (PETDOUBL), etc in the petition    */
%let petsingl=%eval(2500*(100-&pmult));

%let petdoubl=%eval(2500*&pmult*&rnum/&rdenom);

%let pettripl=%eval(2500*&pmult*(&rdenom-
&rnum)*&rnum/(&rdenom*&rdenom));
%let pettripl=%eval(&pettripl+1);
/* the rounding down by macro arithmetic means &PETTRIPL */
/* comes out as 593 not ~594                             */

%let petquad=%eval(2500*&pmult*(&rdenom-&rnum)*(&rdenom-
&rnum)*&rnum/(&rdenom*&rdenom*&rdenom));

%let petquin=%eval(2500*&pmult*(&rdenom-&rnum)*(&rdenom-
&rnum)*(&rdenom-&rnum)*&rnum/(&rdenom*&rdenom*&rdenom*&rdenom));
%let petquin=%eval(&petquin+1);
/* the rounding down by macro arithmetic means &PETQUIN comes */
/* out as 1 not ~2                                            */

%let petsex=%eval(2500*&pmult*(&rdenom-&rnum)*(&rdenom-
&rnum)*(&rdenom-&rnum)*(&rdenom-
&rnum)*&rnum/(&rdenom*&rdenom*&rdenom*&rdenom*&rdenom));

/* PUT these numbers out to the Log as a check */
%put petsingl;
%put &petsingl;
%put petdoubl;
%put &petdoubl;
%put pettripl;
%put &pettripl;
%put petquad;
%put &petquad;
```

```
%put petquin;
%put &petquin;
%put petsex;
%put &petsex;

/* calculate total number of signatures in petition */
%let ntotal=%eval(&petsingl + 2*&petdoubl + 3*&pettripl +
4*&petquad + 5*&petquin + 6*&petsex);
%put ntotal;
%put &ntotal;

/* calculate the desired sample size */
%let nsample=%eval(&ntotal*&sampperc/100);

/* clear any previous petitions (data sets called CIR) */
/* out of memory                                       */
proc datasets memtype=data;
delete cir;
run;

/* generate 250000 individuals who have signed new petition */
/* Save as CIR                                              */
/* This is all very orderly - it is randomly sorted in the  */
/* sampling macro                                           */
data cir;
do i=1 to 250000;

/* the first PETSINGL ...                        */
if i<=&petsingl then do;
/* ...have only signed the petition once (n=1) */
 n=1;
 output;
end;

/* the next PETDOUBL ...                      */
if &petsingl<i and i<=(&petsingl + &petdoubl) then do;
/* ...have signed the petition twice (n=2) */
 n=2;
 output;
end;

/* the next PETTRIPL...                              */
if (&petsingl + &petdoubl)<i and i<=(&petsingl + &petdoubl +
&pettripl) then do;
/* ...have signed the petition three times (n=3) */
 n=3;
 output;
end;

/* the next PETQUAD...                                 */
if (&petsingl + &petdoubl + &pettripl)<i and i<=(&petsingl +
&petdoubl + &pettripl + &petquad) then do;
```

```
/* ...have signed the petition four times (n=4) */
 n=4;
 output;
end;

/* the next PETQUIN...                                  */
if (&petsingl + &petdoubl + &pettripl + &petquad)<i and
i<=(&petsingl + &petdoubl + &pettripl + &petquad + &petquin) then
do;
/* ...have signed the petition five times (n=5) */
 n=5;
 output;
end;

/* the last PETSEX...                                  */
if (&petsingl + &petdoubl + &pettripl + &petquad + &petquin)<i
and
 i<=(&petsingl + &petdoubl + &pettripl + &petquad + &petquin +
&petsex) then do;
/* ...have signed the petition six times (n=6) */
 n=6;
 output;
end;
end;
run;

/* as a check, get number who have signed petition once, */
/* twice, etc                                            */
proc freq data=cir;
table n;
run;

/* for individuals who have signed more than once (n>1)*/
/* create n separate rows in data set CIR2             */
/* one for each signature                              */
data cir2;
set cir;
%let counter=n;
do j=1 to &counter;
output;
end;
run;

/* run sampling macro */
%cirloop;

/* save RESULTS to another file before it is overwritten */
/* with the next lot of simulated samples                */
data geo8pc;
set results;
run;
```

```
/*********************************************/
/* Now generate petition with just with singles */
/* and duplicate signatures                    */
/*********************************************/

/* percentage of multiple signatures in petition*/
%let pmult=5;

/* sampling fraction as a percentage*/
%let sampperc=8;

/* calculate number of individuals who have signed once */
/* (PETSINGL), twice (PETDOUBL), etc in the petition    */
%let petsingl=%eval(2500*(100-&pmult));

%let petdoubl=%eval(2500*&pmult);

/* PUT these numbers out to the Log as a check */
%put petsingl;
%put &petsingl;
%put petdoubl;
%put &petdoubl;

/* calculate total number of signatures in petition */
%let ntotal=%eval(&petsingl + 2*&petdoubl);
%put &ntotal;
%put ntotal;

/* calculate the desired sample size */
%let nsample=%eval(&ntotal*&sampperc/100);

/* clear any previous petitions (data sets called CIR) */
/* out of memory                                       */
proc datasets memtype=data;
delete cir;
run;

/* generate 250000 individuals who have signed new petition */
/* Save as CIR                                              */
/* This is all very orderly - it is randomly sorted in the  */
/* sampling macro                                           */
data cir;
do i=1 to 250000;

/* the first PETSINGL ...                  */
if i<=&petsingl then do;
/* ... have signed the petition once (n=1) */
 n=1;
 output;
end;

/* the remaining PETDOUBL ...              */
```

```
if &petsingl<i and i<=(&petsingl + &petdoubl) then do;
/* ... have signed the petition twice (n=2) */
 n=2;
 output;
end;
end;
run;

/* as a check, get number who have signed petition once, */
/* twice, etc                                            */
proc freq data=cir;
table n;
run;

/* for individuals who have signed more than once (n>1)*/
/* create n separate rows in data set CIR2             */
/* one for each signature                              */
data cir2;
set cir;
%let counter=n;
do j=1 to &counter;
output;
end;
run;

/* run sampling macro */
%cirloop;

/* save RESULTS to another file before it is overwritten */
/* with the next lot of simulated samples                */
data sad8pc;
set results;
run;
```

```
/*********************************************/
/* Now generate petition with equal numbers of  */
/* duplicates, triples, quads and quins         */
/*********************************************/

/* percentage of multiple signatures in petition*/
%let pmult=5;

/* sampling fraction as a percentage*/
%let sampperc=8;

/* calculate number of individuals who have signed once */
/* (PETSINGL), twice (PETDOUBL), etc in the petition    */
%let petsingl=%eval(2500*(100-&pmult));

%let petdoubl=%eval(2500*&pmult/4);

%let pettripl=%eval(2500*&pmult/4);

%let petquad=%eval(2500*&pmult/4);

%let petquin=%eval(2500*&pmult/4);

/* PUT these numbers out to the Log as a check */
%put petsingl;
%put &petsingl;
%put petdoubl;
%put &petdoubl;
%put pettripl;
%put &pettripl;
%put petquad;
%put &petquad;
%put petquin;
%put &petquin;

/* calculate total number of signatures in petition */
%let ntotal=%eval(&petsingl + 2*&petdoubl + 3*&pettripl +
4*&petquad + 5*&petquin);
%put &ntotal;
%put ntotal;

/* calculate the desired sample size */
%let nsample=%eval(&ntotal*&sampperc/100);

/* clear any previous petitions (data sets called CIR) */
/* out of memory                                       */
proc datasets memtype=data;
delete cir;
run;

/* generate 250000 individuals who have signed new petition */
/* Save as CIR                                              */
```

```
/* This is all very orderly - it is randomly sorted in the  */
/* sampling macro                                           */
data cir;
do i=1 to 250000;

/* the first PETSINGL ...                    */
if i<=&petsingl then do;
/* ... have signed the petition once (n=1) */
 n=1;
 output;
end;

/* the next PETDOUBL ...                      */
if &petsingl<i and i<=(&petsingl + &petdoubl) then do;
/* ... have signed the petition twice (n=2) */
 n=2;
 output;
end;

/* the next PETTRIPL ...                               */
if (&petsingl + &petdoubl)<i and i<=(&petsingl + &petdoubl +
&pettripl) then do;
/* ... have signed the petition three times (n=3) */
 n=3;
 output;
end;

/* the next PETQUAD ...                               */
if (&petsingl + &petdoubl + &pettripl)<i and i<=(&petsingl +
&petdoubl + &pettripl + &petquad) then do;
/* ... have signed the petition four times (n=4) */
 n=4;
 output;
end;

/* the remaining PETQUIN ...                          */
if (&petsingl + &petdoubl + &pettripl + &petquad)<i and
i<=(&petsingl + &petdoubl + &pettripl + &petquad + &petquin) then
do;
/* ... have signed the petition five times (n=5) */
 n=5;
 output;
end;
end;
run;

/* as a check, get number who have signed petition once, */
/* twice, etc                                            */
proc freq data=cir;
table n;
run;
```

```
/* for individuals who have signed more than once (n>1)*/
/* create n separate rows in data set CIR2             */
/* one for each signature                              */
data cir2;
set cir;
%let counter=n;
do j=1 to &counter;
output;
end;
run;

/* run sampling macro */
%cirloop;

/* save RESULTS to another file before it is overwritten */
/* with the next lot of simulated samples                */
data uni8pc;
set results;
run;
```

## Appendix 3

## Derivation of Bias Adjustment Factor for $D_{ModEsty}$

$$\hat{D}_{\text{Mod Esty}} = \frac{nd}{n - (1-q)d}$$

The expected number of duplicate signatures in the petition (ie the denominator of B) is

$$E[N - D_{\text{Mod Esty}}] = N - E[D_{\text{Mod Esty}}]$$

Since $D_{\text{Mod Esty}}$ is a ratio of two random variables ( nd and {n-(1-q)d} ) we need to use a Taylor series expansion to calculate its expected value. A first order approximation would be

$$E\left[\frac{X}{Y}\right] \simeq \frac{\mu_X}{\mu_Y}$$

(Mood *et al*, 1974, give a second-order approximation, which although more accurate, is also more unwieldy.)

Using the first-order approximation, $E\left[D_{\text{Mod Esty}}\right] \simeq \frac{nE[d]}{n-(1-q)E[d]}$

So the BAF $\simeq$ True No Duplicates $\Big/ \left(N - \frac{nE[d]}{n-(1-q)E[d]}\right)$

$$= \frac{\text{True No Duplicates } (n-(1-q)E[d])}{N(n-(1-q)E[d]) - nE[d]}$$

$$= \frac{\text{True No Duplicates } (qN-(1-q)E[d])}{qN^2 - NE[d]} \quad \text{since } n=qN$$

$$= \frac{\text{True No Duplicates}}{N}\left(1+\frac{qE[d]}{qN\text{-}E[d]}\right)$$

Now the number of unique signatures in the sample $d = \sum_{i=1}^{k} f_i$ , and so

$$E(d) = \sum_{i=1}^{k} F_i(1-(1-q)^i)$$

$$\text{So BAF} \simeq \frac{\sum_{i=1}^{k}(i-1)F_i}{N}\left(1+\frac{q\sum_{i=1}^{k}F_i(1-(1-q)^i)}{qN-\sum_{i=1}^{k}F_i(1-(1-q)^i)}\right)$$

So for example, for the four full-enumerated Oregon petitions given in Smith-Cayama & Thomas, a first-order estimate of the bias adjustment factor would be...

| | Petition A | Petition B | Petition C | Petition D |
|---|---|---|---|---|
| 5% Sample BAF | 1.003 | 0.991 | 1.021 | 1.011 |
| 8% Sample BAF | 1.003 | 0.992 | 1.020 | 1.010 |
| 10% Sample BAF | 1.003 | 0.993 | 1.020 | 1.010 |

As a check, a set of simulations (500 samples from each of the petitions) were run, and the results are given below:

| | Petition A | | Petition B | | Petition C | | Petition D | |
|---|---|---|---|---|---|---|---|---|
| D | 4256 | | 4546 | | 9738 | | 11584 | |
| 5% Sample | Dhat | BAF | Dhat | BAF | Dhat | BAF | Dhat | BAF |
| Mean | 4225.2 | 1.007 | 4629.1 | 0.982 | 10457.2 | 0.931 | 11487.2 | 1.008 |
| 95% Confidence Interval for Mean | 4336.9 | 0.981 | 4752.9 | 0.956 | 10619.1 | 0.917 | 11661.7 | 0.993 |
| | 4113.5 | 1.035 | 4505.2 | 1.009 | 10295.4 | 0.946 | 11312.7 | 1.024 |
| 8% Sample | | | | | | | | |
| Mean | 4294.1 | 0.991 | 4570.4 | 0.995 | 10522.2 | 0.925 | 11446.2 | 1.012 |
| 95% Confidence Interval for Mean | 4361.8 | 0.976 | 4644.7 | 0.979 | 10625.5 | 0.916 | 11555.5 | 1.002 |
| | 4226.3 | 1.007 | 4496.1 | 1.011 | 10418.8 | 0.935 | 11336.8 | 1.022 |
| 10% Sample | | | | | | | | |
| Mean | 4296.9 | 0.990 | 4590.6 | 0.990 | 10515.0 | 0.926 | 11552.4 | 1.003 |
| 95% Confidence Interval for Mean | 4349.3 | 0.979 | 4650.5 | 0.978 | 10599.7 | 0.919 | 11639.9 | 0.995 |
| | 4244.5 | 1.003 | 4530.7 | 1.003 | 10430.4 | 0.934 | 11465.0 | 1.010 |

The first-order estimates are within the 95% confidence interval for petitions A, B and D It is not clear what it is about Petition C that means the first-order estimate of the Bias Adjustment Factor is so different from the observed bias; it may be related to the higher-order terms in the Taylor expansion, or it may be because petition C has less of a 'tail' of high-multiplicity signatures than would be expected from the Geometric distribution that $D_{\text{Mod Esty}}$ assumes.

# Appendix 4

## Sampling Variability of Estimators and Estimated Standard Errors from Simulations

100 simulations per block (row of table), 5 blocks at each sampling fraction

| | Modified Esty's Estimator | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| q | Geometric Petition | | | Singles & Doubles | | | Uniform Petition | | |
| | SD Est | Mean ESE | Var ESEs | SD Est | Mean ESE | Var ESEs | SD Est | Mean ESE | Var ESEs |
| 3% | 3668 | 2770 | 1653 | 3006 | 2788 | 1113 | 5055 | 2476 | 2938 |
| | 3608 | 2772 | 1600 | 3295 | 2777 | 1337 | 5366 | 2468 | 3310 |
| | 3445 | 2773 | 1458 | 3374 | 2787 | 1402 | 5723 | 2473 | 3765 |
| | 3515 | 2767 | 1518 | 3635 | 2783 | 1628 | 5169 | 2484 | 3072 |
| | 3335 | 2775 | 1367 | 3297 | 2785 | 1339 | 5516 | 2474 | 3499 |
| 5% | 2205 | 2125 | 351 | 1999 | 2129 | 289 | 3548 | 1901 | 850 |
| | 1957 | 2127 | 277 | 2054 | 2131 | 305 | 3566 | 1906 | 859 |
| | 1985 | 2121 | 284 | 2023 | 2130 | 296 | 3478 | 1904 | 817 |
| | 1729 | 2125 | 216 | 2097 | 2129 | 318 | 3377 | 1902 | 771 |
| | 2406 | 2125 | 418 | 1997 | 2132 | 289 | 3121 | 1899 | 658 |
| 8% | 1276 | 1653 | 71 | 1252 | 1657 | 69 | 2501 | 1484 | 256 |
| | 1367 | 1652 | 82 | 1239 | 1656 | 67 | 2164 | 1483 | 192 |
| | 1372 | 1653 | 82 | 1264 | 1658 | 70 | 2041 | 1486 | 170 |
| | 1336 | 1653 | 78 | 1259 | 1658 | 69 | 2031 | 1483 | 169 |
| | 1174 | 1654 | 60 | 1404 | 1658 | 86 | 2360 | 1482 | 228 |
| 10% | 1107 | 1462 | 42 | 901 | 1468 | 28 | 1821 | 1316 | 106 |
| | 1169 | 1461 | 47 | 1025 | 1466 | 36 | 1805 | 1314 | 104 |
| | 965 | 1463 | 32 | 1101 | 1467 | 42 | 1651 | 1314 | 87 |
| | 996 | 1462 | 34 | 1029 | 1467 | 36 | 1682 | 1316 | 90 |
| | 1057 | 1462 | 38 | 867 | 1467 | 26 | 1678 | 1313 | 90 |
| 13% | 828 | 1261 | 17 | 747 | 1265 | 14 | 1335 | 1137 | 42 |
| | 851 | 1260 | 18 | 732 | 1265 | 14 | 1431 | 1136 | 49 |
| | 774 | 1261 | 15 | 706 | 1265 | 13 | 1271 | 1136 | 38 |
| | 810 | 1260 | 17 | 761 | 1265 | 15 | 1431 | 1137 | 49 |
| | 847 | 1261 | 18 | 892 | 1265 | 20 | 1352 | 1137 | 44 |
| 15% | 656 | 1161 | 9 | 682 | 1164 | 10 | 1153 | 1049 | 27 |
| | 717 | 1160 | 11 | 637 | 1164 | 9 | 1191 | 1048 | 29 |
| | 781 | 1160 | 13 | 654 | 1163 | 9 | 1065 | 1049 | 23 |
| | 728 | 1160 | 11 | 533 | 1164 | 6 | 1043 | 1048 | 22 |
| | 733 | 1160 | 12 | 609 | 1163 | 8 | 1145 | 1049 | 26 |

q = Sampling Fraction

SD Est = Standard Deviation of actual estimates

Mean ESE, Var ESE= Mean, Variance of Estimated Standard Errors

100 simulations per block (row of table), 5 blocks at each sampling fraction

| | Goodman's Estimator | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| q | Geometric Petition | | | Singles & Doubles | | | Uniform Petition | | |
| | SD Est | Mean ESE | Var ESEs | SD Est | Mean ESE | Var ESEs | SD Est | Mean ESE | Var ESEs |
| 3% | 5846 | 3604 | $1.9x10^7$ | 3291 | 2971 | 249.4 | 39290 | 31852 | $4.4x10^8$ |
| | 6953 | 3912 | $2.8x10^7$ | 3628 | 2976 | 302.4 | 217803 | 72120 | $4.4x10^{10}$ |
| | 5558 | 3293 | $9.6x10^6$ | 3732 | 2971 | 319.5 | 109928 | 38994 | $1.2x10^{10}$ |
| | 8696 | 3745 | $3.0x10^7$ | 3974 | 2973 | 363.9 | 37207 | 28402 | $4.4x10^8$ |
| | 5537 | 3292 | $9.6x10^6$ | 3627 | 2972 | 302.4 | 33025 | 32751 | $3.5x10^8$ |
| 5% | 3257 | 2557 | $1.5x10^6$ | 2184 | 2277 | 53.8 | 51650 | 29071 | $1.5x10^9$ |
| | 3686 | 3074 | $3.3x10^6$ | 2245 | 2276 | 56.8 | 344848 | 76045 | $1.2x10^{11}$ |
| | 3022 | 2680 | $1.8x10^6$ | 2216 | 2276 | 55.3 | 41291 | 28073 | $1.3x10^9$ |
| | 2690 | 2678 | $1.8x10^6$ | 2298 | 2276 | 59.5 | 254002 | 53695 | $6.1x10^{10}$ |
| | 3494 | 2805 | $2.5x10^6$ | 2181 | 2276 | 53.6 | 51019 | 31510 | $1.7x10^9$ |
| 8% | 1717 | 1963 | $9.3x10^4$ | 1368 | 1765 | 9.3 | 27735 | 16970 | $4.1x10^8$ |
| | 1874 | 1988 | $1.3x10^5$ | 1352 | 1766 | 9.1 | 26406 | 16482 | $4.2x10^8$ |
| | 1818 | 1949 | $1.0x10^5$ | 1378 | 1765 | 9.5 | 16013 | 13292 | $5.2x10^7$ |
| | 1737 | 1967 | $1.2x10^5$ | 1370 | 1765 | 9.4 | 18141 | 15731 | $6.9x10^7$ |
| | 1515 | 2002 | $1.2x10^5$ | 1530 | 1765 | 11.7 | 18036 | 14224 | $6.4x10^7$ |
| 10% | 1478 | 1721 | $2.8x10^5$ | 978 | 1559 | 2.9 | 14106 | 10708 | $8.3x10^7$ |
| | 1328 | 1695 | $1.7x10^4$ | 1115 | 1559 | 3.8 | 21210 | 13592 | $2.1x10^8$ |
| | 1169 | 1694 | $2.0x10^4$ | 1198 | 1559 | 4.4 | 14871 | 10800 | $8.6x10^7$ |
| | 1519 | 1750 | $2.9x10^5$ | 1122 | 1559 | 3.9 | 14796 | 10692 | $8.5x10^7$ |
| | 1277 | 1677 | $1.5x10^4$ | 942 | 1559 | 2.7 | 16532 | 11440 | $1.1x10^8$ |
| 13% | 904 | 1391 | $1.5x10^3$ | 810 | 1341 | 1.0 | 6559 | 6480 | $1.0x10^7$ |
| | 1008 | 1407 | $1.2x10^4$ | 794 | 1340 | 0.9 | 7393 | 6713 | $1.4x10^7$ |
| | 987 | 1411 | $2.4x10^4$ | 766 | 1341 | 0.9 | 7747 | 6157 | $8.8x10^6$ |
| | 1022 | 1422 | $3.4x10^4$ | 826 | 1341 | 1.0 | 7421 | 6650 | $1.2x10^7$ |
| | 965 | 1387 | $1.4x10^3$ | 967 | 1341 | 1.4 | 7562 | 6331 | $9.6x10^6$ |
| 15% | 692 | 1271 | $2.0x10^3$ | 738 | 1231 | 0.5 | 5388 | 4666 | $2.2x10^6$ |
| | 774 | 1278 | $5.6x10^3$ | 690 | 1231 | 0.4 | 4796 | 4441 | $2.3x10^6$ |
| | 880 | 1271 | $2.0x10^3$ | 709 | 1231 | 0.4 | 4221 | 4459 | $2.1x10^6$ |
| | 838 | 1270 | $1.8x10^3$ | 578 | 1231 | 0.3 | 4846 | 4475 | $2.2x10^6$ |
| | 805 | 1273 | $3.0x10^3$ | 659 | 1231 | 0.4 | 5385 | 4482 | $2.2x10^6$ |

q = Sampling Fraction

SD Est = Standard Deviation of actual estimates

Mean ESE, Var ESE= Mean, Variance of Estimated Standard Errors

100 simulations per block (row of table), 5 blocks at each sampling fraction

| | D2 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| q | Geometric Petition | | | Singles & Doubles | | | Uniform Petition | | |
| | SD Est | Mean ESE | Var ESEs | SD Est | Mean ESE | Var ESEs | SD Est | Mean ESE | Var ESEs |
| 3% | 4020 | 2984 | 369.2 | 3291 | 2971 | 249.4 | 6781 | 3280 | 890.0 |
| | 3959 | 2983 | 359.0 | 3628 | 2976 | 302.4 | 7687 | 3287 | 1152.2 |
| | 3845 | 2983 | 338.8 | 3732 | 2971 | 319.5 | 8201 | 3285 | 1310.3 |
| | 3903 | 2985 | 348.6 | 3974 | 2973 | 363.9 | 7166 | 3277 | 1003.6 |
| | 3733 | 2982 | 319.3 | 3627 | 2972 | 302.4 | 7509 | 3282 | 1095.2 |
| 5% | 2449 | 2282 | 67.4 | 2184 | 2277 | 53.8 | 4557 | 2489 | 202.7 |
| | 2207 | 2281 | 54.9 | 2245 | 2276 | 56.8 | 4946 | 2486 | 243.0 |
| | 2204 | 2284 | 54.6 | 2216 | 2276 | 55.3 | 4709 | 2488 | 219.0 |
| | 1901 | 2282 | 40.6 | 2298 | 2276 | 59.5 | 4499 | 2489 | 199.8 |
| | 2633 | 2282 | 77.7 | 2181 | 2276 | 53.6 | 4546 | 2489 | 207.4 |
| 8% | 1409 | 1770 | 9.9 | 1368 | 1765 | 9.3 | 3186 | 1907 | 45.8 |
| | 1498 | 1770 | 11.2 | 1352 | 1766 | 9.1 | 3131 | 1907 | 46.8 |
| | 1512 | 1770 | 11.4 | 1378 | 1765 | 9.5 | 2427 | 1907 | 27.0 |
| | 1442 | 1770 | 10.3 | 1370 | 1765 | 9.4 | 2528 | 1907 | 29.5 |
| | 1255 | 1770 | 7.8 | 1530 | 1765 | 11.7 | 2848 | 1908 | 36.6 |
| 10% | 1194 | 1563 | 4.3 | 978 | 1559 | 2.9 | 2187 | 1671 | 14.0 |
| | 1255 | 1563 | 4.8 | 1115 | 1559 | 3.8 | 2149 | 1672 | 13.7 |
| | 1026 | 1563 | 3.2 | 1198 | 1559 | 4.4 | 1961 | 1672 | 11.7 |
| | 1093 | 1563 | 3.7 | 1122 | 1559 | 3.9 | 1983 | 1672 | 11.6 |
| | 1150 | 1563 | 4.0 | 942 | 1559 | 2.7 | 2126 | 1672 | 13.4 |
| 13% | 879 | 1344 | 1.1 | 810 | 1341 | 1.0 | 1521 | 1422 | 3.5 |
| | 923 | 1344 | 1.3 | 794 | 1340 | 0.9 | 1548 | 1422 | 3.6 |
| | 842 | 1343 | 1.1 | 766 | 1341 | 0.9 | 1447 | 1422 | 3.3 |
| | 879 | 1344 | 1.2 | 826 | 1341 | 1.0 | 1562 | 1422 | 3.7 |
| | 912 | 1344 | 1.2 | 967 | 1341 | 1.4 | 1443 | 1422 | 3.3 |
| 15% | 670 | 1234 | 0.4 | 738 | 1231 | 0.5 | 1282 | 1298 | 1.7 |
| | 753 | 1234 | 0.5 | 690 | 1231 | 0.4 | 1119 | 1298 | 1.2 |
| | 839 | 1234 | 0.6 | 709 | 1231 | 0.4 | 1124 | 1298 | 1.4 |
| | 782 | 1234 | 0.5 | 578 | 1231 | 0.3 | 1178 | 1298 | 1.7 |
| | 785 | 1234 | 0.5 | 659 | 1231 | 0.4 | 1391 | 1297 | 2.1 |

q = Sampling Fraction

SD Est = Standard Deviation of actual estimates

Mean ESE, Var ESE= Mean, Variance of Estimated Standard Errors

100 simulations per block (row of table), 5 blocks at each sampling fraction

| | D3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| q | Geometric Petition | | | Singles & Doubles | | | Uniform Petition | | |
| | SD Est | Mean ESE | Var ESEs | SD Est | Mean ESE | Var ESEs | SD Est | Mean ESE | Var ESEs |
| 3% | 5846 | 3604 | $1.9x10^7$ | 3291 | 2971 | 249.4 | 39290 | 31852 | $4.4x10^8$ |
| | 6953 | 3912 | $2.8x10^7$ | 3628 | 2976 | 302.4 | 32540 | 29330 | $3.4x10^8$ |
| | 5558 | 3293 | $9.6x10^6$ | 3732 | 2971 | 319.5 | 31708 | 28523 | $3.6x10^8$ |
| | 8696 | 3745 | $3.0x10^7$ | 3974 | 2973 | 363.9 | 37207 | 28402 | $4.4x10^8$ |
| | 5537 | 3292 | $9.6x10^6$ | 3627 | 2972 | 302.4 | 33025 | 32751 | $3.5x10^8$ |
| 5% | 3257 | 2557 | $1.5x10^6$ | 2184 | 2277 | 53.8 | 17392 | 15698 | $1.4x10^7$ |
| | 3686 | 3074 | $3.3x10^6$ | 2245 | 2276 | 56.8 | 18312 | 15592 | $1.5x10^7$ |
| | 3022 | 2680 | $1.8x10^6$ | 2216 | 2276 | 55.3 | 14732 | 15493 | $1.1x10^7$ |
| | 2690 | 2678 | $1.8x10^6$ | 2298 | 2276 | 59.5 | 14972 | 15316 | $1.2x10^7$ |
| | 3494 | 2805 | $2.5x10^6$ | 2181 | 2276 | 53.6 | 16010 | 16053 | $1.0x10^7$ |
| 8% | 1717 | 1963 | $9.3x10^4$ | 1368 | 1765 | 9.3 | 6171 | 7123 | $3.7x10^5$ |
| | 1874 | 1988 | $1.3x10^5$ | 1352 | 1766 | 9.1 | 8220 | 7201 | $5.7x10^5$ |
| | 1818 | 1949 | $1.0x10^5$ | 1378 | 1765 | 9.5 | 7638 | 7108 | $6.5x10^5$ |
| | 1737 | 1967 | $1.2x10^5$ | 1370 | 1765 | 9.4 | 7068 | 7209 | $5.1x10^5$ |
| | 1515 | 2002 | $1.2x10^5$ | 1530 | 1765 | 11.7 | 7281 | 7152 | $6.0x10^5$ |
| 10% | 1269 | 1670 | $1.5x10^4$ | 978 | 1559 | 2.9 | 5000 | 4849 | $1.3x10^5$ |
| | 1328 | 1695 | $1.7x10^4$ | 1115 | 1559 | 3.8 | 5263 | 4888 | $1.3x10^5$ |
| | 1169 | 1694 | $2.0x10^4$ | 1198 | 1559 | 4.4 | 5481 | 4802 | $1.5x10^5$ |
| | 1362 | 1701 | $2.3x10^4$ | 1122 | 1559 | 3.9 | 4818 | 4800 | $1.2x10^5$ |
| | 1277 | 1677 | $1.5x10^4$ | 942 | 1559 | 2.7 | 4820 | 4865 | $1.0x10^5$ |
| 13% | 904 | 1391 | $1.5x10^3$ | 810 | 1341 | 1.0 | 2813 | 3062 | $1.6x10^4$ |
| | 992 | 1396 | $1.6x10^3$ | 794 | 1340 | 0.9 | 3002 | 3077 | $1.8x10^4$ |
| | 911 | 1390 | $1.4x10^3$ | 766 | 1341 | 0.9 | 2996 | 3084 | $1.7x10^4$ |
| | 946 | 1391 | $1.4x10^3$ | 826 | 1341 | 1.0 | 2870 | 3110 | $1.7x10^4$ |
| | 965 | 1387 | $1.4x10^3$ | 967 | 1341 | 1.4 | 3133 | 3074 | $2.0x10^4$ |
| 15% | 691 | 1267 | $5.7x10^2$ | 738 | 1231 | 0.5 | 2314 | 2395 | $5.6x10^3$ |
| | 750 | 1263 | $3.2x10^2$ | 690 | 1231 | 0.4 | 2113 | 2398 | $6.4x10^3$ |
| | 889 | 1267 | $4.8x10^2$ | 709 | 1231 | 0.4 | 2051 | 2392 | $5.2x10^3$ |
| | 828 | 1267 | $4.7x10^2$ | 578 | 1231 | 0.3 | 2463 | 2397 | $6.7x10^3$ |
| | 817 | 1266 | $3.6x10^2$ | 659 | 1231 | 0.4 | 2539 | 2401 | $6.0x10^3$ |

q = Sampling Fraction

SD Est = Standard Deviation of actual estimates

Mean ESE, Var ESE= Mean, Variance of Estimated Standard Errors

100 simulations per block (row of table), 5 blocks at each sampling fraction

| | D2+ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| q | Geometric Petition | | | Singles & Doubles | | | Uniform Petition | | |
| | SD Est | Mean ESE | Var ESEs | SD Est | Mean ESE | Var ESEs | SD Est | Mean ESE | Var ESEs |
| 3% | 4041 | 2984 | 373.3 | 3291 | 2971 | 249.4 | 7036 | 3287 | 955 |
| | 3967 | 2983 | 360.4 | 3628 | 2976 | 302.4 | 7719 | 3293 | 1156 |
| | 3826 | 2983 | 335.2 | 3732 | 2971 | 319.5 | 8240 | 3290 | 1317 |
| | 3894 | 2986 | 346.7 | 3974 | 2973 | 363.9 | 7248 | 3283 | 1022 |
| | 3710 | 2982 | 315.2 | 3627 | 2972 | 302.4 | 7728 | 3289 | 1158 |
| 5% | 2439 | 2283 | 66.7 | 2184 | 2277 | 53.8 | 4755 | 2497 | 220 |
| | 2179 | 2281 | 53.3 | 2245 | 2276 | 56.8 | 4947 | 2494 | 240 |
| | 2196 | 2284 | 54.1 | 2216 | 2276 | 55.3 | 4788 | 2496 | 225 |
| | 1902 | 2283 | 40.5 | 2298 | 2276 | 59.5 | 4594 | 2497 | 207 |
| | 2639 | 2282 | 78.1 | 2181 | 2276 | 53.6 | 4437 | 2497 | 193 |
| 8% | 1403 | 1770 | 9.8 | 1368 | 1765 | 9.3 | 3306 | 1916 | 49 |
| | 1497 | 1770 | 11.1 | 1352 | 1766 | 9.1 | 3020 | 1916 | 42 |
| | 1507 | 1770 | 11.3 | 1378 | 1765 | 9.5 | 2553 | 1915 | 29 |
| | 1451 | 1770 | 10.4 | 1370 | 1765 | 9.4 | 2611 | 1916 | 31 |
| | 1269 | 1770 | 8.0 | 1530 | 1765 | 11.7 | 3023 | 1917 | 41 |
| 10% | 1203 | 1563 | 4.4 | 978 | 1559 | 2.9 | 2285 | 1672 | 14 |
| | 1267 | 1563 | 4.8 | 1115 | 1559 | 3.8 | 2222 | 1673 | 14 |
| | 1038 | 1563 | 3.2 | 1198 | 1559 | 4.4 | 2019 | 1673 | 12 |
| | 1087 | 1563 | 3.7 | 1122 | 1559 | 3.9 | 2078 | 1673 | 12 |
| | 1153 | 1563 | 4.0 | 942 | 1559 | 2.7 | 2152 | 1673 | 13 |
| 13% | 890 | 1344 | 1.1 | 810 | 1341 | 1.0 | 1600 | 1424 | 3 |
| | 924 | 1344 | 1.3 | 794 | 1340 | 0.9 | 1662 | 1424 | 4 |
| | 842 | 1343 | 1.1 | 766 | 1341 | 0.9 | 1496 | 1424 | 3 |
| | 880 | 1344 | 1.1 | 826 | 1341 | 1.0 | 1678 | 1423 | 4 |
| | 916 | 1344 | 1.2 | 967 | 1341 | 1.4 | 1542 | 1424 | 3 |
| 15% | 687 | 1234 | 0.4 | 738 | 1231 | 0.5 | 1322 | 1299 | 2 |
| | 765 | 1234 | 0.5 | 690 | 1231 | 0.4 | 1282 | 1299 | 1 |
| | 842 | 1234 | 0.6 | 709 | 1231 | 0.4 | 1196 | 1299 | 1 |
| | 785 | 1234 | 0.5 | 578 | 1231 | 0.3 | 1173 | 1299 | 2 |
| | 789 | 1234 | 0.5 | 659 | 1231 | 0.4 | 1372 | 1299 | 2 |

q = Sampling Fraction

SD Est = Standard Deviation of actual estimates

Mean ESE, Var ESE= Mean, Variance of Estimated Standard Errors

100 simulations per block (row of table), 5 blocks at each sampling fraction

| | D Dup | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| q | Geometric Petition | | | Singles & Doubles | | | Uniform Petition | | |
| | SD Est | Mean ESE | Var ESEs | SD Est | Mean ESE | Var ESEs | SD Est | Mean ESE | Var ESEs |
| 3% | 4068 | 2985 | 378.6 | 3291 | 2971 | 249.4 | 7510 | 3293 | 1094 |
| | 3984 | 2984 | 363.6 | 3628 | 2976 | 302.4 | 7896 | 3299 | 1209 |
| | 3811 | 2983 | 332.2 | 3732 | 2971 | 319.5 | 8401 | 3295 | 1370 |
| | 3901 | 2986 | 347.9 | 3974 | 2973 | 363.9 | 7523 | 3289 | 1103 |
| | 3692 | 2982 | 311.7 | 3627 | 2972 | 302.4 | 8090 | 3295 | 1273 |
| 5% | 2434 | 2283 | 66.4 | 2184 | 2277 | 53.8 | 5152 | 2505 | 263 |
| | 2161 | 2281 | 52.3 | 2245 | 2276 | 56.8 | 5152 | 2502 | 262 |
| | 2193 | 2284 | 53.9 | 2216 | 2276 | 55.3 | 5027 | 2503 | 249 |
| | 1908 | 2283 | 40.9 | 2298 | 2276 | 59.5 | 4866 | 2504 | 234 |
| | 2652 | 2283 | 79.0 | 2181 | 2276 | 53.6 | 4520 | 2505 | 201 |
| 8% | 1403 | 1771 | 9.8 | 1368 | 1765 | 9.3 | 3551 | 1924 | 58 |
| | 1504 | 1771 | 11.3 | 1352 | 1766 | 9.1 | 3081 | 1925 | 44 |
| | 1508 | 1770 | 11.3 | 1378 | 1765 | 9.5 | 2892 | 1924 | 39 |
| | 1467 | 1771 | 10.7 | 1370 | 1765 | 9.4 | 2888 | 1925 | 39 |
| | 1291 | 1770 | 8.3 | 1530 | 1765 | 11.7 | 3367 | 1926 | 53 |
| 10% | 1216 | 1563 | 4.4 | 978 | 1559 | 2.9 | 2553 | 1673 | 14 |
| | 1286 | 1563 | 4.8 | 1115 | 1559 | 3.8 | 2540 | 1674 | 14 |
| | 1058 | 1563 | 3.2 | 1198 | 1559 | 4.4 | 2320 | 1674 | 11 |
| | 1092 | 1563 | 3.7 | 1122 | 1559 | 3.9 | 2355 | 1674 | 11 |
| | 1161 | 1563 | 4.0 | 942 | 1559 | 2.7 | 2355 | 1674 | 13 |
| 13% | 906 | 1344 | 1.1 | 810 | 1341 | 1.0 | 1836 | 1425 | 3 |
| | 931 | 1344 | 1.3 | 794 | 1340 | 0.9 | 1979 | 1425 | 4 |
| | 846 | 1343 | 1.1 | 766 | 1341 | 0.9 | 1754 | 1425 | 3 |
| | 886 | 1344 | 1.1 | 826 | 1341 | 1.0 | 1977 | 1424 | 4 |
| | 924 | 1344 | 1.2 | 967 | 1341 | 1.4 | 1869 | 1425 | 3 |
| 15% | 715 | 1234 | 0.4 | 738 | 1231 | 0.5 | 1571 | 1300 | 2 |
| | 782 | 1234 | 0.5 | 690 | 1231 | 0.4 | 1625 | 1300 | 1 |
| | 852 | 1234 | 0.6 | 709 | 1231 | 0.4 | 1454 | 1300 | 1 |
| | 794 | 1234 | 0.5 | 578 | 1231 | 0.3 | 1425 | 1300 | 1 |
| | 799 | 1234 | 0.5 | 659 | 1231 | 0.4 | 1563 | 1300 | 2 |

q = Sampling Fraction

SD Est = Standard Deviation of actual estimates

Mean ESE, Var ESE= Mean, Variance of Estimated Standard Errors

# Appendix 5

## Cov $(\hat{U}, \hat{D})$ for $D_{\text{Mod Esty}}$

Cov( $\hat{U}$ , $\hat{D}$), where $\hat{U}$ is the estimated number of ineligible signatures in the petition, and $\hat{D}$ is the number of duplicated signatures in the petition according to the modified Esty estimator, is

$$\text{Cov}\left(\hat{U}, N - \hat{U} - D_{\text{Mod Esty}}\right) = \text{Cov}\left(\hat{U}, N - \hat{U} - \frac{(n-u)d}{(n-u)-(1-q)d}\right)$$

$$= \text{Cov}\left(\hat{U}, N\right) - \text{Var}\left(\hat{U}\right) - \text{Cov}\left(\hat{U}, \frac{(n-u)d}{(n-u)-(1-q)d}\right)$$

N is a known constant, so Cov( $\hat{U}$ , N) = 0; and Var( $\hat{U}$ ) has already been shown to be $\frac{u(n-u)}{q^2 n}$. Thus we only need to determine $\text{Cov}\left(\hat{U}, \frac{(n-u)d}{(n-u)-(1-q)d}\right)$

$$\text{Cov}\left(\hat{U}, \frac{(n-u)d}{(n-u)-(1-q)d}\right) = \text{Cov}\left(\frac{N}{n}u, \frac{(n-u)d}{(n-u)-(1-q)d}\right) = \frac{1}{q}\text{Cov}\left(u, \frac{(n-u)d}{(n-u)-(1-q)d}\right)$$

It is obvious that the size of the sample must be equal to the number of ineligible signatures in it, plus the number of unique eligible signatures in it, plus the number of eligible but duplicated signatures in it...

ie $n = u + d + \sum_{i=2}^{k}(i-1)f_i = u + \sum_{i=1}^{k} i f_i$

So

$$\frac{1}{q}\text{Cov}\left(u, \frac{(n-u)d}{(n-u)-(1-q)d}\right) = \frac{1}{q}\text{Cov}\left(n - \sum_{i=1}^{k} i f_i, \frac{\sum_{m=1}^{k} m f_m \sum_{r=1}^{k} f_r}{\sum_{j=1}^{k} j f_j - (1-q)\sum_{j=1}^{k} f_j}\right)$$

$$= \frac{-1}{q} \mathrm{Cov}\left( \sum_{i=1}^{k} i f_i, \frac{\sum_{m=1}^{k} m f_m \sum_{r=1}^{k} f_r}{\sum_{j=1}^{k} (j+q-1) f_j} \right)$$

Writing $f_{i|g}$ for the number of eligible individuals who have signed the petition g times, and whose signatures appear i times in the sample, then $f_i = \sum_{g=1}^{k} f_{i|g}$

and $\frac{-1}{q} \mathrm{Cov}\left( \sum_{i=1}^{k} i f_i, \frac{\sum_{m=1}^{k} m f_m \sum_{r=1}^{k} f_r}{\sum_{j=1}^{k} (j+q-1) f_j} \right) =$

$$\frac{-1}{q} \sum_{i=1}^{k} i \, \mathrm{Cov}\left( \sum_{g=i}^{k} f_{i|g}, \frac{\sum_{m=1}^{k} m \left( \sum_{n=m}^{k} \left( f_{m|n}^2 + \sum_{\substack{r=1 \\ r \neq m}}^{k} f_{m|n} \, f_{r|n} + \sum_{r=1}^{k} \sum_{\substack{t=r \\ t \neq n}}^{k} f_{m|n} \, f_{r|t} \right) \right)}{\sum_{j=1}^{k} \left( (j+q-1) \sum_{h=j}^{k} f_{j|h} \right)} \right)$$

$$= \frac{-1}{q} \sum_{i=1}^{k} \sum_{m=1}^{k} im \sum_{g=i}^{k} \sum_{n=m}^{k} \mathrm{Cov}\left( f_{i|g}, \frac{f_{m|n}^2}{\sum_{j=1}^{k} \left( (j+q-1) \sum_{h=j}^{k} f_{j|h} \right)} \right)$$

$$- \frac{1}{q} \sum_{i=1}^{k} \sum_{m=1}^{k} im \sum_{\substack{r=1 \\ r \neq m}}^{k} \sum_{g=i}^{k} \sum_{n=\max(m,r)}^{k} \mathrm{Cov}\left( f_{i|g}, \frac{f_{m|n} \, f_{r|n}}{\sum_{j=1}^{k} \left( (j+q-1) \sum_{h=j}^{k} f_{j|h} \right)} \right)$$

$$-\frac{1}{q}\sum_{i=1}^{k}\sum_{m=1}^{k} im \sum_{r=1}^{k}\sum_{g=i}^{k}\sum_{n=m}^{k}\sum_{\substack{t=r \\ t\neq n}}^{k} \mathrm{Cov}\left( f_{i|g}, \frac{f_{m|n}\, f_{r|t}}{\sum_{j=1}^{k}\left( (j+q-1)\sum_{h=j}^{k} f_{j|h} \right)} \right)$$

Now, Cov(XY) = E[XY] – E[X]E[Y], so the covariance terms above can be written

$$\mathrm{Cov}\left( f_{i|g}, \frac{f_{m|n}^2}{\sum_{j=1}^{k}\left( (j+q-1)\sum_{h=j}^{k} f_{j|h} \right)} \right) = E\left[ \frac{f_{i|g} f_{m|n}^2}{\sum_{j=1}^{k}\left( (j+q-1)\sum_{h=j}^{k} f_{j|h} \right)} \right] - E\left[ f_{i|g} \right] E\left[ \frac{f_{m|n}^2}{\sum_{j=1}^{k}\left( (j+q-1)\sum_{h=j}^{k} f_{j|h} \right)} \right]$$

$$\mathrm{Cov}\left( f_{i|g}, \frac{f_{m|n}\, f_{r|n}}{\sum_{j=1}^{k}\left( (j+q-1)\sum_{h=j}^{k} f_{j|h} \right)} \right) = E\left[ \frac{f_{i|g} f_{m|n}\, f_{r|n}}{\sum_{j=1}^{k}\left( (j+q-1)\sum_{h=j}^{k} f_{j|h} \right)} \right] - E\left[ f_{i|g} \right] E\left[ \frac{f_{m|n}\, f_{r|n}}{\sum_{j=1}^{k}\left( (j+q-1)\sum_{h=j}^{k} f_{j|h} \right)} \right]$$

$$\mathrm{Cov}\left( f_{i|g}, \frac{f_{m|n}\, f_{r|t}}{\sum_{j=1}^{k}\left( (j+q-1)\sum_{h=j}^{k} f_{j|h} \right)} \right) = E\left[ \frac{f_{i|g} f_{m|n}\, f_{r|t}}{\sum_{j=1}^{k}\left( (j+q-1)\sum_{h=j}^{k} f_{j|h} \right)} \right] - E\left[ f_{i|g} \right] E\left[ \frac{f_{m|n}\, f_{r|t}}{\sum_{j=1}^{k}\left( (j+q-1)\sum_{h=j}^{k} f_{j|h} \right)} \right]$$

These expressions could be expanded using the first few terms of Taylor series, eg

$$E\left[\frac{X}{Y}\right] \simeq \frac{\mu_X}{\mu_Y} \text{ or } E\left[\frac{X}{Y}\right] \simeq \frac{\mu_X}{\mu_Y} - \frac{\mathrm{Cov}(X,Y)}{\mu_Y^{\,2}} + \frac{\mu_X}{\mu_Y^{\,3}} \mathrm{Var}(Y) \quad \text{(Mood } et\ al\text{, 1974)}$$

and $E(f_{i|g}) = F_g \left( \frac{g!}{i!(i-g)!} \right) q^i (1-q)^{g-i}$ (Binomial approximation to hypergeometric)

but they become unwieldy. For instance, using the first-order Taylor expansion, the term

$$E\left[\frac{f_{i|g}f_{m|n}^2}{\sum_{j=1}^{k}\left((j+q-1)\sum_{h=j}^{k}f_{j|h}\right)}\right] - E\left[f_{i|g}\right]E\left[\frac{f_{m|n}^2}{\sum_{j=1}^{k}\left((j+q-1)\sum_{h=j}^{k}f_{j|h}\right)}\right]$$

alone is approximately

$$\frac{E\left[f_{i|g}f_{m|n}^2\right] - E\left[f_{i|g}\right]E\left[f_{m|n}^2\right]}{E\left[\sum_{j=1}^{k}\left((j+q-1)\sum_{h=j}^{k}f_{j|h}\right)\right]}$$

Now $E\left[f_{i|g}f_{m|n}^2\right] - E\left[f_{i|g}\right]E\left[f_{m|n}^2\right] = E\left[f_{i|g}\right]E\left[f_{m|n}^2\right] + Cov(f_{i|g}, f_{m|n}^2) - E\left[f_{i|g}\right]E\left[f_{m|n}^2\right]$

$= Cov(f_{i|g}, f_{m|n}^2)$

Applying a similar approach to the other terms, we get $Cov\left(\hat{U}, \frac{(n-u)d}{(n-u)-(1-q)d}\right)$

$$\simeq \frac{-\sum_{i=1}^{k}\sum_{m=1}^{k} im \sum_{g=i}^{k}\sum_{n=m}^{k}\left(Cov\left(f_{i|g}, f_{m|n}^2\right) + + \sum_{r=1}^{k}\sum_{\substack{t=r \\ t\neq n}}^{k} Cov\left(f_{i|g}, f_{m|n}\, f_{r|t}\right)\right)}{q\,E\left[\sum_{j=1}^{k}\left((j+q-1)\sum_{h=j}^{k}f_{j|h}\right)\right]}$$

$$- \frac{\sum_{i=1}^{k}\sum_{m=1}^{k} im \sum_{g=i}^{k}\sum_{\substack{r=1 \\ r\neq m}}^{k}\sum_{n=max(m,r)}^{k} Cov\left(f_{i|g}, f_{m|n}\, f_{r|n}\right)}{q\,E\left[\sum_{j=1}^{k}\left((j+q-1)\sum_{h=j}^{k}f_{j|h}\right)\right]}$$

And $E\left[\sum_{j=1}^{k}\left((j+q-1)\sum_{h=j}^{k}f_{j|h}\right)\right] = \sum_{h=1}^{k}F_h\left(qh - (1-q)(1-(1-q)^h)\right)$ (Lemma 2)

Although it is possible to derive an approximate expression for $Cov(f_{i|g}, f_{h|g})$ by assuming that the $f_{i|g}$ $(0 \leq i \leq g)$ follow a multinomial distribution; and Smith-Cayama & Thomas give an expression for $Cov(f_{i|g}, f_{m|n})$ based on the hypergeometric distribution, we have not been able to derive an expression for $Cov(f_{i|g}, f_{m|n}\, f_{r|t})$. Fortunately, the simulation studies in Chapter 5 offer an alternative source of information about the approximate value of $Cov(\hat{U}, \hat{D})$. They suggest that compared to the variance of $\hat{U}$ and $\hat{D}$, the covariance is small, and correcting for it does not seriously alter the estimated standard errors.

**Lemma 1**

$$E\left[f_{m|n}^2\right] = E\left[f_{m|n}\right] E\left[f_{m|n}\right] + Cov(f_{m|n}, f_{m|n}) = E^2\left[f_{m|n}\right] + Var(f_{m|n})$$

$$\simeq \left(F_n\left(\frac{n!}{m!(n-m)!}\right)q^m(1-q)^{n-m}\right)^2$$

$$+ F_n\left(\left(\frac{n!}{m!(n-m)!}\right)q^m(1-q)^{n-m}\right)\left(1-\left(\frac{n!}{m!(n-m)!}\right)q^m(1-q)^{n-m}\right)$$

$$= \left(F_n\left(\frac{n!}{m!(n-m)!}\right)q^m(1-q)^{n-m}\right)\left((F_n-1)\left(\left(\frac{n!}{m!(n-m)!}\right)q^m(1-q)^{n-m}\right)+1\right)$$

**Lemma 2**

$$E\left[\sum_{j=1}^{k}\left((j+q-1)\sum_{h=j}^{k} f_{j|h}\right)\right] = \sum_{j=1}^{k}\left((j+q-1)\sum_{h=j}^{k} E\left[f_{j|h}\right]\right)$$

$$= (q-1)\sum_{j=1}^{k}\sum_{h=j}^{k} E\left[f_{j|h}\right] + \sum_{j=1}^{k} j\sum_{h=j}^{k} E\left[f_{j|h}\right]$$

$$= -(1-q)\left(\sum_{h=1}^{k} F_h(1\text{-prob}(0 \text{ out of } h \text{ in sample}))\right) + \left(\sum_{h=1}^{k} F_h \sum_{j=1}^{h} j \text{ prob}(j \text{ out of } h \text{ in sample})\right)$$

$$= \sum_{h=1}^{k} F_h \left( qh - (1-q)(1-(1-q)^h) \right)$$

## References


Bunge J & Fitzpatrick M (1993) 'Estimating the number of species in a population: a review' Journal of the American Statistical Association, **88**, 421, pp 364-373

Bunge J A & Handley J C (1991) 'Sampling to Estimate the Number of Duplicates in a Database' Computational Statistics and Data Analysis, **11**, pp 65-74

DeGroot M H (1986) 'Probability and Statistics' Addison-Wesley

Esty W (1985) 'Estimation of the number of classes in a population and the coverage of a sample' Mathematical Scientist, **10**, pp 41-50

Fienberg S E, Johnson M S & Junker B W (1999) 'Classical multilevel and Bayesian approaches to population size estimation using multiple lists' Journal of the Royal Statistical Society, Series A, **162**, 3, pp 383-407

Goodman L (1949) 'On the estimation of the number of classes in a population' Annals of Mathematical Statistics, **20**, pp 572-579

Haas P J & Stokes L (1998) 'Estimating the Number of Classes in a Finite Population' Journal of the American Statistical Association, **93**, 444, pp 1475-1487

Hill B (1968) 'Posterior distribution of percentiles: Bayes' Theorem for sampling from a population' Journal of the American Statistical Association, **63**, pp 677-691

Hill B (1979) 'Posterior moments of the number of species in a finite population and the posterior probability of finding a new species' Journal of the American Statistical Association, **74**, pp 668-673

Kish L (1965) 'Survey Sampling', Wiley

Lindley D V & Scott W F (1984) 'New Cambridge Elementary Statistical Tables' Cambridge University Press

Mood A M, Graybill F A & Boes D C (1974) 'Introduction to the Theory of Statistics', MacGraw-Hill

Shlosser A (1981) 'On estimation of the size of the dictionary of a long text on the basis of a sample' Engineering Cybernetics, **19**, pp 97-102

Shuster J (1974) 'The validation of a petition by a sample survey' Communications in Statistics, **3**, 3, pp 291-296

Smith-Cayama R A & Thomas D R (1999) 'Estimating the number of distinct valid signatures in initiative petitions' presented at American Statistical Association Joint Statistics Meeting, Baltimore, 1999