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.