The Sampling Distribution of the W Estimator of the Number

of Valid Signatures on a Petition

Mark E. Eakin, Ph.D., Professor

Box 19437

Department of Information Systems and Operations Management

The University of Texas at Arlington

Arlington, TX 76019

Eakin@uta.edu

(817) 272-3529

Mary M. Whiteside, Ph.D., Professor

Box 19437

Department of Information Systems and Operations Management

The University of Texas at Arlington

Arlington, TX 76019

whiteside@uta.edu

(817) 272-3517

Prepared for Submission to *Communications in Statistics*

The Sampling Distribution of the W Estimator of the Number

of Valid Signatures on a Petition

**Abstract**

A new nonlinear estimator, W, for the number of valid, unique signatures on a petition has been shown better, for the cases enumerated and with certain restrictions, than a popular Goodman type statistic, G. This paper extends those results with relaxed conditions by developing the exact probability mass function (PMF) and mean of W and a close approximation of the variance (Var(W)). If the proportion of valid signatures among unique and duplicated signatures is the same, then Var(W) is approximately a function of the means and variances of the two sample statistics. Using the Delta method we estimate Var(W) with the resulting approximation shown to be good, even when the condition of equal proportions does not hold. We compare W to G and establish which estimator is preferred for different intervals of the design parameters. Data from a Washington State petition illustrate the findings.

**Key words**: Classes in a finite population; Goodman estimator; Delta method; Variance approximation

The Sampling Distribution of the W Estimator of the Number

of Valid Signatures on a Petition

**1. Introduction**

The purpose of this paper is to extend the results for the W estimator (Whiteside and Eakin, 2008) introduced as a preferred alternative to the older Goodman estimator, G, of the number of valid signatures on a petition for certain conditions. Recent examples of direct democracy in the form of initiatives, referenda, and recalls are numerous. These applications originate with the validation of signatures on a petition. Petitions are not part of national governance in the United States but are in the European Union and many countries including New Zealand, Canada, and Venezuela. Many states and local governmental entities within the United States allow petitions with signatures of qualified voters for ballot access, recall of elected officials, or citizen initiated referenda and propositions. In November 2010 California rejected Proposition 19 to legalize marijuana. The proposition gained ballot access by a petition with more than half a million signatures that was validated using a random sample and a modified (biased) G estimator. In the summer of 2012 Governor Scott Walker of Wisconsin, with international attention, became the first governor to survive a recall vote (Montopoli, 2012). The vote was triggered by a petition with one million signatures (Marley, 2012).

The problem of verifying that a petition presents the required number of valid signatures appeared in the literature by at least 1969 (Owens, 1969). Statistically, this problem is interesting because the number of duplicated (replicated) signatures and the number of signatures invalid for other reasons (the signer is not a registered voter, the address is nonexistent, or some such error) must be estimated differently. In what follows we refer to these as duplicated and invalid signatures, respectively. Both Owens (1969) and Shuster (1974) approach the problem of verifying a petition from a hypothesis testing perspective.

Smith and Thomas (Smith and Thomas 2005, Smith-Cayama and Thomas 1999) consider several linear estimators and one nonlinear estimator of the number of otherwise valid but duplicated signatures on a petition. They provide expressions for the variance of the linear estimators and estimate the root mean squared error for the biased nonlinear estimator adapted from Haas and Stokes (1998). The estimators of the number of distinct valid signatures on a petition are obtained by subtracting from the total number of signatures (1) the estimated number of invalid signatures (estimated in the usual way as the sample proportion of invalid signatures multiplied by the petition size) and then (2) the estimated number of valid but duplicated signatures. The performance of the estimators is evaluated using data from several completely verified Washington State petitions. Smith and Thomas conclude that for sampling fractions less than or equal to 10% “it is difficult to improve much on the Goodman-type estimator” (Goodman 1949) that is unbiased when replicates are at most duplicates. Sample fractions of this size are typical for many state petitions.

Hedderley and Haslett (2005) investigate the performance of the Smith-Cayama and Thomas (1999) estimators with two simulations. The first examines the estimators of the number of duplicated valid signatures by simulated sampling from artificial petitions with known distributions of multiple signatures. Effects of petition size, sample fraction, and distribution of multiple signatures on bias, distribution, variability and estimated variability of the estimators of duplicate pairs are investigated. They find no one estimator performing “clearly better than the others.” In addition Hedderley and Haslett simulate the variance of the estimator of the number of valid signatures, considering the effect of the variance of the estimator of the number of duplicated valid signatures, the variance of the estimator of the invalid signatures, and the covariance of the two estimators. The variance of the estimator of invalid signatures has an appreciable effect in some cases, but the covariance of the two estimators does not.

All estimators of the number of distinct valid signatures considered by Smith and Thomas start with petition size and subtract off both the estimated number of invalid signatures and estimated number of duplicated valid signatures. Estimating the number of duplicated signatures among signatures that are otherwise valid reduces to determining the number of distinct classes in a finite population. As noted by Haas and Stokes (1998), the initial applications in the literature of this general problem appeared in response to a question submitted by Charles Callard in 1949 to Editor Frederick Mosteller’s “Questions and Answers” column in *The American Statistician* (Mosteller 1949). Mosteller summarizes in the column Goodman’s not yet published results (Goodman 1949), along with other answers. The context for Goodman’s initial work was the problem of determining from a sample how many different colors of balls are in an urn. Equivalent to determining the number of classes in a finite population is the problem of matching lists based on samples (Goodman 1952). For the problem of verifying petitions, Smith and Thomas have reduced a petition of signatures to a population of valid signatures with each signatory a distinct class. Physically, the petition may consist of k lists of signatures, gathered at different locations or on different dates. The question becomes, how many distinct valid signatures are in the pooled lists?

The above cited publications either consider only valid signatures or categorize all sample signatures that happen to be both invalid and duplicated only as invalid. Whiteside and Eakin (2008) present a new nonlinear estimator for the number of valid distinct signatures that is a function of all the duplicated signatures in the sample, both valid and invalid. This estimator is shown to be superior to the popular Goodman type estimator referenced above with certain restrictions. The conditions are that replicates are at most duplicates and that the proportion of valid signatures among uniques and duplicates is the same. The values of the parameters used by Whiteside and Eakin are in Table 1.

(Insert Table 1 about here.)

**2. The New Estimator: W**

This paper develops the sampling distribution of W and compares the performance of W, when the original restriction of equal proportions is relaxed, to the performance of the often cited Goodman type estimator, G. The authors are assuming simple random sampling without replacement, that no signature appears more than twice, and that signature entries described as a duplicate pair are identical in all respects. That is, both members of a pair are either valid or invalid.

Using the notation of Table 2, the class of Smith-Cayama estimators is of the form

.

The Goodman-type estimator, G, is of the Smith-Cayama form with

.

The Whiteside-Eakin estimator, W, defined as

,

incorporatesdias well as dv to estimate the number of duplicate pairs.

(Insert Table 2 about here.)

We prove in Theorem 1 of the Appendix that W is unbiased with the assumptions stated above and the additional restriction that the percent of duplicate valids among the total number of valids equals the percent of duplicate invalids among the total number of invalids

Dv/Nv =Di/Ni (1a)

or alternately that the percent of valids among the uniques is the same as the percent of valids among the duplicates,

Uv/U = Dv/D= Nv/N. (1b)

Based on a numeric computation of the exact probability mass function (PMF) of W under (1a), Var(G) is greater than Var(W) by from 1% to as much as 13% for the design parameters considered in Whiteside and Eakin (2008).

Blote and van Leeuwen (1998) establish the difficulty of observing a sufficient number of duplicated signatures in a sample to accurately estimate duplicates on the petition, even with large samples. Thus, intuition suggests that the additional information provided by di is useful if, in fact, di can tell us something about Dv, which is the case if (1a) holds*.* Thispaper will establish the relative performance of the two estimators, G and W, when the proportion of duplicates is not exactly the same for valid and invalid signatures.

**2.1 *Variance, MSE, and Bias of W***

W can be expressed as

. (2)

Now, in the general case, Var(W) can be approximated using the Delta method. See Appendix for the derivatives, variances, and covariances required for this case, to which we refer as the standard approximation.

In the special case of (1a), the factors of (2), though uncorrelated as shown in Proposition 2 of the Appendix, are not independent. However, we illustrate in the Appendix that the joint PMF of mv/m and d is approximately bivariate normal and therefore the variables mv/m and d are approximately independent in this design as a consequence of the large sample sizes. From Goodman (1960), the variance of the product of two independent variables, XY, is represented as

.

Thus, in the special case of W under (1a), we have an approximation for Var(W) as follows.



where

.

Since



(Raj 1961), where

,

it remains only to approximate



using the Delta method in order to approximate Var(W). The derivatives, variances and covariances needed in order to use the Delta method when (1a) holds, to which we refer as the uncorrelated approximation, are found in the Appendix.

For the uncorrelated case when (1a) holds, W is unbiased; so in order to compare the performances of G and W we can compare variances. However, for the correlated case, instead of variance we must compare MSE(W) with Var(G). MSE(W) is Var(W) plus the squared bias. In the process of applying the Delta method above we obtain the following approximation of the expected value of W by replacing nv, dv, and di in the expression (2)with their expected values:

. (3)

Thus, the approximated bias can be derived as the difference between E[W] and M*v* and the MSE of W can also be approximated.

**2.2 *PMF for Numerical Computation of G and W***

To examine the accuracy of our approximations of the expected value, variance, and MSE(W) and to compare the performance of W to G when (1a) does not hold, we require the PMF’s of G and W, neither of which has a general closed form solution but must be derived from the PMF of the joint distribution of nv and dv and a joint distribution containing nv, dv, and di respectively. This latter distribution is an extension of the results in Blote and van Leeuwen (1998) and is developed below.

In any sample of size n, besides the observable variables, dv and di, there are four unobservable variables uv, d'v, ui, d'i where uv is the number of unique valid signatures in the sample, d'v is the number of duplicate valid signatures without their paired values in the sample with ui and d'i defined similarly for the invalid signatures in the sample. The joint PMF of the six variables is obtained analogously to a multivariate hypergeometric where the uniques and the duplicates (with pairs in the sample) use the combinational formula but the formula for the number of duplicates (without pairs in the sample) has to be modified. The number of ways that d'v could occur is then. The number of ways d'i could occur is calculated in the same manner. This leads to the following PMF:

.

In order to obtain the PMF containing nv, dv, and di, rewrite d'v as (nv-uv-2dv) and d'i similarly and then sum over the possible values of uv and ui resulting in the re-parameterization:

. (4)

We obtain the PMF of G by taking the marginal density function of (4) with respect to nv and dv.

**3. Numerical Computation of PMF**

We compute the PMF of both W and G in order to assess the accuracy of our delta approximations of the expected value and Var(W) and to compare MSE(W) to Var(G). The design parameters for this numerical computation appear in Table 3 and are meant to be similar to the range of values reported in actual petitions (Smith and Thomas 2005). To accomplish this numerical calculation we select the likely values of nv, dv, and di given the design parameters. The likely values of nv, dv, and di are those within eight standard errors of the respective means. This procedure yields at least 99.9999% of the total PMF of W in each of the 243 design parameter combinations.

(Insert Table 3 about here.)

The numerical computation of W shows the delta method in (3) gives an excellent approximation of the expected value of W. The approximation differs from the expected value of W by at most one signature in just 10 cases when estimates are rounded to the nearest integer. For all other design combinations the approximation is exact.

The delta method approximation for Var(W) shows reasonable accuracy in most cases. Figure 1 and Table 4 display the percent difference between the actual and approximated Var(W) under both the standard and uncorrelated delta methods. Because the error in approximating Var(W) can be substantial for the standard case with the delta method (as much as a 35% underestimate), a conservative approach is always to approximate Var(W) assuming the uncorrelated case. The delta approximation for the uncorrelated case will be somewhat too large on average (in the worst cases less than 10%) but seldom too small. So, future applications that require an estimate of Var(W), in which the conservative approach is to overestimate, should develop this estimate based on the delta method under the uncorrelated case.

(Insert Figure 1 and Table 4 about here.)

**3.1 *Bias Correction***

We develop a multiple regression model to estimate the ratio of (3) to MV and use this ratio as a multiplier of W to define a new bias adjusted estimator E. The estimated variance of E can be constructed equivalently using the delta method. Under (1a), the mean square errors of the two estimators are almost identical. Where (1a) does not hold, the estimator E has a smaller bias and in a few cases a smaller mean square error. However, it appears that generally any advantage of E over W with respect to mean absolute error is offset by a greater value of MSE. Thus, E exhibits no appreciable gain by estimating and correcting bias.

***3.2 Comparison of MSE (W) and Var(G)***

Using the numerically computed values we compare the MSE(W) and Var(G), depicted in Figure 2, with the following measure of relative efficiency:



This extends the comparisons in Whiteside and Eakin (2008) to also include cases where (1a) does not hold. Figure 2 depicts the relationship of the relative efficiency with the population size for all values of the other parameters of Table 3.

(Insert Figure 2 about here.)

The three superimposed lines in each graph correspond to the levels of the PDi; each column of graphs correspond to a value of PDv; the rows of graphs correspond to the levels of Pv; and Figures 2a, 2b, and 2c correspond to differing sampling fractions. As the sampling fraction increases, the range of the relative efficiencies increases from (-.4, 1) for sampling fractions of 5% to (-.5, 3) for sampling fractions of 25%. For a given sampling fraction, the nine plots depict the same relative patterns for all combinations of other design values. Note that the advantage goes to W when the relative efficiency measure is below zero. As previously shown in Whiteside and Eakin (2008), W is more efficient than G when PDv=PDi in all 27 graphs. On the other hand, G is more efficient than W when |PDv-PDi| is greatest.

From the structure of W, it is clear that the benefit of W over G occurs when the sample information provided by di, pooled with dv, improves the estimate of duplicate valids on the petition. However, as the difference PDv and PDi increases, the pooled estimate of duplicates used by W increasingly contaminates the estimate of the number of duplicates among valid signatures. As seen in Figure 2 as the absolute difference between PDv and PDi increases, the difference in relative efficiency becomes larger.

The mean square error of W is impacted by the interaction of the sampling fraction and the absolute difference between the PDv and PDi since increasing the sample size increases the variance of counts, and we are dealing with counts in the estimator. PDv and PDi are set in this numerical computation to take on the same three values yielding nine factor level combinations. When PDv is at its smallest value and the difference between PDv and PDi is at its largest, then the influence of PDi dominates PDv and the relative efficiency increases. When PDv assumes its largest value, and the absolute difference between PDv and PDi is also at its largest, the weighting of di relative to dv is not so great as the first case. This pattern is observed in all plots with the sampling fraction proportionally increasing the range of the relative efficiencies across the patterns. Finally, it would seem that when PDv assumes a middle value, the impact of the absolute difference would be less because this difference itself is smaller. However this pattern is seen only for the two largest sampling fractions, indicating the interaction between the sampling faction and the absolute difference.

Two other effects are noted in the graphs of Figure 2. As the population size increases, the relative efficiency becomes larger but the effect appears to be mitigated by the percent of valids in the population and by the sampling faction.

In an attempt to approximate the relative efficiency as a simplified function of the numerous effects noted in the graphs, a model building approach was applied to help choose a parsimonious subset of these effects. A logistic regression model was developed in which the response variable is 1 if MSE (W) > Var(G); 0 otherwise. We examined numerous alternative models. One, though very simple, closely approximated the patterns observed in the graph. Independent variables are the absolute difference between PDv and PDi, the proportion of valid signatures, the proportion of valid duplicate signatures and a three way interaction of the sampling fraction, the population size and |PDv \_ PDi|. Note that the three way interaction is the two way interaction of the sample size and the absolute difference.

The signs of the regression coefficients are consistent with the discussion of Figure 2 and explanation of the effects provided above. As |PDv \_ PDi| increases, the odds that MSE(W)>Var(G) increases; but this interacts with the sampling fraction and population size (note sf\*N=n);i.e., increasing sample size increases the effect. As PDv increases, the odds that MSE(W)>Var(G) decreases. As Pv increases, the odds that MSE(W) >Var(G) decreases.

(Insert Table 5 about here.)

To investigate the predictive ability of the logistic model, we randomly divided the enumerated observations into two approximately equal sized sets. One set was used to fit the predictive model and then to predict the logistic response variable for the holdout set. We repeated this process 10 times with the lowest hit ratio being 90%, the highest 97%, with an average hit ratio of 93%.

This logistic model seems to reinforce the discussion of Figure 2 and allows for interpolation within the design values of Table 3 for assessing the relative merits of W compared to G for estimating the number of valid signatures on a petition. The smallest sample size for our numerical computation is 2500; the largest is 62,500. For these two sample sizes we perform a sensitivity analysis of the classification which indicates the more efficient estimator. For percent valid of 75%, a difference of 2% or more between duplicate valid and invalid signatures always favors G. (However, there is a slight advantage for W with a 2% difference if the percent valid is 85%, the sample size is 2500, and the percent duplicate valid signatures is 5%.) For a 1% difference between duplicate valid and invalid signatures with 75% valid signatures on the petition and sample size of 2500, W is preferred if the percent of duplicate valids is 5%; G if percent of duplicate valids is 2%. For sample sizes of 62,500, G is preferred within the parameters of our numerical computation for differences in percent duplicates between valid and invalid signatures of even 0.5%.

**4. Application of Relative Efficiency to a Washington State Petition**

This example of a Washington state petition comes from Smith & Thomas 2005. Their petition D consists of 228,148 signatures of which 34,542 were found to be invalid. Of the valid signatures, 11,584 were replicated. The replications ranged from signatures repeated twice to signatures repeated up to six times. Smith and Thomas used the rule that when an estimator is used that just assumes all replicates are duplicates, the other copies should be ignored. Using this rule, 550 signatures are removed from both the total number of signatures and the number of duplicated valid signatures. The parameter values for the Washington State petition with these 550 signatures removed are in Table 6.

(Insert Table 6 about here.)

Using the earlier numerical computation approach, we examine the relative efficiency of W and G using the MSE criterion for a range of the proportion of duplicate invalid signatures from .03 to .09, centered on the observed value of PDv =.06. The results are in shown in Figure 3 and Table 7.

(Insert Table 7 and Figure 3 about here.)

If the sampling fraction is less than 25% and the proportion of duplicate invalids is within ± 0.03 of the proportion of duplicate valids, W is generally (but not always) more efficient. However as the sampling fraction increases to 25%, W is only better in the non-correlated case. This is consistent with the prior results for MSE.

To illustrate the absolute accuracy of the two estimators, consider the case where the percent duplicate invalids is hypothesized to be 4% and the sampling fraction is small (5%). Note the observed percent of duplicate valids is 6%. The two standard deviation intervals for W and G about their expected values are from 167,760 to 175,606 and from 166,567 to 175,409, respectively. The percentage errors from 170,988, the observed number of valid unique signatures, are from -1.89% to 2.7% for (biased) W and from -2.59% to 2.59% for G. In the case of a 10% sampling fraction, the respective intervals are reduced: from 169,569 to 173,727 and from 168,689 to 173,287 with the percent errors from -0.83% to 1.6% and from -1.34% to 1.34%. For a 25% sampling fraction, the respective intervals are further reduced: from 170,622 to 172,464 and from 170,004 to 171,972 with the percent errors from -0.21% to 0.86% and from -0.58% to 0.58%.

For the 3% range of hypothesized differences between percent duplicates among valid and invalid signatures, the absolute percent error for W’s interval endpoint ranges from .05% (sampling fraction = 25%, percent duplicate invalids = 3%) to 3.08% (sampling fraction = 5%, percent duplicate invalids = 9%). If we take the average absolute error for the end points of each interval (average of |lower limit – 170,988| and |upper limit – 170,988|), we get 3,932 signatures for W and 4,441 signatures for G in the 5% sampling fraction case. These numbers are reduced to 2,079 and 2,299 for a 10% sampling fraction, and 921 and 984 for a 25% sampling fraction. By this criterion of average absolute error for the endpoints of a two standard deviation interval about the expected value, W is always more efficient than G for the range of sampling fractions and percent duplicate invalids considered.

**5. Conclusion**

Based on the properties exhibited by W and G for the cases considered in our numerical computations, the sensitivity analysis of the logistic model, and the analysis of the Washington State petition; we recommend W over G for relatively small sample sizes. As the petition size and the sampling fraction decrease, the MSE advantage of W relative to G increases. This MSE advantage for W compared to G also improves as the proportion of valid signatures exceeds 75%, the difference in the proportion of duplicates between valid and invalid signatures falls below 3%, and duplicate valid signatures of approximately 5% are reasonable.

**Appendix: Proofs and Approximated Variances**

***Proposition 1***: Under condition (1b)



where Eu,d[.] is the expectation over the joint PMF of u and d.

Proof:



where





.

Under condition (1b) the following results:



Therefore E[mv] is

.

***Proposition 2***: Under condition (1b)



Proof:

.

 Examining the product of the expectations first:

.

Using Proposition 1,



.

Next, examining the expectation of the product term:

.

Again using the result of Proposition 1:

.

Therefore, since the expectation of the product term and the product of the two expectations equal to the same value,

.

***Theorem 1: W is unbiased under condition (1b)***

***Proof:*** 



by Proposition 1



and by covariance of zero in Proposition 2





***Approximation of Var(W) Standard Case:*** By the delta method, the variance of a function of three variables can be approximated as



where the derivatives are evaluated at the expected values of x, y, and z.

The Whiteside-Eakin estimator is of the form:

.

It can be shown that







The derivatives are then evaluated at their expected values:





The variances and covariances are













where



***Justification for using Goodman’s variance of the product of independent variables with d and mv/m*:** If the joint distribution mv/m and d is approximately bivariate normal with zero covariance then mv/m and d are approximately independent. To illustrate that this joint is approximately bivariate normal for a large sample size, the numerically constructed PMF is depicted for the worst (i.e. largest relative Cov((mv/m)2,d2)) case in this study (N=50,000, n=2,500, PDv = 0.02, PDi=0.02 and Pv=0.95). Figure 4a depicts the numerically constructed PMF while Figure 4b is a spline-smoothed version of the PMF. Both Figures depict an approximate bivariate normal distribution.

 If X and Y are independent then Cov(X,Y) = Cov(X2,Y2) = 0 and Goodman’s formula holds. Cov(mv/m,d) is shown in Proposition 2 to be zero under condition (1b). Thus, in using Goodman’s (1960) expression to approximate Var(W), we are ignoring only Cov((mv/m)2,d2). The numerical calculation of Cov((mv/m)2,d2) in our illustration is zero to the fourth decimal place, 0.007% of the smallest expected value in the expression for variance:



=7.877378844 – (0.902510718)\*(8.728222113) = -0.000065.

(Insert Figures 4a and 4b here)

***Approximation of Var(W) Uncorrelated Case*:** It was shown in Section 2.1 that in order to approximate Var(W) under the uncorrelated case it is required to only approximate Var(mv/m) since Var(d) is known. Using a similar approach to the one above, the approximation requires derivatives, variances and covariances. For the following function

,

the derivatives are







with the variances and covariances given above in the standard case.

**References**

Blote, H. W. J. and van Leeuwen, J. M. J. (1998). On the Verification of a Large Petionnement. Retrieved November 5, 2012 from <http://www.lorentz.leidenuniv.nl/~jmjvanl/nrc.pdf>

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

\_\_\_\_\_\_\_\_\_\_\_(1952). On the Analysis of Samples from k Lists. *Annals of Mathematical Statistics* 23: 632 – 634.

\_\_\_\_\_\_\_\_\_\_\_(1960). On the Exact Variance of Products. *Journal of the American Statistical Association* 55: 708 – 713.

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

Hedderley, D. I. and Haslett, S. J. (2005). The Accuracy of Estimators of Number of Signatories to a Petition Based on a Sample. *Journal of Official Statistics* 21: 21 – 38.

Marley, P. (3/29/2012). GAB staff finds more than 900,000 valid signatures to recall Walker. *Milwaukee, Wisconsin, Journal Sentinel*. Retrieved October 29, 2012 from <http://www.jsonline.com/news/statepolitics/gab-staff-finds-more-than-900000-valid-signatures-to-recall-walker-ah4ptmf-144948425.html>

Montopoli, B. (6/5/2012). *CBS News*. Retrieved October 29, 2012, from <http://www.cbsnews.com/8301-503544_162-57447954-503544/wisconsin-recall-walker-opens-slight-lead-as-votes-are-counted/>

Mosteller, F., Ed. (1949). Questions and Answers. *The American Statistician*. 3:12 – 13.

Owens, B. W. (1969). An Acceptance Sampling Procedure for Validating Petitions. *AIIE Transactions* 1: 193 – 198.

Raj, D. (1961). On Matching Lists by Samples. *Journal of the American Statistical Association* 56: 151 – 155.

Shuster, J. J. (1974). The Validation of a Petition by a Sample Survey. *Communications in Statistics* 3:291 – 296.

Smith, R. A. and Thomas, D. R. (2005). Estimating the Number of Distinct Valid Signatures in Initiative Petitions. *Journal of Official Statistics* 21: 489 – 504.

Smith-Cayama, R. A. and Thomas, D. R. (1999). Estimating the Number of Distinct Valid Signatures in Initiative Petitions. *Proceedings of the American Statistical Association, Section on Survey Research Methods* 238 – 243.

Whiteside, M. M. and Eakin, M. (2008). A Better Estimate of the Number of Valid Signatures on a Petition. *The American Statistician* 62: 17 – 21.

\* Proportion of duplicate invalids is set to the same value as the proportion of duplicate valids

|  |
| --- |
| Table 1. Whiteside and Eakin (2008) Design Parameters  |
| Factors | Levels |
| Number of signatures on the petition  | 100,000 | 250,000 | 500,000 |
| Sampling fraction  | 0.03 | 0.10 | 0.25 |
| Proportion of duplicate valids\* | 0.02 | 0.035 | 0.05 |
| Proportion of valids to population size | 0.75 | 0.85 | 0.95 |

|  |
| --- |
| Table 2. Frequency Notation |
|  | Petition Signatures |  | Sample Signatures |
|  | Valid | Invalid | Totals |  | Valid | Invalid | Totals |
| Uniques and 'original' of duplicate pairs | MV | Mi | M |  | mv | mi | m=n-d |
| Redundant copies of duplicate pairs | Dv | Di | D |  | dv | di | d |
| Totals | NV | Ni | N |  | nv | ni | n |

Note: In the first row M = U + D where U is the number of unique signatures on the petition.

Similarly, m = u’ + d where u’ consists of true uniques from U and unpaired (in the sample) d’ from D

|  |
| --- |
| Table 3. Design Parameters  |
| Factors | Levels |
| Number of signatures on the petition (N)  | 50,000 | 100,000 | 250,000 |
| Sampling fraction (n / N) | 0.05 | 0.10 | 0.25 |
| Proportion of duplicate valids, PDv =(Dv / Nv) | 0.02 | 0.03 | 0.05 |
| Proportion of duplicate invalids, PDi = (Di / Ni) | 0.02 | 0.03 | 0.05 |
| Proportion of valids to population size, Pv = (Nv / N) | 0.75 | 0.85 | 0.95 |

|  |
| --- |
| Table 4. ((Approximated Var(W) – Var(W))/Var(W))\*100for Table 3 values |
|   |   | Number | Mean | Median | St. Dev. | Min | Max |
| Standard method |  |  |  |  |  |  |
|  | Eq. (1) holds |  81 | -10.32 | -7.95 | 7.61 | -31.56 | -3.14 |
|  | Eq. (1) does not hold | 162 | -10.13 | -8.01 | 8.44 | -34.56 |  3.83 |
| Uncorrelated method |  |  |  |  |  |  |
|  | Eq. (1) holds |  81 |  2.31 |  1.13 |  2.24 |  0.28 |  6.86 |
|   | Eq. (1) does not hold | 162 |  2.24 |  1.32 |  2.55 |  -1.09 |  8.41 |

|  |
| --- |
| Table 5. Logistic Estimate to Predict if MSE(W)>Var(G): 1 if Yes, 0 if No |
| Parameter | Estimate | St. Error | Z-Test | P-Value |
| Intercept |  17.85 |  4.88 |  3.66 | 0.0002 |
| |PDv-PDi| |  577.86 | 148.92 |  3.88 | 0.0001 |
| PDv | -434.04 | 108.67 | -3.99 | 0.0001 |
| Pv |  -20.46 |  5.59 | -3.66 | 0.0002 |
| sf\*N\*|PDv-PDi| |  0.19 |  0.04 |  4.32 | 0.0000 |

|  |
| --- |
| Table 6. Washington State Petition Values with Replications Beyond Duplicate Ignored |
|  | Frequency |  | Proportions |
| Pop Size | N | 227,598 |  |  |
| Invalid | Ni | 34,542 |  |  |
| Dup Valid | Dv | 11,034 | PDv | 0.060619 |
| Valid | Nv | 182,022 | Pv | 0.799752 |
| Unique Valid | Mv | 170,988 |  |  |

|  |
| --- |
| Table 7: Relative Efficiency of W and G for Washington State Example |
|  | Sampling Fraction |
| Di/Ni | 0.05 | 0.10 | 0.25 |
| 0.03063 | -0.02271 | 0.61811 | 1.89292 |
| 0.04062 | -0.17451 | 0.19833 | 1.10378 |
| 0.05061 | -0.25692 | -0.11443 | 0.28149 |
| 0.06062 | -0.25679 | -0.21041 | -0.12862 |
| 0.07061 | -0.17411 | -0.04436 | 0.32366 |
| 0.08063 | -0.02116 | 0.31169 | 1.15753 |
| 0.09061 | 0.18076 | 0.74540 | 1.94330 |
| Relative Efficiency = Log2(MSE(W)/Var(G) |



 

Figures 1a and 1b

 PDi = 0.02, PDi = 0.03, PDi = 0.05



Pv = 0.75



Pv = 0.85



Pv = 0.95

PDv = 0.02 PDv = 0.03 PDv = 0.05

Figure 2a. Log2 of MSE(W) / Var(G) Versus the Population Size for a Sampling Fraction of 0.05. Each line in a graph is for a different value of the proportion of invalid signatures that are duplicates. Each row of graphs corresponds to a value of the proportion of valid signatures in the population and each column of graphs corresponds to a proportion of valid signatures that are duplicates.

 PDi = 0.02, PDi = 0.03, PDi = 0.05



Pv = 0.75



Pv = 0.85



Pv = 0.95

 PDv = 0.02 PDv = 0.03 PDv = 0.05

Figure 2b. Log2 of MSE(W) / Var(G) Versus the Population Size for a Sampling Fraction of 0.10. Each line in a graph is for a different value of the proportion of invalid signatures that are duplicates. Each row of graphs corresponds to a value of the proportion of valid signatures in the population and each column of graphs corresponds to a proportion of valid signatures that are duplicates.

 PDi = 0.02, PDi = 0.03, PDi = 0.05



Pv = 0.75



Pv = 0.85



Pv = 0.95

 PDv = 0.02 PDv = 0.03 PDv = 0.05

Figure 2c. Log2 of MSE(W) / Var(G) Versus the Population Size for a Sampling Fraction of 0.25. Each line in a graph is for a different value of the proportion of invalid signatures that are duplicates. Each row of graphs corresponds to a value of the proportion of valid signatures in the population and each column of graphs corresponds to a proportion of valid signatures that are duplicates



|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

Figure 3.

 

Figures 4a and 4b