# A Better Estimate of the Number of Valid Signatures on a Petition

**Mary M. Whiteside, Ph.D., and Mark E. Eakin, Ph.D**

**The University of Texas at Arlington**

**Department of Information Systems and Operations Management**

**Box 19437**

**Arlington, TX 76019**

**817 272 3517  817 272 3529**

**whhiteside@uta.edu  eakin@uta.edu**

**Prepared for presentation at the 2004 Joint Statistical Meetings**

**Toronto, Canada**

# A Better Estimate of the Number of Valid Signatures on a Petition


## Abstract

The 2003 recall of California Governor Gray Davis, the repeated attempts in 2003 to recall Mayor Laura Miller in Dallas, Texas, and the 2002 Austin, Texas, City Council elections are among the many current examples that illustrate the use of voter petitions in city and state government in the United States. A better method for estimating the number of valid signatures on a petition could be used in many of these cases. Signatures may be invalid for several reasons: not a registered voter, incorrect address, replicate, etc. The statistical problem is interesting because replicated signatures must be estimated differently than signatures invalid for other reasons. Goodman's seminal work and subsequent Goodman-type estimators first estimate the total number of valid signatures on a petition, ignoring initially whether or not they are replicated. Then the problem reduces to estimating the number of classes in a finite population. We first estimate the number of unique, original signatures, and then the proportion that are valid. The result is a non-biased nonlinear estimator with smaller variance than the Goodman-type statistics for the case where the proportion of duplicate signatures is the same for valid and invalid signatures.


## Introduction

Many existing democracies allow for petitions of registered voters. The process of ballot access, including signature validation, is one carefully explored by emerging democracies across the globe, in particular Russia and other former Soviet states. (Ballot Access 1997). California required 897,158 valid signatures to put Democratic Governor Gray Davis's fate back in the hand of voters in 2003, less than a year after he was re-elected. To achieve this number of valid signatures, recall leaders targeted 1.2 million petition signatures, "a target that would

ensure that there are enough registered voters on the list to fend off a time-intensive examination of all the signatures.“ (Nissenbaum, p. 1) The 1.2 million signatures on the petition allowed for random sampling to validate. These petitioners were successful.

The drive to recall Dallas, Texas, Mayor Laura Miller was unsuccessful. Organizers fell short of the 72,873 signatures required to force a recall election in Dallas in November 2003. Although supporters had gathered an estimated 85,000 or so signatures, they had time to validate only 47,260 of these (Nelson and Levinthal). Why didn’t they sample?

Supporters of Jackie Goodman in Austin were successful in submitting 23,282 signatures. From a sample of 6,038 signatures, city officials declared that the minimum number of valid signatures required by law, 18,263, had been achieved on the petition. (Greenberg, p.1.)

The website for National Voter Outreach discusses the issue of petition validation within the United States. “Out of the 24 states that allow citizen sponsored initiatives, just four of them – Arkansas, Colorado, Oklahoma and South Dakota – presume signatures to be valid.” (Arnold, R. and Johnson, S. 2003). Even so, Colorado and Arkansas still validate signatures. In Colorado, signatures won’t be considered valid unless 147% of the legal requirement is submitted.

Eight states do random sample checks; checking between 2% and 10% of the petition population. California is one of these. The remaining states check each and every signature. Cities and other governmental entities reflect this variety of methodology. For example, Austin samples randomly; Dallas seemingly requires validation of every signature.

States also require signature validation of petitions to qualify minor party and independent candidates for the ballot who do not represent major political parties, such as Ralph Nader for U.S. President in 2004. “Nearly all approaches rely on some variety of sampling techniques for assessing signatures. California uses a random sample to include 100 signatures for petitions containing between 100 and 2,000 signatures and 5% in case of petitions of more than 2,000 signatures. South Carolina uses a sliding scale of actual checks on the first 500 signatures with one signature in 10 being checked after that. Texas provides for checks of either 25% or 1,000 signatures, whichever is more, and qualifies any candidate whose petition passes a test of statistical significance.” (Ballot Access)

## Background

A large number of signatures on a petition (whether for a citizen’s’ legislative initiative, recall of an official, or ballot access for a candidate) may contain a combination of honest mistakes, incomplete information, non-registered voters and replication. Replication means that one individual’s signature appears more than once on the same petition. Replication is fraudulent if individuals knowingly sign the same petition multiple times. Blote and van Leeuwen (1998) discuss the difficult problem of determining “how large a sample must be in order to check reliably whether such malpractices have taken place.” (1998, p.1)

The 2002 Austin, Texas, petition to allow another City Council term for Jackie Goodman had 23,828 signatures. A 25% random sample was selected for signature validation. The reasons a signature was invalid and frequency in the sample of 6,038 signatures were these:

| Category | Number | % of Sample | % of Invalid Signatures |
|---|---|---|---|
| Date of birth missing | 39 | 0.65% | 3.10% |
| Signed for other candidate | 1 | 0.02% | 0.08% |
| Signed for same issue | 52 | 0.86% | 4.13% |
| Invalid date | 5 | 0.08% | 0.40% |
| No address | 19 | 0.31% | 1.51% |
| Not on voter registration list | 1133 | 18.76% | 89.92% |
| Other | 7 | 0.12% | 0.56% |
| Printed name missing | 3 | 0.05% | 0.24% |
| Signature missing | 1 | 0.02% | 0.08% |
| Total | 1260 | 20.87% | 100.00% |

Statistically, the problem of estimating the total number of invalid signatures that are replicates is different and more interesting than the estimate of signatures invalid for other reasons, where the sample proportion of invalid signatures multiplied by the population size gives the optimal estimate. In what follows, invalid signatures will refer only to those invalid for reasons other than replication. Here replicated is categorized ‘Signed for the same issue.’ For example, the proportion of sample signatures invalid for any reason other than replication is (1208/6038)=20%. This leads to an estimated .2*23,828 = 4,767 invalid signatures on the petition. However, this procedure cannot be used to estimate the number of replicates in the petition.

For simplicity, assume that multiples are only duplicates. The 52 signatures that have a duplicate match in the 25% sample suggest another 52*3 signatures **in the sample** that have a

duplicate match in the remaining three-fourths of the population but not observed in the sample. In other words, the observation of 52 duplicate pairs in the sample, leads to the inference of 4*52 = 108 signatures **in the sample** that have a duplicate match in the population. (Sager 2002)

In the absence of invalid signatures, the problem of estimating the number of replicates reduces to estimation of the number of classes in a finite population, where each individual is considered a class and the petition signatures comprise the finite population. Goodman (1949) shows that the linear unbiased estimator defined below for the total number of classes in a finite population is unique under the assumption that the sample size is no smaller than the maximum number of elements in any class. In this application, the sample size must be greater than the number of signatures from any one person. Hass and Stokes (1998) propose an unbiased non-linear estimate of the number of classes in a finite population based on the generalized jackknife technique.

Smith-Cayama and Thomas (1999) compare several linear and one non-linear estimate for the number of replicates, Dhat below. In all cases the estimated number of valid signatures in the population, Mhat, from a sample of size n, is of the form

$$Mhat = N - Uhat - Dhat$$

where N = the number of signatures on the petition

Uhat = the estimate of the invalid signatures = (N/n)*u,

where u=the number of invalid signatures in the sample, and

Dhat = the estimated number of replicates (signatures are assumed to appear no more than 2 or 3 times).

By implication, Smith-Cayama and Thomas categorize signatures in exactly one way. Invalid signatures are counted first. Then, as previously mentioned, among valid signatures, the problem reduces to estimating the number of classes.

## Estimators of Valid Signatures

The question is how best to combine two estimators of invalid and replicated signatures to infer the number of valid signatures in a petition. As a result of the Austin petition, Professor Thomas Sager of The University of Texas at Austin (2002) suggests an estimator when duplicate signatures are first counted in the sample and categorized as duplicates, whether valid or not (although this is not stated explicitly). Because of the great difficulty of estimating duplicates (Blote and van Leeuwen 1998) superficially it seems reasonable to count as a duplicate any signature that is both invalid for other reasons and duplicated. With this distinction the Sager estimate is similar to the general Smith-Cayama type. To clarify the distinction, however, consider the following cross tabulations:

| | **Petition Signatures** | | | | **Sample Signatures** | | |
|---|---|---|---|---|---|---|---|
| | **Valid** | **Invalid** | **Totals** | | **Valid** | **Invalid** | **Totals** |
| **Uniques and 'original' of duplicate pairs** | Mv | Mu | M | | $m_v$ | $m_u$ | n-d |
| **Redundant copies of duplicate pairs** | Dv | Du | D | | $d_v$ | $d_u$ | d |
| **Totals** | | | N | | | | n |

What I call the Goodman-type estimator, in the class of the Smith-Cayama estimators, is defined as G, where

$$G = N - (N/n)*(d_u+m_u) - [N*(N-1)/n*(n-1)]*d_v.$$

Note that duplicate, invalid, signatures are counted as invalid.

Whereas the Sager estimate, where duplicates are first categorized, is defined as S where

$$S = N - (N/n)*m_u - (N/n)^2*(d_v+d_u).$$

Note that duplicate, invalid, signatures are counted as duplicate. Either of these estimates can be applied whether invalid signatures are counted first or duplicate signatures are counted first.

The new estimators introduced in this paper, like Sager's, first estimate the number of duplicates in the population. Replicates beyond duplication are not considered The first Whiteside estimator, W, uses Sager's estimate of duplicates:

$$Dhat_W = (N/n)^2*(d_v+d_u).$$

The population size N is then reduced by this amount to $(N-(N/n)^2*(d_v+d_u))$. Similarly, the sample size is reduced to (n-d). Finally, the proportion of valid signatures on the petition, M, is estimated.

$$W = [(N - Dhat_W)/(n-d)]* m_v.$$

A second form of this estimator, adjW, is adjusted using Goodman's unbiased estimate of population duplicates, $UDhat_W$.

$$UDhat_W = [N*(N-1)/n*(n-1)]*( d_v+d_u).$$

$$adjW = [(N - UDhat_W)/(n-d)]* m_v.$$

## Simulation and Results

A sample of size 200 is randomly selected from a population of 1,000 signatures, 750 of which are valid. The sampling is repeated 5,000 times for each of the following levels of design Factor A.

Factor A:

```
DU = 17, DV = 125, DV/MV = 0.1667, DU/MU = 0.1574
DU = 17, DV = 100, DV/MV = 0.1333, DU/MU = 0.1278
DU = 15, DV =  75, DV/MV = 0.1000, DU/MU = 0.0938
DU = 12, DV =  50, DV/MV = 0.0667, DU/MU = 0.0638
DU =  7, DV =  25, DV/MV = 0.0333, DU/MU = 0.0321
DU =  0, DV =   0, DV/MV = 0.0000, DU/MU = 0.0000
```

Thus, a total of 30,000 samples are selected. For each random sample (nested Factor B), the estimators, G, S, W, and adjW (Factor C) are computed along with their absolute difference from 750, i.e. absolute error. We perform a three factor nested analysis of variance F test (A, B(A), C, A*C) followed by Tukey's test for all pairwise comparisons to examine differences in the mean absolute error of these estimates.

Recall that Goodman's estimator is unbiased. Differences in the means between the Goodman estimate and each Whiteside estimate are not statistically significant for any level of Factor A with an ordinary two factor analysis of variance for Factors A and C for this simulation. When we replicated the simulation design with 20,000 repeats (rather than the 5,000 reported here) and used the three factor analysis with sample Factor B(A) nested in Factor A, the differences in mean estimates between Goodman and Whiteside are statistically significant but less than 1 signature in size (750 vs. 750.6). Also, note that it is true for all factor level combinations that Du*Mv and Dv*Mu are approximately equal. The results for average absolute error appear in Tables 1 and 2.

**Table 1. Average Absolute Error**

```
1. DU = 17, DV = 125, DV/MV = 0.1667, DU/MU = 0.1574

               46.3790   S
               43.3187   G
                41.2844   W
               41.1644   Wa

2. DU = 17, DV = 100, DV/MV = 0.1333, DU/MU = 0.1278

               43.2960   S
               39.5208   G
               37.6863   W
               37.5751   Wa

3. DU = 15, DV =  75, DV/MV = 0.1000, DU/MU = 0.0938

               40.1170   S
               36.1795   G
               34.4305   W
               34.3376   Wa


4. DU = 12, DV = 50, DV/MV = 0.0667, DU/MU = 0.0638

               35.2830   S
               32.2974   G
               30.6298   W
             |  30.5665   Wa

5. DU =  7, DV =  25, DV/MV = 0.0333, DU/MU = 0.0321

               30.0200   S
               27.9160   G
               26.9971   W
               26.9568   Wa

6. DU =  0, DV =   0, DV/MV = 0.0000, DU/MU = 0.0000

               22.0850   G
               22.0850   Wa
               22.0850   W
               22.0850   S
```

**Table 2. The ANOVA Procedure**

Dependent Variable: Absolute Error

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 5022 | 12067743.46 | 2402.98 | 3.54 | <.0001 |
| Error | 114977 | 77986990.03 | 678.28 | | |
| Corrected Total | 119999 | 90054733.49 | | | |

| R-Square | Coeff Var | Root MSE | bias Mean |
|---|---|---|---|
| 0.134005 | 77.71431 | 26.04387 | 33.51233 |

| Source | DF | Anova SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| didv | 5 | 5947878.892 | 1189575.778 | 1753.80 | <.0001 |
| row(didv) | 4999 | 5709955.500 | 1142.220 | 1.68 | <.0001 |
| type | 3 | 327672.613 | 109224.204 | 161.03 | <.0001 |
| type*didv | 15 | 82236.459 | 5482.431 | 8.08 | <.0001 |

The ANOVA Procedure

Tukey's Studentized Range (HSD) Test for bias

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

| | |
|---|---|
| Alpha | 0.05 |
| Error Degrees of Freedom | 114977 |
| Error Mean Square | 678.2834 |
| Critical Value of Studentized Range | 3.63321 |
| Minimum Significant Difference | 0.5463 |

Means with the same letter are not significantly different.

| Tukey Grouping | Mean | N | type |
|---|---|---|---|
| A | 36.1967 | 30000 | S |
| B | 33.5529 | 30000 | G |
| C | 32.1855 | 30000 | W |
| C | | | |
| C | 32.1142 | 30000 | Wa |

## Conclusion

In all cases where duplicate signatures appear on the petition, the average absolute error is smaller for the non-linear Whiteside estimates than for the linear unbiased Goodman estimate or the biased Sager estimate. As the proportion of duplicates increases from about 3.2% to 16%, this difference becomes larger. Moreover, the differences in means for the unbiased Goodman estimate and Whiteside estimates are less than one signature when the simulation is repeated 20,000 times. Thus, when the assumption of equal proportions of duplicates among valid and invalid signatures appears to be reasonable, the Whiteside estimators are preferable with smaller absolute error and negligible bias. For the Austin City Council petition, the values of the estimators (assuming $d_u$=0, making it irrelevant whether duplicates or invalid signatures are counted first) were 18,251 valid signatures with both the Goodman and Sager estimates and 18,373 valid signatures with both Whiteside estimates. The

requirement was for 18,263 valid signatures. Clearly, the selection of the estimator for signature validation of petitions can have unexpectedly far reaching consequences.

## References

Arnold, R. and Johnson, S. (2003), National Voter Outreach, www.directdemocracy.com, Validating signatures.

Ballot Access for American Political Parties More Generally (1997), *Questions of Ballot Access in the Russian Federation Balancing Democratic Values, Common – Sense Approaches to Russian Political Conditions and Respect for Legality*, www.democracy.ru/english/

Blote, H.W.J. and van Leeuwen, J. M. J. (1998), On the Verification of a Large Petionnement, www.lorentz.leidenuniv.nl/~jmjvanl/

Goodman, L. A. (1949), On the Estimation of the Number of Classes in a Population, Annals of Mathematical Statistics, 20, 572-579.

Greenberg, B. (2002), Estimating the Number of Valid Unduplicated Signatures in a Petition. Unpublished consulting report for the city of Austin, TX

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

Nelson, C. M. and Levinthal, D. (2003), Mayor‘s Opponents Vow to Continue with New Signature Drive, The Dallas Morning News, November 25, 2003, page B1.

Nissenbaum, D. (2003), Davis Recall Effort Remains Short of Goal, *Mercury News*, Sacramento Bureau, June 25, 2003, p.1.

Sager, T. (2002), Affidavit and Exhibits, April 5, 2002, Austin, Texas.

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