Title

Resolving Contested Elections: The Limited Power of Post-Vote Vote-Choice Data

Permalink
https://escholarship.org/uc/item/4bm6t84p

Journal
Journal of the American Statistical Association, 105(489)

ISSN
0162-1459

Authors
Glynn, Adam N
Richardson, Thomas S
Handcock, Mark S

Publication Date
2010-03-01

DOI
10.1198/jasa.2009.ap08640

Peer reviewed

Resolving Contested Elections: The Limited Power of Post-Vote Vote-Choice Data[*]

Adam N. Glynn[†] Thomas S. Richardson[‡]
Mark S. Handcock[‡]

March 9, 2009

Abstract

In close elections, the losing side has an incentive to obtain evidence that the election result is incorrect. Sometimes this evidence comes in the form of court testimony from a sample of invalid voters, and this testimony is used to adjust vote totals (Borders v King County, 2005; Belcher v Mayor of Ann Arbor, 1978). However, while courts may be reluctant to make explicit findings about out-of-sample data (e.g. invalid voters that do not testify), when samples are used to adjust vote totals, the court is implicitly making findings about this out-of-sample data. In this paper, we show that the practice of

[*] The authors thank Jon Wakefield, Jeff Gill, D. James Greiner, Sven G. Peterson, Anton Westveld, and the participants of the 2007 meetings of the Midwest Political Science Association for their helpful comments and suggestions. The usual caveat applies.

[†] Department of Government and The Institute for Quantitative Social Sciences Harvard University, 1737 Cambridge Street, Cambridge, MA 02138. aglynn@iq.harvard.edu

[‡] Department of Statistics, University of Washington, Seattle, WA 98195.

adjusting vote totals on the basis of potentially unrepresentative samples can lead to incorrectly voided election results. More generally, we show that given the difficulties of sampling and non-response in this context, even when frame error is minimal and if voter testimony is accurate, such testimony has limited power to detect incorrect election results without precinct level polarization or the acceptance of large Type I error rates. Therefore in U.S. election disputes, even high quality post-vote vote-choice data will often not be sufficient to resolve contested elections without modeling assumptions (whether or not these assumptions are acknowledged).

1 Introduction

After the first machine count in the 2004 Washington State Gubernatorial election, Dino Rossi led Christine Gregoire by 261 votes. Because this margin represented less than one half of one percent of the total votes cast, the Secretary of State ordered a mandatory machine recount, after which Rossi's lead had shrunk to 42. The Washington State Democratic Central Committee requested a hand recount, and as a result of this final count, Gregoire was declared the winner over Rossi by a margin of 129 votes. A number of parties petitioned the Washington State Superior Court in Chelan County, attempting to void this election result on the basis of hundreds of discovered invalid ballots from precincts that voted heavily for Gregoire. The defendants countered by producing hundreds of invalid ballots from precincts that voted heavily for Rossi. In Borders v King County, 2005, the court found that the petitioners had not presented sufficient evidence to overturn the 2004 gubernatorial election. The basis for this finding was grounded in two facts. First, the plaintiff's sample of invalid votes was unscientific and comprised neither a census, nor a representative sample of all invalid votes. Second, even if the sample were scientific, the counts for each candidate and for the number of invalid votes were aggregated to the precinct level, and therefore, any attempt to estimate the number of invalid votes for a particular candidate would be an example of ecological inference. Furthermore, due to the small percentage of invalid votes in each precinct, it is possible that all of the invalid votes were cast for Rossi or that all of the invalid votes were cast for Gregoire (Adolph, 2005). The result in this case therefore hinged on a debate over the proper way to model the votes of invalid voters (Gill, 2005; Katz, 2005; Handcock, 2005). However, in addressing the issues raised in Borders v King County, 2005, the court defined procedures by which sufficient evidence might be obtained to overturn a close election result *without modeling assumptions*. In particular, the court ruled that persons who had cast invalid votes could testify and that their testimony could be used to adjust vote counts. During the trial, the defendants produced four invalid voters who testified that they had voted for Rossi and the court ruled that Rossi's vote total should be reduced by four. This ruling raises a number of pressing questions. First, what would have happened if the court had heard testimony from 130 more invalid Gregoire voters than invalid Rossi voters?

Second, does the state really want to allow samples of this type? There is no evidence that the sample of invalid voters in this case was representative, and one can imagine future scenarios in which the defendants and plaintiffs race to find invalid voters who are willing to testify that they voted for the opposing candidate. Alternatively, one could consider obtaining a complete list of all invalid voters, calling them each to testify on their ballots and adjusting the vote totals accordingly. However, as in many sampling problems, obtaining a census may prove impossible due to time and money constraints. In Borders v King County, 2005, the plaintiffs and the defendants found a total of $1,439$ invalid voters, and this combined sample was not a census of all invalid voters in the state, just a lower bound that was "cherry picked" from the precincts with high Gregoire margins (found by the plaintiffs) and high Rossi margins (found by the defendants). In principle, the parties could have attempted to find all the invalid voters. However, this is of course only feasible for the classes of invalid voters that we know: felons, the deceased, non-citizens, etc. Furthermore, even if the plaintiffs and defendants had collected a census at the first stage, the second stage (calling over one thousand witnesses) would be extremely costly and time consuming. Even if the state (or possibly the parties) were willing to foot the bill, a governor is elected for a four year term, and it is not possible to know ahead of time how long it would take to call all the witnesses (especially when the defendants in the case have an incentive to delay). Finally, even if the parties could find all the invalid voters (as we will assume throughout the rest of this paper) and call them all as witnesses, the court still faces problems. Some invalid voters might be deceased, or forget whom they supported, or refuse to testify (nonresponse) and some invalid voters might lie or misremember (measurement error). The likely magnitude of measurement error is difficult to assess. An invalid voter may misremember and also has an incentive to lie, because telling the truth would hurt their preferred candidate. However, an invalid voter who testifies is sure to be cross examined harshly by the side that is losing a vote, and the risk of a perjury charge is a large cost in comparison to the relatively small chance that your lie will get your preferred candidate elected. In light of this, we will consider situations where the measurement error can be assumed to be negligible. Nonresponse is likely to be a larger problem, even in the courtroom setting. The assumed sampling frame is the list of invalid voters, and some of the classes of invalid voters (e.g. the deceased) will be unable to testify. Additionally, those invalid voters that are able to testify (e.g. released felons) may be unavailable, forget their vote choices, or refuse to testify. (In Belcher v Mayor of Ann Arbor, 1978, some invalid voters refused to testify and were held in contempt of court, but the Michigan Supreme Court eventually upheld their right to refuse to testify.) Nonresponse in samples has been studied extensively. However, most of the techniques that have been developed for nonresponse involve creating a model for the nonresponders based on the responders.

In this paper, we explore the limits of design-based inference without modeling assumptions, and hence avoid the debates over modeling the probability of election reversal (Finkelstein and Robbins, 1973; Downs et al., 1978; Gilliland and Meier, 1986; Robbins, 1986; Harris, 1988) and the traditional weighting or imputation based approaches to nonresponse (Groves et al., 2002; Little and Rubin, 1986). Instead we utilize the discrete and finite nature of the problem to form bounds without assuming any similarity between the responders and the nonresponders. The outline of the paper is as follows. In Section 2 we describe the court's decision problem for both simple random samples and stratified random samples. Taking into account observed characteristics of voters, we apply an exact likelihood ratio testing procedure to address the decision problem, providing power analyses for both types of samples, with and without nonresponse. Section 3 discusses the implications of these results for legal challenges to election results.

2 The Decision Problem for Design Based Inference

We illustrate the general decision problem using the 2004 gubernatorial case. The final judgement in Borders v King County, 2005 summarizes the relevant statute:

> *RCW 29A.68.110 provides that no election may be set aside on account of illegal votes unless it appears that an amount of illegal*

> *votes has been given to the person whose right is being contested that, if taken from that person, would reduce the number of the person's legal votes below the number of votes given to some other person for the same office after deducting therefrom the illegal votes that may be shown to have been given to the other person.*
>
> (Borders v King County, 2005, Final Judgement, p. 21)

Following the law as specified in this judgement, we formulate the corresponding decision problem in the Neyman-Pearson framework with the composite null hypothesis as a tie or a victory for the putative winner in an election with only the valid ballots considered. It may seem more natural to start with the explicit definition of a loss function and decision rule based on the precepts of frequentist decision theory. However, a Neyman-Pearson decision rule is admissible under 0-1 loss, and while not minimax (Chernoff and Moses, 1986), it has a lower risk of incorrectly voiding an election result than does a minimax rule. Hence the Neyman-Pearson rule implicitly protects against the costly legal remedy of conducting a new election, this being the only realistic remedy (Harvard Law Review , 1975). One could argue that such protection should be built into the decision procedure explicitly through an asymmetric loss function, but it would be difficult for the court to decide on a particular loss function, and the Neyman-Pearson rule can be easily related to p-values which more closely reflect the evidentiary nature of court proceedings.

In Borders v King County, 2005, the plaintiffs and defendants together identified 1,439 invalid voters. If we make the working (albeit incorrect) assumptions that this number represents a census of all the invalid voters and that there were two candidates (that we call Gregoire and Rossi without loss of generality), then we can get some insight into the decision problem. (We also assume each voter cast exactly one ballot, it was for one of these candidates, and we exclude spoiled ballots.) Since the margin of victory among all ballots (valid and invalid) for Gregoire was 129 voters, the null hypothesis of a tie or Gregoire victory among valid ballots corresponds to 784 (or greater) of the invalid voters for Gregoire and 655 (or fewer) of the invalid voters for Rossi (this of course involves the simplification to two candidates). Note that 784 is greater than half of the 1,439 invalid voters, hence, if the invalid voters voted for Gregoire and Rossi in the same proportions as valid voters, then the null hypothesis is true (i.e. Gregoire was the rightful winner). Furthermore, this demonstrates that the one sided alternative in the direction of a Rossi win indicates that Gregoire had a proportion of the support among the invalid voters larger than 54%. The court would therefore rule to void the election result only when there was sufficient evidence to reject the null hypothesis of Gregoire support less than or equal to 54%. (Note that this test of the composite null hypothesis is equivalent to the test of the simple null hypothesis of a tie because if we can reject the hypothesis of a tie in the direction of a Rossi victory, we can also reject any hypothesis of a Gregoire victory.)

2.1 The Decision Problem for Simple Random Sampling from a Sampling Frame of Invalid Voters

With a simple random sample, it is relatively straightforward to form decision rules and conduct exact power calculations based on the hypergeometric distribution. If we denote the total number of invalid voters as N and we denote the total number of invalid Gregoire voters as G, we can write the null and alternative hypotheses as the following:

$$\begin{aligned} H_0 : \frac{G}{N} &\leq \frac{784}{1439} \approx .545 \\ H_a : \frac{G}{N} &> \frac{784}{1439} \approx .545 \end{aligned}$$

Given a simple random sample of size n from this finite population, we define g to be the sampled number of Gregoire invalid voters and q_α to be the $1 - \alpha$ quantile from a hypergeometric distribution with $N = 1439$ and $G = 784$. The decision rule that rejects the null for $g > q_\alpha$ is a likelihood ratio test and is uniformly most powerful against all alternatives (see Lehmann and Romano (2005)). Figure 1 shows the exact power calculations for an $\alpha \leq .05$ level test of this type with sample sizes between 130 and 1400, and alternative values of G from 784 to 850. We have chosen $\alpha \leq 5\%$ in this paper, not because we endorse the use of this percentage, but because we must choose some level of α in order to conduct power analyses, and this is a standard level. The $(- - -)$ curve in Figure 1 corresponds to a sample

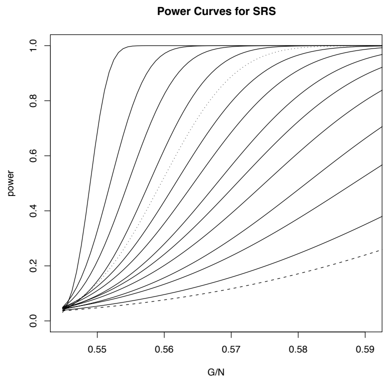


Figure 1: Exact power calculations for a simple random sample with a null hypothesis of an election tie ($N = 1439$ and $G = 784$), $\alpha \leq .05$, and the one sided alternative of $G > 784$ corresponding to a Rossi win. The $(---)$ curve corresponds to a sample size of 130. The next highest curve corresponds to sample size of 200, and each subsequent curve is an additional 100. The $(\cdots)$ curve corresponds to a sample size of 1000.

size of 130, with the next highest curve corresponding to sample size of 200, and each subsequent curve represents an additional 100 observations. The $(\cdots)$ curve corresponds to a sample size of 1000. The sample size of 130 is important, because this is the minimum sample size necessary to change the election result if the sample consisted of only illegal voters who voted for Gregoire and a decision rule was employed that merely deducts the sampled invalid voters from the vote totals (as was done in Borders v King County, 2005). To state this explicitly in the language of hypothesis testing, note that a selective sample can be viewed as a census in which all those who are not in the sample did not respond (see the solid curves of Figure 3). Therefore, a test that rejects the null hypothesis based on an attempt to selectively sample 130 invalid Gregoire voters would have 100% power for all alternative values of G, but unfortunately would also be guaranteed to incorrectly reject the null when $130 \leq G \leq 784$ and would therefore have $\alpha = 1$ (i.e. would be certain to reject the null even if it were true).

We see from the plot that the most powerful $\alpha \leq .05$ level test based on a simple random sample of 130 invalid voters would have very little power against mild to moderate alternatives. In fact, the power doesn't break 50% until $G = 890$ or when the true Gregoire proportion of invalids reaches 62%. Hence, unless we believe that invalid voters heavily favor Gregoire, an exact level $\alpha = 5\%$ likelihood ratio test based on a random sample of this size would be unlikely to generate enough evidence of a Rossi victory. Additionally, Figure 1 shows that to achieve high levels of power for reasonable proportions of Gregoire invalid voters, we need extremely large sample sizes. Even with a sample size of 1000, we only break 50% on power when the Gregoire proportion of invalid voters breaks 56%. Therefore, anyone contemplating the implementation of this sampling scheme and decision rule, should acknowledge the possible necessity of sampling nearly all invalid voters.

The power gets worse if we allow for nonresponse. In this scenario, the current null hypothesis does not provide enough restrictions to allow the exact derivation of the probability of Type I error because under everything but the most extreme hypotheses for G the non-responders could have all been Gregoire voters or all Rossi voters. To state this explicitly, let n_{nr} be the number of nonresponders, and let g_r be the number of observed Gregoire invalid voters among the responders. Then both of these variables are observable, while the original g (the total number of Gregoire invalid voters in the sample) is now unobservable. We know that $g_r \leq g \leq g_r + n_{nr}$, but without further assumptions or a combination of additional data and assumptions, we have *no* means of estimating a distribution over this range. Therefore, in the absence of such assumptions, the *only* way to ensure an $\alpha \leq .05$ level test is to treat all invalid non-responders as if they were invalid Rossi responders when calculating the rejection region, and to reject only when the number of observed Gregoire responders (g_r) exceeds the critical value. Since $max(g - n_{nr}, 0) \leq g_r \leq min(g, n - n_{nr})$, we can determine upper and lower bounds for the probability that g_r will exceed the threshold set by q_α. In our example with a target α of 5%, this effectively means that for large sample sizes and any substantial proportion of nonresponse, the true probability of Type I error could be anywhere between .05 and zero. The actual α bounds are plotted for an example with 1% nonresponse in Figure 2.

The power of this test is also bounded by using the distributions of $\max(g - n_{nr}, 0)$ and $\min(g, n - n_{nr})$. In Figure 3, we present exact power bounds for sample sizes of 130 $(- - -)$, 1000 $(\cdots)$ and the whole population (—) against different alternative values and for different levels of nonresponse. These plots make explicit the following two important points. First, an α level test in the presence of nonresponse is necessarily quite conservative and may have very low power because we must assume that the nonresponders voted for Rossi if we want to maintain the α level of the test.

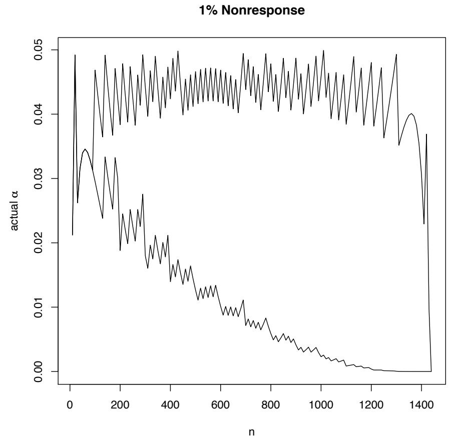


Figure 2: Exact bounds for the actual probability of Type I error in the presence of 1% nonresponse and various sample sizes. The decision rule (reject only if the responding Gregoire invalids exceed the critical value) was chosen to ensure that the probability of Type I error did not exceed 5%.

Second, even if we sample *all the invalid voters*, the presence of nonresponse will create the possibility that we will not be able to reject the null for mild to moderate values of the alternative.

2.2 The Decision Problem for Stratified Random Sampling from a Sampling Frame of Invalid Voters

Stratified random sampling often increases the power of a test, and for this application, we might hope that stratification would increase power for two reasons:

First, the plaintiffs and defendants presented ballot counts for both valid and invalid voters at the precinct level, and therefore, if there were extreme

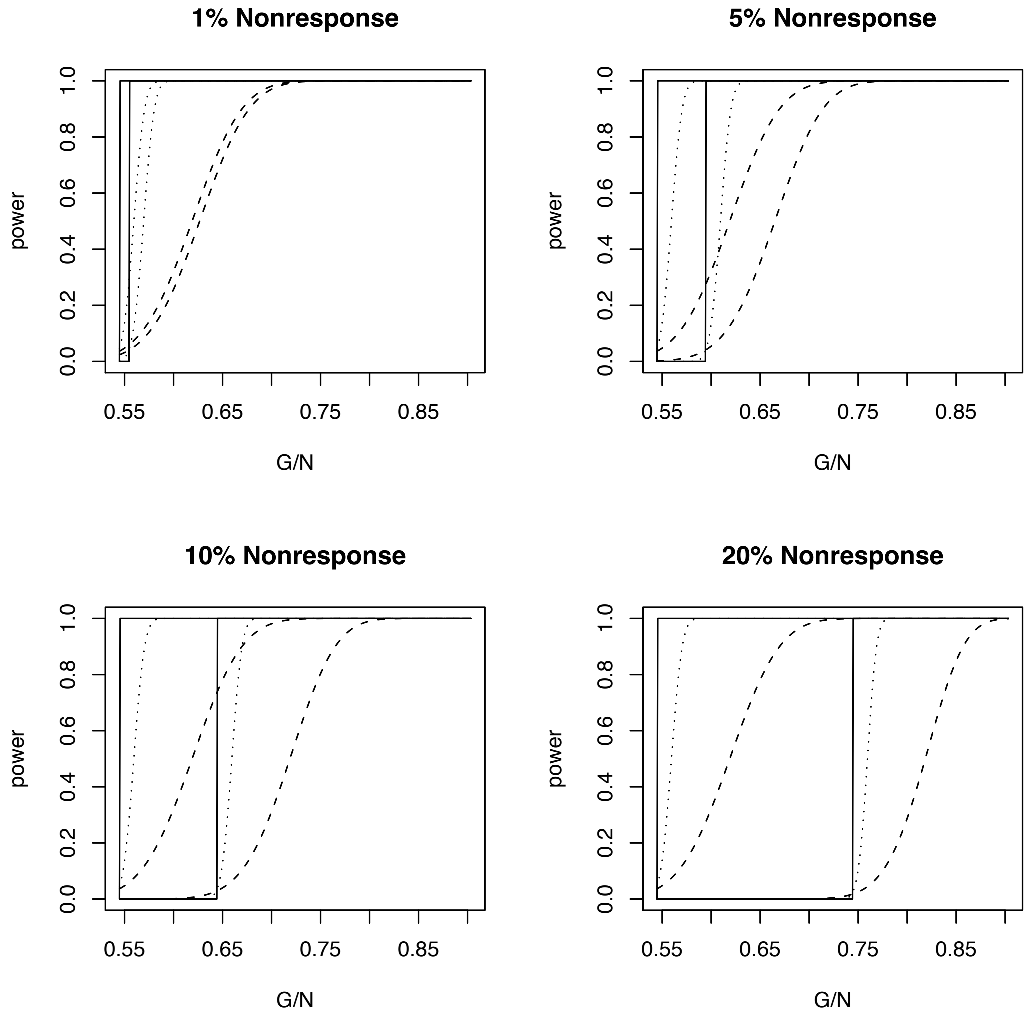
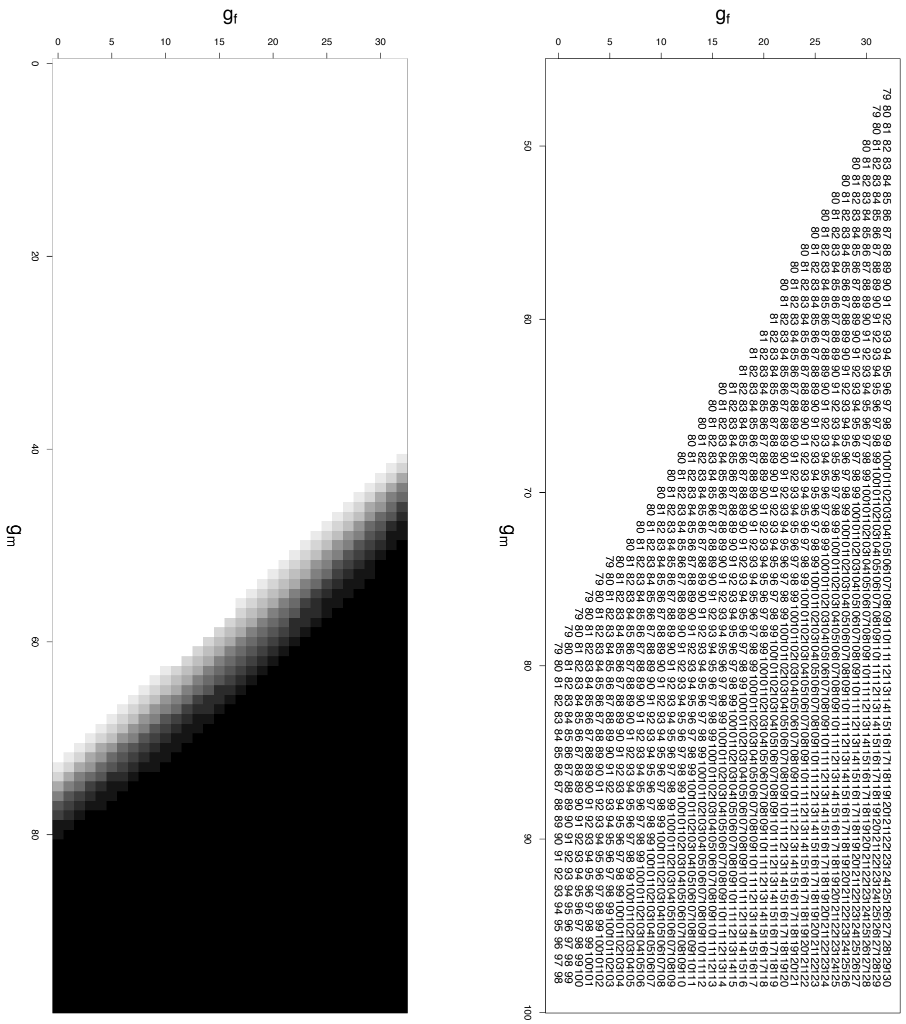


Figure 3: Exact bounds for power in the presence of 1%, 5%, 10%, and 20% nonresponse and with n=130 (− − −), n=1000 (· · ·), and n=1439, or the whole population (—). The decision rule (reject only if the responding Gregoire invalids exceed the critical value) was chosen to ensure that the probability of Type I error did not exceed 5%.

polarization at the precinct level, there would be constraints on the number of possible invalid Gregoire voters. For example, in a given precinct, if there were more invalid voters than the total number of valid and invalid Rossi voters then the remaining invalid voters are Gregoire voters. However, due to the lack of precinct level polarization in U.S. elections, these extra constraints do not come into play because the number of invalid ballots will always be much smaller than the total number of Gregoire ballots or Rossi ballots.

A second reason to think that stratification may increase the power of the test is that it may reduce the variance of the sampled number of Gregoire invalid voters within each stratum. Because we are forming a decision rule based on counts of invalid voters, and the majority of these invalid voters are felons (who are disproportionately male), an obvious stratification is on sex. This is possible (for the most part) because we have sex information on most of the invalid voters. Among the 1,439 invalid voters in Borders v King County, 2005, there were 1,082 invalid male voters, 356 invalid female voters, and one invalid voter of unknown sex. To simplify the presentation, we will assume that the invalid voter of unknown sex was female.

However, if we partition the population into male and female invalid voters, then the parameter space has two dimensions, and the decision problem becomes more complicated. The composite null hypothesis (election tie or Gregoire victory) can still be rejected by considering only the rejection of the election tie hypothesis. However, with more than one stratum, the election tie hypothesis does not completely specify the probability distribution. If we stratify on sex (male and female), then the parameter space has support described by $0 \leq G_m \leq N_m = 1,082$ and $0 \leq G_f \leq N_f = 357$, and the election tie hypothesis states that $G_m + G_f \leq 784$. Clearly a number of (G_m, G_f) pairs will satisfy even the equality portion of this statement, and in order to reject the null, we need simultaneous evidence against all of them. We can formulate an exact test against all (G_m, G_f) pairs in the null region by using the likelihood ratio. Figure 4 (a) presents likelihood ratio values for all points in the sample space where we have sampled 130 individuals and the within strata sample size ratio is approximately equal to the ratio of the dimensions of the parameter space (i.e. $\frac{n_m}{n_f} \approx \frac{N_m}{N_f}$). (It can be shown that proportional sampling tends to minimize the variability in the total number of sampled Gregoire invalid voters needed for rejecting the null (across different combinations of sampled male and female invalid Gregoire voters), and therefore proportional sampling would likely be more palatable to the Court and to the general public than a non-proportional sampling scheme.) The dark regions correspond to the low values of the likelihood ratio, and the white regions correspond to likelihood ratios of one. In order to form a rejection region, we include points of the sample space with small likelihood ratio values and check the probability of this region for every (G_m, G_f) pair such that $G_m + G_f <= 784$. Points are added to the rejection region in the order of the likelihood ratio values until the maximum probability of rejection over all null (G_m, G_f) pairs is as close to .05 as possible without going over.



(a) Likelihood ratios for a sample of 130 individuals proportionally stratified on sex. Dark regions indicate low values of the likelihood ratio. White regions represent likelihood ratio values of one.

(b) Rejection region for a sample of 130 individuals proportionally stratified on sex. The numbers in the plot are the sum of the g_m and g_f indices for the points in the rejection region and therefore represent the total number of invalid Gregoire voters (out of 130) necessary to reject for different numbers of men and women.

Figure 4: Likelihood ratios and rejection region for a sample of 130 individuals proportionally stratified on sex, $\alpha \leq .05$, and a null of $G = G_m + G_f \leq 784$.

Using proportional sampling, we would like to compare the power of the stratified random sample test to the power of the simple random sample test. However, this comparison is made difficult by the extra dimension in the parameter space for the stratified random sample. Each alternative value of G for the simple random sample will correspond to many possible (G_m, G_f) pairs for the stratified random sample. In Figure 5 (a) we show a comparison of the power curve for the likelihood ratio test based on a simple random sample of 130 invalid voters (solid curve) and the minimum and maximum power curves (dashed curves) for the likelihood ratio test based on a stratified proportional random sample with 98 men and 32 women. It is clear from the plot that the stratified LR test is at least as powerful as simple random sample LR test against alternative values of G because the minimum and maximum stratified power curves are greater than or equal the power curve for the simple random sample. Furthermore, the difference between upper and lower bounds on power for fixed G/N in Figure 5 (a) shows that the power of the stratified test can provide substantially greater power for some alternative (G_m, G_f) pairs than others. The benefits of stratification appear to be maximized when $G/N \approx 0.63$ and when $G_m/N_m >> G_f/N_f$. Hence, for alternative values of G where the power of the stratified test can provide substantially greater power over the SRS test, the benefit is maximized when the true Gregoire percentages are very different within each stratum, and the power of the stratified test is nearly identical to the power of the simple random sample test when $G_m/N_m \approx G_f/N_f$.

As shown in Figure 5 (a) the possible benefits from stratification depend on the heterogeneity between the strata, but they may also depend on the level of nonresponse. In the previous section, we presented α level power bounds for the LR test based on a simple random sample with nonresponse. We can calculate similar power bounds for our test based on a stratified random sample by using inequalities analogous to those of the last section.

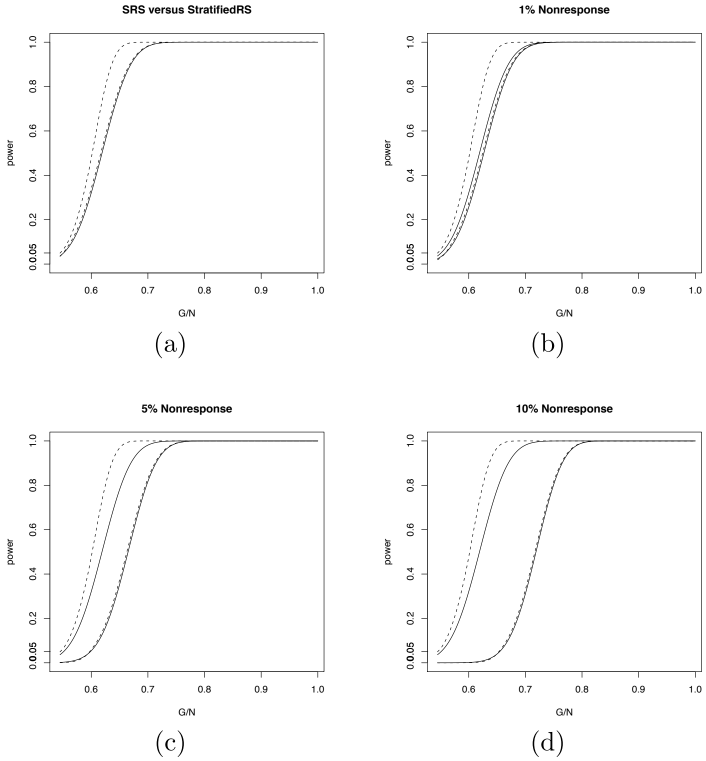
The rejection region in Figure 4 (b) can be represented by the following decision rule:

$$\begin{array}{llcll} \text{Reject if:} & (g_m > q_{1m} & \text{and} & g_f > q_{1f}) & \text{or} \\ & (g_m > q_{2m} & \text{and} & g_f > q_{2f}) & \text{or} \\ & \quad\vdots & & \quad\vdots & \vdots \\ & & & & \text{or} \\ & (g_m > q_{km} & \text{and} & g_f > q_{kf}). & \end{array}$$

where $(q_{1m}, q_{1f}), \ldots, (q_{km}, q_{kf})$ represent points just outside the boundary of the rejection region. In the presence of nonresponse among the male and female invalid voters, we know only that $g_{mr} \leq g_m \leq g_{mr} + n_{mnr}$ and $g_{fr} \leq g_f \leq g_{fr} + n_{fnr}$, where g_{mr} and g_{fr} are the male and female Gregoire responders and n_{mnr} and n_{fnr} are the number of male and female nonresponders. Since g_m and g_f can be anywhere within these bounds, we must use a decision rule which replaces g_m and g_f with g_{mr} and g_{fr} in order to maintain the α level of the test. From the above bounds we know that $\max(g_m - n_{mnr}, 0) \leq g_{mr} \leq \min(g_m, n_m - n_{mnr})$ and $\max(g_f - n_{fnr}, 0) \leq$

$g_{fr} \leq \min(g_f, n_f - n_{fnr})$, and therefore we can obtain bounds for the probability of rejection by considering the joint distribution of the male and female lower and upper bounds.

The joint distributions for the lower and upper bounds allow us to compare the probability bounds for a simple random sample and a stratified random sample under various levels of nonresponse. In Figure 5 we present power bounds for the simple random sample and a sample proportionally stratified on sex with overall sample sizes of 130 and a null of $G = G_m + G_f \leq 784$. The solid curves represent the power bounds from a simple random sample with differing levels of nonresponse (these bounds collapse to a single curve when there is no nonresponse). The dashed curves represent the power bounds from a proportionally stratified random sample with differing levels of nonresponse (the nonresponse is assumed to be proportional between males and females). Figure 5 shows that under a decision rule that bounds α at 5%, substantial nonresponse can seriously reduce the power of the stratified test to the point that there is no guarantee that the stratified test will produce any benefit. For example, with 10% nonresponse, and when $G/N = .63$ and $G_m/N_m >> G_f/N_f$, the power benefit from stratification over SRS can be as great as 0.24 but it can also be as small as 0.01. Hence in the case were stratification promises to provide the greatest benefit, substantial nonresponse *may* effectively eliminate these gains.



(a)

(b)

(c)

(d)

Figure 5: Exact power calculations under (a) 0% (b) 1%, (c) 5%, and (d) 10% nonresponse for a simple random sample and a sample proportionally stratified on sex with overall sample sizes of 130, a null of $G = G_m + G_f \leq 784$. Nonresponse is assumed to be independent of sex, but the numbers of Gregoire and Rossi nonresponders are allowed to vary within their deterministic bounds. The solid curves represent the power bounds from a simple random sample. These bounds collapse to a single curve under 0% non-response. The dashed curves represent the power bounds from a proportionally stratified random sample.

3 Conclusion

As this paper has shown, classical testing within the sampling design-based framework and a Type I error rate of 5% will not have high power unless the sample size is a large proportion of the population or the true parameter value is quite extreme. One can attempt to increase the power through stratification, but the gains are not assured, and will depend on the heterogeneity between the groups. Furthermore, substantial nonresponse may decrease the power of either procedure and may eliminate the gains from a stratified test. In simple terms, in order to maintain the level of the test at α in the presence of nonresponse, we must treat the nonresponders as if they would have given evidence for the null hypothesis, and therefore we may not be able to reject the null for moderate values of the alternative hypothesis, even if we sample the entire population.

There are three major criticisms that can be leveled against the analysis in this paper. First, the testing framework that bounds the maximum probability of Type I error at α is conservative, and the choice of $\alpha \leq .05$ is arbitrary. We are not endorsing this testing procedure or a Type I error of 5%, however, we wanted to make explicit the ramifications for a popular statistical decision procedure when either nonresponse is a concern, or when an unrepresentative sample is used. The second major criticism can be leveled at our treatment of nonresponse. Treating all nonresponders as if they would have provided evidence for the null hypothesis is certainly conservative, but it is the *only* way to maintain the Type I error level of the test at α without making modeling assumptions. In a particular application, there may be reasonable modeling assumptions to be made, but this paper presents the limits of design-based inference for the standard asymmetric testing procedure. Third, in our attempt to produce a model-free statistical analysis, we have still made a number of assumptions. In particular, we have assumed only two candidates, we have ignored the possibility of measurement error or spoiled ballots, and perhaps most importantly the possibility of frame error (which was almost certainly present in Borders v King County, 2005 frame of 1,439 invalid voters). Relaxing any of these assumptions will either increase the Type I error level of the test or decrease the power. Hence, the results presented here represent a form of best case scenario.

The implications of this work for legal challenges to election results are clear. Given the fact that post-vote vote-choice data will often be plagued by nonresponse (and other issues assumed away in this paper), there is little chance of obtaining a representative sample, and court decisions on the basis of such data will likely depend on strong modeling assumptions (whether or not these are acknowledged). More broadly speaking, while courts may be reluctant to make explicit findings about out-of-sample data (e.g. invalid voters who did not testify), courts should note that when they conduct procedures (e.g. the adjustment of vote totals) on the basis of a potentially unrepresentative sample (e.g. invalid voters that testify), they may be implicitly making findings about this out-of-sample data.

References

Adolph, 2005. Report of WSDCC expert Christopher Adolph, Ph.D., Borders v King County, No. 05-2-00027-3 (Wash. Super. Ct., Chelan Co.).

Belcher v Mayor of Ann Arbor, 1978. 402 Mich. Rep. 132-134.

Borders v King County, 2005. No. 05-2-00027-3 (Wash. Super. Ct., Chelan Co.).

H. Chernoff and L. Moses. *Elementary Decision Theory I.* Dover, 1986.

T. Downs, D.C. Gilliland, and L. Katz. Probability in a Contested Election. *The American Statistician*, 32(4):122–125, 1978.

M.O. Finkelstein and H.E. Robbins. Mathematical probability in election challenges. *Columbia Law Review*, 73:241–248, 1973.

Gill, 2005. Report and Supplemental Report Regarding Invalid Ballots Cast in the 2004 Washington State Gubernatorial Race, Anthony Gill, Ph.D., Borders v King County, No. 05-2-00027-3 (Wash. Super. Ct., Chelan Co.).

D.C. Gilliland and P. Meier. The probability of reversal in contested elections. In M.H. DeGroot, S.E. Fienberg, and J.B. Kadane, editors, *Statistics and the Law*. John Wiley & Sons, New York, 1986.

R.M. Groves et al. *Survey nonresponse.* Wiley New York, 2002.

Handcock, 2005. Report of WSDCC expert Mark S. Handcock, Ph.D., Borders v King County, No. 05-2-00027-3 (Wash. Super. Ct., Chelan Co.).

B. Harris. Election Recounting. *The American Statistician*, 42(1):66–68, 1988.

Harvard Law Review (1975). "developments in the law – elections". 88, 1111-1339.

Katz, 2005. Report and Supplemental Reports on 2004 Washington Gubernatorial Election, Jonathan N. Katz, Ph.D., Borders v King County, No. 05-2-00027-3 (Wash. Super. Ct., Chelan Co.).

E.H. Lehmann and J.P. Romano. *Testing Statistical Hypotheses.* Springer, 2005.

R.J.A. Little and D.B. Rubin. *Statistical analysis with missing data.* John Wiley & Sons, Inc. New York, NY, USA, 1986.

H.E. Robbins. Comment on the probability of reversal in contested elections. In M.H. DeGroot, S.E. Fienberg, and J.B. Kadane, editors, *Statistics and the Law.* John Wiley & Sons, New York, 1986.