Post-election audits: statistical power based methods and a trigger for further auditing

Author: Katie Lim

Advisor: Vittorio Addona

Macalester College

Mathematics and Computer Science Department

Abstract

Two important components of an audit procedure are the sample size and the decision rule for expanding the audit. This paper describes a method for determining the audit size that ensures a high probability of detecting miscounts if election altering ones exist. Then, two possible triggers for further auditing are developed: a modified bootstrap confidence interval and a modified Hoeffding bound. Both estimate the net gain in votes for the originally reported loser if a full audit were conducted. In simulations, both methods maintained high power for various miscount situations, with the Hoeffding bound having slightly lower false positive rates.

1 Introduction

Fair elections represent one of the central pillars of a democratic society. Elections enable citizens to decide how their nation is governed, and the fair election of representatives grants the government legitimacy both domestically and internationally. In order to preserve the credibility of this system, it must be ensured that the outcomes of elections truly reflect the voters' intentions, and the public must feel confident in the results. In recent years, computerized voting has become increasingly prevalent in U.S. elections, but it is often viewed as less transparent than hand counted paper ballots. This has lead to increased concern regarding the validity of the results from these machines, and subsequent interest in post-election audits as a method of determining whether the computerized vote tallies match voters' intended outcome. The use of computerized voting varies by state. Some states use optical scanners to count paper ballots and other states use direct-recording electronic (DRE) voting machines, which allow voters to select a candidate using buttons or a touchscreen and record the vote electronically. Since Minnesota uses optical scanners to count paper ballots, this paper focuses on that method of voting, however, the analysis can be applied to DRE machines that have a voter verified paper audit trail (VVPAT). Computerized voting has introduced new possibilities for miscounts in the vote totals that, if large enough, could change the outcome of an election. Possible sources of error could include: programming mistakes, hardware malfunction, or malicious attacks.

The post-election audit is an important part of the election process that offers a final verification of the election outcome. These audits can have several goals. Their purpose may be to verify the accuracy of the voting machines or to provide election officials with information to improve the voting process for future elections. For example, mistakes or ambiguities found in an audit may influence the future design of ballots or improve poll worker training. An alternative goal of an audit is to detect possible election altering miscounts in order to provide the public and candidates with confidence in the election results. The methods discussed in this paper aim to satisfy the last goal, winner verification. The primary concern is designing an audit to ensure that the correct winner of the election is declared. This paper's discussion is restricted to elections between two candidates, although including more than two candidates should not substantially change the analysis.

Currently, only 16 states conduct any sort of post-election audit, but this number is increasing. Most of these states audit a set percentage of precincts in a race without regard to the originally reported margin of victory in that particular race. This method will be referred to as a fixed percentage audit. As this paper will show, for close races this design does not guarantee a high probability of detecting an election altering miscount should one exist because too few precincts are audited. Furthermore, for races with large margins of victory, auditing a fixed percentage of precincts often results in auditing more precincts than necessary to verify the winner with a high probability (McCarthy et al. 2008). This over auditing in races with large margins and under auditing in races with close margins is an inefficient use of a state's resources. The inadequacies of fixed percentage audit plans have been gaining some national attention. A bill that sets audit levels based loosely on the margin of the race has been proposed to the United States Congress by Congressman Rush Holt (D, New Jersey). Congressman Holt's plan uses a tiered approach, which requires a minimum 3% audit for all federal races. If the margin of the race is less than 2% but greater than 1%, a 5% audit is required. For races with margins of 1% or less, election officials must conduct a 10% audit. This new bill modifies the current fixed percentage audit procedure to require more auditing in close races without unduly increasing the burden on election officials for races with landslide victories. The state of Oregon has passed a law using a similar tiered approach and the New Jersey legislature passed a bill that would implement some of the ideas discussed in section 2 of this paper.

The structure of the paper is as follows. Section 2 describes a method developed to determine the sample size of an audit using the criteria of statistical power, and it provides the notation used throughout the remainder of the paper. Section 3 phrases the post-election audit in terms of a hypothesis test and discusses statistical power and false positive rates associated with the hypothesis test. Next, section 4, proposes two triggers for further auditing. The first is based on a modified bootstrap method and the second utilizes Clayton's (1994) approach based on the Hoeffding upper bound for errors in a population. Section 5 describes and discusses the results of a simulation conducted to evaluate the performance of the two triggers. Then, section 6 applies the method for choosing the audit sample size described in section 2 to previous Minnesota elections and compares the results to those under the current Minnesota audit law. Section 7 discusses the possibility of using individual ballots as a sampling unit instead of precincts. Finally, section 8 provides concluding remarks.

2 A power-based audit

This section begins with a quick review of the hypergeometric distribution, which will be used in developing the power-based audit. Next, the relevant literature is discussed and the power-based method is fully explained.

2.1 The hypergeometric distribution

Consider a finite population that contains N objects of two different types. Suppose there are k type 1 objects and $(N-k)$ type 2 objects. Let the random variable X be the number of type 1 objects obtained from a sample of size n drawn without replacement from the population. The random variable X is said to have a hypergeometric distribution. The probability of drawing x type 1 objects in n draws without replacement is given by the following formula:

$$P(X=x)=\frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} \tag{1}$$

$$\textit{where} \quad \binom{k}{x}=\frac{k!}{x!\;(k-x)!}$$

It is not difficult to show that the expected value and variance of X are given by $\left(\frac{kn}{N}\right)$ and $\left(\frac{kn}{N}\right)\left(1-\frac{k}{N}\right)\left(\frac{N-n}{N-1}\right)$, respectively. The hypergeometric distribution is used in section 2.2.1 to determine the audit sample size according to the power-based method.

2.2 Literature review

Saltman (1975) was the first to recognize the shortcomings of the fixed percentage audit strategy. Its weaknesses can best be illustrated through a hypothetical example. Consider an election that involves 1000 precincts each consisting of 1000 votes, and suppose that the race has a 10,000 vote (1%) margin according to the counts obtained from the optical scanners. If there had been a vote shift of at least 5,000 votes from the originally reported loser to the originally reported winner, the initial election outcome would be incorrect. Assume that this error is caused by at least a 5% vote shift (50 votes) in 100 of the precincts (Saltman 1975). In 1975, California had a set 1% audit, which would call for $n = 10$ of the 1000 precincts to be audited. The population of $N = 1000$ precincts can be separated into the $k = 100$ miscounted precincts (type 1 objects) and the $N - k = 900$ correctly counted precincts (type 2 objects). Then, the hypergeometric distribution can be used to calculate the probability of finding at least one miscounted precinct in the audit, which in this case is 0.655 or 65.5%. Saltman concludes that this value is far too low and suggests that the number of precincts audited, n, be based on some pre-specified probability, p, of finding a miscounted precinct, if enough exist in the population of all precincts to alter the election outcome. The next section describes the details of this approach.

2.2.1 Explanation of the model

The development of a model that uses the hypergeometric distribution to determine the number of precincts to audit (n) in order to ensure detection of a possible miscount with probability p depends crucially on the number of miscounted precincts present in the population (k). As k increases, it becomes more likely that a miscounted precinct will be found with a given audit sample size (n). In order to guard against the "worst-case scenario", the minimum value of k that could alter the outcome of the election must be determined. Denote this minimum value by k_{min}. An audit procedure that ensures a probability, p, of finding a miscount when $k = k_{min}$ will have an even higher probability of detecting discrepancies when $k > k_{min}$.

The value of k_{min} depends on the originally reported margin in the race. Suppose candidate A is declared the winner of an election over candidate B, and the margin of victory reported is very large. If the incorrect winner of the election was declared and, in fact, B obtained the most votes, there would have to be many miscounted precincts. On the contrary, in a very close race, only a few miscounted precincts could have caused a mistake in the declaration of a winning candidate. The extent to which precincts are miscounted also affects the value of k_{min}. It is assumed that very large percentage shifts from one candidate to another are detectable by "inspection". In calculating k_{min}, it is therefore important to identify or choose the largest shift that could reasonably go unnoticed to candidates or election officials. This maximum assumed percentage of votes shifted per precinct from one candidate to another will be denoted s_{max}. The larger the value of s_{max}, the smaller the value of k_{min} since a larger s_{max} allows for the possibility that more votes could be changed in each precinct. Fewer miscounted precincts would thus be required to alter the election outcome.

Saltman sets $s_{max} = 5\%$ in his example and recommends using a value between 5-10%. This actually represents an overall 10-20% change in the margin because one vote taken from Candidate A and given to Candidate B will change the difference in their vote totals by two votes. Some more recent papers suggest using $s_{max} = 15\%$ (Dopp and Stenger 2006) or 20% (Holt Bill 2007, Rivest 2006, Stanislevic 2006) signifying a 30% or 40% change, respectively, in the vote counts between the two candidates. It may be that some individuals will not feel comfortable placing an upper bound on the percentage of miscounted votes. To address the possibility that miscounts larger than s_{max} could occur, random audits can be supplemented with targeted audits. These audits would examine precincts selected by candidates as suspicious and in need of further inspection. Allowing candidates to pick a set number of precincts that they would like audited ensures that precincts with unexpected results are examined. This option would also incorporate the research and knowledge of candidates and their parties into the audit procedure.

Another consideration in determining k_{min} is the precinct size distribution. Saltman assumes that all of the precincts contain the same number of votes but, for most races, precincts vary in size. In the 2006 Minnesota governor's race, out of 4123 precincts the largest precinct contained 4088 votes while 9 precincts had fewer than 3 ballots cast. For races with precincts that vary greatly in size, the value of k_{min} depends substantially on the size of the largest precincts. If all of the miscounts occur in large precincts, then fewer precincts will need to contain miscounts in order to have announced the wrong winner. Stanislevic (2006) is the first to consider different sized precincts in his calculation of k_{min}. He assumes that all of the discrepancies happen in the largest precincts, which truly minimizes the value of k_{min}.

These steps describe the procedure for finding k_{min}.

1. Determine the margin, m in votes. The number of votes that would need to have been switched from the originally reported loser to the originally reported winner in order to change the outcome of the race is $\frac{m}{2}$. [1]

2. Using s_{max}, determine the number of votes t that need to be involved in the shift by solving the following equation:

$$ts_{max} = \frac{m}{2}$$

$$t = \frac{m}{2(s_{max})}$$

3. Order the precincts by descending size and calculate the cumulative sum of votes.

[1] Technically the number of votes that would need to have been switched to change the outcome is the ceiling of $\frac{m+1}{2}$. If m is odd, then $\frac{m}{2}$ will not be an integer and one must be added to the margin before dividing by two. If m is even, then the number of votes that would have to be switched in order to reverse the election outcome would actually be one more than $\frac{m}{2}$, which is the same as taking the ceiling of $\frac{m+1}{2}$.

4. Find the fewest number of precincts that will include t votes by finding the precinct at which the number of cumulative votes is at or above t. Call this number k_{min}.

Next, the audit sample size, n, must be selected such that there is at least probability, p, of finding one or more miscounted precincts if there are k_{min} miscounted precincts in total. The model then calls for an audit of n precincts obtained as the smallest integer n that satisfies the following, equivalent, equations:

$$P(X \geq 1) \geq p$$

$$1 - P(X = 0) \geq p$$

$$1 - \frac{\binom{k}{0}\binom{N-k}{n}}{\binom{N}{n}} \geq p$$

The method just described is referred to as a "power-based" audit because, unlike the fixed percentage approach, the sample size, n, is determined to ensure power, p.

2.2.2 Other work

Although the method described in section 2.2.1 was originally proposed by Saltman (1975), other authors have used this model when developing and analyzing post-election audit procedures. Dopp and Stenger (2006), for example, apply the Saltman model to races in Lake County, OH, and Multnomah County, OR.

The hypergeometic distribution calculations used in the power-based model require the use of a computer and some knowledge of statistical software. If the public or election officials want to calculate the number of precincts to audit based on the margin of a race, they may only have access to a hand-held calculator. Rivest (2006) develops a formula that yields an approximation for n that is almost never an underestimate and is always within four of the value found using the hypergeometric distribution. Election officials and the public may view this calculation as more transparent because they can replicate it easily. The calculations and simulation in this paper, however, use statistical software to implement the power-based method.

McCarthy et al. (2008) compare the power-based method with the tiered approach proposed in the Holt Bill. The tiered approach requires that a strict percentage of precincts be audited based loosely on the margin of victory, but this percentage is independent of the total number of precincts involved in a race. According to the power-based audit, as the total number of precincts decreases, the number of precincts necessary to audit does not decrease linearly. The percentage of the total number of precincts that must be audited, in order to obtain a desired value of p, increases when as the total number of precincts decreases. The power-based model calculates the audit size while ensuring a high value of p regardless of the total number of precincts. The authors conclude that the power-based model is more efficient and effective than the tiered approach because it incorporates both the margin of the race and the number of precincts in the race when determining the audit sample size n.

In a recent paper, Aslam, Popa, and Rivest develop alternative approaches to choosing the sample size n that still maintain high power but result in far smaller values of n than Saltman's method. Their major contribution is to demonstrate that when precincts vary in size, there are efficiency gains to assigning a probability distribution of selection that is weighted towards the larger ones. They argue that making it more likely that larger precincts are selected in the auditing process better reflects the "value" of these precincts because attacking one large precinct allows the adversary access to altering far more votes than one small precinct. In the two examples that the authors present, the power-based method requires an n that is over twice as big as their methods. This paper is promising because decreasing the size of n without reducing the statistical power of the audit will result in lower costs without reductions in effectiveness.

This paper will use the power-based method and will extend the audit procedure by developing a trigger for further auditing, which is a necessary next consideration after conducting the initial audit of n precincts. This is also the first paper to provide an analysis of Minnesota's 2006 audit and to offer a detailed discussion of the audit in terms of a hypothesis test. The next section describes two alternative ways of phrasing the audit as a hypothesis test and discusses the power and false positive rate associated with each test.

3 Framing the audit in the context of a hypothesis test

Hypothesis testing is used to make inference on two competing conjectures. The null hypothesis is assumed to be true unless evidence is collected that supports the alternative hypothesis. Once data is collected, a test statistic that is some function of the data can be calculated under the null hypothesis. The decision to reject or not to reject the null hypothesis is made based on the probability of observing a test statistic as or more extreme than the one observed given that the null hypothesis is true. This probability is called a p-value. The null hypothesis is rejected if the p-value is below a certain threshold, α, typically taken to be 0.05. A small α implies that it is not likely that rejection of the null hypothesis occurs when the null hypothesis is true. A second measure of a hypothesis test's effectiveness is its 'power'. Power is the probability of rejecting the null hypothesis when the alternative is true. Attaining high power ensures that failure to reject the null hypothesis does not occur simply because the test has difficulty detecting that the alternative is true. A rejection rule can also be based on attaining a certain power under a specific alternative hypothesis. An ideal hypothesis test has a small α and high power. In the context of an audit, there are several ways that these hypotheses could be defined.

One possibility is to consider testing whether there are enough miscounted precincts in the race to change the result. This is equivalent to testing whether there are at least k_{min} miscounted precincts. The associated null and alternative hypotheses would be:

$$H_0 : k \leq k_{min} - 1 \qquad vs. \qquad H_a : k \geq k_{min} \tag{2}$$

where k is the true number of miscounted precincts

A natural test statistic for this hypothesis test is the number of miscounted precincts found during the audit, denoted as X. X is known to have a hypergeometric distribution with parameters N, n, and k. One possible rejection rule would be to reject the null hypothesis if one or more miscounted precincts is found during the audit. The associated α is the following:

$$\alpha = P(X \geq 1 | k = k_{min} - 1) \tag{3}$$

$$= 1 - \frac{\binom{k_{min}-1}{0}\binom{N-k_{min}+1}{n}}{\binom{N}{n}} \tag{4}$$

The minimum power of this hypothesis test is:

$$
\begin{aligned}
Power &= P(X \geq 1 | k = k_{min}) & (5) \\
&= 1 - \frac{\binom{k_{min}}{0}\binom{N-k_{min}}{n}}{\binom{N}{n}} & (6)
\end{aligned}
$$

Note that both the power and α depend on the sample size of the audit, n. The power-based audit calculates n for a certain value of p, the probability of finding at least one miscounted precinct if k_{min} exist. By definition, p is the minimal power for this hypothesis test. It is the probability of rejecting the null hypothesis (finding one or more miscounts) if the alternative is true (there are at least k_{min} miscounted precincts). Using this method to calculate the sample size of the audit guarantees high power for the hypothesis test described in (2). Achieving high power is especially of interest in a post-election audit because it is imperative that no election altering miscounts go undetected. The power-based method guards against "false negatives" meaning that failure to reject the null hypothesis implies that it is unlikely that enough miscounts exist to change the election outcome. In other words, if zero miscounted precincts are found in the initial audit, this provides confidence in the declared results.

As previously mentioned, the goal of the audit procedure is to ensure that the correct winner has been announced. The hypothesis test to determine if the correct winner was declared is not equivalent to testing whether there are enough miscounted precincts to change the election outcome. It is possible that k_{min} or more miscounted precincts exist without their miscounts having altered the election winner. The k_{min} discrepancies could have occurred in some of the smaller precincts or the miscounts may have been smaller than s_{max}. The hypothesis test that is actually of interest is the following:

$$H_0 : \text{Correct winner was declared} \quad vs. \quad H_a : \text{Incorrect winner was declared} \tag{7}$$

(7) can be rephrased as:

$$H_0 : P_w \geq 0.5 \quad vs. \quad H_a : \; P_w < 0.5 \tag{8}$$

$$where \quad P_w \; = \; true\ proportion\ of\ votes\ for\ the\ originally\ reported\ winner$$

The hypothesis test as phrased in (8) will be discussed in section 7 during the discussion of ballot auditing.

The first hypothesis test as stated in (2) has an easily measured test statistic with a known distribution making it possible to calculate α and the power of the test. One could propose to use the same test statistic and rejection rule for the hypothesis test (7), but there is no method to directly calculate the α or power associated with this test. The important values are the α and power for test (7) because in order for the audit to be effective it must ensure a high level of power for the hypothesis of interest. A statement can be made about α and the power of test (7) in comparison to test (2) by investigating the relationship between the two hypothesis tests. In order to better understand the two tests, it is useful to examine a diagram. Figure 1 in the appendix depicts the two hypothesis tests and rejection regions. There are six areas in the diagram and each can be described by three characterisitcs: (1) was the correct winner declared, (2) were there at least k_{min} miscounted precincts, and (3) was a miscounted precinct revealed in the audit resulting in a rejection of the null hypothesis. The characteristics of each section are summarized in Table 1 of the appendix.

First, consider the null hypotheses for the two tests. If H_0 (2) is true, then H_0 of (7) is also true. This can be seen in Figure 1 because the area where H_0 (7) is true is a subset of the area where H_0 (2) is true. The reverse is not necessarily the case. It is possible to have declared the correct winner when there are more than k_{min} miscounted precincts. When comparing the alternative hypotheses of the two tests, it is clear that if the incorrect winner was declared there must be at least k_{min} miscounted precincts. This can also be determined from Figure 1 since the area where H_a (7) is true is a subset of the area where H_a (2) is true. The converse is not always the case. If there are at least k_{min} miscounted precincts, the correct winner may or may not have been announced. The relationship between the alternative hypotheses is also one way: H_a of (7) implies H_a of (2).

These relationships allow for the comparison of power for the two hypothesis tests when they use the same rejection rule. The power for test (2) is at least p by design. This value corresponds to $P(A \cup B | A \cup B \cup F \cup E)$ in Figure 1. The power for test (7) corresponds to $P(A|A \cup F)$ in Figure 1. The next two proofs will show that the power of test (7) is equal to the power for test (2).

Claim 1:

$$P(B|B \cup E) \;=\; P(A|A \cup F)$$

Proof:

$$P(B|B \cup E) = P(reject\ H_0 | k = k^* \cap correct\ winner\ was\ declared)$$

$$P(A|A \cup F) = P(reject\ H_0 | k = k^* \cap incorrect\ winner\ was\ declared)$$

$$where\ k^*\ is\ any\ value\ \geq k_{min}$$

For both hypothesis tests the null is rejected if one or more miscounted precincts are found. If the audited precincts are randomly selected, then the probability of finding at least one miscounted precinct depends only on k, the total number of miscounted precincts. The value of $k = k^*$ in both of the situations above. The determination of whether the correct winner was announced depends on the size of the miscounts and which precincts are miscounted. Neither of these considerations affect the probability of finding at least one miscounted precinct when k^* exist. Thus, the probability of rejecting the null hypothesis given that there are k^* miscounted precincts and the correct winner is declared is equal to the probability of rejecting the null hypothesis given that there are k^* miscounted precincts and the incorrect winner is declared.

Claim 2:

$$P(A \cup B | A \cup B \cup F \cup E) = P(A|A \cup F)$$

Proof:

$$P(A \cup B | A \cup B \cup F \cup E = P(A|A \cup B \cup F \cup E) + P(B|A \cup B \cup F \cup E)$$

$$= \frac{P(A)}{P(A \cup B \cup F \cup E)} + \frac{P(B)}{P(A \cup B \cup E \cup F)}$$

$$= \frac{P(A)}{P(A \cup F)} \frac{P(A \cup F)}{P(A \cup B \cup F \cup E)} + \frac{P(B)}{P(B \cup E)} \frac{P(B \cup E)}{P(A \cup B \cup F \cup E)}$$

From Claim 1: $P(B|B \cup E) = P(A|A \cup F)$ which can be rewritten as $\frac{P(A)}{P(A \cup F)} = \frac{P(B)}{P(B \cup E)}$. This can be used to simplify the previous equation to:

$$= \frac{P(A)}{P(A \cup F)}(1)$$

and Claim 2 is shown. Thus, the power associated with test (7) is the same as the power for test (2).

A similar process can be used to examine α for test (7). The α for test (2) can be calculated and is equal to $P(D|D \cup C)$. The α of the test in (7) can be represented by $P(D \cup B|D \cup B \cup C \cup E)$. The following two proofs will show that α of test (7) is greater than α of test (2).

Claim 3:

$$P(B|B \cup E) > P(D|D \cup C)$$

Proof:

$$P(D|D \cup C) = P(reject\ H_0|k \leq k_{min} - 1 \cap correct\ winner\ was\ declared)$$

$$P(B|B \cup E) = P(reject\ H_0|k \geq k_{min} \cap correct\ winner\ was\ declared)$$

Again for both hypotheses the null is rejected if at least one of the audited precincts contains a miscount. The two above probabilities in this case will be different because the number of miscounted precincts k is different. The probability of finding a miscounted precinct is higher when there are more of them. Therefore, the second statement has a higher probability than the first statement.

Claim 4:

$$P(D \cup B|D \cup B \cup C \cup E) > P(D|D \cup C)$$

Proof: This follows from Claim 3 by the same logic as was used to prove Claim 2. This implies that the α for test (7) is larger than the α for test (2).

This section has shown that conducting an audit of size n and using the detection of at least one miscounted precinct as a rejection rule yields power of at least p for hypothesis test (7). This implies that choosing a high value for p when calculating n ensures that it is very unlikely that an election altering miscount goes undetected. The “false positive rate” (probability of finding at least one miscounted precinct when the correct winner was declared) is represented by α. In the context of an audit, the priority is to ensure high power, however, in consideration of limited time and resources the value of α is also of concern. When α is large, additional auditing will frequently be required when the correct winner has been announced. As (4) and (6) show, for hypothesis test (2) maintaining a high value for power implies that α will also be quite high. This occurs because if there is a high probability of finding a miscounted precinct when there are k_{min} of them, then there will also be a high probability of finding one when there are $k_{min}-1$ of them. This is an extreme example but it is important to note that for values of k less than, but close to, k_{min} the audit will have a relatively high false positive rate. As discussed previously, this means that the α for test (7) will be even higher.

In this section, a simple rejection rule that consisted of rejecting the null hypothesis if at least one precinct is deemed miscounted was examined. In section (4), possible triggers for further auditing are discussed. In a sense a trigger for further auditing is analogous to a rejection rule because calling for more auditing is like rejecting a null hypothesis. The false positive rate and statistical power associated with these triggers will be considered when evaluating their performance.

4 Triggers for Further Auditing

An audit of size n, selected according to the power-based method, that reveals no miscounted precincts provides strong evidence that the correct winner in the election was declared. Finding a precinct that contains miscounts, however, should not automatically lead to further auditing for the purpose of winner verification, though further investigation into what led to the miscount may be deemed valuable for other auditing goals. There may be instances in which miscounts are found that actually increase the winner's lead or cases where the miscounts found are too small to put the race results in question. A complete audit procedure must contain some rule that is able to distinguish between miscounts that require further auditing and those that do not. One possible approach to developing a trigger is to find an estimate of the net gain in votes that would be obtained by the originally reported loser should a full recount be conducted. The net gain in votes is a sensible value to estimate because it directly determines if the correct winner was declared. This section begins by developing a trigger based on a nonparametric bootstrap and then discusses some of its weaknesses. Next, two alternative triggers for further auditing are offered that may avoid the shortcomings of the traditional bootstrap trigger.

4.1 Trigger based on a bootstrap confidence interval

Most of the previous literature, and section 2.2.1, considers a finite set of N precincts (objects) which are each one of *two* types, correctly counted or miscounted. This simple distinction provides an adequate way of selecting the audit sample size. However, it is an oversimplification of the reality that precincts can be miscounted to varying degrees (e.g. 0.1% discrepancy, 1% discrepancy, 10% discrepancy) and are of varying sizes (50 votes, 500 votes, 5000 votes). The trigger proposed here attempts to more fully utilize the extent to which different precincts are miscounted. Let X_1, X_2, ..., X_N represent the finite population of precincts, where X_i = the net gain in votes for the originally reported loser over the originally reported winner should the i^{th} precinct be audited. Note that in this context the size of the finite population, N, is a known quantity. Let $T = \sum_{i=1}^{N} X_i$, meaning that T is the net gain in votes for the originally reported loser should a full recount be conducted. Recall that m is the originally reported margin of victory in the race, and denote the audit sample by Y_1, Y_2, ... Y_n, where Y_j represents the net gain in votes for the originally reported loser in the j^{th} audited precinct. The problem of interest is to estimate T using Y_1, Y_2, ... Y_n (Horvitz and Thompson 1952). Consider the following straightforward estimator of the population total, T:

$$\widehat{T} \; = \; \frac{N}{n} \sum_{j=1}^{n} Y_j \; = \; N \, \overline{Y} \tag{9}$$

$\widehat{T}$ is an adequate point estimator for T but it is useful to obtain an interval estimator of T in order to provide some measure for the reliability of the estimate. Because the population X_1, X_2, ..., X_N has potentially substantial departures from Normality, and because the audit sometimes has a quite modest sample size, n, the distribution of $\widehat{T}$ may still exhibit skewness. Traditional confidence intervals based on the Central Limit Theorem, therefore, cannot be constructed. Consequently, a nonparametric bootstrap is used to obtain replicates of the estimator, $\widehat{T}$, and then the percentiles of this set of replicates is used to obtain an interval estimate of T. Specifically, the procedure is:

1. Sample *with* replacement from the original audit sample, Y_1, Y_2, ... Y_n, to obtain a bootstrap sample Y_1^*, Y_2^*, ... Y_n^*, of the same size.

2. From the bootstrap sample obtained in **1.** calculate an estimate of T using (9), which is denoted by $\widehat{T}^*$

3. Repeat step **1.** and **2.** a large number of times, say, 1000, to obtain the bootstrap replicates, $\widehat{T}_1^*$, $\widehat{T}_2^*$, ..., $\widehat{T}_{1000}^*$.

4. Sort $\widehat{T}_1^*$, $\widehat{T}_2^*$, ..., $\widehat{T}_{1000}^*$ and form a one-sided 99% confidence interval for T by selecting the 99^{th} percentile of these sorted bootstrap replicates. Denote this value by $\widehat{T}_{(990)}^*$.

Having obtained an interval estimator for T, the next step is to develop a trigger mechanism that will call for further auditing. This can be accomplished by comparing $\widehat{T}_{(990)}^*$ to the originally reported margin of victory, m. If $\widehat{T}_{(990)}^*$ lies below m then the audit should stop since there is no evidence that T could be as large as m. If $\widehat{T}_{(990)}^*$ is larger than m this indicates the plausibility that T is at least as large as m and should trigger further auditing.

In order to test the effectiveness of this trigger, a computer simulation of elections was conducted with various degrees of miscounts. The precinct size distribution used in the simulation was patterned after Minnesota's. There were 4180 precincts containing 2,238,750 votes with individual precinct sizes ranging from 50-3500 votes. It was assumed that races were comprised of two candidates, A and B, and votes were assigned to candidate A in each precinct by generating a proportion of votes from a Beta(α, β) distribution parameterized as:

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\ x^{\alpha-1}\ (1-x)^{\beta-1} \quad for \quad 0 < x < 1$$

Three pairs, (α, β), were used: (3, 2.95), (3, 2.85), and (3, 2.75) to indicate varying strengths of preference among the population for candidate A. The remaining votes in each precinct were assigned to candidate B, and these totals represented the true votes cast for each candidate. To obtain the vote totals "observed" for each candidate from the machine tallies, it was assumed that k precincts were miscounted, each with a miscount or "shift" percentage of s. The different values of k used were: 10, 20, 30, 50, 100, and 200, and the different values of s used were: 0.5, 1, 3, 5, 10, 20.

The miscounts were weighted toward the larger precincts by randomly selecting the k miscounted precincts from the 830 largest precincts, which ranged from 1250 votes to 3500 votes. Scenarios where votes were shifted from candidate A to candidate B and vice versa were both simulated. Each simulation scenario was repeated 5000 times and on each occasion it was determined whether the correct winner of the election would have been declared based on the observed vote counts. With the audit sample size determined as outlined in section 2.2.1, the bootstrap trigger introduced above was applied to ascertain whether it would call for further auditing. In this way, both the **power** and **false positive rate** associated with the trigger could be determined.

Preliminary results showed that this trigger maintained high power in the situations tested and had a fairly low false positive rate.[2] The simulation, however, only examined situations in which precincts were either correctly counted or were miscounted by the same percentage s. Further investigation of more complicated miscount patterns revealed instances in which the trigger failed to achieve high power. For example, when there are 40 miscounted precincts, 35 of which have 1% miscounts and 5 of which contain 20% miscounts this trigger fails to maintain 99% power. The reason for the trigger's poor performance can be found by examining the rationale behind choosing the audit sample size n. Recall that when n precincts are audited there is at least a 99% chance of finding at least one miscounted precinct if enough exist to have altered the election. The problem that arises when there are precincts miscounted by varying degrees is that the audit may only catch one or two miscounted precincts that have relatively minor miscounts in comparison to others. In the example above, if the audit sample only detects one or two miscounted precincts with a 1% discrepancy, the bootstrap samples can only contain precincts miscounted at the

[2] These results were not included in this paper because this is not the preferred trigger, however, they are available to anyone interested. Please email klim@macalester.edu to request a copy.

1% level. The audit and trigger would, in effect, never realize or account for the 20% miscounts that exist in the population. In this case, $\widehat{T}^*_{(990)}$ will often be too small and so the power of the trigger will decrease as it fails to call for further auditing even when the wrong winner has been declared.

In order to guard against the situation described above, the trigger can be modified to include the possibility that there exist heavily miscounted precincts that are not represented in the audit sample. That is, after the sample Y_1, Y_2, ... Y_n is obtained, an artificial data point, Y_{n+1}, can be added whose value is equal to 20% of the votes in the largest precinct (since, under the assumptions, this is the largest miscount that could occur). Including this last data point in the sample, from which the bootstrap samples are taken, will make $\widehat{T}^*_{(990)}$ reflect the possibility that the audit may have only found the least serious miscounts that exist in the population X_1, X_2, ..., X_N. This modified trigger proceeds exactly as before. Bootstrap samples are obtained from Y_1, Y_2, ... Y_n, Y_{n+1}, then $\widehat{T}^*_{(990)}$ is determined and if it is larger than m further auditing is conducted. This trigger is called the modified bootstrap trigger.

A second alternative draws on financial auditing literature, which seeks to find an upper bound for the error in a population from a sample. Clayton (1994) develops an estimation technique for this upper bound based on a modification to the Hoeffding bound. The Hoeffding bound is a specific version of Chebychev's Inequality that assumes that the values of the population X_1,...,X_N are bounded. Chebychev's inequality places an upper bound on the probability that a random variable is a certain distance from its mean. Previous work has shown the Hoeffding bound to be very conservative, meaning that it produces fairly high estimates for the upper bound on error (Hoeffding, 1963) and (Bickel, 1992). The bootstrap technique has been shown to provide "tighter" estimates. By combining the bootstrap and Hoeffding bound, Clayton hopes to provide an estimate that inherits “some of the reliability of the Hoeffding and some of the bootstrap’s tightness.”

He begins with the inequality developed by Hoeffding (1963). Imagine a sample of size n from a population X_1, X_2,...,X_N where X_i is the proportion of items correctly counted in unit i. Then $0 \leq X_i \leq 1$ and the sample mean of the proportion correct is denoted by $\bar{R}$, which estimates the population mean, ρ. It is important to note that in this method the precincts (or sampling units) must be the same size because the population sampled are the proportion of correctly counted votes in a precinct. If the precincts were different sizes, it would not be possible to determine the number of miscounted votes from, say, a 75% correctly counted precicnt. Then for some constant c, where $0 \leq c \leq 1 - \rho$:

$$P\{\bar{R} - \rho \geq c\} \leq \left\{ \left(\frac{\rho}{\rho + c}\right)^{\rho + c} \; \left(\frac{1 - \rho}{1 - \rho - c}\right)^{1 - \rho - c} \right\}^n \tag{10}$$

The traditional Hoeffding bound is a $1-\alpha$ lower bound, L_ρ, on the proportion of the population that is correct, ρ. This can be found by replacing $\rho + c$ with $\bar{R}$ on the right hand side of (10), setting the equation equal to α, and solving for ρ (Clayton, 1994). It is easy to find an upper bound, $u_{1-\rho}$, for the error in the population by taking $1 - L_\rho$. This estimate is the regular Hoeffding bound that is considered to be very conservative (Clayton, 1994). For the trigger developed here, both power and the false positive rate are of concern. The ideal trigger will almost never underestimate T but will also not drastically overestimate T. Otherwise, even when T is far below m, the trigger would often still call for further auditing.

Clayton incorporates the bootstrap in his estimation by proposing that bootstrap samples are obtained through resampling with replacement from the original observations. From these samples, the mean proportion of items correct in each bootstrap sample, $\bar{R}^*$, can be calculated. Now the Hoeffding bound of $P\{\bar{R}^* - \bar{R} \geq c\}$ can be found by replacing ρ by $\bar{R}$ in the right hand side of (10), setting the equation equal to α, and solving for c. The $P\{\bar{R}^* - \bar{R} \geq c\}$ approximates $Pr\{\bar{R} - \rho \geq c\}$ and so L_ρ can be approximated with $L^*_\rho = \bar{R} - c$. The upper bound, $u^*_{1-\rho}$ on the proportion of errors in the population can be found by:

$$U^*_{1-\rho} = 1 - L^*_\rho = 1 - \bar{R} + c \tag{11}$$

An upper bound for the absolute number of errors in the population can be found by multiplying the number of items in the population by this proportion. Once this upper estimate for total error (number of miscounted votes) is obtained, further auditing is triggered if $(U^*_{1-\rho})(V) > \frac{m}{2}$ where V=total number of votes cast.

There are instances in which it is not possible to solve (10) for c. This occurs when $\bar{R}$ is close to 1 because there are very few miscounts found in the audit. If the Hoeffding equation is unsolvable for the observed $\bar{R}$, then the c that is associated with the largest $\bar{R}$ that is solvable for that audit size n is used. Once c is obtained, (11) is solved, using the observed $\bar{R}$, and $(U^*_{1-\rho})(V)$ is compared to $\frac{m}{2}$ as before.

In this section, two different triggers for further auditing have been developed. They are both based on techniques to obtain an estimate of T that can be compared to m in order to make a decision whether to conduct additional auditing. The modified bootstrap builds on the original bootstrap trigger described in 4.1. It incorporates the possibility that the audit fails to detect large miscounts and results in a more conservative estimate for the 99^{th} percentile of $\hat{T}$ that hopefully will maintain power in all situations. The second trigger, the Hoeffding trigger, uses Clayton's (1994) modification to the Hoeffding bound to find an upper bound for the error in the electronic vote tallies. The number of votes counted in error for the winner is equal to the net gain in votes for the losing candidate and so the Hoeffding trigger can be compared to $\frac{m}{2}$ to make a decision on further auditing. A key difference between these two triggers is that the Hoeffding bound as implemented here cannot incorporate understatements. That is, if the winner actually received more votes in a precinct than was originally observed, the method would record that 100% of the votes for the observed winner were correctly counted. This means that the modified Hoeffding bound is conservative in the sense that when the observed winner loses votes in the audit it is noted but when the observed winner gains votes in the audit it is not recorded. The bootstrap method can incorporate both understatements and overstatements. Next, section (5) describes a simulation study conducted to compare the performance of these two triggers.

5 Simulation

The creation of a computer simulation was necessary in order to evaluate the triggers. It was not possible to use data from previous elections because there is no way to determine if the correct winner was actually declared. Without knowing if the correct or incorrect winner has been declared it is impossible to know whether the trigger should call for further auditing or not. By creating a computer simulation, each trigger can be tested in situations when it is known that the correct winner was declared as well as scenarios created in which the incorrect winner is declared. This allows for the calculation of statistical power and false positive rates of the proposed triggers. In this section, the simulation design is described and a brief analysis of the results is provided.

5.1 Description

The simulation examines a variety of election scenarios to ensure the generality of the results. The basic idea behind the simulation is similar to the one described in section 4, but some of the situations that are tested are different. There are 4000 precincts each with 500 votes, and this is unchanged throughout all simulation situations. Same size precincts were used because, as mentioned previously, the Hoeffding bound cannot be used with different sized precincts, and the author has no reason to believe that the modified bootstrap trigger would perform differently if the precincts were various sizes. Next, a margin of victory was chosen for the race and vote tallies were assigned to the two candidates accordingly. If the margin were 53% for example, each of the 4000 precincts would have 53% of its votes go to one candidate and 47% to the other. These represent the "true" vote counts. Then miscounts are introduced into the election by switching votes in a number of precincts. The percentage of votes that were switched and the number of precincts that were affected varied throughout the simulation. The exact values that were examined are discussed below. The vote tallies after the miscounts are referred to as the "observed" vote counts. These represent what the election officials see from the electronic machines and in reality the "true" vote tallies would be unknown without a complete recount. Next, the power-based method developed in section 2.2.1 was used to determine the sample size of the audit. The observed margin is used when calculating the audit sample size because that is how it would be determined in a real election. It is assumed that the maximum percentage of votes shifted in a precinct is 20%, which follows closely with the other literature. For the audited precincts, the observed vote tallies can be compared with the true counts. From these observations, the modified bootstrap and Hoeffding trigger are implemented as described in section 4 to see whether the triggers call for further auditing based on the data from the audit. Finally, the performance of each trigger is evaluated by noting how often it called for auditing when the incorrect winner was declared (statistical power) and how frequently it called for further auditing when the correct winner was observed (false positive rate).

The specific situations that were observed and their results are shown in Tables 2-7 of the appendix. For each situation, the simulation was run 1000 times and the performance of the modified bootstrap trigger and the Hoeffding trigger was recorded. First, the false positive rates of the triggers were examined. Their performance was examined for margins equal to 2%, 6%, 10%, and 14%, with k equal to 20, 50, and 100, and miscounted by s values of 1%, 3%, 5%, and 20%. For these situations, any miscounted precincts were miscounted by the same specified percentage. For example, when the margin was 10% and 20 precincts were miscounted by 3%, all 20 had 3% of votes shifted. The shifted votes were taken from the true loser and given to the true winner, artificially increasing the observed margin. In all of these situations, the correct winner was observed. The number of times the trigger called for further auditing when it was unnecessary was recorded.

The simulation also tested the power of the triggers by looking at situations when the incorrect winner was declared. The first category of situations that was examined were extremely close races when the true margin was 1 vote. After votes were switched from the true winner to the true loser, the observed winner was incorrect. The simulation was designed to examine close races because then the observed margin (m) will be extremely close to T. When the incorrect winner has been declared, T is equal to the sum of the true margin and the observed margin. This is because the total number of votes that were shifted includes those that erased the original margin of victory as well as those that increased the observed margin in favor of the incorrect winner. Given a particular observed margin (m), T will be closer to m in races when the true margin is small. If the triggers maintain high statistical power when T is close to m and the true margin is small, then they will also provide high power for situations when m is much smaller than T (i.e. the true margin is larger). In this sense, the simulation tests the most difficult situation in terms of power to perform a rigorous evaluation of the triggers. The simulation included scenarios with k values of 20, 50 and 100 and s values of 1%, 3%, 5%, and 20%. These results are presented in Table 6, and as in the false positive testing situations, all miscounted precincts are miscounted by that specified percentage. It is also important to determine the power for the two triggers under a more complicated miscount pattern like the one described in section 4. A miscount pattern where 90% of the miscounted precincts had 3% miscounts and the other 10% were miscounted by 20% was examined. The original bootstrap trigger failed to maintain high power in situations with these types of miscounts so it is especially important to ensure that the modified bootstrap and Hoeffding triggers perform better in the same scenarios. This miscount pattern was examined for k equal to 20, 50, and 100 miscounted precincts and the results are presented in Table 7.

The simulation parameter values that were tested represent a variety of different scenarios. The false positive rates were examined for various margins and miscount levels. In the test of statistical power, races when the incorrect winner was declared are created, but the true margin of the race was made to be very small. This represents a "worst-case scenario" and so if the triggers perform well throughout the simulation it is expected that they will maintain power in almost all other situations.

5.2 Results

Throughout the majority of situations both triggers performed well, however, overall the Hoeffding trigger seems to have lower false positive rates and higher power. The results of the simulation are in Tables 2-7 in the appendix. Table 2 shows that the Hoeffding trigger has a 100% false positive rate when the margin is 2%. As problematic as this seems at first glance, there is a tradeoff between power and false positive rates and so a trigger that provides high power may not always be able to have a low false positive rate. The Hoeffding trigger does maintain a very low false positive rate for margins higher than 2% as seen in Tables 3, 4 and 5. In these situations, the Hoeffding trigger far outperforms the modified bootstrap trigger in terms of the false positive rate. Interestingly, the Hoeffding bound goes from a 100% false positive rate to nearly a 0% false positive rate when the true margin of the race moves from 51% to 53%.

The modified bootstrap trigger, on the other hand, generally has an increasing false positive rate as the margin of the race increases. For example, when 50 precincts are miscounted at 3% with a 14% margin the false positive rate for the modified bootstrap trigger is 0.413. The same situation with a 10% margin has a false positive rate of 0.021. This seems counter-intuitive at first because for races with larger margins, the estimate of T would have to be much higher than the true value of T in order to call for further auditing when the correct winner was declared. As the true margin and consequently the observed margin increases, however, the sample size decreases. When the true margin is 14%, the sample size is only about 10 precincts. Examining the construction of the modified bootstrap offers some understanding of what is causing this increasing false positive rate. The modified bootstrap adds a term to the observations Y_1, Y_2, ... Y_n represents that a miscount of 20% in the largest precinct. As the sample size of the audit decreases, that added term is relatively more important. Since the term represents a very large miscount, it may be the reason for the high value of T^*_{990} and, therefore, the cause of the high false positive rates observed in Table 5 of the appendix.

Both the Hoeffding and modified bootstrap triggers provide high power in all instances tested. The Hoeffding bound maintains 100% power, whereas the modified bootstrap method falls to 99% in some cases. For the case in Table 6, when the modified bootstrap trigger falls slightly below 99%, the same situation was rerun 10,000 times and the power was 99.2% providing evidence that overall the modified bootstrap trigger maintains 99% statistical power for all of the situations that tested.

The simulation results provide support for both the modified bootstrap and Hoeffding trigger. Both approaches maintain high power throughout the "worst-case scenario" situations in the simulation, which is the primary concern for a trigger. The Hoeffding bound trigger tends to have a lower false positive in the majority of the scenarios. It does, however, have a 100% false positive rate for races when the true margin is 2%, so it may call for unnecessary auditing more often when races are close. Overall, the results suggest that either trigger could be used to determine whether to continue auditing. The slightly higher statistical power and generally lower false positive rate of the Hoeffding trigger will have to be weighed against the much lower false positive rate of the modified bootstrap for relatively close races. It is possible that auditing more for close races, even when unnecessary is not problematic because the public will be most skeptical of those races. However, if the majority of races are close the Hoeffding trigger may require significantly more resources than the modified boostrap. It is promising, nonetheless, that there are two possibilities to choose from for triggers for further auditing. This secondary decision in the audit process is no less important that the preliminary determination of the audit sample size and so the selection of an appropriate trigger will be important when designing post-election audit procedures. The next section will describe Minnesota's first ever post-election audit conducted in 2006. This audit selected a fixed percentage of precincts to audit and did not use the power-based method or any of the triggers discussed here. It is interesting, nevertheless, to compare the audit that was conducted with what would have been required if the power-based method had been used to decide on the audit sample size.

6 Analysis of recent Minnesota elections

Minnesota conducted its first post-election audit in 2006. Under the current law, 2-4 precincts are audited in each county based on population size. The precincts are randomly selected and, in 2006, this audit procedure resulted in about a 5% statewide audit level. At present, the number of precincts audited is independent of the margins of the races, although there is a recount law in place for races decided by less than a 0.5% margin[3]. Minnesota's current law essentially calls for a fixed percentage audit. This analysis seeks to evaluate how well the 2006 audit performed in achieving power, p, and determine how switching to the power-based audit would affect the demand on Minnesota's resources by comparing the size of the actual audit with the sample size, n, obtained using the power-based method.

The results of the analysis are in Table 8 of the appendix. The decision regarding the specific value of p that is acceptable should be made by election officials. Calcu-

[3] In statewide races there is a automatic full recount paid for by the state if the margin is less than 0.5% and the losing candidate can request and pay for a recount if the margin is less than 1%. In local races the candidate can request a free recount if the margin is less than 0.5%. If the margin is greater than 0.5% the candidate must pay for the recount.

lations of n using 0.95 and 0.99 as two likely values of p were included. It should be noted that the number of audited precincts required to achieve p does not increase linearly. It will require a larger increase of precincts to move from a $p = 0.98$ to $p = 0.99$ than from $p = 0.95$ to $p = 0.96$. In order to assess the performance of the 2006 audit, the power, p, associated with the number of precincts audited in that race was calculated for each race. Only three races had $p < 0.95$, the 2006 Governor's race and the races for the U.S. House of Representative in districts 1 and 6. In order to extend the analysis to the 2002 and 2004 elections, the 2006 numbers of precincts actually audited in each race were used as hypothetical values in order to calculate what p would have been achieved if Minnesota had conducted a similar audit in those years. Out of 30 races, 6 had $p < 0.95$ and 11 had $p < 0.99$. It is clear from Table 8 that there were some inefficiencies in the 2006 audit. In the U.S. Senate race, 202 precincts were audited when only 23 were necessary to achieve $p = 0.99$. These resources could have been put toward more auditing in closer races such as the Governor's race. Based on this analysis, the majority of the time Minnesota's fixed percentage audit achieved a high value of p, however, there were instances when p is unacceptably low. Using the power-based method to determine n would allocate limited resources more efficiently as well as ensure high power for every race.

As clearly demonstrated by the 2006 Governor's race, races with very close margins will require much more auditing than decisive races. This does not necessarily mean that using the power-based method to decide how many precincts to audit requires more auditing overall. Tables 9 and 10 of the appendix illustrate the required number of precincts to be audited in Minnesota's 2002 and 2004 elections using the power-based approach and the 5% fixed audit. Based on Minnesota's 2002-2006 elections, the power-based audit with $p = 0.99$ would require 390 more precincts to be audited compared to the 5% fixed audit. For $p = 0.95$, Minnesota could reduce the total number of precincts audited under the current law by 271. In reality, it may be useful to require that some minimum number of precincts be audited per race regardless of the margin (Norden 2006). This would increase the total number of precincts necessary to audit under the power-based method because races with large margins, such as the 2006 Congressional race in district 7, would require the minimum number of precincts to be audited instead of the 1 called for by the power-based method. Based on the previous three elections, however, it does not seem as though using the power-based method will require substantially more auditing in the long term.

The power-based method of selecting the audit size has recently gained more recognition. Last year, a group of Minnesota election officials and advocates considered revising the current audit law to incorporate some aspects of the power-based audit. As mentioned previously, Oregon has passed a law using a tiered approach and the New Jersey legislature passed a bill that uses the power-based method to calculate the audit size. There is growing awareness and enthusiasm for this method because it represents an improvement over the fixed percentage audit. In the fall of 2007, election officials, advocates and statisticians met for an audit summit in Minneapolis, Minnesota. This conference brought together a diverse group of people all working on aspects of post-election audits and the power-based method was explained in depth. As more research comes out on audit design and triggers for further auditing, existing legislation will be revised and improved, and states without an audit may decide to implement one. A relatively new area of audit research examines the possibility of using individual ballots as the sampling unit in an audit instead of the precinct. The next section offers an overview of this research.

7 Ballot auditing

Currently, audits are conducted at the precinct or election district level, but a possible alternative is to audit individual ballots. There are certain advantages to sampling ballots instead of precincts. Ballot auditing has the potential to save on audit costs by reducing the overall number of ballots audited, due to the efficiency gains of moving to a smaller auditing unit. An additional benefit of using the ballot as the sampling unit, is that it distributes the burden of the audit more evenly across all precincts. This section will describe ballot auditing, outline one proposed method for ballot auditing, and offer an alternative decision rule for further auditing.

In a ballot audit, the sampling units are individual votes. A certain number of ballots are selected to be audited, and the share of votes for each candidate in the audited votes are determined by a hand recount. The percentage of votes for the winner obtained from the audited ballots, v, is an estimate for the the true vote share that would be obtained if a full recount were done. In order to determine the reliability of the estimate, a confidence interval for v can be constructed. Imagine a sample of n ballots drawn without replacement from a population of N total votes. This population contains w ballots cast for the originally reported winner (type A objects) and l ballots cast for the losing candidate (type B objects). The number of w ballots that are found in the audit follows a hypergeometric distribution. A confidence interval for $v = \frac{w}{n}$, the proportion of votes in the sample that were for the winner, can be constructed using the normal approximation to the hypergeometric distribution. The larger the value of n the smaller the confidence interval will be, providing more evidence that the true proportion of votes for the winning candidate is close to the point estimate v.

Ballot auditing encounters some of the same issues as precinct auditing. There are a number of ways to determine both the size of the audit and what the trigger should be for further auditing. The value of n will likely be determined by the desired width of the confidence interval for v. As the size of a race increases, the number of votes necessary to audit to obtain a certain confidence interval width increases only slightly. For example, obtaining a confidence interval of width 2% in a race with 150,000 votes would require an audit of 14,936 ballots or roughly 10%. A race with 10,000,000 votes would require an audit of only 16,560 ballots or 0.17% to achieve the same 2% wide confidence interval (Simon and O'Dell 2006). Thus, auditing a fixed percentage of votes would suffer from the same inefficiencies with respect to the size of the race as auditing a fixed percentage of precincts. Simon and O'Dell (2006) advocate a 10% audit for all races, for the sake of simplicity. They argue that the set percentage makes it easy to choose the sample and ensures that for any U.S. House or Senate race, the estimate will be within 1% of the true value at least 99% of the time. Ten percent is sufficiently large to be certain that there will be almost no under-auditing in small races, with respect to attaining a confidence interval of a certain width, but their method would result in a large amount of over-auditing for big races. A uniform 10% audit represents a large increase in auditing when compared to other audit proposals and may not be feasible due to resource constraints. An alternative to the fixed percentage approach is to choose a fixed number for n, the number of ballots audited. Since the total number of ballots to audit does not change much between race sizes, it may be useful to always audit 16,560 votes, say, for each race. For small federal races of 150,000 votes, the audit would be 1,500 votes larger than necessary, but the audit would never be hundreds of thousands of votes larger than required as a 10% audit may be in very large races. In addition, auditing 16,560 ballots will attain a 99% confidence interval of width 2% or less for all races.

The actual selection of the ballots within precincts could be done using a random number generator. The generator could select a random number from 1-10 and starting at that number ballot in the stack the precinct could audit every tenth ballot. When implementing this method, it would be important to ensure that the ballots from each precinct are still randomly selected. Regardless of how the sampling is done Simon and O'Dell's method does not require unique identification numbers for each ballot so it could be implemented using the current voting system. This is important because many individuals view ballot identification numbers as a violation of privacy.

If, in the future, a method is developed to individually number the ballots without violating anonymity, ballots could be randomly selected for each race at the state level. Each precinct would receive a list of ballots to audit and then would report their totals for each race to the state. Next, the state official would construct confidence intervals for the winners' vote shares based on the sample sizes and results of the preliminary audit.

Once n has been determined and ballots selected, a trigger for further auditing can be developed based on the confidence interval for v. Simon and O'Dell (2006) compare this confidence interval to the winner's originally reported vote tally. If the interval encompasses the original value, then the audit is concluded, otherwise additional auditing is necessary. With respect to the goal of winner verification, there are some situations in which Simon and O'Dell's method calls for unnecessary auditing and others when it fails to provide confidence in the results. For example, if the originally reported tally is 70% to 30% and the 99% confidence interval for v obtained from the audit was 67%-69% their method would call for further auditing even though the winner of the race was almost certainly correct the first time. A more serious problem may occur when the race is close. Suppose, for example, that the originally reported vote counts were 51.5% to 48.5%, and the confidence interval for v obtained from the audit was 49.7%-51.7%. Under Simon and O'Dell's method the election results should be verified and the audit concluded because 51.5% is inside the confidence interval. Since this interval includes 50%, however, it does not provide confidence that the correct winner was declared and further auditing should be conducted. Obviously this sort of problem will arise in close races, but in the previous example the margin was 3%, which is not considered a very close race. A margin of 3% falls far above the threshold of most automatic recount laws. For the majority of races, Simon and O'Dell's method will be adequate to verify the winner, but there are instances in which the audit is terminated prematurely.

A possible alternative trigger could compare the confidence interval for v to 50% rather than to the originally reported results. If the interval is completely above 50%, then the originally declared winner was correct and the election results should be certified, otherwise additional auditing should be conducted until the 99% confidence interval for v is either completely above 50% or a full recount has been performed. This trigger directly gets at winner verification because it only looks at whether the vote share for the winner is above 50% and is unconcerned with how far v is from the original tallies. Recall the hypothesis test (7) mentioned in section (3). Comparing a confidence interval to 50% is essentially equivalent to testing the following hypothesis:

$$H_0 : P_w \geq 0.5 \qquad vs. \qquad H_a : \; P_w < 0.5 \tag{12}$$

$$where \quad P_w \;=\; true\ proportion\ of\ votes\ for\ the\ originally\ reported\ winner$$

Constructing a 99% confidence interval for the proportion of votes that the winner obtained implies that it will be completely above 50% when H_a is true at most 1% of the time. That is, the power for the this method regarding the hypothesis of interest is at least 99%. The causes of discrepancies between v and the original vote shares should be investigated for the purpose of improving the voting system, but this can be done after the race is certified. This proposed trigger will always call for further auditing for instances when the confidence interval includes 50%, and it expedites the process of certifying the election by not requiring further auditing when unnecessary.

8 Conclusion

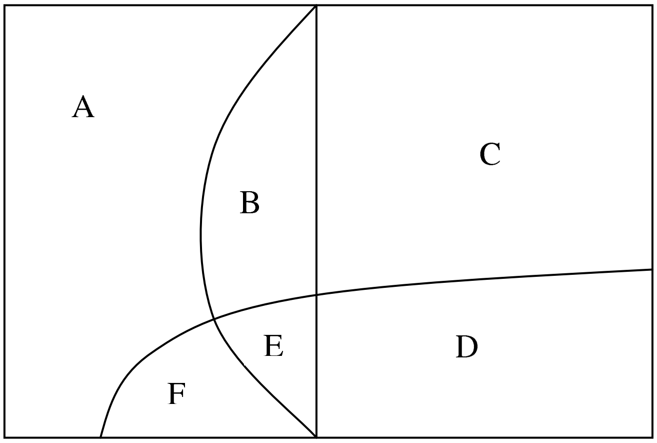
This paper addresses the incorporation of statistical principles into the design of a post-election audit. Specifically, it focuses on choosing a sample size n for the audit and developing a systematic method to determine when the audit should be expanded. Currently, most states choose the sample size by selecting a fixed percentage of precincts. This method has been shown to have low power in certain situations, and an alternative approach has been developed. This new power-based method uses the level of power as a criteria for selecting n, and, therefore, always ensures high power. This paper applies the new method to the 2002, 2004, and 2006 Minnesota elections to compare the power-based audit sample sizes with the ones required under Minnesota's current audit law. Overall, it does not seem as though the power-based audit would require significantly more resources than the fixed percentage audit, however, the required funding would be more uncertain and less evenly spread between election years. This paper is the first to frame an audit in terms of a hypothesis test, and the relationship between the hypothesis test actually conducted and the true test of interest is discussed. Another major contribution of this paper is the development of two possible triggers for further auditing that use the data from the n precincts audited to estimate the net gain in votes for the losing candidate should a full recount be conducted. Then, this value is compared to the margin of the race to determine whether it is necessary to expand the audit. A simulation study was conducted to evaluate the performance of both the modified bootstrap trigger and the modified Hoeffding bound trigger. Both triggers maintained high statistical power throughout a variety of situations, but overall the Hoeffding trigger provided slightly higher power and lower false positive rates. Based on these results, both triggers remain options for states to choose from when designing their audits. When developing a post-election audit, officials are very concerned with transparency. The methods discussed in this paper require some understanding of statistical practices, and election officials will have to decide whether the benefits of achieving high statistical power outweigh the simplicity of alternative procedures. In the future, audits may use individual ballots as a sampling unit instead of precincts due to efficiency gains of using a smaller sampling unit. This paper offers a discussion of the sparse literature on the topic, and suggests a trigger that compares the confidence interval for the proportion of votes for the winner to 50% instead of the originally reported tally for the winner as others have recommended.

This is an exciting time to study the design of post-election audits. Recently there has been increasing awareness of the flaws of the current practice, which has provided motivation for states to re-evaluate their own audits. Public doubt in the election system, arising from problems in previous elections, has increased the importance of the role of audits. One possibility for further research is to determine what the next step for further auditing should be. This paper discusses triggers for further auditing, but does not offer suggestions for how the additional auditing should proceed. The New Jersey bill offers one possibility; it requires the second level of auditing to be another sample of size n. This may be the most appropriate next step, but the validity of alternatives should be examined. The development of a fully specified audit procedure, that maintains high statistical power throughout every stage, will be imperative before it can be implemented into legislation.

Acknowledgments:
I would like to thank Vittorio Addona for his guidance and assistance throughout this project. I would also like to thank Mark Halvorson and Laura Wolff of Citizens for Election Integrity Minnesota for introducing me to this important topic and for their hard work advocating for audit reform.

APPENDIX

Incorrect winner was declared in A, F   Correct winner was declared in B, C, D, E



$K \geq K_{min}$ in A, B, E, F $K \leq K_{min}$-1 in C, D

Reject H_0 in A, B, D   Fail to reject H_0 in C, E, F

Figure 1: Graphic Representation of Alternative Hypothesis Tests

Area	# of miscounted precincts	Declared winner	Audit revealed miscount
A	$\geq k_{min}$	incorrect	Yes
B	$\geq k_{min}$	correct	Yes
C	$< k_{min}$	correct	No
D	$< k_{min}$	correct	Yes
E	$\geq k_{min}$	correct	No
F	$\geq k_{min}$	incorrect	No

Table 1: Explanation of Areas on Diagram

# Miscounted	**Miscount Percentage**							
	1		**3**		**5**		**20**	
20	0.000	1.000	0.005	1.000	0.081	1.000	0.341	1.000
50	0.001	1.000	0.035	1.000	0.264	1.000	0.580	1.000
100	0.001	1.000	0.154	1.000	0.549	1.000	0.789	1.000

Table 2: False Positives, Margin=2%, Bootstrap then Hoeffding

# Miscounted	**Miscount Percentage**							
	1		**3**		**5**		**20**	
20	0.000	0.000	0.001	0.000	0.055	0.000	0.141	0.000
50	0.000	0.000	0.022	0.000	0.025	0.000	0.268	0.000
100	0.001	0.000	0.01	0.000	0.085	0.000	0.460	0.013

Table 3: False Positives, Margin=6%, Bootstrap then Hoeffding

# Miscounted	**Miscount Percentage**							
	1		**3**		**5**		**20**	
20	0.012	0.000	0.011	0.000	0.008	0.000	0.071	0.002
50	0.037	0.000	0.021	0.000	0.027	0.000	0.170	0.014
100	0.070	0.000	0.066	0.000	0.073	0.000	0.282	0.041

Table 4: False Positives, Margin=10%, Bootstrap then Hoeffding

# Miscounted	**Miscount Percentage**							
	1		**3**		**5**		**20**	
20	0.0402	0.000	0.398	0.000	0.400	0.000	0.442	0.002
50	0.392	0.000	0.413	0.000	0.413	0.000	0.473	0.004
100	0.409	0.000	0.368	0.000	0.415	0.000	0.489	0.025

Table 5: False Positives, Margin=14%, Bootstrap then Hoeffding

# Miscounted	Miscount Percentage							
	1		3		5		20	
20	1.000	1.000	1.000	1.000	1.000	1.000	0.986	1.000
50	1.000	1.000	1.000	1.000	1.000	1.000	0.991	1.000
100	1.000	1.000	1.000	1.000	1.000	1.000	0.998	1.000

Table 6: Power, Bootstrap then Hoeffding

Number Miscounted	Power	
20	1.000	1.000
50	1.000	1.000
100	1.000	1.000

Table 7: Power, 10% of miscounted precincts have 20% miscounts and 90% have 3% miscounts, Bootstrap then Hoeffding

Race	Margin(%)	Audit for 99%	Audit for 95%	# Actually Audited
Governor	1	886	601	202
District 1	5.6	136	92	51
District 2	16.2	19	12	16
District 3	29.8	5	4	4
District 4	39.3	2	2	8
District 5	34.9	4	3	4
District 6	8	51	35	16
District 7	41.2	1	1	68
District 8	30	12	8	37
Governor	20.3	23	15	202

Table 8: Minnesota's 2006 Audit

Race	**Margin(%)**	**Audit for 99%**	**Audit for 95%**	**# for 5% Audit**
President	3.5	220	145	202
District 1	24.1	17	12	51
District 2	16.1	19	13	16
District 3	29.3	5	4	4
District 4	24.2	8	6	8
District 5	45.2	1	1	4
District 6	8.0	48	33	16
District 7	32.2	9	6	68
District 8	33.0	9	6	37

Table 9: Minnesota's 2004 Election

Race	**Margin(%)**	**Audit for 99%**	**Audit for 95%**	**# for 5% Audit**
U.S. Senate	2.2	367	243	202
District 1	26.9	13	9	51
District 2	11.1	30	20	16
District 3	44.1	1	1	4
District 4	28.3	6	4	8
District 5	41.1	1	1	4
District 6	22.3	12	8	16
District 7	30.6	10	7	68
District 8	37.4	5	4	37
Governor	7.9	84	55	202

Table 10: Minnesota's 2002 Election

References

[1] Aslam, Javed A., Raluca A. Popa and Ronald L. Rivest. "On Auditing Elections When Precincts Have Different Sizes." January, 2008. www.mit.edu/˜ralucap/AslamPopaRivest-OnEstimatingTheSizeAndConfidenceOfAStatisticalAudit.pdf.

[2] Bentkus, V. and M. van Zuijlen. "On Conservative Confidence Intervals." *Lithuanian Mathematical Journal.* April, 2003. 43(2). 141-160.

[3] Bickel, Peter J. "Inference and Auditing: The Stringer Bound." *International Statistical Review.* August, 1992. 60(2). 197-209.

[4] Clayton, Howard R. "A Combined Bound for Errors in Auditing on Hoeffding's Inequality and the Bootstrap." *Journal of Business and Economic Statistics.* October, 1994. 12(4). 437-448.

[5] Cordero, Arel, David Wagner, and David Dill. "The Role of Dice in Election Audits Extended Abstract." June, 2006. www.cs.berkeley.edu/˜daw/papers/dice-wote06.pdf.

[6] Dopp, Kathy and Frank Stenger. "The Election Integrity Audit." National Election Data Archive. September, 2006. http://vote.nist.gov/ElectionIntegrityAudit.pdf.

[7] New Jersey Senate No. 507. Approved January 14, 2008. http://www.njleg.state.nj.us/bills/BillView.asp.

[8] Hoeffding, Wassily. "Probability Inequalities for Sums of Bounded Random Variables." *Journal of the American Statistical Association.* March, 1963. 58(301). 13-30.

[9] Holt, Rush. Proposed bill H.R. 811 before U.S. House of Representatives 2007. http://thomas.loc.gov/cgi-bin/query/z?c110:H.R.811.RH:.

[10] Horvitz, D. G. and D. J. Thompson. "A Generalization of Sampling Without Replacement From a Finite Universe." *Journal of the American Statistical Association.* December, 1952. 47(260). 663-685.

[11] Lobdill, Jerry. "Election Audit Sampling Plan Design-It's Not Just About Sampling Without Replacement." October, 2006. vote.nist.gov/Considering-Vote-Count-Distribution-in-Designing-Election-Audits-Rev-2-11-26-06.pdf.

[12] McCarthy, John, Howard Stanislevic, Mark Lindeman, Arlene Ash, Vittorio Addona, and Mary Batcher. "Precentage-Based versus Power-Based Vote Tabulation Statistical Audits." *The American Statistician.* February, 2008. 62(1). 11-16.

[13] Norden, Lawrence. "The Machinery of Democracy: Voting System Security, Accessibility, Usability, and Cost." The Brennan Center for Justice. October, 2006. http://www.brennancenter.org/content/resource/the_machinery_of_democracy _voting_system_security_accessibility_usability_a.

[14] Oregon House Bill 3270. Effective January 1, 2008. http://www.leg.state.or.us/cgi-bin/searchMeas.pl.

[15] Rivest, Ronald L. "On Estimating the Size of a Statistical Audit." MIT Computer Science and Aritificial Intelligence Labo-

ratory. November, 2006. http://people.csail.mit.edu/rivest/Rivest-OnEstimatingTheSizeOfAStatisticalAudit.pdf.

[16] Saltman, Roy G. "Effective Use of Computing Technology in Vote-tallying." National Bureau of Standards. March, 1975. csrc.nist.gov/publications/nistpubs/NBS_SP_500-30.pdf.

[17] Stanislevic, Howard. "Random Auditing of E-Voting Systems: How Much is Enough?" Vote TrustUSA. August, 2006. www.votetrustusa.org/pdfs/VTTF/EVEPAuditing.pdf.

[18] Simon, Jonathan D. and Bruce O'Dell. "An End To "Faith-Based" Voting: Universal Precinct-based Handcount Sampling To Check Computerized Vote Counts In Federal and Statewide Elections." Election Defense Alliance. September, 2006. http://www.electiondefensealliance.org/files/New_UBS_811Update_061707.pdf.