Super-Simple Simultaneous Single-Ballot Risk-Limiting Audits

Philip B. Stark
Department of Statistics
University of California, Berkeley

Abstract

Simultaneous risk-limiting audits of a collection of contests have a known minimum chance of leading to a full hand count if the outcome of any of those contests is wrong. Risk-limiting audits are generally performed in stages. Each stage involves drawing a sample of ballots, comparing a hand count of the votes on those ballots with the original count, and assessing the evidence that the original outcomes agree with the outcomes that a full hand count would show. If the evidence is sufficiently strong, the audit can stop; if not, more ballots are counted by hand and the new evidence is assessed. This paper derives simple rules to determine how many ballots must be audited to allow a simultaneous risk-limiting audit to stop at the first stage if the error rate in the sample is sufficiently low. The rules are of the form "audit at least ρ/μ ballots selected at random." The value of ρ depends on the simultaneous risk limit and the amount of error to be tolerated in the first stage without expanding the audit. It can be calculated once and for all without knowing anything about the contests. The number μ is the "diluted margin": the smallest margin of victory in votes among the contests, divided by the total number of ballots cast across all the contests. The initial sample size does not depend on any details of the contests, just the diluted margin. This is far simpler than previous methods.

For instance, suppose we are auditing a collection of contests at simultaneous risk limit 10%. In all, N ballots were cast in those contests. The smallest margin is V votes: The diluted margin is $\mu = V/N$. We want the audit to stop at the first stage provided the fraction of ballots in the sample that overstated the margin of some winner over some loser by one vote is no more than $\mu/2$ and no ballot overstates any margin by two votes. Then an initial sample of $15.2/\mu$ ballots suffices. If the sample shows any two-vote overstatements or more than 7 ballots with one-vote overstatements, more sampling might be required, depending on which margins have errors. If so, simple rules that involving only addition, subtraction, multiplication, and division can be used to determine when to stop.

1 Introduction

This paper presents some extremely simple methods for conducting the first stage of risk-limiting audits of a collection of contests. The methods allow most contests in an election to be confirmed with a *single* audit sample of fewer than 1,000 ballots, at a low risk that any of the apparent outcomes differs from the outcome a full hand count would show—unless the audit finds many errors that caused an apparent margin to appear larger than a hand-count margin.

The *outcome* of a contest is the set of winners, not the exact vote totals. The outcome of a collection of contests is the set of winners of all the contests. The *machine-count outcome* or *apparent outcome* is the outcome that will become officially final unless an audit or other action intervenes. The *hand-count outcome* or *true outcome* is the outcome that a full manual tally of the audit trail would show. Generally, as a matter of legal definition, the hand-count outcome is correct—even though hand counting is not perfect, and even though the audit trail might not be complete and accurate, so the outcome a hand count shows might not reflect the will of the voters.

A *risk-limiting audit* has a guaranteed minimum chance of progressing to a full hand count if the apparent outcome is incorrect [7, 8, 10, 12, 9, 11, 6], thereby correcting the apparent outcome. The *risk* is the maximum chance that the audit fails to correct an apparent outcome that is incorrect, no matter what caused the outcome to be incorrect. Risk-limiting audits generally count votes by hand until there is strong evidence that the reported outcome is correct, or until all the votes have been counted by hand and the correct outcome is known.

Risk-limiting audits have been endorsed by the American Statistical Association [14] and a number of election integrity groups [4].

A *simultaneous risk-limiting audit* of a collection of contests has a guaranteed minimum chance of progressing to a full hand count of *all* of the contests that have incorrect apparent outcomes. [9, 11]. The *simultaneous risk* of a simultaneous risk-limiting audit is the maximum chance that the audit will fail to correct one or more of the apparent outcomes that are incorrect, no matter what caused them to be incorrect.

A *risk-measuring audit* is one that reports the strength of the evidence that the outcome is correct, but does not necessarily continue to count votes until that evidence is strong or all votes have been counted by hand. In statistical language, the *measured risk* is the P-value of the hypothesis that the outcome is incorrect, given the data collected by the audit [12].

Stark and his collaborators have developed several methods for risk-limiting and risk-measuring audits and applied those methods to audit six election contests in California [3, 5, 7, 8, 9, 10, 11, 12]. This paper develops a special case of methods in [12, 9, 11] to give extremely simple rules to calculate how large a sample to draw initially so that the audit can stop without additional counting provided the number of ballots in the sample with errors that overstate a margin by one vote is not too large, and no ballot in the sample overstates any margin by two votes. If there are too many errors in the sample, to control the simultaneous risk will require expanding the sample, possibly to a full hand count; formulae in [9, 11] (reproduced below) determine when sampling can stop.

Among the benefits of the method presented here are:

1. The entire collection of contests is audited at once, rather than having to draw separate samples for each contest under audit. This decreases logistical complexity. Moreover, the simultaneous risk is limited for the set of contests.

2. If a ballot is selected for audit, every contest on that ballot is audited. This decreases the number of pieces of paper that must be handled.

3. The rule for selecting the initial sample size is extremely simple: divide a constant by the "diluted margin." Computing the constant involves taking logarithms, but it only needs to be computed once. It does not depend on the particulars of the contests, their margins, or the audit results.

4. The conditions under which the audit progresses beyond the first stage are simple and make sense intuitively: too many ballots with errors that overstate a margin by one vote, or any ballots that overstate a margin by two votes.

5. If the audit does have to progress beyond the first stage, the calculations to determine when to stop are simple.

6. The audit really limits the simultaneous risk: The chance of a full hand count if *any* of the outcomes is wrong is guaranteed to be at least as high as claimed.

The methods presented here trade simplicity for efficiency: There are methods that can limit risk by counting fewer ballots when the apparent outcomes are correct (e.g., [9, 11, 2]), but the calculations are more complicated. The methods presented here are derived from more efficient methods by applying a series of simplifying approximations that guarantee that there is a known large chance of correcting any incorrect apparent outcomes—the approximations are conservative.

Despite the inefficiency, very few ballots need to be audited to limit the simultaneous risk when the apparent outcome is in fact correct. (When one or more apparent outcomes are incorrect, the goal is to count *all* the ballots in those contests by hand to correct the apparent outcomes.) That is because the audit sample is a simple random sample of ballots, rather than a sample of precincts, for instance. For a heuristic explanation of the statistical advantage of sampling individual ballots rather than clusters of ballots, see [13].

The approach taken here involves comparing the machine interpretation of an individual ballot (cast vote record, CVR) with a human interpretation of the same ballot, for a random sample of individual ballots. Current federally certified vote tabulation systems do not make it easy to see how the machine interpreted any particular cast ballot, but this sort of "single-ballot auditing" has been performed in a small contest [11]. There are ballot scanning and vote tabulation systems offered by the Humboldt Transparency Project, Clear Ballot Group, and TrueBallot that make it easy to associate CVRs with individual physical ballots. The next generation of official vote tabulation systems could be designed to make such single-ballot auditing trivial.

2 Terminology and Conventions

When the CVR and human reading of a ballot differ, by definition, the human reading is correct, even if the difference results from voter error. For instance, a voter might use an inappropriate pen, make an inadequate mark, mark outside the target area, or mark the ballot for a listed candidate and also vote for that candidate as a write-in.

An *apparent winner* of a contest is a candidate who won according to the apparent outcome. The other candidates are *apparent losers*. (To keep the language simple, a position on a measure, such as "yes" or "no," will be called a candidate and referred to as if it were a person. The math is the same, but the margin needs to be computed differently for measures that require a supermajority. See [7]. We do not consider instant-runoff voting (IRV) or other preference voting schemes.) A *true winner* is a candidate who would be declared a winner on the basis of a full hand count of the audit trail, if there were a full hand count. The other candidates are *true losers*. Within each contest, the machine count of the votes for each apparent winner is greater than the machine count for each apparent loser, by the *apparent margin* between those two candidates. Errors do not necessarily affect any margin. For instance, if there are two light marks in a vote-for-one contest, the CVR might show that as an undervote while a human might see it as an overvote. The interpretations differ, but the difference does not change any of the margins, so it cannot cause the apparent outcome to differ from the true outcome.

An error that increases an apparent margin is an *overstatement*. For instance, if a mark that the machine counted as an undervote is interpreted by a human as a vote for an apparent loser, that is an overstatement of one vote. Similarly, if the machine interprets a hesitation mark as an overvote and a human reader interprets it as a vote for an apparent loser, that is a one-vote overstatement. An error that decreases an apparent margin is an *understatement* (or a *negative overstatement*). If the CVR shows an overvote where a human would see a vote for an apparent winner, that is a one-vote understatement.

A single ballot can understate or overstate one or more margins by up to two votes in each contest. For instance, if the CVR shows a vote for an apparent winner while a human would see a vote for an apparent loser, that is a two-vote overstatement. Such errors are expected to be quite rare. Generally, a two-vote overstatement indicates a programming error (such as a ballot definition error), fraud, or other serious problem. If the audit finds a two-vote overstatement, additional hand counting might well be justified even if Statistics does not require it.

The apparent outcome of a given contest is correct if, for all contests, a hand count would show that every apparent winner of that contest got more votes than every apparent loser of that contest. If, for some apparent winner and some apparent loser, the apparent margin is less than the overstatement errors minus the understatement errors, summed over all the ballots in the contest, the apparent outcome of that contest is wrong. Conversely, if, for every winner and loser, the overstatement errors minus the understatement errors amount to less than 100% of the margin between that pair of candidates, all the apparent outcomes are correct.

The MACRO (maximum across-race relative overstatement) [9, 11] combines the overstatement errors within contests and across different contests into a single summary. To compute the MACRO for a single ballot, first divide each overstatement error on the ballot by the reported margin (in votes) that it affects. That gives a number no bigger than 100% for each margin—each (winner, loser) pair in each contest on the ballot. The MACRO is the largest of those numbers. Only the largest number counts, even if more than one contest or more than one margin in a contest has an error. If the sum of the MACRO over all the ballots in all the contests is less than 100%, the apparent outcomes of all the contests must be correct.

The methods presented here use a simplified version of MACRO: Instead of dividing each overstatement error on the ballot by the margin it affects, it divides each overstatement error by the smallest of the margins in any of the contests. That amounts to pretending that every margin is equal to the smallest margin, which errs on the side of safety. It makes the true simultaneous risk smaller than the nominal simultaneous risk limit.

To make the MACRO concrete, suppose that there are five contests under audit. Not all ballots contain all five contests—some of the contests are jurisdiction-wide and some are smaller. We consider two hypothetical ballots. The first ballot, summarized in table 1, includes three of those contests. The CVR for that ballot shows an undervote for the first contest, a vote for one of the apparent winners of the second contest, and a vote for one of the apparent losers of the third contest. A human interprets the marks as a vote for one of the apparent losers of the first contest, a vote for one of the apparent losers of the second contest, and a vote for one of the apparent winners of the third contest. Then there was a one-vote overstatement in the first contest, a two-vote overstatement in the second contest, and a two-vote understatement in the third contest. There are three errors, but the maximum overstatement is two votes.

The second ballot, described in table 2, includes four of those contests. The CVR for that ballot shows an undervote for the first contest, a vote for one of the apparent winners of the second contest, a vote for one of the apparent losers of the third contest, and a vote for one of the apparent winners of the fourth contest. A human interprets the marks as a vote for one of the apparent losers of the first contest, an overvote in the second contest, a vote for the same apparent loser of the third contest as the CVR, and a vote for the same apparent winner of the fourth contest as the CVR. Then there was a one-vote overstatement in the first contest, a one-vote overstatement in the second contest, and a zero-vote overstatement in the third and fourth contests. There are two er-

	contest				
	1	2	3	4	5
CVR	undervote	winner	loser	not on ballot	not on ballot
Hand	loser	loser	winner	not on ballot	not on ballot
overstatement	1	2	-2	0	0

Table 1: Hypothetical CVR and hand interpretation of a ballot that contains three of five contests under audit. "Winner" and "loser" denote an apparent winner and an apparent loser, respectively. The maximum overstatement is two votes.

rors of a single vote: the maximum overstatement is one vote.

The *diluted margin* μ is the smallest margin in votes among the contests under audit, divided by the total number of ballots cast across all the contests under audit. So, for example, if we are auditing five contests in a jurisdiction where 100,000 ballots were cast in all, and the smallest margin among those five contests is 2,000 votes, the diluted margin is $\mu = (2,000/100,000) \times 100\% = 2\%$.[1] The diluted margin plays an important role in the new procedure: The sample size for the first stage is inversely proportional to the diluted margin.

One version of the super-simple simultaneous audit works as follows. It requires picking three numbers: the simultaneous risk limit α, the "error inflation factor" $\gamma \geq 100\%$, and the "error tolerance" $\lambda < 100\%$, all of which are described below. The simultaneous risk limit α might be set in legislation. The values of γ and λ are operational choices that affect efficiency but not risk.

1. Pick the simultaneous risk limit α, e.g., 10%. This is the largest chance that an incorrect outcome will not be corrected by the audit.

2. Pick an error inflation factor $\gamma \geq 100\%$. Any value of γ greater than or equal to 100% works, but γ controls a tradeoff between initial sample size and the amount of additional counting required when the sample finds too many overstatements, especially two-vote overstatements. If $\gamma = 100\%$, a two-vote overstatement may trigger a full hand count (depending on which margin is overstated by two votes). If $\gamma > 100\%$, a two-vote overstatement in the sample generally will require more hand counting, but not necessarily a full hand count. The larger γ is, the larger the initial sample needs to be, but the less additional counting will be required if the sample finds a two-vote overstatement or a large number of one-vote maximum overstatements. For concreteness, take $\gamma = 110\%$.

3. Pick a tolerance $\lambda < 100\%$ for one-vote maximum overstatements in the initial sample as a percentage of the diluted margin μ. If the percentage of ballots in the sample with of one-vote maximum overstatements is no more than $\lambda\mu$ and no ballot in the sample has a two-vote overstatement, the audit can stop. For instance, if we take $\lambda = 50\%$ and the diluted margin is 2%, the audit will be able to stop at the first stage if, in the initial sample, the percentage of ballots that have one-vote maximum overstatements is not more than $\lambda\mu = 50\% \times 2\% = 1\%$, and no ballots in the sample have two-vote overstatements. The larger λ is, the larger the initial sample size will have to be to give high confidence that even though the error rate in the sample is a large fraction of the diluted margin, the error rate for the contests as a whole still is less than the diluted margin.

4. Calculate the sample-size multiplier ρ, which depends on α, γ, and λ through the formula

$$\rho = \frac{-\log \alpha}{\frac{1}{2\gamma} + \lambda \log(1 - \frac{1}{2\gamma})}.$$

For $\alpha = 10\%$, $\gamma = 110\%$ and $\lambda = 50\%$, the value of ρ is 15.2. However they are set, the values of α, γ and λ, determine ρ once and for all, so even though the formula for ρ looks complicated and involves logarithms, it only needs to be computed once, before the audit starts. It does not depend on the margins, the number or sizes of the individual contests, or on the audit data.

5. Find the diluted margin μ.

6. Draw at least ρ/μ ballots at random and audit them. If the percentage of ballots in the sample with one-vote maximum overstatements is not more than $\lambda\mu$ and no ballot in the sample has a two-vote overstatement, the audit can stop: All contests are confirmed at simultaneous risk no greater than α. In the example, the diluted margin is 2% and $\rho = 15.2$, so we would audit a random sample of $15.2/2\% = 760$ ballots. If fewer than 8 of those ($\lambda\mu = 1\%$; 1% of 760 is 7.6) have a maximum one-vote overstatement and none has a two-vote overstatement, we can stop. Otherwise, the sample might need to be

	contest				
	1	2	3	4	5
CVR	undervote	winner	loser	winner	not on ballot
Hand	loser	overvote	loser	winner	not on ballot
overstatement	1	1	0	0	0

Table 2: Hypothetical CVR and hand interpretation of a ballot that contains four of five contests under audit. "Winner" and "loser" denote an apparent winner and an apparent loser, respectively. In contest 3, the CVR and hand count found votes for one and the same apparent loser, and in contest 4, the CVR and hand count found votes for one and the same apparent winner. There are two overstatement errors, but the maximum overstatement is one vote.

expanded, potentially to a full hand count. The methods in [9, 11] determine how much additional auditing is required; simple formulae are given below in equations 9 and 10.

3 The Math

We combine the Kaplan-Markov method and the MACRO test statistic of [9, 11, 12] with worst-case upper bounds on the effect that error in the interpretation of any individual ballot can have on any of the reported margins.

We generally follow the notation of [12, 9, 11]. There are C contests under audit; N ballots were cast in all. There might not be any contest that appears on all N ballots. Contest c appears on N_c of the N cast ballots. The numbers N and $\{N_c\}_{c=1}^{C}$ are known. Let $\mathcal{W}_c$ denote the set of reported winners of contest c and let $\mathcal{L}_c$ denote the set of reported losers of contest c. Let $v_{pi} \in \{0, 1\}$ denote the reported votes for candidate i on ballot p, and let $a_{pi} \in \{0, 1\}$ denote the actual votes for candidate i on ballot p, that is, the vote as a human auditor would interpret the ballot. If contest c does not appear on ballot p then $v_{pi} = a_{pi} = 0$.

The reported margin of reported winner $w \in \mathcal{W}_c$ over reported loser $\ell \in \mathcal{L}_c$ in contest c is

$$V_{w\ell} \equiv \sum_{p=1}^{N} (v_{pw} - v_{p\ell}) > 0. \tag{1}$$

Let V be the smallest reported margin among all C contests:

$$V \equiv \min_c \min_{w\in\mathcal{W}_c, \ell\in\mathcal{L}_c} V_{w\ell}. \tag{2}$$

The actual margin of reported winner $w \in \mathcal{W}_c$ over reported loser $\ell \in \mathcal{L}_c$ in contest c is

$$A_{w\ell} \equiv \sum_{p=1}^{N} (a_{pw} - a_{p\ell}). \tag{3}$$

The reported winners of all C contests are the actual winners of those contests if

$$\min_c \min_{w\in\mathcal{W}_c, \ell\in\mathcal{L}_c} A_{w\ell} > 0. \tag{4}$$

Otherwise, at least one reported electoral outcome is wrong.

Risk-limiting audits generally do not test directly whether inequality 4 holds. Instead, they test a condition that is sufficient but not necessary for inequality 4 to hold. The reduction to a sufficient condition produces a computationally simple test that is still conservative; i.e., the simultaneous risk remains below its nominal limit. One such reduction relies on the maximum across-contest relative overstatement (MACRO [9, 11]). The MACRO for ballot p is the largest percentage by which difference between the CVR and hand interpretation of that ballot resulted in overstating any margin in any of the c contests:

$$e_p \equiv \max_c \max_{w\in\mathcal{W}_c \ell\in\mathcal{L}_c} (v_{pw} - a_{pw} - v_{p\ell} + a_{p\ell})/V_{w\ell}. \tag{5}$$

The outcomes of all the contests must be correct if $E \equiv \sum_{p=1}^{N} e_p < 1$. Thus a risk-limiting audit can rely on testing whether $E \geq 1$.

Testing whether $E \geq 1$ would always require a very large sample if we knew nothing at all about e_p without auditing ballot p. Fortunately, there is an a priori upper bound for e_p. At worst, the CVR for ballot p shows a vote for the "least-winning" apparent winner of the contest with the smallest margin, but a hand interpretation shows a vote for the runner-up in that contest:

$$\begin{aligned} \tilde{u}_p &\equiv \max_c \max_{w\in\mathcal{W}_c \ell\in\mathcal{L}_c} (v_{pw} - v_{p\ell} + 1)/V_{w\ell} \\ &\leq \max_c \max_{w\in\mathcal{W}_c \ell\in\mathcal{L}_c} 2/V_{w\ell} \\ &\leq 2/V. \end{aligned} \tag{6}$$

Knowing that $e_p \leq \tilde{u}_p$ can make it possible to conclude reliably that $E < 1$ by examining only a small fraction of the ballots—depending on the values $\{\tilde{u}_p\}_{p=1}^{N}$ and on the values of e_p for the audited ballots.

The Kaplan-Markov method [12, 9, 11]—applied to sampling individual ballots—will not stop short of a full hand count if the ratio of e_p to its upper bound is equal to 1 for any ballot in the sample, no matter how many other ballots show no error or understatement errors. The need for a full hand count can sometimes be avoided by increasing the upper bound so that the bound cannot be attained, for instance, by inflating it by a small percentage. The simultaneous risk remains strictly controlled.

To that end, we take the error bound for each ballot to be

$$u_p \equiv \gamma 2/V > \tilde{u}_p \tag{7}$$

where the "inflator" $\gamma > 1$. That ensures that e_p/u_p cannot be larger than $1/\gamma < 1$. The cost of inflating the upper bound in this way is that a larger sample will be needed than if $\{\tilde{u}_p\}$ were used as the bounds and the sample did not happen to include any ballots with e_p equal to $\tilde{u}_p$. On the other hand, inflating the error bounds can help avoid a full count when that full count would merely confirm that all the apparent outcomes are correct. The larger the value of γ, the larger the initial sample needs to be to allow the audit to stop if at most a given number of ballots overstated one or more margins by one vote, but the less the sample will need to be expanded if ballots in the sample overstate any margin by two votes–unless a full hand count is required.

With u_p defined by equation 7, the total error bound across all N ballots is

$$U \equiv 2\gamma N/V = 2\gamma/\mu, \tag{8}$$

where μ is the diluted margin V/N. The diluted margin plays an important role in determining the sample size: The initial sample size is $1/\mu$ multiplied by a constant that depends on the desired simultaneous risk limit, the number of errors to be tolerated without expanding the audit, and the inflator γ. Note that $U > 2\gamma > 2$.

Suppose that n of the N ballots are drawn with replacement with equal probability. Let e_r be the value of the error e_p as defined in equation 5 for the rth randomly selected ballot. The Kaplan-Markov MACRO P-value is [9, 11]

$$P_{KM} = \prod_{r=1}^{n} \frac{1 - 1/U}{1 - \frac{e_r}{2\gamma/V}}. \tag{9}$$

An audit with simultaneous risk limit α can be conducted by continuing to hand count the votes on ballots selected at random until $P_{KM} \le \alpha$ or until the votes on all the ballots have been counted by hand; see [11].

The Kaplan-Markov P-value depends on which margins in which contests are affected by error. But P_{KM} can be bounded in a simple way that depends only on the number of ballots in the sample that overstate one or more margins by one vote but no margin by two votes, and the number of ballots in the sample that overstate one or more margins by two votes. This is the main contribution of this paper.

Suppose that of the n ballots in the sample, the audit finds that n_1 ballots overstate at least one margin by one vote but none by two votes, and that n_2 ballots overstate at least one margin by two votes. The remaining $n - n_1 - n_2$ ballots in the sample do not overstate any margin. Then

$$\begin{aligned} P_{KM} &\le P(n, n_1, n_2; U, \gamma) \\ &\equiv [1 - 1/U]^{n-n_1-n_2} \times \left[\frac{1 - 1/U}{1 - 1/(2\gamma)}\right]^{n_1} \times \\ &\quad \times \left[\frac{1 - 1/U}{1 - 2/(2\gamma)}\right]^{n_2} \\ &= [1 - 1/U]^n \times [1 - 1/(2\gamma)]^{-n_1} \times \\ &\quad \times [1 - 1/\gamma]^{-n_2}. \end{aligned} \tag{10}$$

4 Special cases

Table 3 shows some special cases of the P-value bound $P(n, n_1, n_2; U, \gamma)$ of equation 10 for margins of 2%, 1%, and 0.5%; $\gamma = 101\%$ and $\gamma = 110\%$; sample sizes between 500 and 2000 ballots; and 0–5 ballots showing errors that overstated at least one margin by one vote or by two votes.

The next two subsections develop rules of thumb for computing initial sample sizes. The rules ensure that if those samples have sufficiently few ballots that overstate one or more margins by one vote and no ballots that overstate any margin by two votes, all the contests can be certified at simultaneous risk limit α without counting any more ballots. If there are too many ballots with errors in the initial sample, the sample might need to be enlarged to limit the simultaneous risk; the Kaplan-Markov P-value of equation 9 or the upper bound $P(n, n_1, n_2; U, \gamma)$ of equation 10 can be used to determine when counting can stop.

4.1 Sample finds no more than k ballots that overstate any margin by 1 vote and no ballot that overstates any margin by 2 votes

Suppose we would like to be able to stop the audit at the first stage provided no more than k ballots in the sample overstate any margin by one one vote and no ballot in the sample overstates any margin by two votes. That is, we would like to find the smallest sample size n so that $P(n, k, 0; U, \gamma) \le \alpha$. Note that

$$\frac{x}{1+x} \le \log(1+x) \le x, \quad x > -1. \tag{11}$$

diluted			$P(n, n_1, n_2; U, \gamma)$			
margin		ballots w/	inflator $\gamma = 101\%$		inflator $\gamma = 110\%$	
μ	draws	errors	1-vote errors	2-vote errors	1-vote errors	2-vote errors
2%	500	0	0.7%	0.7%	1.0%	1.0%
		1	1.4%	69.8%	1.9%	11.4%
		2	2.7%	100.0%	3.5%	100.0%
		3	5.4%	100.0%	6.4%	100.0%
	750	0	0.1%	0.1%	0.1%	0.1%
		1	0.1%	5.8%	0.2%	1.2%
		2	0.2%	100.0%	0.4%	12.8%
		3	0.4%	100.0%	0.7%	100.0%
		4	0.9%	100.0%	1.2%	100.0%
		5	1.7%	100.0%	2.2%	100.0
1%	750	0	2.4%	2.4%	3.3%	3.3%
		1	4.8%	100.0%	6.0%	36.1%
		2	9.5%	100.0%	11.0%	100.0%
	1000	0	0.7%	0.7%	1.1%	1.1%
		1	1.4%	70.6%	1.9%	11.6%
		2	2.7%	100.0%	3.5%	100.0%
		3	5.4%	100.0%	6.5%	100.0%
0.5%	1000	0	8.4%	8.4%	10.3%	10.3%
	1250	0	4.5%	4.5%	5.8%	5.8%
		1	8.9%	100.0%	10.7%	64.0%
	1500	0	2.4%	2.4%	3.3%	3.3%
		1	4.8%	100.0%	6.0%	36.2%
		2	9.5%	100.0%	11.1%	100.0%
	2000	0	0.7%	0.7%	1.1%	1.1%
		1	1.4%	71.1%	1.9%	11.6%
		2	2.8%	100.0%	3.5%	100.0%
		3	5.5%	100.0%	6.5%	100.0%

Table 3: Upper bounds $P(n, n_1, n_2; U, \gamma)$ on the Kaplan-Markov P-value for various margins and sample sizes for a random sample of individual ballots.

Column 1: diluted margin μ. Column 2: sample size n. Column 3: number of ballots that show one or more errors that overstated a margin. Column 4: Bound on the P-value if those errors overstated margins by at most one vote, for error bound inflator $\gamma = 101\%$. Column 5: Bound on the P-value if error overstated at least one margin by two votes on each ballot with an error, for error bound inflator $\gamma = 101\%$. Columns 6, 7: same as columns 4, 5, but for error bound inflator $\gamma = 110\%$.

Since $U > 2$, it follows that $-1/U > -1/2 > -1$, and 11 implies that

$$\frac{-1}{U-1} \leq \log(1-1/U) \leq -1/U. \tag{12}$$

Take the logarithm of both sides of equation 10:

$$\begin{aligned}\log P &= n\log(1-1/U) - n_1\log(1-1/(2\gamma)) - \\ &\quad -n_2\log(1-1/\gamma)\end{aligned} \tag{13}$$

If $P \leq \alpha$ then $P_{KM} \leq \alpha$, so we seek the smallest sample size n such that

$$n\log(1-1/U) - k\log\left(1-\frac{1}{2\gamma}\right) \leq \alpha. \tag{14}$$

I.e.,

$$\log\alpha + k\log\left(1-\frac{1}{2\gamma}\right) \geq n\log(1-1/U). \tag{15}$$

By applying 12, we can see that it suffices to take

$$\log\alpha + k\log\left(1-\frac{1}{2\gamma}\right) > -n/U = -\frac{n}{2\gamma/\mu}. \tag{16}$$

Thus we can stop the audit and confirm the outcomes of all the contests at simultaneous risk limit α if a random sample of size

$$n \geq -2\gamma\left(\log\alpha + k\log\left(1-\frac{1}{2\gamma}\right)\right)\cdot\frac{1}{\mu} \tag{17}$$

ballots contains at most k ballots that overstate one or more margins by one vote and no ballots that overstate any margin by two votes.

This initial sample size n is a constant that depends on α, k, and γ, divided by the diluted margin μ: The initial sample size is inversely proportional to the diluted margin. This sort of simplicity seems desirable, even at the expense of a bit of extra counting. The extreme efficiency of single-ballot auditing keeps the burden manageable, despite the slack in the inequalities.

For $\gamma = 110\%$, $k = 3$ and $\alpha = 10\%$, inequality 17 says that if the sample size n is at least 9.06 divided by the diluted margin $\mu = V/N$, we can stop the audit if $n_1 \leq 3$ and $n_2 = 0$. If $n_1 > 3$ or $n_2 > 0$, we can use the Kaplan-Markov P-value in equation 9 to decide whether to count more votes by hand and to determine when the audit can stop: We continue to sample until $P_{KM} \leq \alpha$. Calculating P_{KM} requires nothing more complicated than arithmetic.

4.2 Sample percentage of ballots that overstate one or more margins by one vote is no more than a fraction λ of the diluted margin μ and no sampled ballot overstates any margin by two votes

Suppose we would like to be able to stop the audit at the first stage provided the sample percentage of ballots that overstate a margin by one vote is no more than than a fraction λ of the diluted margin $\mu = V/N$ and no ballot in the sample shows an overstatement of two votes. Then the initial sample size n must be large enough that $P(n, \lfloor n\mu\lambda\rfloor, 0; U, \gamma) \leq \alpha$:

$$\log\alpha \geq n\log(1-1/U) - \lfloor n\mu\lambda\rfloor\log\left(1-\frac{1}{2\gamma}\right). \tag{18}$$

Now $n\mu\lambda \geq \lfloor n\mu\lambda\rfloor$ and $\log\left(1-\frac{1}{2\gamma}\right) < 0$, so

$$-n\mu\lambda\log\left(1-\frac{1}{2\gamma}\right) \geq -\lfloor n\mu\lambda\rfloor\log\left(1-\frac{1}{2\gamma}\right). \tag{19}$$

Hence, if n is large enough that

$$\begin{aligned}\log\alpha &\geq n\log(1-1/U) - n\mu\lambda\log\left(1-\frac{1}{2\gamma}\right) \\ &= n\left[\log(1-1/U) - \mu\lambda\log\left(1-\frac{1}{2\gamma}\right)\right]\end{aligned} \tag{20}$$

then inequality 18 must also hold. This leads us to the condition

$$n \geq \frac{\log\alpha}{\log(1-1/U) - \mu\lambda\log\left(1-\frac{1}{2\gamma}\right)}. \tag{21}$$

By 12, it is enough to take

$$n \geq \frac{-\log\alpha}{\frac{1}{U-1} + \mu\lambda\log\left(1-\frac{1}{2\gamma}\right)}. \tag{22}$$

The term $U-1$ in the denominator can be replaced with U to simplify the approximation even more conservatively; substituting $U = 2\gamma/\mu$ then shows that

$$n \geq \frac{1}{\mu}\cdot\frac{-\log\alpha}{\frac{1}{2\gamma} + \lambda\log\left(1-\frac{1}{2\gamma}\right)} \tag{23}$$

suffices. Let

$$\rho = \rho(\alpha,\gamma,\lambda) \equiv \frac{-\log\alpha}{\frac{1}{2\gamma} + \lambda\log\left(1-\frac{1}{2\gamma}\right)}. \tag{24}$$

The constant ρ is the "sample-size multiplier": Given the values of of α, γ and λ, we can calculate ρ once and for all. We can take the initial sample size to be $n = \rho/\mu$, where μ is the diluted margin, and stop the audit provided no more than $n\lambda\mu$ of the ballots in the sample have one-vote maximum overstatements and none has a two-vote overstatement. As before, the initial sample size n is inversely proportional to the diluted margin, and the diluted margin is the *only* property of the collection of contests that enters the sample-size calculation. This makes calculating an adequate initial sample size extremely simple.

As a special case of inequality 23, consider a simultaneous risk limit $\alpha = 10\%$, an inflator $\gamma = 110\%$, and $\lambda = 10\%$; i.e., we want to be able to stop the audit at stage 1 if no more than a fraction $\lambda\mu$ of the ballots in the sample have errors that overstate the margin of one or more contests by one vote, but we are willing to expand the sample if more ballots than that overstate a margin by one vote or if any ballot overstates a margin by two votes. We calculate $\rho(10\%, 110\%, 10\%) = 5.85$, so a sample of size $5.85/\mu$ suffices to confirm all the contest outcomes at simultaneous risk limit 10%, provided the percentage of ballots with 1-vote overstatements is not more than 10% of the diluted margin and there are no ballots with 2-vote overstatements of any margin. In particular, if the diluted margin is $\mu = 2\%$, a sample of 293 ballots suffices. (Note that $\lambda\mu = 0.2\%$ in that case, and that $\lfloor 0.2\% \times 293 \rfloor = 0$, so if the sample had any overstatements at all, the audit might have to progress to the second stage.)

If $\lambda = 50\%$ but the other numbers in the previous example stay the same, we find $\rho(10\%, 110\%, 50\%) = 15.2$, so we would need an initial sample of $15.2/\mu = 761$ ballots, but we could stop the audit at the first stage provided no more than 7 of the ballots in the sample overstate one or more margins by at most one vote, and none overstates any margin by two votes. If any ballot in the sample overstates one margin by two votes, or more than 7 ballots in the sample overstate a margin by one vote, it might be necessary to expand the audit to limit the simultaneous risk to $\alpha = 10\%$: The audit should continue until either the actual Kaplan-Markov P-value in equation 9 (or its upper-bound $P(n, n_1, n_2; U, \gamma)$ of inequality 10) is less than $\alpha = 10\%$, or until all ballots have been tallied by hand and the correct outcomes of the contests are known.

Table 4 gives exemplar initial sample sizes for simultaneous risk limits α of 10%, 5% and 1% and diluted margins μ of 5%, 2%, 1%, and 0.5% and error fraction tolerances λ of 50% and 20%. The multiplier ρ grows as the risk limit α shrinks, because it takes a larger sample to have higher "confidence" that $E < 1$. Similarly, ρ grows as λ grows: The larger λ is, the more error we are tolerating in the sample; to ensure that $E < 1$, we need to know that E is not much larger than the sample error rate. But to estimate E more precisely requires a larger sample.

Setting λ large demands quite a bit of the sample: We are asking to be able to conclude that the total error is less than the diluted margin when the error in the sample is a substantial fraction of the diluted margin. That can lead to extremely large initial samples; combined with the slack in the inequalities, ρ can be infinite. This is readily avoided by choosing a more reasonable value of λ, such as 50%.

It is hard to give universal guidelines for selecting λ and γ. There are tradeoffs that will vary with the machine-counting technology used to count votes, the length of the canvass or the time allowed to complete the audit, the amount of public notice required, the difficulty of retrieving individual ballots, the cost of labor, and so on. If $\lambda\mu$ is less than the "benign" error rate of the machine-counting technology (in my experience, on the order of a tenth of a percent for central-count optical scan, primarily because of voter error), it is likely that the audit will progress beyond the first stage.

Both contests with extremely small margins and contests with larger margins that appear on only a small fraction of ballots can cause μ to be small. Separating them from the rest could reduce the overall workload, especially if including them would cause $\lambda\mu$ to be below the benign error rate of the machine-counting technology. This suggests a three-tier strategy: Collect all contests that, as a group, have $\lambda\mu$ rather larger than the benign error rate of the vote tabulation technology and audit them simultaneously. Audit contests with very small margins individually, or count them by hand entirely if their margins are on the order of the natural error rate of the machine-counting technology. Audit the remaining small contests with larger margins in groups that keep $\lambda\mu$ reasonably large for each group.

5 Conclusions

The MACRO method [9, 11] applied to single ballot audits can yield simple, conservative rules for determining the initial sample size of simultaneous risk-limiting audits. For a given desired simultaneous risk limit α and tolerance for the percentage of ballots that overstate one or more margins by one vote, the initial sample size is a constant divided by the "diluted margin," the smallest margin in votes divided by the total number of ballots cast in all the contests. The constant depends on α and the error tolerance, but not on anything to do with the contests, so the constant can be computed once and for all. The initial sample size depends on the details of the contests only through the diluted margin.

	$\lambda = 50\%$			$\lambda = 20\%$		
diluted	risk limit α			risk limit α		
margin μ	10%	5%	1%	10%	5%	1%
5%	305	396	609	139	180	277
2%	761	989	1521	346	450	691
1%	1521	1978	3041	691	899	1382
0.5%	3041	3956	6081	1382	1798	2764
multiplier ρ	15.20	19.78	30.40	6.91	8.99	13.82

Table 4: Initial sample sizes n and sample-size multipliers ρ for various simultaneous risk limits and tolerances for the percentage of ballots that overstate one or more margins by one vote, inflator $\gamma = 110\%$.

Column 1: diluted margin of victory μ. Columns 2–4: initial sample sizes n for various simultaneous risk limits if the audit is to stop when the percentage of ballots in the sample that overstate a margin by one vote is not more than 50% of the diluted margin. Columns 5–7: initial sample sizes n for various simultaneous risk limits if audit is to stop when the percentage of ballots in the sample that overstate a margin by one vote is not more than 20% of the diluted margin. Last row: In columns 2–7, the sample sizes n are equal to these "multipliers" divided by the diluted margins μ. The values of n are computed using inequality 23. The values of the simultaneous risk bound $P(n, n_1, n_2; U, \gamma)$ are generally on the order of $2/3$ of the nominal values in the column headings.

If any ballot in the initial sample overstates some margin by two votes, or if more than the tolerated number of ballots overstate one or more margins by one vote, the sample might need to be expanded, potentially progressing to a full hand count. When the sample has more error than the tolerance the design contemplated, either the exact Kaplan-Markov MACRO P-value or a simple upper bound on that P-value can be used to determine when to stop counting more ballots by hand. The stopping rule involves only simple arithmetic: addition, subtraction, multiplication, and division.

The method presented here has the advantage of simplicity. The cost of its extreme simplicity is some statistical inefficiency: More ballots have to be counted by hand than if sharper bounds were used. However, single-ballot audits are so efficient that this additional cost might easily be worthwhile. Unfortunately, to implement single-ballot audits on a wide scale may require changes to vote tabulation systems, because it is necessary to associate individual cast vote records (CVRs) with individual physical ballots. To my knowledge, no federally certified vote tabulation system makes that association possible. Most do not even store CVRs. Auditing by using an unofficial vote tabulation system that does produce CVRs—such as those of Clear Ballot Group, the Humboldt Transparency Project, or TrueBallot—and confirming transitively that the system of record is correct, might be the best interim option [1].

Another advantage of the method presented here is that the CVRs are not needed to determine the sampling probabilities: The same upper bound on error, and hence the same sampling probability, is used for every ballot, regardless of which contests appear on the ballot and regardless of how the vote-tabulation system interpreted the ballot. However, once the sample is drawn, it is necessary to determine how the voting system interpreted the ballots in the sample. This is essentially how the first single-ballot risk-limiting audit was performed, in Yolo County, CA, in November 2009 [11].

6 Acknowledgments

I thank Joseph Lorenzo Hall, Mark Lindeman, and Eric K. Rescorla for helpful comments and conversations.

References

[1] Calandrino, J., Halderman, J., and Felten, E. Machine-assisted election auditing. In *Proc. 2007 USENIX/ACCURATE Electronic Voting Technology Workshop (EVT 07)* (August 2007), USENIX.

[2] Checkoway, S., Sarwate, A., and Shacham, H. Single-ballot risk-limiting audits using convex optimization. In *Proceedings of the 2010 Electronic Voting Technology Workshop / Workshop on Trustworthy Elections (EVT/WOTE '10)* (2010), D. Jones, J.-J. Quisquater, and E. Rescorla, Eds., USENIX.

[3] Hall, J. L., Miratrix, L. W., Stark, P. B., Briones, M., Ginnold, E., Oakley, F., Peaden, M., Pellerin, G., Stanionis, T., and Webber, T. Implementing risk-limiting post-election audits in California. In *Proc. 2009 Electronic Voting Technology Workshop/Workshop on Trustworthy Elections (EVT/WOTE '09)* (Montreal, Canada, August 2009), USENIX.

[4] Lindeman, M., Halvorson, M., Smith, P., Garland, L., Addona, V., and McCrea, D. Principles and best practices for post-election audits. `www.electionaudits.org/files/best%20practices%20final_0.pdf`, 2008.

[5] Miratrix, L., and Stark, P. The trinomial bound for post-election audits. *IEEE Transactions on Information Forensics and Security 4* (2009), 974–981.

[6] Saldaña, L. California assembly bill 2023. www.leginfo.ca.gov/pub/09-10/bill/asm/ab_2001-2050/ab_2023_bill_20100325_amended_asm_v98.html, 2010.

[7] Stark, P. Conservative statistical post-election audits. *Ann. Appl. Stat. 2* (2008), 550–581.

[8] Stark, P. A sharper discrepancy measure for post-election audits. *Ann. Appl. Stat. 2* (2008), 982–985.

[9] Stark, P. Auditing a collection of races simultaneously. Tech. rep., arXiv.org, 2009.

[10] Stark, P. CAST: Canvass audits by sampling and testing. *IEEE Transactions on Information Forensics and Security, Special Issue on Electronic Voting 4* (2009), 708–717.

[11] Stark, P. Efficient post-election audits of multiple contests: 2009 California tests. Tech. rep., Social Science Research Network, 2009. 2009 Conference on Empirical Legal Studies.

[12] Stark, P. Risk-limiting post-election audits: P-values from common probability inequalities. *IEEE Transactions on Information Forensics and Security 4* (2009), 1005–1014.

[13] Stark, P. Risk-limiting vote-tabulation audits: The importance of cluster size. *Chance* (2010), in press.

[14] Statistical Association, A. American Statistical Association statement on risk-limiting post-election audits. www.amstat.org/outreach/pdfs/Risk-Limiting_Endorsement.pdf, 2010.

Notes

[1] The denominator of the diluted margin is the total number of ballots cast across all contests, not the votes cast in the particular contest. So, for instance, that margin of 2,000 votes might be in a contest that appeared on only 12,000 of the 100,000 ballots, and there might have been only 8,000 votes cast in that contest: 5,000 for the winner and 3,000 for the loser. The diluted margin is $2,000/100,000 = 2\%$, not $2,000/8,000 = 25\%$.