Efficient post-election audits of multiple contests: 2009 California tests*

Philip B. Stark

Department of Statistics, Code 3860
Berkeley, CA 94720-3860
e-mail: stark@stat.berkeley.edu

Abstract: Risk-limiting post-election audits have a pre-specified minimum chance of requiring a full hand count if the outcome of the contest is not the outcome that a full hand count of the audit trail would show. The first risk-limiting audits were performed in 2008 in California. Two refinements to increase efficiency will be tested in Marin and Santa Cruz counties, California, in November 2009. The first refinement is to audit a collection of contests as a group by auditing a random sample of batches of ballots and combining observed discrepancies in the contests represented in those batches in a particular way: the maximum across-contest relative overstatements (MACRO). MACRO audits control the familywise error rate (the chance that one or more incorrect outcomes fails to be corrected by a full hand count) at a cost that can be lower than that of controlling the per-comparison error rate with independent audits. A risk-limiting audit for the entire collection of contests can be built on MACRO using a variety of probability sampling schemes and ways of combining MACRO across batches. The second refinement is to base the test on the Kaplan-Markov confidence bound, drawing batches with probability proportional to an error bound (PPEB) on the MACRO. The Kaplan-Markov bound is especially well suited to sequential testing: After each batch is audited, a simple calculation—a product of fractions—determines whether to audit another batch or to stop the audit and confirm the apparent outcomes.

Keywords: familywise error rate, Kaplan-Markov martingale confidence bound, nonnegative random variable, per-comparison error rate, probability proportional to size, sequential test, simultaneous test, statistical audit.
JEL Codes: C12, C14, K39, M40

1. Introduction

This paper is about statistical methods for testing whether errors in counting votes—no matter what caused them—changed the apparent outcome of one or more contests in an election. The methods have a big chance of catching and fixing wrong outcomes before the winners are certified.

The *apparent outcome* or *semi-official outcome* of a contest is the electoral outcome election officials are prepared to report at the end of the canvass, the

*Paper to be presented at the 2009 Conference on Empirical Legal Studies (CELS 2009), University of Southern California Gould School of Law, Los Angeles, CA. I thank Mike Higgins, Mark Lindeman, Luke Miratrix and Ron Rivest for helpful conversations.

outcome that will be certified unless an audit or something else intervenes. An apparent outcome is *wrong* if it disagrees with the outcome that a full hand count of the audit trail would show: The outcome shown by a full hand count of the audit trail is the *correct* outcome, by definition. Of course, if the audit trail is inaccurate or incomplete, the outcome of a hand count of the audit trail might not reflect how the votes were actually cast.

Suitably designed post election audits can control the risk of certifying an election outcome that is wrong. That risk is limited to α if the audit ensures that the chance of a full hand count is at least $1 - \alpha$ whenever the apparent outcome is wrong, whatever the reason. The outcome of the full hand count is then reported as the official outcome of the contest, thereby correcting any error in the apparent outcome.

Risk-limiting audit has become a term of art. The consensus definition[1] is that an audit is risk-limiting if and only if it has a known minimum probability of requiring a full manual count whenever the apparent outcome is wrong. A risk-limiting audit ends in one of two ways: It leads to a full hand count and the result of that hand count is reported as the official result, or it ends without a full hand count and the apparent outcome is reported the official result. A risk-limiting audit is designed so that that when the outcome is wrong, the first possibility is likely (has chance at least $1 - \alpha$) and the second is unlikely (has chance at most α). A risk-limiting audit is efficient if, when the outcome is right, the first possibility is unlikely and the second is likely to occur after the smallest possible amount of hand counting.

As of this writing, no jurisdiction in the U.S. requires risk-limiting audits. For discussion and a summary of U.S. audit laws as of mid-2009, see [5]. A list of current audit legislation is maintained by Verified Voting.[2] In 2007 the National Association of Secretaries of State surveyed states' practices for post-election audits.[3] See also [13, 6].

Some proposed legislation appears to allow audits to report as the official outcome of the contest an outcome other than the apparent "semi-final" outcome—without hand counting the entire audit trail.[4] This seems unwise, if not unconstitutional. Using statistical evidence to confirm an outcome is desirable because it is economical. But using statistical evidence to overturn an apparent outcome without a full hand count is not desirable because it introduces the possibility that the statistical calculation will disenfranchise the majority of voters.

A statistical test of the hypothesis that the outcome is wrong can err in two

[1] See `http://www.electionaudits.org/principles.html` and [17, 18, 19, 20, 5, 11].

[2] `http://verifiedvoting.org/article.php?id=5816`

[3] `http://nass.org/index.php?option=com_docman&task=doc_download&gid=54`

[4] H.R. 2894, *Voter Confidence and Increased Accessibility Act of 2009*, § 322(b)(2)(B) allows NIST to approve an audit method if "the reported election outcome will have at least a 95 percent chance of being consistent with the election outcome that would be obtained by a full recount." This provision of H.R. 2894 appears to allow the audit to alter a correct apparent outcome into an incorrect reported outcome as long as it does not do so too frequently. H.R. 2894 has other shortcomings in its audit provisions that preclude it from limiting the risk that an incorrect outcome will be certified. For instance, when an audit finds errors, the bill leaves it to the discretion of the state whether to count more batches by hand. See [5].

ways: It can conclude that the outcome is right when it is wrong, or conclude that the outcome is wrong when it is right. To eliminate the possibility of committing the first kind of error requires hand counting every contest in its entirety, exactly what auditing tries to avoid: Indeed, the risk is defined to be the chance of the first kind of error when the outcome is wrong. But the second kind of error can be eliminated by requiring a full hand count to set the record straight whenever the audit does not provide strong evidence that the outcome is correct. That is how risk-limiting audit methods work.

Constructing a risk-limiting audit is easy—at least, in theory. For instance, if the audit trail of each contest is counted by hand with probability $1 - \alpha$ no matter what, then it is counted by hand with probability $1-\alpha$ when the outcome is wrong. Such a rule is inefficient, because if there is little error, counting a small percentage of audit records selected randomly could give strong evidence that the apparent outcome is right, obviating the need to count the rest of the audit trail. Efficiency in post-election auditing comes from devising methods that have probability $1 - \alpha$ of counting all the audit records by hand when the outcome is wrong, but count as few ballots as possible when the outcome is right.

Single ballot auditing—where the reported interpretation of the votes on a ballot is compared with auditors' interpretation of the votes for the same ballot—could be quite efficient [3], but currently no jurisdiction has the technology, processes, and procedures in place to support auditing individual ballots and still maintain the secrecy of the ballot.[5] Instead, current audit methods compare hand counts of the audit trail for a random sample of *batches* of ballots with the reported vote totals for those same batches.[6] Obtaining timely vote reports in machine-readable form for moderate size batches, such as precincts, is a bottleneck for post election auditing [5, 11].

Pilot studies in three California counties have shown that risk-limiting audits of individual contests that range in size from about 10 precincts to about 200 precincts can be conducted economically, within the canvass period, at a cost of about \$0.35–\$0.50 per audited ballot [5, 11]. However, it is cumbersome to audit a large number of contests in a single election by repeating the audit process independently for each of those contests. The difficulty of auditing a collection of contests is a logistical barrier to wider use of post-election audits to control risk.

[5]The Humboldt County Election Transparency Project `http://humtp.com/` takes a step in that direction, making available to the public ballot images and software to tally the votes, but it currently lacks safeguards on voter privacy and on the chain of custody of ballot images.

[6]Although laws requiring audits have been around at least since the 1960s, [16] appears to be the first to consider the probability that a random sample of batches of ballots will find one or more errors if the outcome of the contest is wrong. His analysis is predicated on simple random sampling (generally, laws mandate stratified random sampling for contests that cross jurisdictional boundaries) and on the assumption that at most a given fraction of ballots could have been miscounted in each batch. Most work on election auditing since then has followed suit by focusing on the chance of finding one or more errors (and making the same assumption about the maximum error that each batch can hold), e.g., [15, 1, 9]. As discussed by [5, 17, 19], audits of voter-marked ballots almost always find at least one error, so the probability of detecting error is not as important as the strength of the evidence that the outcome is correct, given the level of error the audit finds.

This paper presents two theoretical advances in post election auditing and tests of those advances in practice. The first is an approach to auditing an arbitrarily large number of contests in an election by hand-counting the votes in a random sample of batches of ballots for every contest subject to audit that appears on those ballots. Auditing every contest on a random sample of batches of ballots is built into some state audit laws, such as California's "1% audit."[7]

In this approach, for each batch of ballots in the sample, the discrepancies in the votes in the contests represented in that batch are combined into a single summary statistic, the maximum across-contest relative overstatement (MACRO). This is a straightforward extension of the approach in [18] to cover more than one contest. Error in the apparent margin between each winner and each loser in a given contest is normalized by the apparent margin between that pair of candidates. The largest normalized error in a batch—maximized first across pairs of apparent winners and losers in a given contest and then across contests—summarizes the error in the batch. This maximum across-contest relative overstatement can then be used with existing methods designed for auditing individual contests to limit the risk of certifying an incorrect outcome to α, for instance, the methods introduced in [17, 19, 20, 11]. The result is a simultaneous risk-limiting audit of all the contests: The audit limits the chance that one or more incorrect outcomes will go uncorrected to at most α.

The second advance to be tested in the November 2009 pilot is a new method for deciding whether to stop the audit or audit more batches, given the error the audit has found. The method is based on the Kaplan-Markov confidence bound for the mean of a nonnegative random variable [7]. The bound can be used to calculate a P-value for the hypothesis that the electoral outcome [20] is wrong when the audit sample is drawn with probability proportional to a bound on the error [1, 11, 20]. In the 2009 pilot, error will be measured by MACRO, so the bound on the error is a bound on the MACRO. In simulations, this method is rather more efficient than existing methods: It generally requires less auditing than other known risk-limiting methods when the outcome is correct [10].

This paper introduces MACRO and presents the sequential test based on the Kaplan-Markov bound, which is remarkably simple to compute. It gives a cartoon application to a set of three contests in an election in a jurisdiction roughly the size of a county. When the pilot audit is completed in November 2009, the paper will be revised to report the results, including details of the contests audited and audit costs.

2. Maximum Across-Contest Relative Overstatement (MACRO)

As [18] notes, for the apparent outcome of a contest to be wrong, the margin between some apparent winner and some apparent loser of the contest must be overstated by at least 100% of the margin between that pair of candidates. Scaling errors by the margins they affect makes them commensurable. This idea extends to multiple contests in the same election: For the apparent outcome of

[7] California Elections Code §15360.

any of those contests to be wrong, for some contest, the margin between some apparent winner and some apparent loser must be overstated by at least 100% of the apparent margin between them.

Suppose there are N batches of ballots that together cover C contests. Not every contest is represented on every ballot, but together the N batches include every ballot for all C contests. Contest c has K_c "candidates," which could be politicians or positions on an issue. For instance, the "candidates" for a ballot measure might be "yes on Measure A" and "no on measure A." The total number of candidates or positions in all contests is $K = \sum_{c=1}^{C} K_c$. We take those K candidates to be enumerated in some canonical order, for instance, alphabetically.

Voters eligible to vote in contest c may vote for up to f_c candidates in that contest (contest c can have up to f_c winners). The f_c candidates who apparently won contest c are those in $\mathcal{W}_c$. Those who apparently lost contest c are in $\mathcal{L}_c$. The apparent vote for candidate k in batch p is v_{kp}. (If ballots in batch p do not include the contest c in which candidate k is competing, $v_{kp} \equiv 0$.) The apparent vote for candidate k is $V_k \equiv \sum_{p=1}^{N} v_{kp}$. If candidates w and ℓ are contestants in the same contest c, the reported margin of apparent winner $w \in \mathcal{W}_c$ over apparent loser $\ell \in \mathcal{L}_c$ is

$$V_{w\ell} \equiv V_w - V_\ell > 0. \tag{1}$$

The actual vote for candidate k in batch p—the number of votes for k that an audit would find—is a_{kp}. If the ballots in batch p do not include the contest in which candidate k is competing, $a_{kp} \equiv 0$. The actual vote for candidate k is $A_k \equiv \sum_{p=1}^{N} a_{kp}$. If candidates w and ℓ are contestants in the same contest c, the actual margin of candidate $w \in \mathcal{W}_c$ over candidate $\ell \in \mathcal{L}_c$ is

$$A_{w\ell} \equiv A_w - A_\ell. \tag{2}$$

The apparent winners of all C contests are the true winners of those contests if

$$\min_{c \in \{1,\ldots,C\}} \min_{w \in \mathcal{W}_c, \ell \in \mathcal{L}_c} A_{w\ell} > 0. \tag{3}$$

If $w \in \mathcal{W}_c$ and $\ell \in \mathcal{L}_c$, define

$$e_{pw\ell} \equiv \begin{cases} \frac{(v_{wp} - v_{\ell p}) - (a_{wp} - a_{\ell p})}{V_{w\ell}}, & \text{if ballots in batch } p \text{ contain contest } c \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

For the true outcome of any of the C contests to differ from its apparent outcome, there must exist $c \in \{1, \ldots, C\}$, $w \in \mathcal{W}_c$ and $\ell \in \mathcal{L}_c$ for which $\sum_{p=1}^{N} e_{pw\ell} \geq 1$. The *maximum across-contest relative overstatement in batch* p (the MACRO in batch p) is

$$e_p \equiv \max_{c \in \{1,\ldots,C\}} \max_{w \in \mathcal{W}_c, \ell \in \mathcal{L}_c} e_{pw\ell}. \tag{5}$$

Now

$$\max_{c \in \{1,\ldots,C\}} \max_{w \in \mathcal{W}_c, \ell \in \mathcal{L}_c} \sum_{p=1}^{N} e_{pw\ell} \leq \sum_{p=1}^{N} \max_{c \in \{1,\ldots,C\}} \max_{w \in \mathcal{W}_c, \ell \in \mathcal{L}_c} e_{pw\ell} = \sum_{p=1}^{N} e_p \equiv E. \tag{6}$$

The sum on the right, E, is the *maximum across-contest relative overstatement* (MACRO). If $E < 1$, the apparent electoral outcome of each of the C contests is the same outcome that a full hand count would show. (For techniques to deal with contests that require a super-majority, see [17].)

Consider the family of C null hypotheses, *the outcome of contest c is incorrect*. The condition $E < 1$ is sufficient for the entire family of C null hypotheses to be false. If an audit gives strong statistical evidence that $E < 1$, we can safely conclude that the apparent outcomes of *all* C contests are correct. If we test the hypothesis $E \geq 1$ at significance level α, that gives a test of the family of C hypotheses with familywise error rate no larger than α.

Suppose the number of valid ballots cast in batch p for contest c is at most b_{cp}.[8] Clearly $a_{wp} \geq 0$ and $a_{\ell p} \leq b_{cp}$, if ℓ is a candidate in contest c. Hence, $e_{pw\ell} \leq (v_{wp} - v_{\ell p} + b_{cp})/V_{w\ell}$, and so

$$e_p \leq \max_{c \in \{1,\ldots,C\}} \max_{w \in \mathcal{W}_c, \ell \in \mathcal{L}_c} \frac{v_{wp} - v_{\ell p} + b_{cp}}{V_{w\ell}} \equiv u_p. \tag{7}$$

The bound u_p is a limit on the relative overstatement of *any* margin that can be concealed in batch p, the MACRO in batch p. If $U \equiv \sum_p u_p < 1$, the outcome of the election must be correct so no audit is needed.

Otherwise, if the values of u_p are generally small, error sufficient to cause the wrong candidate to appear to win any of the contests must be spread out across many batches, while if some of the values of u_p are large, outcomes can be wrong even if very few batches have any error at all. The values of u_p can be used to adjust sampling probabilities to give greater scrutiny to batches that can conceal larger errors (e.g., NEGEXP or PPEB sampling [1]).

The values of e_p observed in a random sample (including simple, stratified, NEGEXP, and PPEB random samples) can be used to calculate a P-value for the compound null hypothesis that one or more of the apparent outcomes of the C contests differs from the outcome that a full hand count of all the ballots in that contest would show. See [20] for details. Those P-value calculations can be embedded in a sequential procedure for testing whether one or more of the outcomes is wrong, using approaches like that described by [19]. The resulting test controls the *familywise error rate* (FWER) for testing the collection of hypotheses that the outcome of each contest is correct. That is, the test keeps small the chance of incorrectly concluding that the outcomes are correct when any of the outcomes is wrong.

In the pilot audits in November 2009, we will use a method described in section 3, based on PPEB sampling and measuring error as taint: The *taint* of batch p is

$$\tau_p = \frac{e_p}{u_p} \leq 1. \tag{8}$$

[8] If the batches are homogeneous with respect to ballot style, then $b_{1p} = b_{2p} = \ldots = b_{Cp}$ except for contests c for which $b_{cp} = 0$. In some jurisdictions, however, VBM ballots are counted in "decks" that bear no special relationship to geography. Then, the values of b_{cp} for a single batch p can depend on the contest c.

For PPEB samples, it is mathematically convenient—and efficient—to work with taint τ_p rather than with error e_p, because the expected value of the taint in a batch drawn by PPEB is E/U [20, 11].

3. A sequential test using the Kaplan-Markov bound

We draw batches at random from the N batches of ballots in the contest. In each draw, the chance of selecting batch p is u_p/U. The draws are independent, so a given batch can be drawn more than once. This is an example of sampling with probability proportional to an error bound (PPEB) [1], called *dollar unit sampling* or *monetary unit sampling* in the financial auditing literature [21, 8, 4, 12, 14].

Let T_j be the taint of the jth PPEB draw; that is, $T_j = \tau_p \equiv e_p/u_p$ for the batch p that is selected in the jth draw. Then $\mathbb{E}T_j = E/U$ [20, 11].

If the expected value of the taint of the MACRO is less than $1/U$, the apparent outcomes of all C contests are correct. Hence, if the P-value of the hypothesis that $\mathbb{E}T_j \geq 1/U$ is less than α, we can stop the audit. Otherwise, we need to count more ballots by hand—possibly all of them. (The P-value depends on the sampling scheme, the test statistic, the batch error bounds, the observed error, and so on.)

[7] presents two methods for constructing a confidence bound for a nonnegative random variable, based on a martingale optional stopping theorem and Markov's inequality in [2]. [20] points out that these *Kaplan-Markov* bounds can be used to calculate a P-value for the hypothesis that the outcome of a contest is wrong if the sample is drawn with probability proportional to an error bound (PPEB). The Kaplan-Markov approach can be used to audit a collection of contests by substituting the taint of the MACRO for the taint of the maximum pairwise overstatement within a single contest.

The second of the two Kaplan-Markov bounds yields a very clean and simple technique for sequential auditing. Suppose we have drawn n batches by PPEB using the upper bounds $\{u_p\}$ for MACRO. Let T_j be the taint of the batch drawn on the jth trial. Define

$$P_n \equiv \prod_{i=1}^{n} \frac{1 - 1/U}{1 - T_i}. \tag{9}$$

We can stop the audit as soon as $P_n < \alpha$ [20].[9] If any of the apparent outcomes is wrong, the probability that the audit stops before every batch has been audited is at most α, If we do end up auditing every batch, we know the correct outcomes of the contests. Hence, this sequential audit limits the risk to at most α.

The following procedure is a risk-limiting sequential audit:

[9] For technical reasons, we need to impose a finite bound M on the number of draws. If the audit has not stopped by the time we have drawn M times, the remaining batches are counted by hand.

1. Pick the risk limit $\alpha \in (0,1)$ and the maximum number of draws $M > 0$. Calculate the batch error bounds $\{u_p\}$ for the MACRO and $U = \sum_p u_p$. Set $n = 1$.
2. Draw a batch using PPEB: Select batch p with probability u_p/U, independently of previous draws. Let q denote the index of batch that is drawn. Audit batch q if it was not audited previously.
3. Find $T_n \equiv t_q \equiv e_q/u_q$, taint of the batch q just drawn.
4. Compute
$$P_n \equiv \prod_{j=1}^{n} \frac{1-1/U}{1-T_j}. \tag{10}$$
5. If $P_n < \alpha$, stop; certify all S apparent outcomes. If $n = M$, audit all unaudited batches. If all batches have been audited, stop; report the outcomes according to the hand counts. Otherwise, $n \leftarrow n+1$ and go to step 2.

Although the Kaplan-Markov bound allows us to audit one batch at a time, compute T_j and P_n after each batch is audited, and stop when $P_n < \alpha$, we know ahead of time that a very small number of draws cannot give strong evidence that $\mathbb{E}T_j < 1/U$. To find the minimum number of draws that could possibly allow us to certify the election with risk-limit α, we need a lower bound on e_p in each batch. The lower bound is not in general $-u_p$, both because e_p involves a maximum across contests, and because within a given contest, the amount by which the reported vote might have overstated the margin generally is not equal to the amount by which the reported vote might have understated the margin.

If at most b_{cp} votes were cast in contest c in batch p, the reported margin $v_{wp} - v_{\ell p}$ between apparent winner w and apparent loser ℓ in contest c must have overstated the true margin between those candidates by at least $v_{wp} - v_{\ell p} - b_{cp} \leq 0$. That is,
$$e_{pw\ell} \geq v_{wp} - v_{\ell p} - b_{cp}. \tag{11}$$
The MACRO in batch p is the maximum of the relative overstatements across candidates and across contests, so
$$\begin{aligned} e_p &\equiv \max_{c\in\{1,\ldots,C\}} \max_{w\in\mathcal{W}_c, \ell\in\mathcal{L}_c} e_{pw\ell} \\ &\geq \max_{c\in\{1,\ldots,C\}} \max_{w\in\mathcal{W}_c, \ell\in\mathcal{L}_c} \frac{v_{wp} - v_{\ell p} - b_{cp}}{V_{w\ell}} \equiv u_p^-. \end{aligned} \tag{12}$$
Even if every observed batch MACRO is equal to its lower bound u_p^-, the Kaplan-Markov P-value 9 will be larger than α unless n is sufficiently large. That can be used to set an initial sample size for the audit. Let τ^- be the smallest value of u_p^-/u_p. The audit cannot stop before drawing
$$n_0(\tau^-) \equiv \min\left\{k : \left(\frac{1-1/U}{1-\tau^-}\right)^k < \alpha\right\} = \left\lceil \frac{\ln\alpha}{\ln(1-1/U) - \ln(1-\tau^-)} \right\rceil \tag{13}$$
times. This minimum number is extremely optimistic: It is based on the assumption that the errors favored the losers to the maximum extent possible in the

Contest	precincts	batches	ballots	winner	loser	margin	IP batches winner	IP batches loser	VBM batches winner	VBM batches loser
A	200	400	120,000	60,000	54,000	6,000	200	180	100	90
B	100	200	60,000	30,000	24,000	6,000	200	160	100	80
C	60	120	36,000	18,000	12,600	5,400	200	140	100	70

TABLE 1

Hypothetical reported results for an election with three overlapping contests.

Contest A spans the entire jurisdiction, 200 precincts. Contest B includes 100 of the precincts in the jurisdiction. Contest C includes 60 of the precincts in the jurisdiction; 30 of those are also in contest B. Each precinct is divided into two batches of ballots, 400 ballots cast in-precinct (IP) and 200 ballots cast by mail (VBM). In addition to valid votes for the candidates, there are undervotes and invalid ballots.

batch that could have the biggest negative taint (understatement), and that the PPEB draw selected that batch every time. Generally, rather more draws will be required. If the audited batches have no errors at all—neither overstatements or understatements of any reported margin—then the number of PPEB draws required to certify the outcome is $n_0(0) = \lceil \ln \alpha / \ln(1 - U^{-1}) \rceil$.[10]

In auditing a single contest whose outcome is correct, negative taints are to be expected occasionally: Sometimes error will overstate the margin and sometimes it will understate the margin. But in using MACRO to audit a collection of contests, negative taints will tend to be rarer, because the MACRO e_p for batch p is negative only if every margin in *every* contest under audit was understated in batch p.

4. Illustration

This section presents a cartoon of an election with $C = 3$ contests in a jurisdiction that has 200 precincts. Each of the three contests has only two contestants. Contest A is jurisdiction-wide; the reported result is 50% for the apparent winner, 45% for the apparent loser, and 5% undervotes and invalid ballots. Contest B involves half the precincts in the jurisdiction; the reported result for this contest is 50% for the apparent winner, 40% for the apparent loser, and 10% undervotes and invalid ballots. Contest C involves 60 of the precincts in the jurisdiction, of which 30 overlap with the second contest. The reported result for this contest is 50% for the apparent winner, 35% for the apparent loser, and 15% undervotes and invalid ballots.

The auditable batches of ballots comprise ballots cast either in-precinct (IP) or by mail (VBM) for each of the 200 precincts in the jurisdiction; thus there are $N = 400$ auditable batches of ballots in all. For the sake of illustration, we take the IP batches to contain 400 ballots each and the VBM batches to contain 200 ballots each, and we assume that, for each contest, the reported margins are the same in all 400 batches. A summary is given in table 1.

There are eight situations to consider in calculating u_p: IP versus VBM batches and batches where voters can vote only in contest A, in contests A

[10] Note that this is equivalent to the minimum number given by [1], since the margin has been re-normalized to 1.

batch type	batches	u_p
IP–Contest A only	70	0.0700
VBM–Contest A only	70	0.0350
IP–Contests A and B	70	0.0733
VBM–Contests A and B	70	0.0367
IP–Contests A and C	30	0.0852
VBM–Contests A and C	30	0.0426
IP–Contests A, B & C	30	0.0852
VBM–Contests A, B & C	30	0.0426

TABLE 2

Upper bounds on the MACRO in each batch for the eight kinds of batches of ballots in a hypothetical contest.

and B, in contests A and C, and in all three contests. Consider an IP batch in which voters can vote in all three contests ($b_{cp} = 400$):

$$\begin{aligned} u_p &= \max\left\{\frac{200-180+400}{6,000}, \frac{200-160+400}{6,000}, \frac{200-140+400}{5,400}\right\} \\ &= \max\{0.0700, 0.0733, 0.0852\} = 0.0852. \end{aligned} \tag{14}$$

For a VBM batch in which voters were eligible to vote in all three contests ($b_{cp} = 200$),

$$\begin{aligned} u_p &= \max\left\{\frac{100-90+200}{6,000}, \frac{100-80+200}{6,000}, \frac{100-70+200}{5,400}\right\} \\ &= \max\{0.0350, 0.0367, 0.0426\} = 0.0426. \end{aligned} \tag{15}$$

Table 2 lists the values of u_p for all eight cases. The total of the MACRO error bounds $\{u_p\}$ for all $N = 400$ batches is

$$U = 70\times(0.0700+0.0350+0.0733+0.0367)+2\times30\times(0.0852+0.0426) = 22.718. \tag{16}$$

Suppose we want to design a PPEB-based audit that has at least a 75% chance of requiring a full hand count if a full hand count would show a different outcome for any of the three contests. That controls the risk (that an incorrect result will not be corrected by a full hand count) to be at most $\alpha = 0.25$. We can base such an audit on the Kaplan-Markov approach described in section 3.

Suppose we make $n = 36$ PPEB draws, 5 of which show taint $\tau_p = 0.04$ and the rest of which show $\tau_p = 0$.[11] Then $P = 0.243$: We could stop the audit without a full hand count. The risk that the outcome of any of the three contests is wrong is at most 25% (and plausibly far lower, since this approach makes a number of very conservative choices).

[11] Taint of 0.04 corresponds to a different number of errors in different batches, depending on the value of u_p in the batch and the margin that the error affects. In an IP batch of ballots that includes contest C, an error that overstates the margin in contest A or contest B by 20 votes is a taint of just under 0.04, while in a VBM batch of ballots that includes only contest A, an error that overstates the margin in contest A by 8 votes is a taint of just under 0.04.

Note that the expected number of distinct batches drawn in the $n = 36$ draws is

$$\sum_{p=1}^{400}[1 - (1 - u_p/U)^{36}] = 34.3, \tag{17}$$

about 8.6% of the 400 auditable batches. However, those batches would tend to be the larger (IP) batches. Let b_p denote the number of ballots in batch p ($b_p = 400$ for IP batches and $b_p = 200$ for VBM batches). The expected number of ballots audited is

$$\sum_{p=1}^{400} b_p[1 - (1 - u_p/U)^{36}] = 11,387.3, \tag{18}$$

about 9.5% of the 120,000 ballots. The expected number of votes audited, 20,617.68, can be calculated analogously: Substitute in place of b_p the number of voting possibilities in batch p (from 200 for VBM batches that include only contest A up to 1,200 for IP batches that include all three contests).

In contrast, suppose we were auditing only contest A. Then the error bounds would be $u_p = 0.07$ for the 200 IP batches and $u_p = 0.035$ for the 200 VBM batches; The total error bound would be $U_A = 21$, a bit smaller than the previous value, $U = 22.718$. If the sample taints in $n = 36$ draws were as before—five equal to 0.04 and 31 equal to 0—the value of P would be 0.212. This is a bit smaller than the value 0.243 for auditing all three contests, stronger evidence that the outcome of that single contest was correct: The sequential audit might have stopped before $n = 36$ draws.

If we had made only $n = 33$ draws and had seen five taints equal to 0.04 and 28 equal to zero, the value of P would be 0.245, and we would be able to confirm the outcome of that single contest with risk no greater than $\alpha = 0.25$. The workload would be somewhat lower, both because we would be counting only one contest on each ballot and because the number of batches drawn would be lower. The expected number of batches audited would be 31.6 versus 34.3, and the expected number of ballots audited would be 9,778 versus 11,387. But we would only be testing the outcome of contest A.

Suppose we audited all three contests independently. We have a choice to make about multiplicity—the fact that we are testing more than one hypothesis. The simultaneous audit based on MACRO has the property that there is at least 75% chance of a full hand count of every contest that has an incorrect outcome, i.e., risk at most $\alpha = 0.25$ that one or more incorrect outcomes will be certified. Suppose we choose to maintain this property—keeping the familywise error rate (FWER) at most $\alpha = 0.25$. We split the risk across the three audits by requiring each to have chance at least $0.75^{1/3} = 0.909$ of a full count if the outcome is incorrect. The chance all three will progress to full counts if all three outcomes are incorrect is then at least $0.909^3 = 0.75$.

We could instead control the per-comparison error rate (PCER) to be at most $\alpha = 0.25$. That would mean that for each audited contest, the chance of a full hand count if the outcome is wrong is at least 75%. However, the chance

		FWER				PCER		
Contest	U	n	expected batches	expected ballots	expected votes	n	expected batches	expected ballots	expected votes
A	21.00	52	48.49	16,074.23	16,074.23	33	31.58	10,488.77	10,488.77
B	11.00	28	26.01	8,615.69	8,615.69	17	16.27	5,402.16	5,402.16
C	7.67	19	17.50	5,795.81	5,795.81	12	11.41	3,787.51	3,787.51
all			85.13	28,038.26	30,485.73		56.38	18,649.98	19,678.44
MACRO	22.72	36	34.30	11,387.29	20,617.68				

TABLE 3

Comparison of independent and simultaneous audits controlling FWER and PCER.

The familywise error rate (FWER) of a collection of audits is the chance that one or more fails to result in a hand count when the corresponding outcome is incorrect. If the FWER is at most 0.25, the chance that there is a full hand count of every contest with an incorrect outcome is at least 75%. The per-comparison error rate (PCER) of a collection of audits is the chance that each audit fails to result in a hand count when the outcome of the contest under audit is incorrect. If the PCER is at most 0.25, then, for each contest, if the outcome is wrong, there is at least a 75% chance of a full hand count. However, the chance that there is a full hand count of every contest with an incorrect outcome could be less than 75%: PCER is less stringent than FWER. The total bounds on the error are given in column 2. Suppose we design the audits to stop if no more than five nonzero taints of no more than 0.04 are observed; otherwise, the audit progresses to a full hand count. (This way of setting the initial sample size generalizes equation 13; it results in a "staged" audit as proposed in [19].) To control the FWER, the number of draws required is in column 3; the expected number of distinct batches audited in column 4; the expected number of distinct ballots audited in column 5; and the expected number of votes audited is in column 6. To control the PCER, the number of draws is in column 7; the expected number of distinct batches audited in column 8; the expected number of distinct ballots audited in column 9; and the expected number of votes audited is in column 10. The row labeled "all" gives the overall expected number of distinct batches and ballots audited in the three independent audits to control the FWER or the PCER. The row labeled "MACRO" gives the values for a simultaneous audit of all three contests using the maximum across contest relative overstatement of margins, which controls the FWER to be 0.25 or below. Far less work is required than using independent audits the risk to the same level, measured by expected ballots or batches. The expected number of votes is far less than required to control the FWER using independent audits, and only slightly higher than required to control the PCER—even though the MACRO audit controls FWER.

that one or more of the three contests escapes a full hand count can be greater than 0.25 when more than one outcome is wrong. This way of dealing with multiplicity is unfair to MACRO, because MACRO in fact has a lower error rate. Keeping the PCER below 0.25 requires rather smaller sample sizes than keeping the FWER below 0.25.

Table 3 lays out the total error bounds for auditing the three contests separately and the sample sizes that would be needed to stop the audits without a full count if the corresponding samples had at most five taints no larger than 0.04 and the rest of the taints were zero, while keeping the familywise error rate (FWER) or the per-comparison error rate (PCER) under $\alpha = 0.25$.

How much work should we expect to do to audit all three contests separately? Let u_{Ap} denote the error bound for batch p if only contest A is audited. Let $U_A = \sum_{p=1}^{N} u_{Ap}$. Define u_{Bp}, U_B, u_{Cp} and U_C analogously. The expected number of

distinct batches that would be audited in all is

$$\sum_{p=1}^{N}[1-(1-u_{Ap}/U_A)^{n_A}(1-u_{Bp}/U_B)^{n_B}(1-u_{Cp}/U_C)^{n_C}], \tag{19}$$

and the expected number of distinct ballots audited would be

$$\sum_{p=1}^{N} b_p[1-(1-u_{Ap}/U_A)^{n_A}(1-u_{Bp}/U_B)^{n_B}(1-u_{Cp}/U_C)^{n_C}]. \tag{20}$$

For some of those batches of ballots, only one contest would be audited; for some, two contests; and for some, all three. See table 3 for numerical comparisons of MACRO against independent audits that control FWER or PCER. MACRO is much more efficient in this example. Even though MACRO controls FWER, in this example the workload is lower than for independent audits that only control PCER—a less stringent criterion—if work is measured by the number of batches or ballots audited. (The number of votes audited is a bit higher than for independent audits that control PCER, but far lower than for independent audits that control FWER.)

The simultaneous audit using MACRO controls the FWER risk with potentially much less auditing effort than independent audits. The savings will vary by jurisdiction, depending on how ballots are organized and stored, on the margins and error bounds for each contest, on the number of batches, and on the number of votes, because there are logistical costs associated with pulling batches of ballots together for counting, and there are economies in counting all the contests on a single ballot.

5. Application: Contests in Marin and Santa Cruz counties, California, November 2009

This section will be written when the audits are complete, in late November 2009. As of this writing, the contests to be audited have not been selected, but Elaine Ginnold, Registrar of Voters for Marin County, and Gail Pellerin, County Clerk for Santa Cruz County, have confirmed their willingness to participate.

The audits will use the sequential auditing technique presented above in section 3 based on sampling with probability proportional to a bound (PPEB sampling [1]) on the MACRO, defined above in section 2.

6. Summary

A collection of contests can be audited simultaneously using the maximum across-contest relative overstatement (MACRO). Drawing batches using probability proportional to an upper bound on the MACRO—a form of PPEB sampling [1]—and analyzing the results using the Kaplan-Markov bound [20] can lead to reasonably efficient and economical control of the familywise error rate:

the risk that one or more incorrect election outcomes will escape a full hand count. Compared with auditing contests independently to control the risk to the same level, the MACRO approach can reduce the expected number of batches, ballots, and votes that need to be audited. The Kaplan-Markov bound allows batches to be drawn sequentially with no "penalty" for breaking the audit into many stages. That can greatly decrease the auditing burden when the apparent outcomes of the contests are correct.

References

[1] J.A. Aslam, R.A. Popa, and R.L. Rivest. On auditing elections when precincts have different sizes. In *2008 USENIX/ACCURATE Electronic Voting Technology Workshop, San Jose, CA, 28–29 July*, 2008.

[2] L. Breiman. *Probability*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1992.

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

[4] S.E. Fienberg, J. Neter, and R.A. Leitch. Estimating total overstatement error in accounting populations. *J. Am. Stat. Assoc.*, 72:295–302, 1977.

[5] Joseph Lorenzo Hall, Luke W. Miratrix, Philip B. Stark, Melvin Briones, Elaine Ginnold, Freddie Oakley, Martin Peaden, Gail Pellerin, Tom Stanionis, and Tricia Webber. 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.

[6] D. Jefferson, K.Alexander, E. Ginnold, A. Lehmkuhl, K. Midstokke, and P.B. Stark. Post election audit standards report–evaluation of audit sampling models and options for strengthening California's manual count. www.sos.ca.gov/elections/peas/final_peaswg_report.pdf, 2007.

[7] H.M. Kaplan. A method of one-sided nonparametric inference for the mean of a nonnegative population. *The American Statistician*, 41:157–158, 1987.

[8] R.S. Kaplan. Sample size computations for dollar-unit sampling. *J. Accounting Res.*, 13:126–133, 1975.

[9] J. McCarthy, H.I. Stanislevic, M. Lindeman, A. Ash, V. Addona, and M. Batcher. Percentage-based versus statistical-power-based vote tabulation audits. *The American Statistician*, 62:11–16, 2008.

[10] L. Miratrix and P.B. Stark. Sequential audits using the Kaplan-Markov bound. Technical report, Dept. Statistics, Univ. of Calif., Berkeley, 2009.

[11] L. Miratrix and P.B. Stark. The trinomial bound for post-election audits. *IEEE Transactions on Information Forensics and Security*, accepted:tbd, 2009.

[12] J. Neter, R.A. Leitch, and S.E. Fienberg. Dollar unit sampling: Multinomial bounds for total overstatement and understatement errors. *The Accounting Review*, 53:77–93, 1978.

[13] L. Norden, A. Burstein, J.L. Hall, and M. Chen. Post-election audits: restoring trust in elections. Technical report, Brennan Center for Justice, New York University and Samuelson Law, Technology & Public Policy Clinic at University of California, Berkeley School of Law (Boalt Hall), New York, 2007.

[14] Panel on Nonstandard Mixtures of Distributions. *Statistical models and analysis in auditing: A study of statistical models and methods for analyzing nonstandard mixtures of distributions in auditing*. National Academy Press, Washington, D.C., 1988.

[15] R.L. Rivest. On estimating the size of a statistical audit. people.csail.mit.edu/rivest/Rivest-OnEstimatingTheSizeOfAStatisticalAudit.pdf, 2006.

[16] R.G. Saltman. Effective use of computing technology in vote-tallying. Technical Report NBSIR 75-687, National Bureau of Standards, Washington, DC, 1975.

[17] P.B. Stark. Conservative statistical post-election audits. *Ann. Appl. Stat.*, 2:550–581, 2008.

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

[19] P.B. Stark. CAST: Canvass audits by sampling and testing. *IEEE Transactions on Information Forensics and Security, Special Issue on Electronic Voting*, accepted:tbd, 2009.

[20] P.B. Stark. Risk-limiting post-election audits: p-values from common probability inequalities. *IEEE Transactions on Information Forensics and Security*, accepted:tbd, 2009.

[21] K.W. Stringer. Practical aspects of statistical sampling in auditing. In *Proceedings of the Business and Economic Statistics Section*, pages 405–411, Washington, D.C., 1963. American Statistical Association.