A Gentle Introduction to Risk-Limiting Audits

Mark Lindeman
Philip B. Stark | University of California, Berkeley

Risk-limiting audits provide statistical assurance that election outcomes are correct by manually examining portions of the audit trail—paper ballots or voter-verifiable paper records. This article sketches two types of risk-limiting audits, ballot-polling audits and comparison audits, and gives example computations. These audits do not require in-house statistical expertise.

A risk-limiting audit is a method to confirm that the hardware, software, and procedures used to tally votes in an election found the real winners. Risk-limiting audits don't guarantee that the electoral outcome is right, but they have a large chance of correcting the outcome if it's wrong. They involve manually examining portions of an audit trail of (generally paper) records that are voter verifiable: voters had the opportunity to verify that the records recorded their selections accurately. Systems that don't produce voter-verifiable paper records (VVPRs), such as paperless touchscreen voting systems, can't be audited this way.

Risk-limiting audits address the limitations and vulnerabilities of voting technology, including possible flaws in algorithms used to infer voter intent, configuration and programming errors, mechanical problems, and malicious subversion. Computer software can't be guaranteed to be perfect or secure, so voting systems should be *software independent*: an undetected change or error in voting system software should be incapable of causing an undetectable change or error in an election outcome.[1] Well-curated audit trails provide software independence; risk-limiting audits leverage this software independence by checking the audit trails strategically. Indeed, risk-limiting audits can correct erroneous outcomes, no matter what caused the error, whenever the audit trail reflects the correct outcome. There is extensive literature on postelection audits; we don't summarize it here. And we omit important implementation details. Our point is merely that efficient risk-limiting audits don't require complicated calculations or in-house statistical expertise.

Risk Limits

The simplest risk-limiting audit is a full hand tally of a reliable audit trail; such a count, if accurate, reveals the correct outcome. However, a full hand count generally wastes resources; examining far fewer ballots can provide strong evidence that the outcome is correct, if those ballots are chosen at random by suitable means. Hence, to keep the counting burden as low as possible, the methods we describe here conduct an "intelligent" incremental recount that stops when the audit provides sufficiently strong evidence that a full hand count would confirm the outcome produced by the voting system. As long as the audit doesn't yield sufficiently strong evidence, more

Copublished by the IEEE Computer and Reliability Societies

ballots are manually examined, potentially progressing to a full hand tally of all the ballots. (The full hand count can be part of the audit or a separate process.) The criterion "sufficiently strong" is quantified by the *risk limit*—the largest chance that the audit will stop short of a full hand tally when the original outcome is in fact wrong, no matter why it's wrong, including voter errors, configuration errors, bugs, equipment failures, and deliberate fraud.

Smaller risk limits entail stronger evidence that the outcome is correct. All else equal, a risk-limiting audit examines more ballots if the risk limit is 1 percent than if it's 10 percent. Smaller (percentage) margins between candidates require more evidence because there's less room for error. All else equal, the audit examines more ballots if the margin is 1 percent than if it's 10 percent.

The risk limit is sometimes misconstrued as the chance that the final outcome (after auditing) is wrong. A risk-limiting audit emends the outcome only if it leads to a full hand tally that disagrees with the original outcome. Hence, risk-limiting audits can't harm correct outcomes. But if the original outcome is wrong, there's a chance the audit won't correct it. The risk limit is the largest such chance. If the risk limit is 10 percent and the outcome is wrong, there is at most a 10 percent chance (and typically much less) that the audit won't correct the outcome—at least a 90 percent chance (and typically much more) that the audit will correct the outcome.

Audit Trails

Risk-limiting audits involve manually interpreting votes in portions of the audit trail. The best audit trail is voter-marked paper ballots. VVPRs printed by voting machines aren't as good. Voters might not actually inspect VVPRs. Printers can jam or run out of paper. VVPRs can be fragile and cumbersome to audit. (As we noted, paperless touchscreen voting machines don't provide a suitable audit trail.) In this article, we call audit trail entries "ballots" regardless of how they were created. Like a recount, a risk-limiting audit assumes there is a correct interpretation of each ballot. Rules for interpreting ballots must be established before the audit starts.

Ballot-Level Audits

States that mandate hand counting as part of audits generally require counting votes in selected ballot *clusters*. For instance, under California law, each county counts the votes in 1 percent of precincts; each cluster comprises the ballots cast in one precinct.

The smaller the clusters, the less counting a risk-limiting audit requires, if the outcome is correct. (If the outcome is wrong, the audit has a large chance of counting all the votes, regardless of the cluster sizes.) A random sample of 100 individual ballots can be almost as informative as a random sample of 100 entire precincts! Hand counting is minimized when clusters consist of one ballot each, yielding *ballot-level audits*.[2]

Ballot-level audits save work, but finding individual ballots among millions stored in numerous physical *batches*, such as boxes or bags, is challenging. It requires knowing the number of ballots in each batch (that is, having a *manifest*), how to locate each batch, and how to identify each ballot in each batch uniquely. Labeling each ballot helps but is prohibited in some jurisdictions. Ballot-level auditing elevates privacy concerns. The most efficient ballot-level audits—*comparison audits*—require the voting system interpretation of every ballot, which no federally certified vote-tabulation system reports.[3]

If a voting system doesn't report its interpretation of each ballot, auditors can perform a *transitive audit*, using an unofficial system that does.[4] If the two systems show different outcomes, all votes should be hand counted. If the systems show the same outcome, a risk-limiting audit of the unofficial system checks the outcome of the system of record: both are right or both are wrong. If both are wrong, the risk-limiting audit has a large chance of requiring a full hand count.[5]

Before the Audit

Because a risk-limiting audit relies on the audit trail, preserving the audit trail complete and intact is crucial. If a jurisdiction's procedures for protecting the audit trail are adequate in principle, ensuring compliance with those procedures (possibly as part of a comprehensive canvass or a separate compliance audit) can provide strong evidence that the audit trail is trustworthy. If the compliance audit doesn't generate convincing affirmative evidence that ballots haven't been altered, added, or lost, a risk-limiting audit might be mere theater.[3,5]

Sampling ballots efficiently requires a ballot manifest that describes in detail how the ballots are organized and stored. For instance, a jurisdiction might keep cast ballots in 350 batches, labeled 1 to 350. The manifest might say, "There are 71,026 ballots in 350 batches: batch 1 has 227 ballots; batch 2 has 903 ballots; … ; and batch 350 has 114 ballots." If the jurisdiction numbers its ballots, the manifest might say, "Batch 1 contains ballots 1–227; batch 2 contains ballots 228–1,130; … ; and batch 350 contains ballots 70,913–71,026."

Auditors should verify that the number of ballots in the manifest matches the total according to the election results. It's good practice to count the ballots in the batches containing the ballots selected for audit to check whether the manifest is accurate. If the manifest is inaccurate, the risk limit might be incorrect.

Two Simple Risk-Limiting Audits

We present simple examples of two kinds of risk-limiting audits: ballot-polling and comparison audits. (Kenneth

Johnson distinguishes between two similar kinds of audits, although he doesn't address risk-limiting audits per se.[6]) "Simple" means that the calculations are easy, even with a pencil and paper, so observers can check the auditors' work. Tools that perform these calculations are available at http://statistics.berkeley.edu/~stark/Vote/auditTools.htm.

Ballot-Polling Audits

A ballot-polling audit examines a random sample of ballots. When the vote shares in the sample give sufficiently strong evidence that the reported winner really won, the audit stops. Ballot-polling audits require knowing who reportedly won, but no other data from the vote-tabulation system. They are best when the vote-tabulation system can't export vote counts for individual ballots or clusters of ballots, or when retrieving the ballots that correspond to such counts is impractical. The following ballot-polling audit, which relies on Abraham Wald's sequential probability ratio test,[7] has a risk limit of 10 percent: there's at least a 90 percent chance it will require a full hand count if the reported winner actually lost. It assumes that the winner's reported share of the valid votes is greater than 50 percent—a majority rather than a mere plurality:

1. Let s be the winner's share of the valid votes according to the vote-tabulation system; this procedure requires s to be greater than 50 percent. Let t be a positive number small enough that when t is subtracted from s, the difference is still greater than 50 percent. (Increasing t reduces the chance of a full hand count if the voting system outcome is correct but increases the expected number of ballots to be counted during the audit.) Set the test statistic T to 1. The audit ends when T becomes large enough or small enough.
2. Select a ballot at random (a ballot can be selected more than once). The following steps apply each time.
3. If the ballot doesn't show a valid vote, return to step 2.
4. If the ballot shows a valid vote for the winner, multiply T by $2(s - t)$.
5. If the ballot shows a valid vote for anyone else, multiply T by $2(1 - (s - t))$.
6. If T is greater than 9.9, the audit has provided strong evidence that the reported outcome is correct. Stop.
7. If T is less than 0.011, perform a full hand count to determine who won. Otherwise, return to step 2.

If the reported winner's true share of the vote is at least $s - t$, there is at most a 1 percent chance that this procedure will lead to a full hand count; that chance and the risk limit can be altered by adjusting the comparisons in steps 6 and 7.

As a numerical example, suppose one candidate reportedly received 60 percent of the valid votes. Set t to 1 percent. If the reported winner really received at least $s - t = 59$ percent of the vote, there is at most a 1 percent chance that the procedure will lead to a (pointless) full hand count. Note that $1 - (s - t) = 1 - 0.59 = 41$ percent. To audit, repeat steps 2–7, drawing ballots at random and updating T until it's either greater than 9.9 or less than 0.011.

The number of ballots audited depends on the vote shares and on which ballots happen to be selected. If the first 14 ballots drawn all show votes for the winner, then

$$T = 1 \times (2 \times 0.59) \times (2 \times 0.59) \times \ldots \times (2 \times 0.59) = (2 \times 0.59)^{14} = 10.15,$$

and the audit stops.

If the reported winner's true vote share is 60 percent, the audit is expected to examine 120 ballots; for a 55 percent share, 480; and for a 52 percent share, 3,860. The expected workload grows quickly as the reported winner's share decreases.

When the outcome is correct, the number of ballots the audit examines depends only weakly on the number of ballots cast, so the percentage of ballots examined in large contests can be quite small. For example, in the 2008 presidential election, 13.7 million ballots were cast in California; Barack Obama was reported to have received 61.1 percent of the vote. A ballot-polling audit could confirm that Obama won California at 10 percent risk (with t equal to 1 percent) by auditing roughly 97 ballots—0.0007 of 1 percent of the ballots cast—if Obama really received more than 61 percent of the votes.

Each county's expected auditing workload is proportional to the percentage of ballots cast in the county. Almost 25 percent of the ballots were cast in Los Angeles County, the largest (in ballots cast) of California's 58 counties. More than 75 percent of the ballots were cast in the largest 12 counties. The 14 smallest counties together accounted for less than 1 percent of ballots cast. So, approximately 24 of the 97 ballots would be from Los Angeles; 73 from the 12 largest counties, including Los Angeles; and perhaps one ballot total from the 14 smallest counties.

If the winner's share were 52 percent rather than 61.1 percent, the expected number of ballots to examine would be 3,860—far more than 97 but still less than 0.0003 percent of the ballots cast. Of those, Los Angeles would have expected to examine approximately 946; the 12 largest counties, approximately 2,922; and the 14 least populous counties, approximately 35. Because ballot-polling audits don't require data from the vote-tabulation system, they're an immediate practical option for auditing large contests. Indeed, auditors could confirm all statewide contests with a single ballot-polling audit expected to examine 3,860

ballots if the winner of every contest actually received at least 52 percent of the valid votes. Comparison audits generally involve examining fewer ballots but require much more of the vote-tabulation system.

Comparison Audits

Comparison audits check outcomes by comparing hand counts to voting system counts of the votes in ballot clusters. In ballot-level comparison audits, each cluster is one ballot. Comparison audits can be thought of as having two phases. First, auditors check whether the voting system subtotals for every cluster of ballots sum to the contest totals for every candidate. If the subtotals don't add up to the contest totals, the reported results are inconsistent; the audit can't proceed. Second, auditors spot-check the voting system subtotals against hand counts for randomly selected clusters to assess whether the subtotals are sufficiently accurate to determine who won. If not, the audit has a large chance of requiring a full hand count.

This section uses a variant of the "super simple" ballot-level risk-limiting comparison audit.[8] It presumes the auditors know how the vote-tabulation system (or, for transitive audits, an unofficial system) interpreted each ballot. The audit compares a manual interpretation of ballots selected at random to the voting system interpretation of those ballots, continuing until there is strong evidence that the outcome is correct—or leading to a full hand count that determines the outcome.

Suppose the manual interpretation of a ballot differs from the voting system interpretation. If changing the voting system interpretation to match the manual interpretation would increase the margins between the winner and every loser, the ballot has an *understatement*. (There are as many margins as there are losers; an understatement, by definition, affects every margin.) For instance, if the voting system records an overvote but the manual interpretation shows a vote for the winner, the ballot has an understatement. Understatements don't call the outcome into question because correcting them benefits the winner.

If changing the voting system interpretation to match the manual interpretation would decrease the margin between the winner and any loser, the ballot has an *overstatement* equal to the maximum number of votes by which any margin would decrease. If the voting system records an undervote but the manual interpretation finds a vote for one of the losers, the ballot has an overstatement of one vote. If the voting system records a vote for the winner but the manual interpretation finds an overvote, that ballot has an overstatement of one vote.

If the voting system interprets a ballot as a vote for the winner, but a manual interpretation finds a vote for one of the losers, that ballot has an overstatement of two votes. For voter-marked paper ballots, occasional one-vote misstatements are expected, owing to the vagaries of how voters mark their ballots. From time to time, the system will interpret a light mark as an undervote or a hesitation mark as an overvote. But two-vote overstatements should be quite rare—a properly functioning voting system should not award a vote for one candidate to a different candidate.

To have an overstatement, it is enough for the margin between the winner and *any* loser to have been reported incorrectly, but to have an understatement, the margins between the winner and *every* loser need to have been reported incorrectly.

We present a simple rule for a risk-limiting comparison audit with a 10 percent risk limit. The rule depends on the diluted margin, m—the smallest reported margin (in votes) divided by the number of ballots cast. Dividing by the number of ballots, rather than by the number of valid votes, allows for the possibility that the vote-tabulation system mistook an undervote or overvote for a valid vote, or vice versa. Suppose the audit has examined n ballots (see the "Random Selection" section in this article). Let u_1 and o_1 be the number of one-vote understatements and overstatements, respectively, among those n ballots; similarly, let u_2 and o_2 be the number of two-vote understatements and overstatements. The audit can stop if

$$n \geq (4.8 + 1.4\,(o_1 + 5o_2 - 0.6\,u_1 - 4.4\,u_2))/m. \qquad (1)$$

(This follows from Equation 9 in "Super-Simple Simultaneous Single-Ballot Risk-Limiting Audits," with a risk limit of $\alpha = 10$ percent and $\gamma = 1.03905$, by the same conservative approximation used to derive Inequality 17 there, with a bit of rounding.[8])

Overstatements increase the required sample size and understatements decrease it, but not by equal amounts. We have more confidence in the outcome if the sample shows no misstatements than if it shows large but equal numbers of understatements and overstatements. In Inequality 1, a one-vote understatement offsets 60 percent of a one-vote overstatement, and a two-vote understatement offsets 88 percent of a two-vote overstatement.

If the diluted margin is 10 percent, each one-vote overstatement increases the required sample size by 1.4/.10, or 14 ballots, and each one-vote understatement decreases the required sample size by 1.4 × 0.6/.10, or 8.4 ballots. Each two-vote overstatement increases the required sample size by 1.4 × 5/.10, or 70 ballots, and each two-vote understatement decreases the required sample size by 1.4 × 4.4/.10, or 61.6 ballots. For a diluted margin of 5 percent, these numbers double; for 2 percent, they quintuple.

With this method, auditors can check one ballot at a time against its voting system interpretation sequentially, or check a larger number in parallel. Moreover, auditors can decide at any point to abort the audit and require a full hand count. The risk limit will be 10 percent, provided the audit continues either until Inequality 1 is satisfied or until there is a full hand count. If there is a full hand count, the hand count outcome replaces the reported outcome.

Suppose 10,000 ballots were cast in a particular contest. According to the vote-tabulation system, the reported winner received 4,000 votes and the runner-up received 3,500 votes. Then the diluted margin is $m = (4{,}000 - 3{,}500)/10{,}000$, or 5 percent. We consider sampling ballots incrementally and in stages.

Sampling incrementally. In an incremental audit, auditors draw a ballot at random and check by hand whether the voting system interpretation of that ballot is correct before drawing the next ballot. If there is a one-vote understatement and no other misstatements among the first 80 ballots examined, u_1 is 1 and o_1, u_2, and o_2 are all 0, and the audit can stop, because

$$80 \geq (4.8 - 1.4 \times 0.6 \times 1)/0.05. \quad (2)$$

In our example, if there are no overstatements or understatements among the first 96 ballots examined, u_1, o_1, u_2, and o_2 are all 0, and the audit can stop because

$$96 \geq 4.8/0.05. \quad (3)$$

Sampling in stages. To simplify logistics, auditors might draw many ballots at once, with replacement (see the "Random Selection" section), and then compare each to its voting system interpretation. If Inequality 1 doesn't hold, the auditors draw and compare another set of ballots. Each set of draws and comparisons is a *stage*.

If auditors expect errors at a particular rate, they can select the first-stage sample size so the audit stops there if the expectation proves correct or pessimistic. Suppose they expect a one-vote overstatement and a one-vote understatement per thousand ballots (0.001 per ballot) and expect two-vote misstatements to be negligibly rare. In our example with a diluted margin of 5 percent, auditors could use an initial sample of $4.8/m$ ballots (rounded up) or 96 ballots. If overstatements are as infrequent as expected, there are unlikely to be any among the first 96 ballots: the audit will stop at the first stage. More conservatively, an initial sample of $6.2/m$ ballots (in our example, 124 ballots) allows the audit to stop at the first stage if it shows a one-vote overstatement.

Sorting the sample (for instance, by precinct) before retrieving the ballots and checking their interpretation can save some effort. But then all ballots drawn in the stage should be checked before determining whether to stop. Otherwise, the procedure is biased in favor of ballots from precincts that are early in the sorted order.

Table 1 gives stopping sample sizes for various diluted margins and numbers of overstatements and understatements, for 10 percent risk. It can help auditors select the first-stage sample size for different expected rates of error.

Random Selection

Risk-limiting audits rely on random sampling. (Random samples can be augmented with "targeted" samples chosen by other means; see "CAST: Canvass Audits by Sampling and Testing."[9]) If the sample isn't drawn appropriately, the risk limit will be wrong. The risk-limiting methods we describe rely on drawing a random sample of ballots with replacement. This is like putting all the ballots into an enormous mixer, stirring them thoroughly, and drawing a ballot without looking. The ballot is returned to the mixer, the ballots are mixed again, and another ballot is drawn (possibly the same ballot). This procedure is repeated until the audit stops. If a ballot is drawn more than once, it enters the calculations as many times as it is drawn. So a sample of 200 ballots might contain 198 different ballots, two of which are counted twice in the calculations.

Public confidence requires that observers can verify that the selection is fair—that all ballots are equally likely to be selected in each draw. This speaks against a number of common methods for selecting samples, including "arbitrary" selection by election officials; drawing slips of paper, where there is little hope of confirming that each ballot is represented by exactly one slip and that the slips have been adequately mixed; using proprietary software such as Excel; or using any source of putative randomness that can't readily be checked.

Trustworthy methods of generating random numbers often have two features: a physical source of randomness and input from multiple parties (so that even if some parties collude, any noncolluding party could foil an attempt to rig the sample). An efficient, effective, and transparent approach is to use a simple mechanical method—such as rolling dice[10]—to generate a "seed" for a well-designed pseudorandom number generator (PRNG). PRNGs can generate arbitrarily many pseudorandom numbers from a single seed. PRNG output is deterministic given the seed, but the numbers produced by good PRNGs have many of the desirable properties of random sequences. And any observer who knows the seed and the PRNG can check the output. For good PRNGs, small changes in the seed yield very different sequences, so starting with a random seed makes it effectively impossible for anyone to render the audit less effective by anticipating which ballots will be examined.

Table 1. Sample sizes for ballot-level comparison audits at a 10 percent risk limit.

Diluted margin (%)	0 understatements					1 one-vote understatement				
	Number of one-vote overstatements					Number of one-vote overstatements				
	0	1	2	3	4	0	1	2	3	4
0.2	2,400	3,100	3,800	4,500	5,200	1,980	2,680	3,380	4,080	4,780
0.5	960	1,240	1,520	1,800	2,080	792	1,072	1,352	1,632	1,912
1	480	620	760	900	1,040	396	536	676	816	956
2	240	310	380	450	520	198	268	338	408	478
5	96	124	152	180	208	80	108	136	164	192
10	48	62	76	90	104	40	54	68	82	96
20	24	31	38	45	52	20	27	34	41	48

The auditTools page (http://statistics.berkeley.edu/~stark/Vote/auditTools.htm) provides a good PRNG suggested by Ronald L. Rivest. It relies on the SHA-256 cryptographic hash function, which is in the public domain and has been implemented in many programming languages. This allows observers to confirm that the sequence of pseudorandom numbers is correct, given the seed.

A ballot manifest can be used to identify the particular ballots that correspond to the random (or pseudorandom) numbers. Before the audit, auditors use the manifest to assign a unique number to each ballot, if the ballots aren't already marked uniquely. Suppose the manifest lists 822 ballots in three batches, numbered 1 through 3; the batches contain 230, 312, and 280 ballots, respectively. Auditors can consider the 230 ballots in batch 1 to be ballots 1 through 230, the 312 ballots in batch 2 to be ballots 231 through 542, and the 280 ballots in batch 3 to be ballots 543 through 822. Ballot 254 is the 24th ballot in batch 2. We assume that the ballots are stored in an order that doesn't change during the audit, so that "the 24th ballot in batch 2" uniquely identifies a particular ballot.

To draw the audit sample, auditors generate random numbers between 1 and 822 and retrieve the corresponding ballot. If 254 is generated, they retrieve batch 2, find the 24th ballot in that batch, and audit that ballot.

More Complicated Situations

We've discussed only contests in which the candidate with the most votes wins. The methods can be extended to audit a collection of contests simultaneously with a single sample, and to audit contests that require a supermajority, contests with more than one winner, cross-jurisdictional contests, and ranked-choice voting contests.

Contests with more than one winner and collections of contests can be audited with a comparison audit on the basis of the maximum relative overstatement (MRO) of pairwise margins.[11,12] A pairwise margin is the margin in votes between any winner and any loser in a given contest. An overstatement of a pairwise margin, divided by that margin, is the relative overstatement of the pairwise margin. A one-vote overstatement of a wide margin casts less doubt on the outcome than a one-vote overstatement of a narrow margin; relative overstatements take this into account. The MRO is the maximum relative overstatement on each audited ballot. The arithmetic can be simplified by treating all overstatements as if they affected the smallest diluted margin. This is conservative, but if overstatements are rare, the workload remains manageable. That simplification is the heart of the "super simple" audit method.[8]

For simultaneous audits of multiple contests, the diluted margin is the smallest reported margin in votes, divided by the total number of ballots on which at least one of the contests appears. If a contest appears on only a small fraction of ballots, it might take less work to audit it separately so that its diluted margin considers only the ballots containing the contest.

If comparison audits are infeasible, contests with more than one winner and collections of contests can be audited with ballot-polling audits.[13]

Auditing contests that cross jurisdictional boundaries is straightforward if all the results are available before the audit starts and the sample can be drawn from all ballots as a pool. If the jurisdictions draw samples independently, the computations are complicated.[14] Auditing instant-runoff or ranked-choice contests is a topic of research: even computing the margin of victory is difficult.[15]

A Practical Example: Merced County, California

The methods we describe have been used to audit elections in California, including the November 2011

election in Merced County. That audit, authorized by California's 2010 law AB 2023 and funded by a grant from the US Election Assistance Commission, was a comparison audit that used a single sample to confirm two City of Merced contests: the mayoral contest, and the (vote-for-three) councilmember contest. In the mayoral contest, which had five candidates, the voting system reported that Stan Thurston received 2,231 votes and runner-up Bill Blake received 2,037—a margin of 194 votes, or 2.79 percent of valid votes cast. In the councilmember contest, the margin of decision (between the third- and fourth-place candidates) was wider—959 votes.

Because Merced's voting system can't report its interpretation of individual ballots, a transitive audit was conducted: the county captured digital images of the 7,120 cast ballots and prepared a ballot manifest. Kai Wang, PhD student at the University of California, San Diego, interpreted the images using software he wrote, spot-checking difficult cases by hand. His vote totals were slightly higher than the official totals but gave the same winners. The margin he found for the mayoral contest was 192 votes, a diluted margin of approximately 2.70 percent. Before the audit started, the unofficial interpretations were posted to a website so that anyone interested could verify that those interpretations didn't change during the audit.

The initial sample was large enough to confirm the original results at a 10 percent risk limit if it revealed few overstatements. The minimum sample size if there were no misstatements would be $4.8/m$, or 178. The initial sample size was chosen on the assumptions that the rates of one-vote overstatements and understatements would be 0.001, rounded up to the nearest whole number, and that the rates of two-vote overstatements and understatements would be negligible. This led the auditors to anticipate a one-vote overstatement and a one-vote understatement in the sample. Inequality 1 with $o_1 = 1$ and $u_1 = 1$ yields

$$n \geq (4.8 + 1.4 \times (1 - 0.6 \times 1))/0.027 = 198.5. \quad (4)$$

Inequality 1 rounds to the nearest tenth, but the auditTools page does not; the initial sample was 198 ballots. To allow for a one-vote overstatement without any compensating one-vote understatement, the initial sample size would be 230 instead: when o_1 is 1 and u_1, o_1, and o_2 are 0, we need $n \geq (4.8 + 1.4 \times 1)/0.027 = 229.6$.

Each of the four people present contributed two digits to a seed, which was used with the PRNG on the auditTools page to generate 198 numbers between 1 and 7,120, the number of ballots. Auditors retrieved each of the corresponding ballots using the manifest and the lookup tool on the auditTools page. Their manual interpretation of each ballot matched Kai Wang's interpretation, so the audit stopped, transitively confirming the official winners of both contests at a 10 percent risk limit by looking at 198 ballots.

While the mathematics that underlies risk-limiting audits might be daunting, the calculations required to conduct the audit can be extremely simple: arithmetic that could easily be done with pencil and paper or a four-function calculator. Simplicity improves transparency and can increase public confidence by allowing anyone interested to check the calculations.

Seventeen states will conduct statewide audits of the 2012 presidential election; several more states will conduct partial audits. These audits vary widely in quality; no state currently requires risk-limiting audits. (Colorado law requires risk-limiting audits beginning in 2014.) Many states need to upgrade their voting systems to provide audit trails, preferably auditable at the ballot level. By 2016, all states could implement risk-limiting audits of the presidential election and, at least, other major contests. That step would be a giant leap for election verification in the US. ■

Acknowledgments

We're grateful to Jennie Bretschneider, Ronald L. Rivest, and Barbara Simons for their helpful comments.

References

1. R. Rivest, "On the Notion of 'Software Independence' in Voting Systems," *Philosophical Trans. Royal Soc. A*, vol. 366, no. 1881, 2008, pp. 3759–3767.
2. P. Stark, "Risk-Limiting Vote-Tabulation Audits: The Importance of Cluster Size," *Chance*, vol. 23, no. 3, 2010, pp. 9–12.
3. P.B. Stark and D.A. Wagner, "Evidence-Based Elections," *IEEE Security & Privacy*, vol. 10, no. 5, 2012, pp. 33–41.
4. J. Calandrino, J. Halderman, and E. Felten, "Machine-Assisted Election Auditing," *Proc. 2007 Usenix/Accurate Electronic Voting Technology Workshop* (EVT 07), Usenix Assoc., 2007; www.usenix.org/event/evt07/tech/full_papers/calandrino/calandrino.pdf.
5. J. Benaloh et al., "SOBA: Secrecy-Preserving Observable Ballot-Level Audit," *Proc. 2011 Electronic Voting Technology Workshop/Workshop Trustworthy Elections* (EVT/WOTE 11), Usenix Assoc., 2011; www.usenix.org/event/evtwote11/tech/final_files/Benaloh.pdf.
6. K. Johnson, "Election Certification by Statistical Audit of Voter-Verified Paper Ballots," 2004; http://papers.ssrn.com/sol3/papers.cfm?abstract_id=640943.
7. A. Wald, "Sequential Tests of Statistical Hypotheses," *Annals of Mathematical Statistics*, vol. 16, no. 2, 1945, pp. 117–186.
8. P. Stark, "Super-Simple Simultaneous Single-Ballot

Risk-Limiting Audits," *Proc. 2010 Electronic Voting Technology Workshop/Workshop Trustworthy Elections* (EVT/WOTE 10), Usenix Assoc., 2010; www.usenix.org/events/evtwote10/tech/full_papers/Stark.pdf.
9. P. Stark, "CAST: Canvass Audits by Sampling and Testing," *IEEE Trans. Information Forensics and Security*, vol. 4, no. 4, 2009, pp. 708–717.
10. A. Cordero, D. Wagner, and D. Dill, "The Role of Dice in Election Audits—Extended Abstract," IAVoSS Workshop Trustworthy Elections (WOTE 06), 2006; www.cs.berkeley.edu/~daw/papers/dice-wote06.pdf.
11. P. Stark, "A Sharper Discrepancy Measure for Post-Election Audits," *Annals of Applied Statistics*, vol. 2, no. 3, 2008, pp. 982–985.
12. P. Stark, "Efficient Post-Election Audits of Multiple Contests: 2009 California Tests," *Proc. 2009 Conf. Empirical Legal Studies*, 3 Aug. 2009; http://ssrn.com/abstract=1443314.
13. M. Lindeman, P. Stark, and V. Yates, "BRAVO: Ballot-Polling Risk-Limiting Audits to Verify Outcomes," *Proc. 2012 Electronic Voting Technology Workshop/Workshop Trustworthy Elections* (EVT/WOTE 12), Usenix Assoc., 2012; www.usenix.org/system/files/conference/evtwote12/evtwote12-final27.pdf.
14. M. Higgins, R. Rivest, and P. Stark, "Sharper P-Values for Stratified Post-Election Audits," *Statistics, Politics, and Policy*, vol. 2, no. 1, 2011, art. 7.
15. T. Magrino et al., "Computing the Margin of Victory in IRV Elections," *Proc. 2011 Electronic Voting Technology Workshop/Workshop Trustworthy Elections* (EVT/WOTE 11), Usenix Assoc., 2001; www.usenix.org/event/evtwote11/tech/final_files/Magrino.pdf.

Mark Lindeman is an unaffiliated political scientist. His research interests include public opinion, political behavior, and election verification issues. Lindeman has a PhD in political science from Columbia University. He was an executive editor of *Principles and Best Practices for Post-Election Audits* and coauthor of *Public Opinion* (Westview). Contact him at taxshift@gmail.com.

Philip B. Stark is a professor and chair of the Department of Statistics at the University of California, Berkeley, and is working with the California and Colorado Secretaries of State on pilot risk-limiting audits. His research interests include uncertainty quantification, risk assessment, and nonparametric inference. Stark has a PhD in Earth science from the University of California, San Diego. He's a member of the Institute for Mathematical Statistics, the American Statistical Association, the Bernoulli Society, the Institute of Physics, and the Royal Astronomical Society. Contact him at stark@stat.berkeley.edu.