CAST: Canvass Audits by Sampling and Testing

Philip B. Stark

Abstract—Canvass Audit by Sampling and Testing (CAST) is a method to decide on the basis of a hand count of a random sample of batches of ballots whether to count the entire audit trail of a contest by hand. Data are collected in stages. At stage s, one computes P_s, an upper bound on the P-value of the hypothesis that the apparent outcome is wrong, conditional on the results of previous audit stages. If P_s is less than a threshold α_s—strong evidence that the apparent outcome is correct—the audit stops. Otherwise, the audit advances to stage $s+1$. If the audit has not stopped by stage $s = S$, there is a full hand count. The thresholds $\{\alpha_s\}_{s=1}^{S}$ are chosen to guarantee that whenever the apparent outcome is wrong, the chance of a full hand count is at least $1 - \alpha$: The risk of certifying an incorrect outcome is at most α. Conditioning allows deliberate selection of some batches of ballots for audit. CAST can use any sampling and testing scheme for which upper bounds on conditional P-values can be calculated. A turn-key implementation for proportionally stratified random samples is given, using the maximum relative overstatement of pairwise margins as the test statistic. This implementation works with contests that cross jurisdictional boundaries and contests that allow voters to select more than one candidate.

Index Terms—Auditing, elections, hypothesis test, P-value, risk-limiting audit, sequential test, stratified sampling.

I. INTRODUCTION

E LECTION systems are complex: They can involve voting machines, central tabulators, optical scanners, memory cards and readers, software, designers, printers, programmers, operators, and pollworkers, elaborate chains of custody—and human voters. Error is inevitable. In a given contest, did a losing candidate or position appear to win? This question can be answered by a postelection audit, a hand count of the audit trail for a random sample of batches of ballots.

"Apparent outcome," "semiofficial outcome," and "preliminary official outcome" all mean the outcome once the election officials represent that all votes have been counted. "Wrong" means that the apparent outcome disagrees with the outcome a full hand count of the audit trail would show. Ensuring the integrity of the audit trail is vital: A full hand count need not show the true outcome if the audit trail is incomplete or inaccurate, or if the hand count has errors. Nonetheless, a hand count is generally the legal standard—by definition, a full hand count gives the true outcome.[1]

[1]Agreement between hand counts of audit batches and semiofficial results does not ensure that the batch results were aggregated correctly. It is important to ensure that the numbers add up. Checking those sums is not required by any current audit law, as far as I know.

A post-election audit can be designed so that if the apparent outcome is wrong for any reason—including human error, bugs, mechanical failures, or fraud—the probability that the audit will require a full hand count is at least $1 - \alpha$. That limits the risk of certifying an incorrect outcome to at most α.

This paper gives a general recipe for risk-limiting audits: Canvass Audit by Sampling and Testing (CAST). It refines the approach of [1] and [2], the first method for risk-limiting post election audits; see Section V-C. CAST was used to audit contests in Marin County, CA, in February 2008, and contests in Marin, Santa Cruz, and Yolo counties, CA in November 2008 [3], [4]. When the apparent outcome is correct, CAST tries to keep the amount of hand counting low.[2] The auditor can select the sampling design, the test statistic, the number of stages, the total permissible risk, the tradeoff of risk at different stages, batches for targeted auditing, and other variables.

Section II lays out the "sample and conditional test" strategy. Section III gives a step-by-step implementation of CAST for proportional stratified random sampling using the maximum relative overstatement of pairwise margins [2] as the test statistic.

Section IV illustrates CAST using a cartoon of a U.S. House of Representatives contest. Section V discusses technical details. Proofs relating to worst-case behavior are in the Appendix.

II. THE STRATEGY

One contest is audited at a time.[3] The auditor wants to guarantee that the chance of a complete hand count is at least $1 - \alpha$ if the apparent outcome is wrong. The auditor is willing to perform at most S stages of sampling before a full hand count—if the audit has not stopped before then.

Suppose that, if the apparent outcome is wrong and the audit gets to stage s, the chance that the audit progresses from stage s to stage $s + 1$ is at least $1 - \alpha_s$, given the results observed at all previous stages. The minimum chance that the audit progresses to a full hand count is

$$(1 - \alpha_1)(1 - \alpha_2) \cdots (1 - \alpha_S). \qquad (1)$$

If $\prod_{s=1}^{S}(1 - \alpha_s) \geq 1 - \alpha$, the chance that the audit will lead to a complete hand count when the apparent outcome is wrong is at least $1 - \alpha$.

Batches of ballots[4] may be selected deliberately for audit ("targeted") provided the conditional probability is at least $1 - \alpha_s$ that the audit will progress from stage s to stage $s + 1$ if the

[2]Randomly counting $1 - \alpha$ of all contests by hand catches all but α of the incorrect outcomes, but is inefficient. See [3].

[3]But see [5].

[4]A batch could comprise all the ballots cast in a precinct, the ballots cast on a particular machine, or other convenient group for which a semiofficial subtotal is available.

outcome is wrong, given the audit data from previous stages.[5] At each stage, the reported votes in a random sample of batches of ballots are compared with hand counts of the audit trail for those batches. Over all ways the apparent outcome could differ from the true outcome, an upper bound is found for the maximum chance that the current stage of audit would discover "as little"[6] error as it did. That bound, P_s, is an upper bound on the P-value of the hypothesis that the apparent outcome is wrong, conditional on the results of the audit at stages prior to stage s. If $P_s < \alpha_s$, the audit stops.

If not, some batches may be selected for targeted auditing; the margin is adjusted for all errors seen so far; s is incremented; a new random sample is drawn from the batches not yet audited; and P_s is calculated.[7] If the outcome has not been certified by stage S, the rest of the audit trail is counted by hand, so the correct outcome is known.

For example, suppose the auditor seeks to guarantee at least a 90% chance (risk $\alpha \leq 10\%$) of a full hand count whenever the count would show that the semiofficial outcome is wrong. The auditor contemplates drawing an initial sample, a second sample if necessary, and then counting all ballots by hand if there is not strong evidence that the outcome is correct. Then $S = 2$ and the auditor could take $1 - \alpha_1 = 1 - \alpha_2 = \sqrt{90\%} = 94.9\%$.

Alternatively, the auditor could set $1 - \alpha_1 = 91\%$ and $1 - \alpha_2 = 98.9\%$, since $0.91 \times 0.989 = 0.9$. Taking α_s larger for early audit stages and smaller for later stages can reduce the overall audit burden when the apparent outcome is correct, by reducing the chance that the audit progresses beyond the first stage.

The larger S is, generally the smaller the individual values of α_s must be to ensure a high chance of a full hand count whenever the apparent outcome is wrong: Each additional stage introduces another opportunity to err by stopping short of full hand count when the outcome is wrong. This is one reason to keep the number of stages relatively small.[8] Moreover, each stage brings administrative and logistical burdens and a reduction in transparency.

This strategy can be used with any sampling design and any measure of vote discrepancies for which upper bounds on the conditional P-values can be calculated.[9]

III. CAST STEP-BY-STEP

This section illustrates CAST using a sample drawn by proportional stratified random sampling. Batches are grouped into strata for convenience. One stratum might comprise ballots cast in-precinct on election day; another might comprise vote-by-mail (VBM) ballots; and a third might comprise provisional ballots. If a contest crosses county lines, batches might be stratified by county to allow counties to audit independently. Simple random samples of batches are drawn independently from different strata. A stratum is sampled only after the counts have been reported for all batches in that stratum. To do otherwise invites fraud.

The test statistic in this example is the maximum relative overstatement of pairwise margins [2]. Within an auditable batch of ballots, the maximum relative overstatement of pairwise margins is the maximum fraction by which error inflated the margin of any apparent winner over any apparent loser. [4], [6] show how to calculate P-values for CAST using sampling proportional to an error bound with functions of the *taint* of the maximum relative overstatement of pairwise margins as the test statistic. (Taint is the ratio of the maximum relative overstatement of pairwise margins to the maximum possible relative overstatement of pairwise margins: the fraction of the possible error that is actually attained.)

If the number of batches drawn from each stratum is proportional to the number of batches in the stratum, the P-value can be bounded by pretending that the sample was an unstratified random sample with replacement [1]. The recipe below uses this bound, which is not sharp when the sampling fraction is appreciable. If there is only one stratum, it is better to work directly with the hypergeometric distribution, as in [1].

Some batches can be selected deliberately rather than randomly. For example, candidates might be allowed to select a few batches they find suspicious; this is sometimes called a *targeted audit* in distinction to a *random audit*. There can be rules such as "if the targeted audit of a batch reveals a discrepancy of more than X votes, count every ballot by hand." That does not increase the risk that the audit stops short of a full manual count when the outcome is wrong, provided the margin is adjusted for the errors observed in the targeted batches when calculating the conditional P-values.

The following subsections give detailed steps, using the following notation: There are N batches of ballots and K candidates. The current stage of the audit is s. The number of batches not yet audited when stage s starts is N_s; $N_1 = N$ unless there is targeted auditing before the first random sample. The number of votes reported for candidate k in batch p is v_{kp}. The number of votes an audit would show for candidate k in batch p is a_{kp}. The *adjusted margin* of candidate w over candidate ℓ (at stage s) is

$$V_{w\ell;s} = \sum_{p \text{ audited}} (a_{wp} - a_{\ell p}) + \sum_{p \text{ unaudited}} (v_{wp} - v_{\ell p}). \quad (2)$$

These sums are over batches p that were or were not audited before stage s, respectively. That is, $V_{w\ell;s}$ is the apparent margin between w and ℓ after correcting any errors found in the batches audited prior to stage s, including targeted batches. The apparent margin before any auditing is $V_{w\ell} \equiv V_{w\ell;1}$. In a winner-take-all contest, the apparent winner is the candidate w for whom $V_{wk} > 0$ for all $k \neq w$; the rest are apparent losers. In a contest of the form *vote for up to f*, candidate w is an apparent winner if $\#\{k : V_{wk} < 0\} < f$. Let $\mathcal{W}$ denote the apparent winners and

---

[5]The "true" vote counts in batches already audited, whether those batches were selected at random or deliberately, are treated as known. The reported margin is adjusted to reflect the hand counts in those batches, and those batches are excluded from future samples.

[6]The measure of error is an important ingredient in the test. In the examples below, the measure is the maximum relative overstatement of pairwise margins. See [2], [4], and [6].

[7]Some batches may be deliberately excluded from audit if one assumes that those batches contribute as much error as they can, and adjusts the P-value accordingly. For some sampling designs, this can improve efficiency.

[8]Some sequential methods avoid this issue. For instance, the Kaplan-Markov bound in [6], [7] allows stages to consist of auditing individual batches, and still controls the risk.

[9]See [4], [6], [7] for bounds on the P-values for other sampling designs, using a related measure of vote discrepancy: the *taint* of the maximum relative overstatement.

$\mathcal{L}$ the apparent losers, so $V_{w\ell} > 0$ for all $w \in \mathcal{W}$ and $\ell \in \mathcal{L}$, and $\#\mathcal{W} = f$.

Step 1: Pick the Tolerable Risk α: The audit guarantees that the probability of a full hand count is at least $1 - \alpha$ when the apparent outcome is wrong. Typically, choosing α will be a matter of legislation or administrative rule. To limit auditing burden, it can be desirable to choose α larger for small contests than for countywide or statewide contests.[10]

Step 2: Pick the Maximum Number of Stages and the Risk Limits: Pick the maximum number of stages S. If the audit does not confirm the apparent outcome at stage $s = 1$, a second sample will be drawn, and so on. If the audit has not confirmed the apparent outcome by stage $s = S$, the entire audit trail will be counted by hand.

Choose $\{\alpha_s\}_{s=1}^{S}$ between 0 and 1 so that

$$\prod_{s=1}^{S}(1 - \alpha_s) \geq 1 - \alpha. \tag{3}$$

For example, one could take $\alpha_s = 1 - (1 - \alpha)^{1/S}$ for all s. Set $s = 1$.

Step 3: Define the Batches and Strata: Define the batches of ballots from which the audit samples will be drawn, and group them into strata. Generally, the fewer votes each auditable batch contains, the smaller the audit effort required to confirm a correct apparent outcome. Batches and strata must satisfy three requirements: (i) Every batch is in exactly one stratum. (ii) Semiofficial counts for every batch in a stratum must be published prior to drawing the sample from that stratum. (iii) There must be an upper bound on the number of valid votes in the batch for any candidate or position in the contest.[11]

Let C denote the total number of strata; and let B_c denote the number of batches in stratum c, for $c = 1, \ldots, C$. Then $\sum_{c=1}^{C} B_c = N$.

Step 4: Find Upper Bounds on the Number of Votes Per Candidate Per Batch: Audits require a limit on the amount that error in each batch might have inflated the apparent margin (see Section III). A limit can be derived from the maximum number of votes b_p any candidate or position can get in batch p; b_p could from voter registrations, poll books, or an accounting of ballots.[12] See [1].

Step 5: Calculate all Adjusted Margins: Let N_s be the number of batches still unaudited when stage s starts. For each apparent winner $w \in \mathcal{W}$ and each apparent loser $\ell \in \mathcal{L}$, calculate the adjusted margin $V_{w\ell;s}$.

In a winner-take-all contest, there are $K - 1$ margins: The apparent winner is paired with each of the remaining candidates in turn. If a contest allows each voter to vote for up to 3 of 7 candidates, then the 3 apparent winners each have a margin over each of the 4 apparent losers: There are $3 \times 4 = 12$ adjusted margins $V_{w\ell;s}$.

Step 6: Bound the Maximum Overstatement of Pairwise Margins: In each batch not yet audited, the semiofficial counts together with the bound on the number of valid votes per candidate (step 3) limit the overstatement the margin between any apparent winner and any apparent loser in that batch.

Let

$$u_{w\ell p} \equiv \frac{v_{wp} - v_{\ell p} + b_p}{V_{w\ell;s}}. \tag{4}$$

For each unaudited precinct p, compute

$$u_p \equiv \max_{w \in \mathcal{W}, \ell \in \mathcal{L}} u_{w\ell p}. \tag{5}$$

Error in batch p could not have overstated the adjusted margin between any apparent winner and any apparent loser by more than u_p times that adjusted margin. See [2].

Step 7: Select Batches for Targeted Audits: When a few unaudited batches p have u_p much larger than the rest, auditing them can reduce the next sample size. If batches are selected for targeted audit, hand count the votes in those batches and return to step 5. Batches can also be *excluded* from audit by assuming that the error in such batches is u_p and adjusting the margin accordingly.

Step 8: Select the Desired Threshold for "Escalation": The next step is to set the tolerable level of error, t, a number between 0 and 1. If any adjusted margin is found to be overstated by t or more, the audit will progress to the next stage. Generally, the larger t is, the larger the sample size needs to be. If t is chosen so large that

$$\sum_{p \text{ unaudited}} \min(t, u_p) \geq 1 \tag{6}$$

a full hand count will be required to confirm the election. But when t is small there tends to be a greater chance that the audit will progress to the next stage. For example, if $t = 0$, the audit will have to go to the next stage if the current stage finds even one discrepancy that overstates any margin. One way to select t is to choose a tolerance for the number of votes by which any margin can be overstated, then express that as a fraction of the smallest margin. For example, Section IV considers a threshold overstatement of 3 votes, expressed as a fraction of the margin of victory. The value of t can be changed at each stage of the audit; the sample size for stage s then depends on the value of t for stage s.

Step 9: Find Sample Sizes for the Next Random Sample: First the total additional sample size is calculated. That number is then allocated across strata in proportion to the number of batches in each. Other choices are possible; see, e.g., [1]. Assume that the semiofficial counts are available for all batches in all strata.[13]

To find the overall sample size, define a new list of numbers. For the N_s batches p not yet audited, let

$$t_p \equiv \min(t, u_p), \quad T \equiv \sum_{p \text{ unaudited}} t_p, \quad \text{and} \tilde{u}_p \equiv u_p - t_p. \tag{7}$$

---

[10]If α is small, it will often be necessary to hand count most of the ballots in small contests, even when the apparent outcome is correct.

[11]Generally, subtotals and bounds are available only at the precinct level and above, although some jurisdictions use "decks" of ballots run through scanners as a group. See Section III and [3], [4].

[12]An accounting of ballots verifies that the number of ballots sent to a precinct equals the total returned voted, unvoted, and spoiled. DREs have no physical ballot: An accounting of ballots impossible. Using the number of registered voters as a bound is conservative if same-day registration is included.

[13]If counts are available only for some strata when the audit starts, margins will not be known. One can proceed using a reasonable initial choice of sample sizes for the strata for which counts have been reported, but to decide whether to stop the audit requires reported counts for all batches. Step 12 then needs to be modified. See [1] for more discussion.

1) Starting with the largest value of $\tilde{u}_p$, add successively smaller values of $\tilde{u}_p$ just until the sum of those values is $1 - T$ or greater. Let q denote the number of terms in the sum.

2) Find the smallest whole number n such that

$$\left(\frac{N_s - q}{N_s}\right)^n \leq \alpha_s. \qquad (8)$$

3) For $c = 1, \ldots, C$, the sample size n_c for stratum c is

$$n \times \frac{\#\text{unaudited batches in stratum } c}{N_s} \qquad (9)$$

rounded up to the nearest whole number. Thus,

$$n^* = n_1 + n_2 + \cdots + n_C \geq n. \qquad (10)$$

[1] proves that these sample sizes guarantee that if the apparent outcome is wrong and we progress to the next audit stage when any batch in the sample has a maximum margin overstatement greater than t, the chance of stopping the audit at stage s is at most α_s. As discussed above, the maximum P-value is less than α_s if the sampling fraction is large.

Step 10: Draw the Next Random Sample and Count Votes: Select batches using a transparent, verifiable source of randomness.[14] For each stratum $c = 1, \ldots, C$, draw a simple random sample of n_c batches from the as-yet-unaudited batches in stratum c, and count the votes for each candidate in each batch in the sample by hand.

Step 11: Calculate the Maximum Relative Overstatement: For each of the n^* batches p just audited in this stage, define

$$e_{w\ell p} = \frac{v_{wp} - v_{\ell p}}{V_{w\ell;s}} - \frac{a_{wp} - a_{\ell p}}{V_{w\ell;s}}. \qquad (11)$$

There are $n^* \times w \times \ell$ such values. Calculate

$$t_s = \max_{p \text{ just audited}} \max_{w \in \mathcal{W}, \ell \in \mathcal{L}} e_{w\ell p}. \qquad (12)$$

Step 12: Stop, Perform a Full Count, or Proceed to the Next Stage: If $t_s \leq t$, stop the audit.[15] If $t_s > t$ and $s = S$, count the rest of the ballots by hand. Otherwise, increment s; perform any additional desired targeted auditing; and return to step 5.

IV. EXAMPLE: CARTOON OF A U.S. HOUSE RACE

This section presents a stylized U.S. House of Representatives contest with an apparent margin of 5.2%. Batches are all the same size and reported votes are the same in all batches, so the calculations can be done by hand.

A. Stage 1

Step 1: The probability of a full hand count is at least $1 - \alpha = 90\%$ if the apparent outcome is wrong.

[14]Pseudorandom numbers generated by computer generally are not appropriate. However, using public die rolls to generate a random seed for a high-quality, open-source, reproducible pseudorandom number generator can be efficient and secure. This is how batches were selected in Marin and Santa Cruz counties, California, for November 2008 risk-limiting audits [3], [4]. The jurisdictions generated 6-digit random numbers by rolling 10-sided dice; those numbers were then used in the R implementation of the Mersenne Twister to generate the sample.

[15]This can be sharpened: What matters is whether $P_s < \alpha_s$, which might hold even if $t_s > t$.

Step 2: The maximum number of stages of auditing is $S = 2$. The chance of error is equal in the two stages: $\alpha_1 = \alpha_2 = 1 - (1 - \alpha)^{1/2} = 1 - \sqrt{0.9} = 5.1\%$. Set $s = 1$.

Step 3: There are 400 precincts, 300 in one county and 100 in another. Votes cast in each precinct are divided into batches of ballots cast in-precinct on election day (IP) and cast by mail (VBM). Strata 1 and 2 correspond to IP and VBM ballots in the first county, respectively, and strata 3 and 4 correspond to IP and VBM ballots in the second county. Thus, $C = 4$, $B_1 = B_2 = 300$, and $B_3 = B_4 = 100$.

Step 4: Each precinct has 255 ballots cast in-precinct and 255 ballots cast by mail. There has been an accounting of ballots: The number of ballots in batch p, $b_p = 255$, is known.

Step 5: $N_1 = 800$. To keep things simple, suppose the semiofficial count in each of the 800 batches is the same: 125 votes for candidate 1, 112 votes for candidate 2, 13 votes for candidate 3, 2 overvoted ballots (ballots marked for two or more of the candidates), and 3 undervoted ballots (ballots with no mark for any of the candidates). The apparent totals are given in Table I.

The apparent margin of candidate 1 over candidate 2 in votes is

$$V_{12} = 100\,000 - 89\,600 = 10\,400 \text{ votes}. \qquad (13)$$

The apparent margin of candidate 1 over candidate 3 is

$$V_{13} = 100\,000 - 10\,400 = 89\,600 \text{ votes}. \qquad (14)$$

As a percentage of votes cast, the margin of victory is

$$\frac{100\,000 - 89\,600}{200\,000} \times 100\% = 5.2\%. \qquad (15)$$

Step 6: In each batch p, the largest fraction by which error could have overstated the margin of candidate 1 over candidate 2 is

$$u_{12p} = \frac{125 - 112 + 255}{10\,400} = 0.02577. \qquad (16)$$

The largest fraction by which error could have overstated the margin of candidate 1 over candidate 3 is

$$u_{13p} = \frac{125 - 13 + 255}{89\,600} = 0.00410. \qquad (17)$$

In each precinct p, the most by which error could have overstated the margin of candidate 1 over either of the other candidates is

$$u_p = \max(0.02577, 0.00410) = 0.02577. \qquad (18)$$

Step 7: There will be no targeted audit.

Step 8: Set the threshold t as follows: The audit will stop if the margin has been overstated by no more than 3 votes in any precinct in the sample. (The larger this number, the larger the initial sample size needs to be, but the smaller the chance of proceeding to the next stage if the election outcome is correct.) Three votes is $3/10\,400 = 0.00029$ of the (smaller) margin of victory. The audit proceeds to the second stage if in any batch p in the first-stage sample, either e_{12p} or e_{13p} is greater than $t = 0.00029$. If that happens, the audit adjusts the margins for the errors the first stage found, takes a new sample, and tests again; see below.

TABLE I
HYPOTHETICAL SEMI-OFFICIAL ELECTION RESULTS

$125 \times 800 = 100,000$	votes for candidate 1
$112 \times 800 = 89,600$	votes for candidate 2
$13 \times 800 = 10,400$	votes for candidate 3
$250 \times 800 = 200,000$	total votes cast
$1,600$	overvotes
$2,400$	undervotes
$255 \times 800 = 204,000$	total ballots

Step 9: Since $u_p = 0.02577$ and $t = 0.00029$, $t_p = t = 0.00029$ in every batch p, and

$$\tilde{u}_p = u_p - t = 0.02548 \qquad (19)$$

in every batch p. All $\tilde{u}_p$ are equal, so sorting is unnecessary. No batches have been audited, so $T = 0.00029 \times 800 = 0.23077$. The smallest number of batches for which the sum of $\tilde{u}_p$ is at least $1 - T$ is thus the smallest whole number q so that $q \times 0.02548 \geq 1 - 0.23077$:

$$\frac{1 - 0.23077}{0.02548} = 30.2 \qquad (20)$$

so $q = 31$.

Find the smallest whole number n so that

$$\left(\frac{800 - 31}{800}\right)^n \leq \alpha_1 = 0.051. \qquad (21)$$

Take the logarithm of both sides:

$$n \log\left(\frac{(800 - 31)}{800}\right) \leq \log(0.051)$$
$$n \times (-0.03952) \leq -2.9759$$
$$n \geq 75.3 \qquad (22)$$

so $n = 76$.

Allocate the sample across the strata in proportion to the number of batches in each stratum, rounding up to the nearest whole number: $76 \times 300/800 = 28.5$, so

$$n_1 = n_2 = 29 \qquad (23)$$

and $76 \times 100/800 = 9.5$, so

$$n_3 = n_4 = 10. \qquad (24)$$

The total first-stage sample size is $n^* = 29+29+10+10 = 78$, 9.75% of the 800 batches.

Step 10: Draw independent simple random samples of 29 batches of IP ballots from county 1, 29 VBM batches from county 1, 10 batches of IP ballots from county 2, and 10 batches of VBM ballots from county 2. Count the votes in the sampled batches by hand.

Step 11: For each audited batch p, compute e_{12p} and e_{13p}. Let t_1 be the largest of those $78 \times 2 = 156$ numbers.

Step 12: If $t_1 \leq 0.00029$, stop the audit. Otherwise, set $s = 2$ and go to Stage 2.

B. Stage 2

If a full hand count would show that candidate 1 is not the winner, there is at least a 94.9% chance that the audit proceeds to the second stage, and then at least a 94.9% chance that the audit proceeds to a full hand count, giving an overall chance of at least $94.9\% \times 94.9\% = 90\%$ that the audit leads to a full hand count. If no batch has errors that overstate any margin by more than t, the audit has no chance of finding $t_1 > 0.00029$ and proceeding to stage 2: The outcome will be certified at stage 1.

What if some batches have errors that overstate the margin of candidate 1 over candidate 2 by more than 3 votes? This section examines what would happen in two scenarios. In the first, the apparent outcome is wrong. In the second, the apparent outcome is right and the net error is zero, but 8 batches overstate the margin by more than t, and 8 have compensating errors that understate the margin.

1) The Outcome is Wrong: Suppose that in 100 of the batches, the true vote was

80	votes for candidate 1
160	votes for candidate 2
13	votes for candidate 3
1	overvote and
1	undervote

so there are still 255 ballots in each batch. In each of the remaining 700 batches, the true vote was

124	votes for candidate 1
113	votes for candidate 2
15	votes for candidate 3
2	overvotes, and
1	undervote.

Then the true total vote is

$$100 \times 80 + 700 \times 124 = 94\,800 \ (\text{candidate 1})$$
$$100 \times 160 + 700 \times 113 = 95\,100 \ (\text{candidate 2})$$
$$100 \times 13 + 700 \times 15 = 11\,800 \ (\text{candidate 3}).$$

The apparent outcome is wrong: Candidate 2 is the rightful winner. The overstatement of the margin of candidate 1 over candidate 2 in the 100 batches with large errors is

$$\frac{(125 - 112) - (80 - 160)}{10\,400} = 0.00894 \qquad (25)$$

and in the 700 batches with smaller errors it is

$$\frac{(125 - 112) - (124 - 113)}{10\,400} = 0.00019. \qquad (26)$$

The overstatement of the margin of candidate 1 over candidate 3 in the 100 batches with large errors is

$$\frac{(125 - 13) - (80 - 13)}{89\,600} = 0.00050 \qquad (27)$$

and in the 700 batches with smaller errors it is

$$\frac{(125 - 13) - (120 - 15)}{89\,600} = 0.00008. \qquad (28)$$

The first and third of these four overstatements—those for the 100 batches with large errors—exceed $t = 0.00029$. The chance the first stage of the audit finds one or more batches with margin

overstatements $e_p > 0.00029$ (so that the audit would progress to the second stage) is at least

$$1 - \left(\frac{800 - 100}{800}\right)^{76} = 99.996\% \qquad (29)$$

substantially greater than the guaranteed minimum of 94.9%.

Absent stratification, the expected number of batches in the sample with large errors would be

$$78 \times \frac{100}{800} = 9.75. \qquad (30)$$

Suppose the sample finds 10 of the batches with large errors and 68 of the batches with small errors. The audit would return to step 5, but with $s = 2$ and $N_2 = 800 - 78 = 722$ as-yet-unaudited precincts.

Step 5: The vote counts, adjusted for the error observed in the first-stage sample, are

$$10 \times 80 + 68 \times 124 + 722 \times 125 = 99\,482 \text{ (candidate 1)}$$
$$10 \times 160 + 68 \times 113 + 722 \times 112 = 90\,148 \text{ (candidate 2)}$$
$$10 \times 13 + 68 \times 15 + 722 \times 13 = 10\,536 \text{ (candidate 3).}$$

Adjusted for the error found at stage 1, the margins are

$$V_{12;2} = 99\,482 - 90\,148 = 9\,334 \ (1\,066 \text{ less})$$
$$V_{13;2} = 99\,482 - 10\,536 = 88\,946 \ (654 \text{ less}). \qquad (31)$$

Step 6: The revised values of $u_{w\ell p}$, taking into account the adjusted margins, are

$$u_{12p} = \frac{125 - 112 + 255}{9\,334} = 0.02871$$
$$u_{13p} = \frac{125 - 13 + 255}{88\,946} = 0.00412. \qquad (32)$$

Thus, in each precinct p, the most by which error could have overstated the margin of candidate 1 over either of the other candidates is

$$u_p = \max(0.02871, 0.00412) = 0.02871. \qquad (33)$$

Step 7: There is no targeted selection.

Step 8: Set the threshold to correspond to a 3 vote overstatement of the adjusted margin: $t = 3/9334 = 0.00032$.

Step 9: As before, $\alpha_2 = 5.1\%$. Since $u_p = 0.02871$ and $t = 0.00032$, $t_p = t = 0.00032$ in every batch p, and

$$\tilde{u}_p = u_p - t_p = 0.02839 \qquad (34)$$

in every batch p. All $\tilde{u}_p$ are equal so sorting is not needed. 722 batches remain unaudited, so $T = 722 \times 3/9334 = 0.23205$. The smallest number of batches for which the sum of $\tilde{u}_p$ is at least $1 - T$ is thus the smallest whole number q so that $q \times 0.02839 \geq 1 - 0.23205$

$$\frac{1 - 0.23205:}{0.02839} = 27.05 \qquad (35)$$

so $q = 28$.

Now find the smallest whole number n so that

$$\left(\frac{722 - 28}{722}\right)^n \leq \alpha_2 = 0.051. \qquad (36)$$

Take the logarithm:

$$n \log\left(\frac{(722 - 28)}{722}\right) \leq \log(0.051)$$
$$n \times (-0.03955) \leq -2.9759$$
$$n \geq 75.1 \qquad (37)$$

so $n = 76$, as before (coincidentally).

Allocate the sample across the strata in proportion to the number of batches in each stratum, rounding up to the nearest whole number. There are 722 unaudited precincts, 271 in strata 1 and 2, and 90 in strata 3 and 4: $76 \times 271/722 = 28.5$, so

$$n_1 = n_2 = 29 \qquad (38)$$

and $76 \times 90/722 = 9.5$, so

$$n_3 = n_4 = 10 \qquad (39)$$

giving a total sample size $n^* = 29 + 29 + 10 + 10 = 78$, as before. Thus, the second stage will audit $78/722 = 10.8\%$ of the remaining 722 batches.

Step 10: Draw independent random samples of 29 batches of IP ballots from county 1, 29 VBM batches from county 1, 10 batches of IP ballots from county 2, and 10 batches of VBM ballots from county 2. Count the votes in each batch in the sample by hand.

Step 11: For each audited batch p, compute e_{12p} and e_{13p}. Let t_2 be the largest of those $78 \times 2 = 156$ numbers.

Step 12: If $t_2 \leq 0.00032$, stop the audit. Otherwise, perform a full hand count.

In this hypothetical, 10 of the 100 batches with large errors were found in stage 1, so 90 of those batches remained among the 722 still unaudited. The chance that the second-stage sample contains at least one of them—so that

$$t_2 = \frac{(125 - 112) - (80 - 160)}{9\,334} = 0.00996 \qquad (40)$$

and a full hand count occurs—is at least

$$1 - \left(\frac{722 - 90}{722}\right)^{76} = 99.996\% \qquad (41)$$

again rather larger than the guarantee of 94.9%.

2) The Outcome is Correct: Suppose 8 batches (1%) have errors that overstate the margin between candidate 1 and candidate 2 by more than t and 8 batches have compensating errors that understate the margin. The outcome of the contest is correct. Nonetheless, the sample could contain one or more of the 8 batches with large overstatements, triggering the audit to progress to stage 2. What is the chance that will occur? The answer depends on how the 8 batches with large errors are spread across the 4 strata. The chance that the audit will go to stage 2 is largest if all 8 are in one of the small strata. The chance is then

$$1 - \frac{\binom{92}{10}}{\binom{100}{10}} = 58.3\%. \qquad (42)$$

If the 8 batches with large overstatements are spread proportionately across the strata, 3 in each of the two large strata and 1 in

each of the two small strata, the chance that the audit will go to stage 2 is a bit smaller:

$$1 - \left(\frac{\binom{297}{29}}{\binom{300}{29}}\right)^2 \times \left(\frac{\binom{99}{10}}{\binom{100}{10}}\right)^2 = 56.1\%. \tag{43}$$

The chance that the audit will progress from stage 2 to a full hand count depends on the number of batches with large overstatement or understatement errors found at stage 1 and which strata they are in. Generally, the larger the fraction of the 8 large overstatement errors discovered in the first stage and the smaller those errors are, the smaller the chance the audit will progress to a full hand count.

To get a feel for the chance of a full hand count, suppose there are 8 overstatement errors and 8 understatement errors distributed as follows: In each of the two large strata, there are 3 batches that overstate the margin between candidate 1 and candidate 2 by 10 votes and 3 batches that understate it by 10 votes, and that in each of the two small strata, there is 1 batch that overstates that margin by 10 votes and one that understates it by 10 votes. In the first-stage sample, the expected number of batches with 10 vote overstatements of the margin is

$$3 \times \frac{29}{300} + 3 \times \frac{29}{300} + 1 \times \frac{10}{100} + 1 \times \frac{10}{100} = 0.78. \tag{44}$$

The expected number with 10 vote understatements of the margin is the same. Suppose the first-stage sample finds 1 batch with a 10 vote overstatement in one of the large strata, and 1 batch with a 10 vote understatement in some stratum. Then $t_1 > t$, so the audit will go to the second stage at step 5.

Step 5: The net error found in the first stage is zero, so the adjusted margins remain equal to the apparent margins:

$$V_{12;2} = 100\,000 - 89\,600 = 10\,400 \text{ votes} \tag{45}$$
$$V_{13;2} = 100\,000 - 10\,400 = 89\,600 \text{ votes}. \tag{46}$$

Step 6: Since the margins have not changed, still

$$u_p = \max(0.02577, 0.00410) = 0.02577. \tag{47}$$

Step 7: There will be no targeted auditing.

Step 8: The margins have not changed, so a 3 vote overstatement of the smaller margin gives $t = 3/10\,400 = 0.00029$.

Step 9: Since $u_p = 0.02577$ and $t = 0.00029$, $t_p = t = 0.00029$ in every batch p, and

$$\tilde{u}_p = u_p - t = 0.02548 \tag{48}$$

in every batch p. These are all equal, so sorting is unnecessary. There are 722 unaudited batches, so $T = 0.00029 \times 722 = 0.20938$. The smallest number of batches for which the sum of $\tilde{u}_p$ is at least $1 - T$ is thus the smallest whole number q so that $q \times 0.02548 \geq 1 - 0.20938$

$$\frac{1 - 0.20938}{0.02548} = 31.03 \tag{49}$$

so $q = 32$.

Now find the smallest whole number n so that

$$\left(\frac{722 - 32}{722}\right)^n \leq \alpha_1 = 0.051 \tag{50}$$

namely, $n = 66$.

There are 722 unaudited precincts, 271 in strata 1 and 2, and 90 in strata 3 and 4: $66 \times 271/722 = 24.8$, so

$$n_1 = n_2 = 25 \tag{51}$$

and $66 \times 90/722 = 8.2$, so

$$n_3 = n_4 = 9 \tag{52}$$

giving a total sample size $n^* = 25 + 25 + 9 + 9 = 68$, Thus, the second stage would audit $68/722 = 9.4\%$ of the remaining 722 batches in this scenario.

Step 10: Draw independent random samples of 25 batches of IP ballots from county 1, 25 VBM batches from county 1, 9 batches of IP ballots from county 2, and 9 batches of VBM ballots from county 2. Count the votes in every batch in the sample by hand.

Step 11: For each audited batch p, compute e_{12p} and e_{13p}. Let t_2 be the largest of those $68 \times 2 = 136$ numbers.

Step 12: If $t_2 \leq 0.00029$, stop the audit. Otherwise, count the remaining $800 - 78 - 68 = 654$ ballots by hand.

In this scenario, the chance that the second stage sample has one or more of the batches with a 10 vote overstatement of the margin between candidates 1 and 2 is

$$1 - \frac{\binom{269}{25}}{\binom{271}{25}} \times \frac{\binom{268}{25}}{\binom{271}{25}} \times \left(\frac{\binom{89}{9}}{\binom{90}{9}}\right)^2 = 50.2\%. \tag{53}$$

The overall chance that the audit would progress to a full hand count is thus

$$56.1\% \times 50.2\% = 28.1\%. \tag{54}$$

If there were fewer batches with large errors, the chance of a full hand count would be smaller.

For example, if only one batch overstates the margin by more than t, the chance of proceeding to the second stage is at most 10%, and the chance of a full hand count is zero. If two batches overstate the margin by more than t, the chance of proceeding to the second stage is at most 19.1%. If the first stage finds both errors, the chance of a full hand count is zero; if not, the chance of proceeding to a full hand count is on the order of 10%. The overall chance of a full hand count would be at most 1.8%, neglecting any adjustments to the margin. Table II presents the maximum probability of an unnecessary full hand count in a variety of scenarios.

C. Varying the Assumptions

In this example—a 5.2% margin, 10% risk split evenly across 2 stages ($\alpha_1 = \alpha_2 = 5.1\%$) and 800 batches—the initial sample size is nearly 10% of the batches. Table II shows how the first-stage sample size depends on the margin, the risk limit, and the allocation of risk across stages. In these hypothetical examples, in each batch the number of votes reported for candidate 3 is fixed at 13, the number of overvotes is fixed at 2, and the number of undervotes is fixed at 3. The first-stage threshold t for escalation is set so that the outcome is certified if the margin between candidate 1 and candidate 2 has been overstated by 3 votes or fewer (because 3 votes is a larger fraction of V_{12} than of V_{13}, a much larger overstatement of the margin between candidate 1 and candidate 3 is required before the audit escalates to stage 2).

TABLE II
HYPOTHETICAL 2-STAGE AUDIT WITH 3 VOTE ERROR THRESHOLD. CONTESTS HAVE 4 STRATA, TWO WITH 300 BATCHES AND TWO WITH 100. EACH BATCH HAS 255 BALLOTS, INCLUDING 13 VOTES REPORTED FOR CANDIDATE 3, 2 OVERVOTES AND 3 UNDERVOTES. COLUMNS: (1) MARGIN BETWEEN CANDIDATE 1 (WINNER) AND CANDIDATE 2 (RUNNER-UP). 5.2% MARGIN IS 125 VOTES FOR CANDIDATE 1 AND 112 FOR CANDIDATE 2 IN EACH BATCH; 10.0% IS 131 VERSUS 106; 19.6% IS 143 VERSUS 94. (2) MINIMUM CHANCE OF A FULL HAND COUNT IF THE OUTCOME IS WRONG. (3) MINIMUM CHANCE THE AUDIT GOES FROM STAGE 1 TO STAGE 2 IF THE OUTCOME IS WRONG. (4) MINIMUM CHANCE THE AUDIT GOES FROM STAGE 2 TO A FULL HAND COUNT IF THE OUTCOME IS WRONG, IF IT GETS TO STAGE 2. (5) STAGE 1 SAMPLE SIZE BEFORE ADJUSTING FOR STRATIFICATION. (6) STAGE 1 SAMPLE SIZE ADJUSTED FOR STRATIFICATION. (7) COLUMN 6 AS A PERCENTAGE OF 800. (8) MAXIMUM CHANCE THE AUDIT PROGRESSES TO STAGE 2 IF 1% OF BATCHES OVERSTATE THE MARGIN BY MORE THAN 3 VOTES. (9) MAXIMUM CHANCE OF A FULL HAND COUNT IF 1% OF BATCHES OVERSTATE THE MARGIN BY MORE THAN 3 VOTES, AND THE STAGE 1 NET ERROR IS ZERO. (10) SAME AS (8), IF 0.5% OF BATCHES HAVING LARGE MARGIN OVERSTATEMENTS. (11) SAME AS (9), IF 0.5% OF BATCHES HAVING LARGE MARGIN OVERSTATEMENTS. (12) STAGE 2 SAMPLE SIZE IF THE NET ERROR IN STAGE 1 IS ZERO

(1) V_{12}	(2) $1-\alpha$	(3) $1-\alpha_1$	(4) $1-\alpha_2$	(5) n	(6) n^*	(7) f	(8) $\gamma_{0.01}$	(9) $\omega_{0.01}$	(10) $\gamma_{0.005}$	(11) $\omega_{0.005}$	(12) n_2^*
5.2%	75%	76.0%	98.9%	37	38	4.75%	34.7%	23.8%	18.8%	7.2%	108
		86.6%	86.6%	51	54	6.75%	45.3%	17.7%	25.5%	4.9%	50
	90%	91.0%	98.9%	61	62	7.75%	50.0%	34.9%	28.7%	11.3%	108
		94.9%	94.9%	76	78	9.75%	58.3%	31.2%	34.8%	9.5%	68
10.0%	75%	76.0%	98.9%	18	20	2.50%	22.3%	9.3%	11.6%	2.4%	54
		86.6%	86.6%	25	28	3.50%	28.7%	6.3%	15.3%	1.5%	26
	90%	91.0%	98.9%	29	30	3.75%	28.7%	12.1%	15.3%	3.1%	54
		94.9%	94.9%	36	38	4.75%	34.7%	11.2%	18.8%	2.8%	36
19.6%	75%	76.0%	98.9%	9	12	1.50%	15.4%	4.0%	7.9%	1.0%	28
		86.6%	86.6%	13	14	1.75%	15.4%	2.1%	7.9%	0.5%	14
	90%	91.0%	98.9%	15	16	2.00%	15.4%	4.0%	7.9%	1.0%	30
		94.9%	94.9%	18	20	2.50%	22.3%	4.5%	11.6%	1.1%	20

Table II also shows the maximum probability of progressing to the second stage if 1% (i.e., 8) or 0.5% (i.e., 4) of the 800 batches have errors that overstate the margin between candidates 1 and 2 by more than 3 votes. These bounds are denoted $\gamma_{0.01}$ and $\gamma_{0.005}$. Finally, Table II gives upper bounds on the probability of a full hand count in those two scenarios, with the additional assumption that the net error uncovered in stage 1 is zero—overstatement errors are balanced by understatement errors in the sample. Those bounds are denoted $\omega_{0.01}$ and $\omega_{0.005}$. The bounds γ and ω assume that the 4 or 8 batches with large errors are all in one stratum, which maximizes the chance that the audit will progress to the next stage and to a full hand count. (If the batches with large errors were distributed randomly, the probability of a full hand count would be lower.) The Appendix discusses how these bounds are calculated.

V. DISCUSSION AND TECHNICAL NOTES

The turn-key implementation is a simple, flexible recipe for risk-limiting audits, but the simplicity comes at the cost of efficiency. There are methods that require less auditing when the apparent outcome is correct.

A. Background Error Rate

The treatment of the threshold margin overstatement t has slack [6]. Each batch is in one of two categories: margin overstatement of t or less, or margin overstatement exceeding t. The test at stage s is based on the number of batches in the sample in the second category—the audit progresses to stage $s+1$ if that number is not zero. To ensure that the test is conservative, batches in the first category are treated as if they have margin overstatements of t, and batches in the second category are treated as if they have the largest margin overstatement their bounds u_p permit. In the example in Section IV, setting t to correspond to a 3-vote overstatement of the margin between the winner and the runner-up means that every batch can overstate the margin by $3/250 = 1.2\%$ of its 250 votes. Hence, if the apparent margin were 1.2% or less, the method would require a full hand count.

A more powerful test can be constructed by binning batches into more than two categories. For example, batches could be grouped into those with no margin overstatement, overstatement between 0 and some threshold t, and audit batches with margin overstatements greater than t [4]; some methods, such as the Kaplan-Markov bound [6] do not require binning the errors. Sampling proportional to an error bound [8] can increase power over simple or stratified sampling [4], [6].

B. Improving the Treatment of Stratification

The sample size calculation at step 9 of the turn-key implementation assumes that the q batches with enough error to alter the apparent outcome could be spread arbitrarily among the strata. But batches with large values of u_p might be concentrated in a relatively small subset of the strata, increasing the chance that the sample would find at least one batch with large errors if the aggregate error is large enough to produce the apparent margin. That is, the "worst case" for which the sample size is calculated might not be feasible for the actual set of error bounds. When the strata are small, the binomial probability bound in the turn-key implementation (see [1]) is weak.

C. Relation to Previous Work

Most previous work on postelection audits has focused on the *detection* question: "If the outcome is wrong, how large must a sample be to guarantee a large chance of finding at least one error?" Since audits routinely find error, it is more useful to answer the *confirmation* question: "Given the amount of error the audit found so far, is it safe to stop counting, or should more ballots be counted by hand?" The only methods I am aware of that address the confirmation question were developed in [1], [2], and [4].

CAST differs from the method of [1] and [2]. CAST controls the overall probability that the audit will stop short of a full hand count when the outcome is wrong by controlling the probability of stopping erroneously at each stage, conditional on the sample at previous stages. In contrast, [1] does not condition on the previous stages; it uses Bonferroni's inequality to bound the overall

probability of stopping short of a full hand count by the sum of the probabilities at each stage.

Conditioning on the audit results at previous stages rather than treating the sample as "telescoping" to include previous stages has several benefits. (1) If an early stage of audit finds a large discrepancy, the test statistic at later stages is not necessarily large, because at each stage only the incremental sample enters the test statistic. (2) Margin overstatement errors discovered at one stage do not lead to more than one step of escalation if there are canceling margin understatement errors: Only overstatement errors are involved in the test statistic at a given stage, but the margin is adjusted sequentially to account for both overstatement and understatement errors. (3) It makes it easy to incorporate "targeted" sampling.

APPENDIX
BOUNDS ON THE PROBABILITY OF ESCALATION

Suppose there are q batches with margin overstatements greater than t. The probability that the sample will contain none of the q batches with a large margin overstatement is minimized when all q of those batches are in one stratum. This result is established using the following lemma.

Lemma 1: Suppose $N > n$, $M > m$, $\min(N, M) \geq v \geq r \geq 0$ are all integers. Then

$$\frac{\binom{N-r}{n}}{\binom{N}{n}} \times \frac{\binom{M-v+r}{m}}{\binom{M}{m}} \geq \min \left\{ \frac{\binom{N-v}{n}}{\binom{N}{n}}, \frac{\binom{M-v}{m}}{\binom{M}{m}} \right\}. \quad (55)$$

Proof: If $v = 0$ or $v = 1$, the result is trivial; it is also clearly true when $M - m < v$ or $N - n < v$, because then the probability is 1. Suppose $v \geq \min(M - m, N - n)$. In general if $\ell < K$,

$$\frac{\binom{K-\ell}{k}}{\binom{K}{k}} = \frac{(K-k)(K-k-1)\cdots(K-k-\ell+1)}{K(K-1)\cdots(K-\ell+1)}. \quad (56)$$

Apply (56) twice to find

$$\frac{\binom{N-r}{n}}{\binom{N}{n}} \times \frac{\binom{M-v+r}{m}}{\binom{M}{m}}$$
$$= \frac{(N-n)(N-n-1)\cdots(N-n-r+1)}{N(N-1)\cdots(N-r+1)}$$
$$\times \frac{(M-m)(M-m-1)\cdots(M-m-v+r+1)}{M(M-1)\cdots(M-v+r+1)}. \quad (57)$$

Identity (56) also shows that

$$\frac{\binom{N-v}{n}}{\binom{N}{n}} = \frac{(N-n)(N-n-1)\cdots(N-n-v+1)}{N(N-1)\cdots(N-v+1)} \quad (58)$$

and

$$\frac{\binom{M-v}{m}}{\binom{M}{m}} = \frac{(M-m)\cdots(M-m-v+1)}{M(M-1)\cdots(M-v+1)}. \quad (59)$$

Thus, (57) is to be compared to

$$\min \left\{ \frac{(N-n)\cdots(N-n-v+1)}{N(N-1)\cdots(N-v+1)}, \right.$$
$$\left. \frac{(M-m)\cdots(M-m-v+1)}{M(M-1)\cdots(M-v+1)} \right\}. \quad (60)$$

Now

$$\frac{(N-n)\cdots(N-n-v+1)}{N(N-1)\cdots(N-v+1)}$$
$$= \frac{(N-n)\cdots(N-n-r+1)}{N(N-1)\cdots(N-r+1)}$$
$$\times \frac{(N-n-r)\cdots(N-n-v+1)}{(N-r)(N-r-1)\cdots(N-v+1)}. \quad (61)$$

Suppose without loss of generality that the first term in the minimum in (60) is the smaller of the two. Then the lemma can be false only if

$$\frac{(N-n)\cdots(N-n-r+1)}{N(N-1)\cdots(N-r+1)}$$
$$\times \frac{(M-m)\cdots(M-m-v+r+1)}{M(M-1)\cdots(M-v+r+1)}$$
$$< \frac{(N-n)(N-n-1)\cdots(N-n-v+1)}{N(N-1)\cdots(N-v+1)};$$

i.e., only if

$$\frac{M-m}{M} \times \cdots \times \frac{M-m-v+r+1}{M-v+r+1}$$
$$< \frac{N-n-r}{N-r} \times \cdots \times \frac{N-n-v+1}{N-v+1}. \quad (62)$$

Both sides of (62) are products of $v - r$ terms; (62) says the geometric mean of the terms on the left is less than the geometric mean of the terms on the right. If so, the smallest term on the left is smaller than the largest term on the right. Since $x < y$ implies that $(x - 1)/(y - 1) < (x/y)$, the smallest term on the left is $(M - m - v + r + 1)/(M - v + r + 1)$ and the largest on the right is $(N - n - r)/(N - r)$. Thus,

$$\frac{M-m-v+1}{M-v+1} < \frac{M-m-v+2}{M-v+2} < \cdots$$
$$< \frac{M-m-v+r}{M-v+r}$$
$$< \frac{M-m-v+r+1}{M-v+r+1} < \frac{N-n-r}{N-r}$$
$$< \frac{N-n-r+1}{N-r+1} < \cdots$$
$$< \frac{N-n-1}{N-1} < \frac{N-n}{N}.$$

Therefore,

$$\frac{M-m-v+1}{M-v+1} \times \frac{M-m-v+2}{M-v+2} \times \cdots$$
$$\times \frac{M-m-v+r}{M-v+r}$$
$$< \frac{N-n-r+1}{N-r+1} \times \frac{N-n-r+2}{N-r+2} \times \cdots$$
$$\times \frac{N-n-1}{N-1} \times \frac{N-n}{N}. \quad (63)$$

Multiplying the left-hand side (LHS) of (62) by the LHS of (63) and the right-hand side (RHS) of (62) by the RHS of (63) shows that if the lemma is false

$$\frac{(M-m)(M-m-1)\cdots(M-m-v+1)}{M(M-1)\cdots(M-v+1)}$$
$$< \frac{(N-n)(N-n-1)\cdots(N-n-v+1)}{N(N-1)\cdots(N-v+1)}$$

contradicting the assumption that the first term in (60) is the smaller of the two. ●

The lemma shows that if two strata have a total of v batches with large overstatements, the chance the sample contains at least one batch with a large overstatement is largest if all v are in one stratum (if one can hold them; if not, the maximum chance is 100%). Using the independence of the samples in different strata and applying the lemma recursively to pairs of strata shows that the probability that the sample contains at least one batch with a large overstatement is largest when all such batches are in the same stratum—if every stratum could hold all of them. Hence, the largest chance of escalation can be found by first checking whether $q \geq \min_c(N_c - n_c)$. If so, the maximum chance of escalation is 100%. If not, the maximum chance can be found by comparing C numbers, the probabilities when all the batches with large overstatements are in stratum c, for $c = 1, \ldots, C$.

The conditional probability of progressing to stage $s+1$ from stage s given that there are v batches with overstatement errors greater than t among the unaudited batches is thus maximized when all v of those batches are in one stratum. The larger v is, the larger that conditional probability, so the conditional probability of passing from stage s to stage $s + 1$ is largest if exactly one batch with a large overstatement was found in the sample at each stage before s. (If none was found the audit would not have progressed.)

Let n_{sc} be the number of batches drawn at stage s from stratum c, and let $n_{0c} \equiv 0$. Suppose q batches in all have overstatement errors greater than t. If $q < S$, the chance of a full hand count is zero—if the audit gets to stage $S - 1$, no batches with large errors are left. Otherwise, the chance of a full hand count is no larger than

$$\prod_{s=1}^{S} \left(1 - \min_c \frac{\dbinom{N_c - \sum_{r=1}^{s-1} n_{rc} - (q-s+1)}{n_{sc}}}{\dbinom{N_c - \sum_{r=1}^{s-1} n_{rc}}{n_{sc}}} \right). \tag{64}$$

REFERENCES

[1] P. Stark, "Conservative statistical post-election audits," *Ann. Appl. Stat.*, vol. 2, pp. 550–581, 2008.
[2] P. Stark, "A sharper discrepancy measure for post-election audits," *Ann. Appl. Stat.*, vol. 2, pp. 982–985, 2008.
[3] J. L. Hall, L. W. Miratrix, P. B. Stark, M. Briones, E. Ginnold, F. Oakley, M. Peaden, G. Pellerin, T. Stanionis, and T. Webber, "Implementing risk-limiting post-election audits in California," presented at the Proc. 2009 Electron. Voting Technol. Workshop/Workshop on Trustworthy Elections (EVT/WOTE '09)., Montreal, Canada, Aug. 2009.
[4] L. W. Miratrix and P. B. Stark, "Election audits using a trinomial bound," *IEEE Trans. Inf. Forensics Security*, vol. 4, no. 4, pp. 974–981, Dec. 2009.
[5] P. Stark, Auditing a Collection of Races Simultaneously 2009 [Online]. Available: http://arxiv.org/abs/0905.1422v1
[6] P. B. Stark, "Risk-limiting postelection audits: Conservative P-values from common probability inequalities," *IEEE Trans. Inf. Forensics Security*, vol. 4, no. 4, pp. 1005–1014, Dec. 2009.
[7] P. Stark, Efficient Post-Election Audits of Multiple Contests: 2009 California Tests Social Sci. Res. Netw., 2009 [Online]. Available: http://ssrn.com/abstract=1443314
[8] J. Aslam, R. Popa, and R. Rivest, "On auditing elections when precincts have different sizes," presented at the Proc. 2008 USENIX/ACCURATE Electron. Voting Technol. Workshop San Jose, CA, Jul. 28–29, 2008.

Philip B. Stark received the A.B. degree in philosophy from Princeton University and the Ph.D. degree in earth science from the University of California, San Diego. He was a postdoctoral fellow at the Scripps Institution of Oceanography, University of California, San Diego, and in the Department of Statistics, University of California, Berkeley.

He is Professor of Statistics, University of California, Berkeley. He served on the 2007 Post Election Audit Standards Working Group for California Secretary of State Debra Bowen, and designed and conducted the first four risk-limiting post election audits ever performed. For a more complete biography, see http://statistics.berkeley.edu/ stark/bio.htm.