CAST: Canvass Audit by Sampling and Testing

Philip B. Stark

Department of Statistics

University of California

Berkeley, CA 94720-3860

DRAFT

27 August 2008

Last revised 10 November 2008, 8:40pm

Abstract

An election audit trail can reflect an electoral outcome that differs from the semi-official outcome. If so, the winner should be determined by a hand count of all the votes as recorded in the audit trail. CAST is a statistical method for deciding whether to certify the outcome of a contest, or to count the entire audit trail by hand. It has a known, pre-specified chance of requiring a full hand count whenever that count would show an outcome different from the semi-

official outcome. This limits the risk of certifying an incorrect outcome erroneously. CAST also allows some ballots to be selected for audit deliberately rather than randomly, "targeted" auditing. CAST requires: (i) the desired minimum chance (e.g., 90%) that if the preliminary outcome is wrong—i.e., disagrees with the audit trail—CAST will require a full manual count; (ii) the maximum number of audit stages permitted before a full manual count; (iii) the semi-official results by precinct or "batch;" (iv) limits on the number of valid votes for any candidate that could have been cast in each batch. CAST collects data incrementally: if the the semi-official results overstate the margin by a sufficiently small amount in a sufficiently large sample, CAST says "certify." Otherwise, CAST says "audit more batches and check again." Eventually, either CAST says "certify," or there has been a full hand count. CAST is a refinement and simplification of the method of Stark [4, 5].

1 Introduction

Election systems are incredibly complex: they can involve voting machines, central tabulators, optical scanners, memory card readers, software, designers, printers, programmers, operators and pollworkers, elaborate chains of custody—and human voters. Errors in counting votes are inevitable. Do those errors affect the electoral outcome? Are they *material*?

If there is a reliable paper audit trail, that question can be answered[1] by a

[1] "Statistics means never having to say you're certain," so in this context answers are

post election audit: a hand count of the paper audit trail in a random sample of batches of ballots. Well designed and executed post election audits can limit the risk of certifying a preliminary official outcome that disagrees with the paper trail: no matter how the errors in the vote counts arose—whether at random, because of human error, or through malice or fraud—the audit can have a pre-determined chance of requiring a full manual count whenever that would show a different outcome.

An increasing number of jurisdictions have laws or regulations mandating post election audits. Some—New Jersey for instance[2]—require audits that limit the chance of certifying an outcome that is incorrect. There is so far only one basic method for conducting a post election audit guaranteed to have a high chance of requiring a full manual count when the paper trail would show a different outcome [4, 5].[3] That method has been used to audit a ballot measure in Marin County in February 2008 [2]. This paper presents a step-by-step recipe called CAST (Canvass Audit by Sampling and Testing) for a risk-limiting post-election audit.[4] CAST is not *optimal* (it is possible

of the form, "either the outcome agrees with what a full manual count of the audit record would show, or a very unlikely event has occurred."

[2] See `http://www.njleg.state.nj.us/2006/Bills/S1000/507_R1.PDF` and `http://www.njleg.state.nj.us/2006/Bills/A3000/2730_R1.PDF`

[3] There are of course trivial methods, such as performing a full manual count of every contest, or performing a full manual count of a large fraction of contests selected at random. Such methods do not use audit data to try to confirm the outcome without a full manual count. The point of the present method is to hand count as few votes as possible when a full hand count would not change the outcome.

[4] CAST is similar to the method of [4, 5] but differs in important ways (see section 4.3).

to derive methods that require less auditing when the outcome is correct, although none currently exists), but it is *conservative* in that it guarantees a pre-specified chance of a full manual count whenever that would change the outcome.

CAST is a refinement and simplification a method introduced by [4, 5] for determining—by auditing a random sample of batches of ballots—whether to certify the semi-official outcome of an election. CAST has a large chance of requiring a full hand count whenever the semi-official outcome is wrong.[5] When a full hand count would not change the outcome, CAST tries to minimize the amount of auditing effort before declaring "certify."

CAST is performed in stages. At each stage, the semi-official vote counts

[5] "Semi-official" results are sometimes called "preliminary official results," the results once the election officials represent that all votes have been counted and the outcome could be certified, but for the possibility that the audit says otherwise. 'Wrong" means that the semi-official outcome disagrees with the outcome a full manual count would show. Of course, even a full manual count might not identify the rightful winner. For example, the audit record could be incomplete or inaccurate: paper ballots can be lost before they are counted, ballot boxes can be "stuffed," etc. Moreover, it is impossible to audit direct recording electronic (DRE) voting machines that do not produce voter verifiable paper audit trails (VVPATs) without relying on electronic records. And agreement between hand counts of audit batches and semi-official results does not ensure that the batch-level results were aggregated correctly across the contest; viz. the recently discovered Diebold/Premier software "glitch" that prematurely aborts the downloading of data from memory cartridges (`http://voices.washingtonpost.com/the-trail/2008/08/21/ohio_voting_machines_contained.html`, `http://www.informationweek.com/news/management/legal/showArticle.jhtml?articleID=210000402`).

in a random sample of batches of votes[6] are compared with hand counts of the paper audit trail for those batches, for each contest under audit. Additional batches may be selected for audit deliberately rather than randomly, either before the random selection begins, or between stages of random selection.[7] For each audited batch, errors that resulted in overstating the margin between any winner and any loser in a given contest are expressed as a percentage of the semi-official margin between those candidates, adjusted for any errors discovered at previous stages. A calculation is performed to see whether error observed in the current stage sample is small enough to give confidence that the total error in the semi-official count did not change the outcome, given the errors observed at previous stages.[8] If it is sufficiently likely that the audit would have uncovered more error at the current stage if the outcome were wrong, the audit stops and the semi-official result is certified. Otherwise, the semi-official margin is adjusted to take into account the net error observed in the audit sample, a new sample is drawn from the batches not yet audited, and the calculation is performed again using the

[6] A batch could comprise all the votes cast in a precinct, the votes cast on a particular machine, or other convenient group for which a semi-official subtotal is available.

[7] The vote counts in any batches already audited, whether at random or deliberately, are treated as known perfectly. The reported margin is adjusted to reflect the manual counts in those batches, and those batches are excluded from future samples.

[8] The calculation answers the question, "If the semi-official outcome disagrees with what a full manual count would show, what is the minimum chance that the current stage of audit would have found at least one batch that overstated the margin by more than it actually did?"

adjusted margin. After a pre-determined number of steps, either the semi-official outcome has been certified, or there has been a complete manual count of the audit trail, and the correct outcome is known.[9]

CAST ensures that the chance of a full manual count is large whenever the audit trail contradicts the semi-official outcome. That controls the risk of erroneously certifying and incorrect outcome: Let β denote the minimum chance of a full manual count when a full manual count would show a different outcome. Then if an outcome is incorrect, the chance that it will be certified anyway is at most $(100\% - \beta)$.

The next section presents the method, step by step. Section 3 illustrates the calculations using a cartoon of a U.S. House of Representatives contest and explores how the initial audit sample size depends on the margin, the desired confidence level and the allocation of error across audit stages. Section 4 discusses some technical details. Section B gives R code to reproduce the examples.

2 CAST Step-by-step

This section gives the steps in CAST. It assumes that semi-official results are available for all batches of ballots before the audit starts. Batches can be grouped into *strata*[10] for convenience, with random samples of batches

[9] "Correct" means what the audit trail would show. This does not take into account the fact that a full manual count is only as good as the audit trail it counts.

[10] Strata must be *mutually exclusive* (no ballot can be in more than one stratum) and *exhaustive* (every ballot must be in some stratum).

drawn independently from different strata.[11] For example, 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. Batches might be stratified by county, if a contest crosses county lines.

Batches should never be selected from a stratum before the semi-official counts have been announced for all batches in that stratum. To do otherwise is to invite fraud.

Some batches might be selected for audit deliberately rather than randomly. For example, laws and regulations might permit the candidates in a contest each to select for audit 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 all votes by hand." A rule of that form can only increase the chance of a full manual count when such a count would show a different outcome from the semi-official outcome, so it can be used with CAST without compromising the guaranteed minimum chance of a full manual count when the semi-official outcome is wrong.

The following subsections list the steps in CAST.

[11] [4] gives two methods for dealing with stratification in the probability calculations. Many others are possible; current research seeks optimal approaches.

Step 1: pick the chance β of catching an incorrect semi-official result

The first step is to pick the minimum chance (e.g., 90%) of a full manual count when the semi-official outcome is wrong—that is, when the semi-official outcome differs from the outcome a full manual count would show. Typically, this is a matter of legislation or administrative rule. To limit auditing burden, it can be desirable to choose β smaller for small contests than for countywide or statewide contests.[12]

Step 2: pick the maximum number of stages S and the "escalation probabilities" $\beta_1, \beta_2, \ldots, \beta_S$

The second step is to pick the maximum number of sampling stages before counting all votes by hand. If the audit cannot confirm the semi-official outcome after the first stage of sampling, a second sample will be drawn, and so on. If the audit cannot confirm the semi-official outcome at stage S (because the audit has found too much error at every stage), every vote will be counted by hand.

Choose a set of S numbers, $\beta_1, \beta_2, \ldots, \beta_S$ between 0 and 1 whose product is β. For example, one could take $\beta_1 = \beta_2 = \cdots = \beta_S = \beta^{1/S}$, the S^{th} root of the desired probability of a full manual count when that would change

[12] If β is large, it will often be necessary to hand count most of the ballots in small contests, even when the semi-official outcome is correct.

the outcome.[13] The value β_s will be the minimum chance that the audit progresses from stage s to stage $s+1$ if the outcome is wrong. The larger the number of stages, generally the larger the individual values of β_s need to be to keep the overall chance of certifying an erroneous outcome under control:

[13] Suppose we want an overall chance β that there will be a complete manual count if the semi-official results are wrong. We contemplate performing at most S audit stages before conducting a full manual count—if the outcome has not been certified by one of those stages. Suppose that at each stage, we test in such a way that, if the semi-official outcome is incorrect, the chance that the audit goes from stage s to stage $s+1$ is at least β_s, given the results observed at all previous stages. Then the overall chance that the audit progresses to a full manual count is

$$\beta_1 \times \beta_2 \times \cdots \times \beta_S. \quad (1)$$

So, if we pick the values $\beta_1, \ldots, \beta_S$ so that their product is β, the chance that the audit will lead to a complete manual count when the outcome is incorrect is at least β. For example, suppose we want a 90% chance ($\beta = 90\%$) of conducting a full manual count whenever it would show that the semi-official outcome is wrong, and we contemplate drawing an initial sample, a second sample if necessary, and then proceeding to a full manual count if we have not yet certified the outcome. Then $S = 2$ and we could choose $\beta_1 = \beta_2 = \sqrt{90\%} = 94.9\%$.

We could instead take $\beta_1 = 91\%$ and $\beta_2 = 98.9\%$, since $0.91 \times 0.989 = 0.9$. Taking β_s smaller for early audit stages and larger for later stages may reduce the overall audit burden when the semi-official outcome is correct, because that makes it less likely that the audit will progress beyond the first stage.

The same argument shows that it is fine to allow targeted auditing before and between stages of random audit, provided that at stage s the test that is used guarantees that the conditional probability is at least β_s that the audit will not certify the outcome if the outcome is wrong, given everything the targeted and random audit has discovered prior to stage s.

each additional stage introduces another possibility of certifying an erroneous outcome. This is one reason to keep the number of stages relatively small. Another is that each stage entails administrative and logistical burdens. And the larger the number of stages, the more difficult it is for the public to observe the audit and verify that it has been carried out correctly.

Step 3: select the subtotals that comprise batches and the strata

The third step is to define the batches of votes from which the audit sample will be drawn, and to partition those batches into strata for convenience. Generally, the fewer votes each auditable batch contains, the smaller the audit effort required to confirm the semi-official outcome if that outcome is correct. Batches must satisfy two requirements, though: (i) semi-official counts for each batch in a stratum must be published prior to drawing the sample from that stratum. (So, if a jurisdiction does not publish semi-official totals by machine within a precinct, but only for precincts as a whole, batches must comprise at least entire precincts.) (ii) There must be an upper bound on the number of valid votes in the batch for any candidate or position in the contest. (Generally, such bounds are available only at the precinct level and above; see section 2.)

The strata must be defined so that every batch is in one and only one stratum. Generally, strata should not cross jurisdictional boundaries, so that a single jurisdiction can carry out the audit of all the batches selected from a given stratum. Let P denote the total number of batches of ballots across all strata; 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

$$B_1 + B_2 + \cdots + B_C = P. \tag{2}$$

Step 4: find upper bounds on the number of votes per candidate per batch

To establish a limit on the extent to which error in each batch could possibly overstate the semi-official margin (see section 2), one needs to know the maximum number of votes any candidate or position could possibly get, batch by batch. Upper bounds on the number of votes a candidate or position could get in any precinct can be derived from voter registrations, pollbooks, or an accounting of ballots.[14] For example, if an accounting of ballots confirms that b_p ballots were voted in precinct p, then any candidate or position could receive at most b_p votes in that precinct. See [4].

[14] An accounting of ballots compares the number of ballots sent to a precinct with the number returned voted, unvoted and spoiled to account for each piece of paper. An accounting of ballots is generally impossible for DRE voting machines, because paper ballots are not sent to polling places for DREs, even if voter-verifiable paper audit trails (VVPATS) are sent back from the polling places, so there is no way to tell whether the number of voted ballots balances. An accounting of ballots should be performed whenever possible, to ensure that ballots have neither disappeared nor materialized. Using the number of registered voters as a bound is extremely conservative and tends to require inordinately large samples unless voter turnout is extremely high.

Step 5: set initial values of the variables

Set $s = 1$; s represents the current stage of the audit. Let P_s be the number of as-yet-unaudited batches at stage s. If there has been no targeted auditing so far, then $P_1 = P$.

Step 6: calculate all pairwise margins

The sixth step is to find the all margins of victory for the contest, according to the semi-official results and what the audit has discovered so far.

For each semi-official winner w and each semi-official loser ℓ in the contest, calculate the margin of victory in votes:

$$V_{w\ell} = \text{ (votes for winner } w) \; - \; \text{(votes for loser } \ell). \tag{3}$$

In this calculation, use the semi-official results for the P_s batches that have not yet been audited, and the audit results for the $P - P_s$ batches that have already been counted by hand.

If any of these margins $V_{w\ell}$ is now zero or negative, the audit has already found so much error that the list of winners has changed. If that occurs, abort the audit and count all the votes by hand.

In a winner-take-all contest with K candidates in all, there are $K-1$ such margins of victory: the apparent winner is paired with each of the remaining candidates in turn. Some contests, such as city council races, can have several winners. For example, suppose a contest allows each voter to vote for up to 3 of 7 candidates, and the three candidates receiving the largest number of votes are the winners. Then the 3 semi-official winners each have a margin of victory over each of the 4 semi-official losers, and there are $3 \times 4 = 12$ margins of victory $V_{w\ell}$.

Step 7: find upper bounds on the maximum overstatement of pairwise margins

In each batch not yet audited, the semi-official counts in each batch, together with the bounds on the number of valid votes per candidate in each batch (section 2), limit the amount that error in a given batch could have overstated the margin between any semi-official winner and any semi-official loser.

Let candidate w be one of the semi-official winners of the contest and let candidate ℓ be one of the semi-official losers of the contest, as above. For each batch p that has not been audited, compute

$$u_{w\ell p} = \frac{(\text{votes for candidate } w \text{ in batch } p) - (\text{votes for candidate } \ell \text{ in batch } p) + b_p}{V_{w\ell}}, \tag{4}$$

according to the semi-official results for batch p. Compute the largest value of $u_{w\ell p}$ for all pairs (w, ℓ) of semi-official winners and losers. Call that number u_p. Then u_p is the most by which error in counting the votes in batch p could have overstated the margin between any apparent winner and any apparent loser, expressed as a fraction of the margin of victory between those two candidates, adjusted for any errors that were discovered in batches already audited. See [5].

Step 8: targeted audits

At this point, if there are a few un-audited batches p for which u_p is much larger than the rest, auditing those batches deliberately can reduce substantially the sample size required in the random audit to follow. If additional batches are selected for targeted audit, count them and return to step 6.

Step 9: 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 margin is overstated by t or more, the audit will progress to the next stage. Generally, the larger the value of t, the larger the sample size will need to be. If t is chosen so large that the sum over all as-yet-unaudited batches of the smaller of u_p and t is 1 or larger, a full manual count will be required to confirm the election. The smaller t is, the smaller the sample size at each stage, but when t is small there is also generally 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 start with 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 3 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, provided the sample size for stage s is calculated using the value of t for stage s. (Again, see section 3.)

Step 10: find sample sizes for the next random sample

The next step is to calculate the incremental number of batches to be selected at random from each stratum. First the total additional sample size is calculated; then that number is allocated across strata in proportion to the number of batches in each stratum. Other choices are possible; see [4]. We shall assume that the semi-official counts are available for all batches in all strata.[15]

To find the overall sample size, we define a new list of numbers. For the P_s batches p not yet audited, let t_p be the smaller of t and u_p, let T be the sum of those values t_p, and let $\tilde{u}_p = u_p - t_p$.

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
$$\left(\frac{P_s - q}{P_s}\right)^n \leq 1 - \beta_s. \tag{5}$$

3. For $c = 1, \ldots, C$, the sample size n_c for stratum c is
$$n \times \frac{\#\text{unaudited batches in stratum } c}{P_s}, \tag{6}$$

[15] If counts are available only for some strata when the audit starts, the overall margins will not be known. It is possible to proceed with some reasonable but arbitrary initial choice of sample sizes for the strata for which semi-official counts are known, although the decision of whether to certify requires semi-official counts for all batches. Step 13 then needs to be modified. If the initial sample size is too small, the audit will proceed to the next stage even if the first stage finds no error whatsoever. See [4] for more discussion.

rounded up to the nearest whole number. Thus,

$$n^* = n_1 + n_2 + \cdots + n_C \geq n. \tag{7}$$

[4] proves that these sample sizes guarantee that the chance of certifying erroneously at stage s is at most $1 - \beta_s$.

Step 11: draw the next sample and count votes

The next step is to draw the samples of batches for audit. Select batches using a transparent, mechanical, verifiable source of randomness, such as fair 10-sided dice. Computer-generated "pseudo-random" numbers are not appropriate, because it is essentially impossible for the public to verify whether the algorithm is fair or has been tampered with. For each stratum $c = 1, \ldots, C$, draw a 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 12: calculate the maximum pairwise overstatement

For each of the n^* batches p just audited in this stage, calculate

$$\begin{aligned} e_{w\ell p} &= \frac{(\text{reported votes for } w \text{ in batch } p) - (\text{reported votes for } \ell \text{ in batch } p)}{V_{w\ell}} \\ &\quad - \frac{(\text{audited votes for } w \text{ in batch } p) - (\text{audited votes for } \ell \text{ in batch } p)}{V_{w\ell}} \end{aligned}$$

for all pairs (w, ℓ) of semi-official winners w and losers ℓ. There are $n^* \times w \times \ell$ of those values. Call the largest of them t_s.

Step 13: certify, perform a full count, or proceed to the next step

If $t_s \leq t$ certify the election and stop.[16] If $t_s > t$ and we are at the final stage $s = S$, count all the votes by hand. Otherwise, add one to s; perform any additional desired targeted auditing; set P_s to be the number of batches not yet audited; and return to step 6.

3 Numerical Example: Cartoon of a U.S. House Race

This section works through an approximation of a U.S. House of Representatives contest with a semi-official margin of about 5%. The batches are all the same size and the reported votes in all the batches are the same so that all the calculations can be done by hand. If the contest is not certified within the first two stages, there will be a complete manual count.

3.1 Stage 1

Step 1.

We require probability at least $\beta = 90\%$ of a full manual count whenever that would show a different outcome.

[16]This can be refined slightly; it can be the case that $t_s > t$ but that the chance of observing a pairwise margin overstatement even larger is at least β_s if the outcome is wrong.

Step 2.

We take the maximum number of stages of auditing to be $S = 2$. We balance the chance of error between the two stages by taking $\beta_1 = \beta_2 = \beta^{1/2} = \sqrt{0.9} = 94.9\%$.[17]

Step 3.

There are 400 precincts, of which 300 are in one county and 100 are 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). We stratify on county and mode of voting, IP versus VBM, so $C = 4$. 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 $B_1 = B_2 = 300$ 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 so that the number of ballots in batch p, $b_p = 255$, can be considered to be known with certainty.

[17] When the outcome of the contest is in fact correct and very few precincts have more than a few errors that overstate the margin, it can be more efficient to allocate more of the error to the first stage, for example, 91% chance of passing from stage 1 to stage 2 if the outcome is wrong, and 98.9% chance of passing from stage 2 to a full manual count if the outcome is wrong. See table 2 below.

Step 5.

We set $s = 1$, $P_1 = 800$.

Step 6.

To keep the example simple, we suppose that the semi-official 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 and 3 undervoted ballots. The semi-official totals are given in table 1.

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

Table 1: Hypothetical semi-official election results.

The semi-official margin of candidate 1 over candidate 2 in votes is

$$V_{12} = 100,000 - 89,600 = 10,400 \text{ votes.} \tag{8}$$

The semi-official margin of candidate 1 over candidate 3 is

$$V_{13} = 100,000 - 10,400 = 89,600 \text{ votes.} \tag{9}$$

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

$$\frac{100,000 - 89,600}{200,000} \times 100\% = 5.2\%. \tag{10}$$

Step 7.

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

$$u_{p12} = \frac{125 - 112 + 255}{10,400} = 0.02577 \tag{11}$$

Similarly, the largest percentage by which error could have overstated the margin of candidate 1 over candidate 3 is

$$u_{p13} = \frac{125 - 13 + 255}{89,600} = 0.00410. \tag{12}$$

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

$$u_p = \max(0.02577, 0.00410) = 0.02577. \tag{13}$$

Step 8.

We will not perform any targeted audits.

Step 9.

We set the threshold for escalation, t, as follows: we would like to be able to certify the election even if the margin has been overstated by up to 3 votes in every precinct. (The larger we pick 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.) Note that 3 votes amounts to $3/10400 = 0.00029$ of the (smaller) margin of victory. We will proceed to the second stage if in any batch p in the first-stage sample, either e_{p12} or e_{p13} is greater than $t = 0.00029$. If that happens, we will adjust the margins to account for the errors the first stage found, take a new sample, and test again; see below.

Step 10.

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

$$\tilde{u}_p = u_p - t = 0.02548 \tag{14}$$

in every batch p. Since these are all equal, there is no sorting to do. We have not audited any batches yet, so $T = 0.00029 \times 800 = 0.23077$. The smallest number of batches for which the sum of $\tilde{u}_p \geq 1-T$ is thus the smallest whole number q so that $q \times 0.02548 \geq 1 - 0.23077$:

$$\frac{1 - 0.23077}{0.02548} = 30.2, \tag{15}$$

so $q = 31$.

Now we find the smallest whole number n so that

$$\left(\frac{800 - 31}{800}\right)^n \leq 1 - \beta_1 = 0.051. \tag{16}$$

If we take the logarithm of both sides, we find

$$n \log((800 - 31)/800) \quad \leq \quad \log(0.051)$$

$$\begin{aligned} n \times (-0.03952) &\leq -2.9759 \\ n &\geq 75.3, \end{aligned} \tag{17}$$

so $n = 76$.

We 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

$$n_1 = n_2 = 29, \tag{18}$$

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

$$n_3 = n_4 = 10, \tag{19}$$

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

Step 11.

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 every batch in the sample by hand.

Step 12.

For each audited batch p, compute e_{p12} and e_{p13}. Let t_1 be the largest of those $78 \times 2 = 156$ numbers.

Step 13.

If $t_1 \leq 0.00029$, certify the election. Otherwise, set $s = 2$ and go to the next section.

3.2 Stage 2

If a full manual count would show that candidate 1 is not the winner, there is at least a 94.9% chance that the audit will proceed to the second stage, and then at least a 94.9% chance that the audit will proceed to a full manual count, giving an overall chance of at least $94.9\% \times 94.9\% = 90\%$ that the audit will lead to a full manual count. Of course, if none of the audit batches 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 audit batches have errors that overstate the margin of candidate 1 over candidate 2 by more than 3 votes? In this section we imagine two scenarios and look at the consequences. In the first, the semi-official outcome is wrong in a particular way. In the second, the election outcome is correct, but 8 audit batches have errors that overstate the margin by more than t, and 8 have compensating errors that understate the margin.

3.2.1 The outcome is wrong

We now consider a scenario in which the outcome of the election is wrong. Suppose that in 100 of the batches, the true vote was:

$$\begin{aligned} 80 &\quad \text{votes for candidate 1,} \\ 160 &\quad \text{votes for candidate 2,} \\ 13 &\quad \text{votes for candidate 3,} \\ 1 &\quad \text{overvote, and} \\ 1 &\quad \text{undervote,} \end{aligned}$$

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

$$\begin{aligned} 124 &\quad \text{votes for candidate 1,} \\ 113 &\quad \text{votes for candidate 2,} \\ 15 &\quad \text{votes for candidate 3,} \\ 2 &\quad \text{overvotes, and} \\ 1 &\quad \text{undervote.} \end{aligned}$$

Then the true total vote is

$$\begin{aligned} 100 \times 80 + 700 \times 124 &= 94,800 \text{ votes for candidate 1,} \\ 100 \times 160 + 700 \times 113 &= 95,100 \text{ votes for candidate 2, and} \\ 100 \times 13 + 700 \times 15 &= 11,800 \text{ votes for candidate 3.} \end{aligned}$$

The semi-official 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

$$\frac{(125-112)-(80-160)}{10,400} = 0.00894, \tag{20}$$

and in the 700 batches with smaller errors it is

$$\frac{(125-112)-(124-113)}{10,400} = 0.00019. \tag{21}$$

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

$$\frac{(125-13)-(80-13)}{89,600} = 0.00050, \tag{22}$$

and in the 700 batches with smaller errors it is

$$\frac{(125-13)-(120-15)}{89,600} = 0.00008. \tag{23}$$

The first and third of these four overstatements are bigger than $t = 0.00029$; they occur in the same batches. In this scenario, the chance that the first stage of the audit would find one or more batches with margin overstatements e_p greater than 0.00029 (so that the audit would progress to the second stage) is at least

$$1 - \left(\frac{800-100}{800}\right)^{76} = 99.996\%, \tag{24}$$

rather greater than the guaranteed minimum of 94.9%.

If the sample were not stratified, the expected number of batches in the sample with large errors would be

$$78 \times \frac{100}{800} = 9.75. \tag{25}$$

Suppose that the sample found 10 of the batches with large errors and 68 of the batches with small errors. We would return to step 6, but with $s = 2$ and $P_2 = 800 - 78 = 722$ as-yet-unaudited precincts.

Step 6.

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

$$\begin{aligned}
10 \times 80 + 68 \times 124 + 722 \times 125 &= 99,482 \text{ votes for candidate 1} \\
10 \times 160 + 68 \times 113 + 722 \times 112 &= 90,148 \text{ votes for candidate 2} \\
10 \times 13 + 68 \times 15 + 722 \times 13 &= 10,536 \text{ votes for candidate 3.}
\end{aligned}$$

The margins, adjusted for the error found at stage 1, are

$$\begin{aligned}
V_{12} &= 99,482 - 90,148 = 9,334 \text{ (a reduction of 1,066 votes)} \\
V_{13} &= 99,482 - 10,536 = 88,946 \text{ (a reduction of 654 votes).} \qquad (26)
\end{aligned}$$

Step 7.

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

$$\begin{aligned}
u_{p12} &= \frac{125 - 112 + 255}{9,334} = 0.02871 \\
u_{p13} &= \frac{125 - 13 + 255}{88,946} = 0.00412. \qquad (27)
\end{aligned}$$

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

$$u_p = \max(0.02871, 0.00412) = 0.02871. \qquad (28)$$

Step 8.

We will not perform any targeted audits.

Step 9.

We adjust the threshold to correspond to a 3 vote overstatement of the margin of victory: $t = 3/9334 = 0.00032$.

Step 10.

As before, $\beta_2 = 94.9\%$. Since $u_p = 0.02871$ and $t = 0.00032$, $t_p = t = 0.00032$ in every batch p, and

$$\tilde{u}_p = u_p - t_p = 0.02839 \tag{29}$$

in every batch p. Since these are all equal, there is no sorting to do. We have 722 batches not yet audited, so $T = 722 \times 3/9334 = 0.23205$. The smallest number of batches for which the sum of $\tilde{u}_p \geq 1 - T$ is thus the smallest whole number q so that $q \times 0.02839 \geq 1 - 0.23205$:

$$\frac{1 - 0.23205}{0.02839} = 27.05, \tag{30}$$

so $q = 28$.

Now we find the smallest whole number n so that

$$\left(\frac{722 - 28}{722}\right)^n \leq 1 - \beta_2 = 0.051. \tag{31}$$

If we take the logarithm of both sides, we find

$$n \log((722 - 28)/722) \quad \leq \quad \log(0.051)$$

$$
\begin{aligned}
n \times (-0.03955) &\leq -2.9759 \\
n &\geq 75.1,
\end{aligned} \tag{32}
$$

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

As before, we 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

$$
n_1 = n_2 = 29, \tag{33}
$$

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

$$
n_3 = n_4 = 10, \tag{34}
$$

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 11.

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 12.

For each audited batch p, compute e_{p12} and e_{p13}. Let t_2 be the largest of those $78 \times 2 = 156$ numbers.

Step 13.

If $t_2 \leq 0.00032$, certify the election. Otherwise, perform a full manual count.

In this hypothetical example, we found 10 of the 100 batches with large errors in stage 1, so 90 of those batches remained among the 722 that were not audited at stage 1. The chance that the second-stage sample contains at least one of them—so that

$$t_2 = \frac{(125 - 112) - (80 - 160)}{9,334} = 0.00996 \tag{35}$$

and a full manual count occurs—is at least

$$1 - \left(\frac{722 - 90}{722}\right)^{76} = 99.996\%, \tag{36}$$

again rather larger than the guarantee of 94.9%.

3.2.2 The outcome is correct

Suppose that 8 audit batches (1%) have errors that overstate the margin between candidate 1 and candidate 2 by more than t, and 8 audit batches have compensating errors that understate the margin. The outcome of the contest is correct. Nonetheless, there is a chance that sample will 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 audit batches with large errors are spread across the 4 strata. If all 8 are in one of the small strata, that maximizes the chance that the audit will go to stage 2. The chance is then

$$1 - \frac{\binom{92}{10}}{\binom{100}{10}} = 58.3\%. \tag{37}$$

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:

$$1 - \left(\frac{\binom{297}{29}}{\binom{300}{29}} \right)^2 \times \left(\frac{\binom{99}{10}}{\binom{100}{10}} \right)^2 = 56.1\%. \tag{38}$$

The chance that the audit will progress from stage 2 to a full manual count depends on the number and location of the audit batches with large overstatement or understatement errors are uncovered at stage 1. 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 manual count.

To get a feel for the probability of a full manual count, suppose that there are 8 overstatement errors and 8 understatement errors distributed in the strata as follows: in each of the two large strata, there are 3 audit batches with errors that overstate the margin between candidate 1 and candidate 2 (V_{12}) by 10 votes and 3 audit batches with errors that understate V_{12} by 10 votes, and that in each of the two small strata, there is 1 audit unit with errors that overstate V_{12} by 10 votes and one with errors that understate V_{12} by 10 votes. In the first-stage sample, the expected number of audit batches with 10 vote overstatements of V_{12} is

$$3 \times \frac{29}{300} + 3 \times \frac{29}{300} + 1 \times \frac{10}{100} + 1 \times \frac{10}{100} = 0.78. \tag{39}$$

The expected number with 10 vote understatements of V_{12} is the same. Suppose that the first stage sample finds 1 batch with a 10 vote overstatement of V_{12} in one of the large strata, and 1 batch with a 10 vote understatement of V_{12} in some stratum. Then $t_1 > t$, so the audit will go to the second stage at step 6.

Step 6

The net error found in the first stage is zero, so the adjusted margins remain equal to the semi-official margins:

$$V_{12} = 100,000 - 89,600 = 10,400 \text{ votes} \tag{40}$$

and

$$V_{13} = 100,000 - 10,400 = 89,600 \text{ votes.} \tag{41}$$

Step 7.

Since the margins have not changed, we still find

$$u_p = \max(0.02577, 0.00410) = 0.02577. \tag{42}$$

Step 8.

We will not perform any targeted audits.

Step 9.

Again, because the margins have not changed, a 3 vote overstatement of V_{12}, the smaller margin, corresponds to $t = 3/10400 = 0.00029$.

Step 10.

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

$$\tilde{u}_p = u_p - t = 0.02548 \tag{43}$$

in every batch p. Since these are all equal, there is no sorting to do. There are 722 unaudited batches, so $T = 0.00029 \times 722 = 0.20938$. The smallest number of batches for which the sum of $\tilde{u}_p \geq 1 - T$ is thus the smallest whole number q so that $q \times 0.02548 \geq 1 - 0.20938$:

$$\frac{1 - 0.20938}{0.02548} = 31.03, \tag{44}$$

so $q = 32$.

Now we find the smallest whole number n so that

$$\left(\frac{722 - 32}{722}\right)^n \leq 1 - \beta_1 = 0.051, \tag{45}$$

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

$$n_1 = n_2 = 25, \tag{46}$$

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

$$n_3 = n_4 = 9, \tag{47}$$

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 11.

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 12.

For each audited batch p, compute e_{p12} and e_{p13}. Let t_2 be the largest of those $68 \times 2 = 136$ numbers.

Step 13.

If $t_2 \leq 0.00029$, certify the election. Otherwise, count the remaining $800 - 78 - 68 = 654$ ballots manually.

In this scenario, the chance that the second stage sample has one or more of the audit batches with a 10 vote overstatement of V_{12} is

$$1 - \frac{\binom{269}{25}}{\binom{271}{25}} \times \frac{\binom{268}{25}}{\binom{271}{25}} \times \left(\frac{\binom{89}{9}}{\binom{90}{9}}\right)^2 = 50.2\%. \tag{48}$$

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

$$56.1\% \times 50.2\% = 28.1\% \tag{49}$$

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

For example, if there is only one audit batch that 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 manual count is zero. If there are two audit batches that 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, there is no chance of proceeding to a full manual count; if not, the chance of proceeding to a full manual count is on the order of 10%. The overall chance of a full manual count would be at most 1.8%, neglecting any adjustments to the margin. Table 2 presents the maximum probability of an unnecessary full manual count in a variety of scenarios.

3.3 Varying the Assumptions

In this hypothetical example—with a 5.2% margin, 90% confidence split evenly across 2 stages ($\beta_1 = \beta_2 = 94.9\%$) and 800 audit batches—the initial sample size is nearly 10% of the audit batches. Table 2 shows how the first-stage sample size changes as some of these choices are varied: the margin, the confidence level, and the allocation of error 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 2 also shows the maximum probability that the audit will progress to the second stage if 1% (i.e., 8) or 0.5% (i.e., 4) of the 800 audit 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 $\gamma_{0.005}$ in the table. Finally, table 2 gives upper bounds on the probability of a full manual 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 probability 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 placed in one stratum, which makes it most likely that the audit will progress to the next stage and to a full manual count. (If they were distributed randomly, the probability of a full manual count would be lower.) Appendix A discusses how these bounds are calculated. Sample R code is given in appendix B.

4 Discussion and technical notes

4.1 Background error rate

The treatment of the background error rate—the threshold margin overstatement t—can be sharpened considerably. Each audit unit is in one of two categories: margin overstatement of t or less, or margin overstatement exceeding t. The test at each stage is based on the number of audit batches in the sample that are in the second category—the audit progresses to the

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
V_{12}	β	β_1	β_2	n	n^*	f	$\gamma_{0.01}$	$\omega_{0.01}$	$\gamma_{0.005}$	$\omega_{0.005}$	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 2: Hypothetical 2-stage audit with 3 vote error threshold. Contests have 4 strata, two with 300 audit 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 manual 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 manual 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 audit batches overstate V_{12} by more than 3 votes. (9) Maximum chance of a full manual count if 1% of audit batches overstate V_{12} by more than 3 votes, and the stage 1 net error is zero. (10) Same as (8), but with 0.5% of batches having large overstatements of V_{12}. (11) Same as (9), but with 0.5% of batches having large overstatements of V_{12}. (12) Stage 2 sample size if the net error in stage 1 is zero.

next stage if that number is not zero. To ensure that the test is conservative, audit batches are presumed to have the largest margin overstatement they can without changing categories. That is, audit batches in the first category are treated as if they have margin overstatements of t, and audit 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 3, setting t to correspond to a 3-vote overstatement of the margin between the winner and the runner-up means that every audit unit can overstate the margin by $3/250 = 1.2\%$ of its 250 votes. Hence, for that value of t, if the semi-official margin were 1.2% or less, the method would require a full manual count.

One can construct a sharper test by grouping audit batches into more than two categories. For example, audit batches could be divided into those with no margin overstatement, overstatement between 0 and some threshold t, and audit batches with margin overstatements greater than t. Using three categories leads to "trinomial bounds;" more generally, one can derive multinomial bounds. (Multinomial bounds have been used in financial auditing [1, 3] using a different sampling design: unstratified sampling with probability proportional to size. The categories are pennies out of each dollar of potential error, 101 categories in all.) Work in progress (Miratrix and Stark, 2008) develops a test based on a trinomial bound. The audit progresses from one stage to the next if the number of audit batches in the sample in either or both of the second two categories is too large.

4.2 Improving the treatment of stratification

The sample size calculation at step 10 assumes that the q batches with enough error to alter the apparent outcome could be distributed arbitrarily among the strata. In fact, unless the error bounds u_p are all equal, batches that can hold large errors might be concentrated in a relatively small subset of the strata, which would make it more likely 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. Work in progress (Higgins and Stark, 2008) sharpens the treatment of stratification by calculating the worst feasible case: the attainable distribution of error across batches that could account for the apparent margin and minimizes the probability of finding any batches with large errors in a stratified sample.

When the strata are small, the probability bound CAST uses (based on sampling with replacement) is weak. Small contests are generally contained within a single jurisdiction, though, so there may be no need for more than two strata—votes cast in precinct and votes cast by mail. In that case, it can be tractable to derive sharper probability bounds using the hypergeometric distribution in each stratum. Work in progress (Higgins and Stark, 2008) sharpens the binomial bound when audit batches in different strata have different error bounds.

4.3 Comparison with previous work

CAST differs from the method of [4, 5] in important ways. Instead of using Bonferroni's inequality to bound the overall chance of certifying an incorrect outcome, CAST controls the probability by designing the test at each stage to control the conditional probability of certifying erroneously.

Second, CAST conditions on the audit results at previous stages, rather than treating the sample at stage s as a "telescoping" sample that includes previous stages as [4] does. This has a number of benefits. First, 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. Second, margin overstatement errors discovered at one stage do not lead to more than one step of escalation if there are canceling margin understatement errors: while only overstatement errors are involved in the test statistic at a given stage, the margin is adjusted sequentially to account for both overstatement and understatement errors. Third, this makes it easy to incorporate "targeted" sampling.

Acknowledgments. I am grateful to Lynn Garland, Edward Gracely, Mike Higgins, Mark Lindeman, John McCarthy, Luke Miratrix, Jasjeet Sekhon and Howard Stanislevic for comments on an earlier draft.

A Bounds on the probability of escalation

This section finds the maximum probability that the audit will progress from one stage to the next if there are a given number q of audit batches with margin overstatements greater than t. It also shows how to use that maximum to bound the probability of an unnecessary full manual count on the assumption that the net error discovered in each stage cancels so that the margin does not change.

The basic result is that the probability that the sample will contain none of the q batches with a large margin overstatement is smallest when all q of those batches are in one stratum (if there is any stratum c for which N_c, the size of the stratum, minus n_c, the size of the sample drawn from that stratum, is less than q, the probability can be 100%). We establish this by means of the following lemma.

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

$$\frac{\binom{N-r}{n}}{\binom{N}{n}} \times \frac{\binom{M-v+r}{m}}{\binom{M}{m}} \geq \min\left\{ \frac{\binom{N-v}{n}}{\binom{N}{n}}, \frac{\binom{M-v}{m}}{\binom{M}{m}} \right\}. \tag{50}$$

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

$$\frac{\binom{K-\ell}{k}}{\binom{K}{k}} = \frac{(K-k)(K-k-1)\cdots(K-k-\ell+1)}{K(K-1)\cdots(K-\ell+1)}. \tag{51}$$

Applying 51 twice, we find

$$\frac{\binom{N-r}{n}}{\binom{N}{n}} \times \frac{\binom{M-v+r}{m}}{\binom{M}{m}} = \frac{(N-n)(N-n-1)\cdots(N-n-r+1)}{N(N-1)\cdots(N-r+1)} \times \frac{(M-m)(M-m-1)\cdots(M-m-v+r+1)}{M(M-1)\cdots(M-v+r+1)} \tag{52}$$

Identity 51 also shows that

$$\frac{\binom{N-v}{n}}{\binom{N}{n}} = \frac{(N-n)(N-n-1)\cdots(N-n-v+1)}{N(N-1)\cdots(N-v+1)} \tag{53}$$

and

$$\frac{\binom{M-v}{m}}{\binom{M}{m}} = \frac{(M-m)\cdots(M-m-v+1)}{M(M-1)\cdots(M-v+1)}. \tag{54}$$

Thus 52 is to be compared with

$$\min\left\{\frac{(N-n)\cdots(N-n-v+1)}{N(N-1)\cdots(N-v+1)}, \frac{(M-m)\cdots(M-m-v+1)}{M(M-1)\cdots(M-v+1)}\right\}. \tag{55}$$

Now

$$\frac{(N-n)\cdots(N-n-v+1)}{N(N-1)\cdots(N-v+1)} = \frac{(N-n)\cdots(N-n-r+1)}{N(N-1)\cdots(N-r+1)} \times \frac{(N-n-r)\cdots(N-n-v+1)}{(N-r)(N-r-1)\cdots(N-v+1)} \tag{56}$$

Suppose without loss of generality that the first term in the minimum in 55 is the smaller of the two. Then the lemma can be false only if

$$\frac{(N-n)\cdots(N-n-r+1)}{N(N-1)\cdots(N-r+1)} \times \frac{(M-m)\cdots(M-m-v+r+1)}{M(M-1)\cdots(M-v+r+1)} \tag{57}$$

$$< \frac{(N-n)(N-n-1)\cdots(N-n-v+1)}{N(N-1)\cdots(N-v+1)}; \tag{58}$$

i.e., only if

$$\frac{M-m}{M}\times\cdots\times\frac{M-m-v+r+1}{M-v+r+1} < \frac{N-n-r}{N-r}\times\cdots\times\frac{N-n-v+1}{N-v+1}. \quad (59)$$

Both sides of 59 are products of $v-r$ terms; inequality 59 is equivalent to the assertion that the geometric mean of the terms on the left is less than the geometric mean of the terms on the right. If so, then the smallest of the terms on the left is smaller than the largest of the terms on the right. Since $x<y$ implies that $\frac{x-1}{y-1} < \frac{x}{y}$, the smallest term on the left is $\frac{M-m-v+r+1}{M-v+r+1}$ and the largest term on the right is $\frac{N-n-r}{N-r}$. Thus

$$\begin{aligned}\frac{M-m-v+1}{M-v+1} &< \frac{M-m-v+2}{M-v+2} < \cdots < \frac{M-m-v+r}{M-v+r} < \frac{M-m-v+r+1}{M-v+r+1} \\ &< \frac{N-n-r}{N-r} < \frac{N-n-r+1}{N-r+1} < \cdots < \frac{N-n-1}{N-1} < \frac{N-n}{N}. \quad (60)\end{aligned}$$

Therefore,

$$\begin{aligned}&\frac{M-m-v+1}{M-v+1}\times\frac{M-m-v+2}{M-v+2}\times \cdots \times\frac{M-m-v+r}{M-v+r} \\ &< \frac{N-n-r+1}{N-r+1}\times\frac{N-n-r+2}{N-r+2}\times \cdots \times\frac{N-n-1}{N-1}\times\frac{N-n}{N}, \quad (61)\end{aligned}$$

Multiplying the left side of 59 times the left side of 61 and the right side of 59 times the right side of 61 shows that if the lemma is false,

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

which contradicts the assumption that the first term in 55 is the smaller of the two. •

The lemma implies that for any pair of strata that have a total of v audit batches with large overstatements, the probability that the sample contains at least one audit batch with a large overstatement is maximized if all v are in the same stratum (provided they all fit; otherwise the probability is 100%). If we use the independence of the samples in different strata and apply the lemma result recursively to pairs of strata, we see that the probability that the sample contains at least one audit batch with a large overstatement is maximized when all the batches with large overstatements are in the same stratum—if every stratum has room for all of them. Hence the maximum probability of escalation can be found by first checking whether the number q of batches with large overstatements is greater than or equal to the smallest value of $N_c - n_c$. If so, the maximum probability of escalation is 100%. If not, the maximum probability 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 that the audit progresses 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. Clearly, the larger the value of v, 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 progress to the next stage.)

Let n_{sc} denote the size of the sample drawn at stage s from stratum c, and define $n_{0c} \equiv 0$. Suppose that there are q batches with overstatement errors greater than t in the original population of batches. If $q < S$, the chance of a full manual count is zero—if the audit gets to stage $S-1$, there are no more batches with large errors left to find. Otherwise,

$$\Pr\{\text{full manual count}\} \leq \prod_{s=1}^{S} \left(1 - \min_c \frac{\binom{N_c - \sum_{r=1}^{s-1} n_{rc} - (q-s+1)}{n_{sc}}}{\binom{N_c - \sum_{r=1}^{s-1} n_{rc}}{n_{sc}}} \right). \qquad (63)$$

B R code to replicate the example

The following R commands calculate the initial stratum sample sizes $n_1, \ldots, n_4$, the total initial sample size n^*, and the maximum probability of escalation to the 2nd stage as given in section 3.

```
####################  R Code for House contest cartoon
samSizes <- function(b, Cv, t, u) {
    # b: the minimum probability of advancing to the next stage
    #     if the outcome is wrong
    # Cv: vector of strata sizes, not-yet-audited batches
    # t: threshold for tolerable error
    # u: upper bound on the margin overstatement in each batch
    tp <- min(t, u); # tolerable overstatement a batch can hold
    utilde <- u-tp;  # overstatement above t a batch can hold
    P <- sum(Cv);    # total batches in the contest
    T <- tp*P;       # bit of margin accounted for by t in each batch
    q <- ceiling((1-T)/utilde);  # minimum no. tainted batches required to
                     #  account for remainder of the margin
    n <- ceiling( log(1-b)/log((P-q)/P) ); # sample size for sampling with
```

```
                     #  replacement, unstratified
    ceiling(n*Cv/P) # sample sizes in each stratum
}

S <- 2;                        # max stages before a full hand count
Cv <- c(300, 300, 100, 100); # strata sizes in batches
P <- sum(Cv);                  # total number of batches
alterFrc <- c(0.01, 0.005);  # fractions of batches with errors,
                               #  alternative hypothesis
alter <- ceiling(alterFrc*sum(Cv)); # batches w/ errors, alternative
                               #  hypothesis
vtol <- 3;                     # tolerable overstatement before escalation, votes

v <- c(125, 112, 13, 2, 3);  # vote in batch: candidates 1, 2, 3, overvotes,
                               #  undervotes 5.2% margin scenario
# v <- c(131, 106, 13, 2, 3); # 10% margin scenario
# v <- c(143, 94, 13, 2, 3); # 19.6% margin scenario

v12 <- P*(v[1]-v[2]);        # margin of candidate 1 vs candidate 2
v13 <- P*(v[1]-v[3]);        # margin of candidate 1 vs candidate 3
margin <- 100*v12/(P*(v[1]+v[2]+v[3])) # margin in percent
t <- vtol/min(v12, v13);     # tolerable overstatement as fraction of smaller
                               #  margin
u12 <- (v[1]-v[2]+sum(v))/v12; # bound on overstatement of v12
```

```
u13 <- (v[1]-v[3]+sum(v))/v13; # bound on overstatement of v13
u <- max( u12, u13 );        # bound on overstatement of either margin

beta <- 0.9;                 # min chance of full hand count for wrong outcome
betav <- rep(beta^(1/S),S);  # min chance of escalation for wrong outcome
# betav <- c(beta+0.01, beta/(beta+0.01)); # allocate more error to 1st stage

nv1 <- samSizes(betav[1], Cv, t, u);

gamma11 <- 1-min(dhyper(0, alter[1], Cv-alter[1], nv1)); # max chance of
                             #  escalation to 2nd stage if alter[1] batches
                             #  have big errors
gamma12 <- 1-min(dhyper(0, alter[2], Cv-alter[2], nv1)); # max chance of
                             #  escalation to 2nd stage if alter[2] batches
                             #  have big errors
Cv2 <- Cv - nv1              # un-audited batches remaining in each
                             #  stratum after stage 1
nv2 <- samSizes(betav[2], Cv2, t, u);
gamma21 <- 1-min(dhyper(0, alter[1]-1, Cv2-alter[1]+1, nv2)); # max chance of
                             #  escalation from stage 2 to full manual count if
                             #  alter[1]-1 batches have big errors
gamma22 <- 1-min(dhyper(0, alter[2]-1, Cv2-alter[2]+1, nv2)); # max chance of
                             #  escalation from stage 2 to full manual count if
                             #  alter[2]-1 batches have big errors
```

```
omega1 <- gamma11*gamma21;   # max chance of full hand count, alter[1] errors
                             # (conditional on no net error in 1st stage,
                             #  1 error found in 1st stage)
omega2 <- gamma12*gamma22;   # max chance of full hand count, alter[2] errors
                             # (conditional on no net error in 1st stage,
                             #  1 error found in 1st stage)

#######  output #######
margin                       # margin in percent
betav                        # min escalation probs, wrong outcome
nv1                          # 1st stage stratum sample sizes
sum(nv1)                     # overall sample size including stratification
sum(nv2)                     # 2nd stage sample size including stratification
100*sum(nv1)/sum(Cv)         # fraction of audit batches in the initial sample
c(gamma11, omega1)           # max escalation & full count prob, 1% errors
c(gamma12, omega2)           # max escalation & full count prob, 0.5% errors
```

References

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

[2] E. Ginnold, J.L. Hall, and P.B. Stark, *A risk-limiting audit of Measure A in the February 2008 Marin County election*, 2008.

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

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

[5] ______, *A sharper discrepancy measure for post-election audits*, Ann. Appl. Stat. **in press** (2008), in press.