# Stratified Risk-Limiting Audits: A Path Forward for Colorado

Kellie Ottoboni[1][0000−0002−9107−3402], Philip B. Stark[1][0000−0002−3771−9604], Mark Lindeman[2][0000−0001−8815−815X], and Neal McBurnett[0000−0001−8667−1830]

[1] Department of Statistics, University of California, Berkeley, CA, USA
[2] Department of Political Science, Columbia University, NY, USA

**Abstract.** Colorado CRS 1-7-515 requires risk-limiting tabulation audits (RLAs) starting in 2017. Most Colorado counties (comprising 98.2% of voters) have voting equipment amenable to ballot-level comparison audits, but some are only able to perform ballot-polling audits. How to combine ballot-polling and ballot-level comparison audits to check outcomes of contests that cross jurisdictional lines has not been addressed. Moreover, Colorado's current audit software (RLATool) does not support audits of cross-jurisdictional contests at all. This paper addresses both gaps, along the way introducing a simple, efficient way to use stratified sampling in RLAs. (Stratification makes it easier to combine ballot-polling and ballot-level comparison, and also is useful to reduce coordination among jurisdictions required to audit cross-jurisdictional contests.) We present simple but inefficient methods, more efficient methods that combine ballot polling and ballot-level comparisons using stratified samples, and methods that combine ballot-level comparison and variable-size batch comparison audits without stratification, noting the changes to RLATool that each of these methods would require. We provide open-source reference implementations of the preferred methods in Jupyter notebooks.

**Keywords:** stratified sampling, Fisher's combining function, sequential hypothesis tests

**Acknowledgements.** We are grateful to Ronald L. Rivest and Steven N. Evans for helpful conversations and suggestions.

## 1 Introduction

A risk-limiting audit (RLA) of an election is a procedure that has a known, prespecified minimum chance of correcting the electoral outcome if the outcome is incorrect—that is, if the reported outcome differs from the outcome that a full manual tabulation of the votes would find. RLAs require a durable, voter-verifiable record of voter intent, such as paper ballots, and they assume that this audit trail is sufficiently complete and accurate that a full hand tally would

show the true electoral outcome. That assumption is not automatically satisfied: a *compliance audit* [13] is required.

Risk-limiting audits are generally (but not necessarily) incremental: they examine more ballots, or batches of ballots, until either (i) there is strong statistical evidence that a full hand tabulation would confirm the outcome, or (ii) the audit has led to a full hand tabulation, the result of which should become the official result.

RLAs have been piloted in California, Colorado, and Ohio, and a test of RLA procedures has been conducted in Arizona. RLA bills are being drafted or are already under consideration in California, Virginia, Washington, and other states. A number of laws have either allowed or mandated risk-limiting audits, including California AB 2023 (Saldaña), SB 360 (Padilla), and AB 44 (Mullin); Rhode Island SB 413A and HB 5704A; and Colorado Revised Statutes (CRS) 1-7-515. At the time of writing, California is considering another RLA bill, AB 2125.

CRS 1-7-515 requires Colorado to conduct risk-limiting audits beginning in 2017. (There are provisions to allow the Secretary of State to exempt some counties.) The first set of coordinated risk-limiting election audits across the state took place in Colorado in November, 2017.[3] Those audits only covered contests restricted to a single county: counties could audit independently. To audit statewide elections and contests that cross county lines, Colorado will need to implement new approaches and modify RLATool[4], the open-source audit software used for their 2017 audits.

Colorado's "uniform voting system" program[5] led many Colorado counties to purchase (or to plan to purchase) voting systems that are auditable at the ballot level: those systems export cast vote records (CVRs) for individual ballots in a manner that allows the corresponding paper ballot to be identified, and conversely, make it possible to find the CVR corresponding to any particular paper ballot. We call counties that have such systems "CVR" counties. It is estimated that by June, 2018, 98.2% of active Colorado voters will be in CVR counties. CVR counties can perform "ballot-level comparison audits," [6] which are currently the most efficient approach to risk-limiting audits in that they require examining the fewest ballots than other methods do, when the outcome of the contest under audit is in fact correct.

Voting systems in other counties ("legacy" or "no-CVR" counties) do not allow auditors to check how the system interpreted voter intent for individual ballots. Election results involving those counties can still be audited, provided the voting systems create a voter-verifiable paper trail (*e.g.*, voter-marked paper ballots) that is conserved to ensure that it remains accurate and intact, and organized well enough to permit ballots to be selected at random. Pilot audits in California [2] suggest it is more efficient to audit such systems using"ballot-polling" [5,6] than using "batch-level comparisons."

[3] See https://www.sos.state.co.us/pubs/elections/RLA/2017RLABackground.html
[4] https://github.com/FreeAndFair/ColoradoRLA/
[5] https://www.sos.state.co.us/pubs/elections/VotingSystems/UniformVotingSystem.html

There is currently no literature on how to combine ballot polling and ballot-level comparisons in a single risk-limiting audit. Existing methods would either require all counties to use the lowest common denominator (ballot-polling, which does not take advantage of the CVRs, and thus is expected to require more auditing than a method that uses the available CVRs), or would require no-CVR counties to perform batch-level comparisons, which were found in California to be less efficient than ballot-polling audits.[6]

The audit software RLATool needs additional features to be able to audit contests that cross county lines. First, the current version (1.1.0) of RLATool needs to be modified to account for contests that cross jurisdictional boundaries; currently, it treats every contest as if it were entirely contained in a single county. Second, to audit a contest that includes votes in CVR counties and votes in no-CVR counties, new statistical methods are needed to preserve the efficiency of ballot-level comparison audits in the CVR counties.

We focus on near-term requirements for risk-limiting audits in Colorado, developing a new, widely applicable method for using stratified sampling in RLAs. Section 3 presents crude but inefficient approaches that could be implemented easily. Section 4 presents an approach based on comparison audits with different batch sizes. This approach is statistically simple and relatively efficient, but might require changing how counties handle their ballots. Section 5 presents our recommended approach, which combines ballot-level comparisons in counties that can perform them with ballot-polling in the no-CVR counties. All the approaches require new software, including some changes to RLATool. We provide example software implementing the risk calculations for our recommended approach as a Python Jupyter notebook.[7] Sections 6 and 7 explain modifications to ballot-level comparison and ballot-polling audits needed for the stratified audit. Section 8 gives recommendations and considerations for implementation.

## 2 Preliminary notation

Here and generally throughout the paper, we discuss auditing a single plurality contest at a time, although the same sample can be used to audit more than one contest (and super-majority contests), and there are ways of combining audits of different contests into a single process [9,11]. We use terminology drawn from a number of papers; the key reference is Lindeman and Stark, 2012 [6]. An *overstatement error* is an error that caused the margin between *any* reported winner and *any* reported loser to appear larger than it really was. An *understatement error* is an error that caused the margin between *every* reported winner and *every* reported loser to appear to be smaller than it really was.

Throughout, we will refer to a contest between reported winner $w$ and reported loser $\ell$. The total number of reported votes for candidate $w$ is denoted

[6] See [7] for a different (Bayesian) approach to auditing contests that include both CVR counties and no-CVR counties. In general, Bayesian audits are not risk-limiting.

[7] See https://github.com/pbstark/CORLA18.

$V_w$ and the total for candidate $\ell$ is denoted $V_\ell$, so that $V_w > V_\ell$, since $w$ is the reported winner. Additional notation is introduced below as the need arises.

## 3 Simple approaches

### 3.1 Hand count the legacy counties

The simplest approach to combining legacy counties with CVR counties is to require every legacy county to do a full hand count, and to conduct a ballot-level comparison audit in CVR counties, based on contest margins adjusted for the results of the manual tallies in the CVR counties. For instance, imagine a contest with two candidates, reported winner $w$ and reported loser $\ell$. Suppose that a full manual tally of the votes in the legacy counties shows $V'_w$ votes for $w$ and $V'_\ell$ votes for $\ell$. Suppose that a total of $N$ ballots were cast in the CVR counties. Then the *diluted margin* for the comparison audit in the CVR counties is defined to be $[(V_w - V'_w) - (V_\ell - V'_\ell)]/N$. Requiring a full hand count in the legacy counties has obvious disadvantages, but does not force CVR counties to do additional auditing to compensate for the legacy counties.

### 3.2 Treat legacy counties as if every ballot selected from them for audit has a two-vote overstatement

Another simple-but-inefficient approach is to sample uniformly from all counties as if one were performing a ballot-level comparison audit everywhere, but to treat any ballot selected from a legacy county as a two-vote overstatement, essentially following [1].

## 4 Variable batch sizes

Another approach is to perform a comparison audit across all counties, but to use batches consisting of more than one ballot (batch-level comparisons) in legacy counties and batches consisting of a single ballot (ballot-level comparisons) in CVR counties.[8] This requires that the no-CVR counties report vote subtotals for physically identifiable batches. If a county's voting system can only report subtotals by precinct but the county does not sort paper ballots by precinct, this approach might require revising how the county handles its paper; we understand that this is the case in many Colorado counties.

That said, many California counties that do not sort vote-by-mail (VBM) ballots by precinct conduct the statutory 1% audits by manually retrieving the ballots for just those precincts selected for audit from whatever physical batches they happen to be in: the situation is identical to that in Colorado.

[8] For majority and plurality elections, including those in which voters can select more than one candidate, audits can be based on overstatement and understatement errors at the level of batches.

Another tactic is the "Boulder-style" batch-level audit,[9] which requires generating vote subtotals after each physical batch is scanned, and exporting those subtotals in machine-readable form. That in turn may require using extra memory cards, repeatedly initializing and deleting tabulation databases, or other measures that add complexity and opportunity for error.

Those two approaches are labor-intensive but provide a viable short-term solution, especially combined with information from SCORE[10] to check that the reported batch-level results contain the correct number of ballots for each contest under audit. Moreover, this approach does not unduly increase the workload in CVR counties to compensate for legacy equipment.

Variable-batch-size comparison audits require modifying or augmenting RLATool in several ways:

1. The CVR reporting tool would need to be modified to allow no-CVR counties to report batch-level results in a manner analogous to how CVR counties report ballot-level results, or an external tool would need to be provided.
2. The sampling algorithm would have to allow sampling batches with unequal probability. Efficient batch-level audits involve sampling batches with probability proportional to a bound on the possible overstatement error in the batch. It would also need to calculate the appropriate sampling probability for each batch. Again, this could be accommodated using an external tool to draw the sample from legacy counties.
3. The risk calculations would need to be modified. This, too, could be done with external software, with suitable provisions for capturing audit data from RLATool or directly from legacy counties.

None of these changes is enormous; the details are worked out in published papers [9,10,11] including calculating batch-level error bounds, drawing the samples with probability proportional to an error bound, and calculating the attained risk from the sample results. Indeed, this is the method that was used in several of California's pilot audits, including the audit in Orange County. Section 6 derives a method for comparison audits with variable batch sizes.

## 5 Stratified "hybrid" audits

Other approaches involve *stratified sampling*: partitioning the cast ballots into non-overlapping groups and sampling independently from those groups. [8,4] discuss stratified sampling in batch-level comparison audits. One could stratify by county, but in general it is simpler and more efficient statistically (i.e., it requires examining fewer ballots) to minimize the number of strata. We consider methods that use two strata: one comprising the ballots cast in CVR counties and the other comprising the ballots cast in no-CVR counties. Every ballot cast

[9] See http://bcn.boulder.co.us/~neal/elections/boulder-audit-10-11/.

[10] SCORE is Colorado's Statewide Voter Registration System; see https://www.sos.state.co.us/pubs/elections/SCORE/SCOREhome.html (last visited 5 May 2018).

in the contest is in exactly one of the two strata. (Stratification is useful in other circumstances, too, for instance to "decouple" sampling in different counties. The method we develop here works generally to construct a RLA using stratified sampling.)

Let $V_{w\ell} > 0$ denote the contest-wide margin (in votes) of reported winner $w$ over reported loser $\ell$. Let $V_{w\ell,s}$ denote the margin (in votes) of reported winner $w$ over reported loser $\ell$ in stratum $s$. Note that $V_{w\ell,s}$ could be negative in one stratum. Let $A_{w\ell}$ denote the margin (in votes) of reported winner $w$ over reported loser $\ell$ that a full hand count of the entire contest would show, that is, the *actual* margin rather than the *reported* margin. Reported winner $w$ really beat reported loser $\ell$ if and only if $A_{w\ell} > 0$. Define $A_{w\ell,s}$ to be the actual margin (in votes) of $w$ over $\ell$ in stratum $s$; this too may be negative.

Let $\omega_{w\ell,s} \equiv V_{w\ell,s} - A_{w\ell,s}$ be the *overstatement* of the margin of $w$ over $\ell$ in stratum $s$. Reported winner $w$ really beat reported loser $\ell$ if and only if $\omega_{w\ell} \equiv \omega_{w\ell,1} + \omega_{w\ell,2} < V_{w\ell}$.

The null hypothesis $\omega_{w\ell,1} + \omega_{w\ell,2} \geq V_{w\ell}$ is true if and only if there exists *some* $\lambda \in \Re$ such that $\omega_{w\ell,1} \geq \lambda V_{w\ell}$ and $\omega_{w\ell,2} \geq (1-\lambda)V_{w\ell}$.[11] If, for all $\lambda$, we can reject the hypothesis that the overstatement error in stratum 1 is greater than or equal to $\lambda V_{w\ell}$ *and* the overstatement error in stratum 2 is greater than or equal to $(1-\lambda)V_{w\ell}$, then we can conclude that the outcome is correct. (The approach generalizes to $S$ strata: if there is no tuple $(\lambda_s)_{s=1}^S$ such that $\sum_s \lambda_s = 1$ and $\omega_s \geq \lambda_s V_{w\ell}$ for all $s$, then the outcome is correct.)

To test the conjunction hypothesis that both stratum null hypotheses are true, we use Fisher's combining function. Let $p_s(\lambda_s)$ be the $P$-value of the hypothesis $\omega_{w\ell,s} \geq \lambda_s V_{w\ell}$. Define $\lambda_1 \equiv \lambda$ and $\lambda_2 \equiv 1 - \lambda$. If the null hypothesis that $\omega_{w\ell,1} \geq \lambda_1 V_{w\ell}$ and $\omega_{w\ell,2} \geq \lambda_2 V_{w\ell}$ is true, then

$$\chi(\lambda_1, \lambda_2) = -2 \sum_{s=1}^{2} \ln p_s(\lambda_s) \tag{1}$$

has a probability distribution that is dominated by the chi-square distribution with 4 degrees of freedom.[12] Fisher's combined statistic will tend to be small when both null hypotheses are true. If either is false, then as the sample size increases, Fisher's combined statistic will tend to grow.

If, for all $\lambda_1$ and $\lambda_2 = 1 - \lambda_1$, we can reject the conjunction hypothesis at level $\alpha$, the audit can stop. The stratified audit thus involves examining more randomly selected ballots from the two strata until either the minimum value Fisher's combined statistic over all $\lambda$ is larger than the $1-\alpha$ quantile of the chi-square distribution with 4 degrees of freedom, or until both strata have been fully hand tabulated.

[11] Set $\lambda = \frac{\omega_{w\ell,1}}{\omega_{w\ell,1}+\omega_{w\ell,2}}$.

[12] If the two tests had continuously distributed $P$-values, the distribution would be exactly chi-square with 4 degrees of freedom, but if either $P$-value has atoms when the null hypothesis is true, it is in general stochastically smaller. This follows from a coupling argument along the lines of Theorem 4.12.3 in [3].

Calculating $p_s(\cdot)$ is described in sections 6 and 7 for the CVR and no-CVR strata, respectively. In general, $p_s(\lambda)$ could be a $P$-value for the hypothesis $\omega_{w\ell,s} \geq \lambda_s V_{w\ell}$ from any test procedure (although if the audit is to be sequential, the tests in the two strata must be sequential tests or some other method must be used to account for multiplicity). We assume, however, that $p_s$ is based on a one-sided test, and that the tests for different values of $\lambda$ "nest" in the sense that if $a > b$, then $p_s(a) > p_s(b)$. This monotonicity is a reasonable requirement because the evidence that the overstatement is greater than $a$ should be weaker than the evidence that the overstatement is greater than $b$, if $a > b$. In particular, this monotonicity holds for the tests proposed in sections 6 and 7.

### 5.1 Maximizing Fisher's combined *P*-value

The audit can stop if the maximum of Fisher's combined $P$-value over all $\lambda$ is not larger than $\alpha$, the risk limit. For a given set of audit data, finding the maximum $P$-value over all $\lambda$ is a one-dimensional optimization problem, but the objective function is not necessarily concave. We need a computational strategy to ensure that the maximum is small without evaluating the $P$-value for all $\lambda$.

The approach embodied in the software we provide uses a grid search, refining the grid once the maximum has been bracketed. This is not guaranteed to find the global maximum exactly, although it can approximate the maximum as closely as one desires, by refining the mesh.

A more rigorous approach is to find bounds on Fisher's combining function for all $\lambda$. (Lower bounds translate directly into an upper bound on the $P$-value as a function of $\lambda$: if the lower bound is everywhere larger than the $1-\alpha$ quantile of the chi-squared distribution with 4 degrees of freedom, the maximum $P$-value is no larger than $\alpha$.) Let $\lambda_-$ be the smallest possible value of $\lambda$ and $\lambda_+$ be the largest possible value of $\lambda$. Some values of $\lambda$ can be ruled out *a priori*, because (for instance) $\omega_{w\ell,s} \leq V_{w\ell,s} + N_s$, where $N_s$ is the number of ballots cast in stratum $s$, and thus

$$1 - \frac{V_{w\ell,2} + N_2}{V_{w\ell}} \leq \lambda \leq \frac{V_{w\ell,1} + N_1}{V_{w\ell}}. \tag{2}$$

Recall that $p_s(\cdot)$ increases monotonically in its argument, so $p_1(\lambda)$ is monotonically increasing in $\lambda$ and $p_2(1-\lambda)$ is monotonically decreasing in $\lambda$. Suppose $[a,b) \subset [\lambda_-, \lambda_+]$. Then for all $\lambda \in [a,b)$, $-2\ln p_1(\lambda) \geq -2\ln p_1(b)$ and $-2\ln p_2(1-\lambda) \geq -2\ln p_2(1-a)$. Thus

$$\chi(\lambda) = -2(\ln p_1(\lambda) + \ln p_2(1-\lambda)) \geq -2(\ln p_1(b) + \ln p_2(1-a)) \equiv \chi_-[a,b). \tag{3}$$

This gives a (constant) lower bound for $\chi$ on the interval $[a,b)$; the corresponding upper bound is $\chi(\lambda) \leq -2(\ln p_1(a) + \ln p_2(1-b)) \equiv \chi_+[a,b)$. Partitioning $[\lambda_-, \lambda_+]$ into a collection of intervals $[a_k, a_{k+1})$ and finding $\chi_-[a_k, a_{k+1})$ and $\chi_+[a_k, a_{k+1})$ for each yields a piecewise-constant lower and upper bounds for $\chi(\lambda)$; if, for all $\lambda \in [\lambda_-, \lambda_+]$, the lower bound is larger than the $1-\alpha$ quantile of the chi-square distribution with 4 degrees of freedom, the audit can stop; if the upper bound is

anywhere less than the $1-\alpha$ quantile of the chi-square distribution with 4 degrees of freedom, the sample size in one or both strata needs to be larger.

## 6 Batch comparison audits of a tolerable overstatement in votes

In this section we expand previous comparison auditing work (already embodied in RLATool) to handle two new requirements. The first relates to partitioning the permissible overstatement through the parameters $\lambda_s$, as discussed in section 5. The second handles batch-level comparison audits.

The first requirement is to test whether the overstatement of any margin (in votes) exceeds some fraction $\lambda$ of the overall margin $V_{w\ell}$ between reported winner $w$ and reported loser $\ell$. If the stratum contains all the ballots cast in the contest, then for $\lambda = 1$, this would confirm the election outcome. For stratified audits, we might want to test other values of $\lambda$, as described above.

In Colorado, comparison audits so far have been ballot-level comparisons (*i.e.*, batches consisting of a single ballot). This section also addresses the second requirement by deriving a method for batches of arbitrary size, which might be useful for Colorado to audit contests that include CVR counties and legacy counties. We keep the *a priori* error bounds tighter than the "super-simple" method [11]. To keep the notation simpler, we consider only a single contest, but the MACRO test statistic [9,11] automatically extends the result to auditing $C > 1$ contests simultaneously. The derivation is for plurality contests, including "vote-for-$k$" plurality contests. Majority and super-majority contests are a minor modification [8].[13]

### 6.1 Notation

- $\mathcal{W}$: the set of reported winners of the contest
- $\mathcal{L}$: the set of reported losers of the contest
- $N_s$ ballots were cast in all in the stratum. (The contest might not appear on all $N_s$ ballots.)
- $P$ "batches" of ballots are in stratum $s$. A batch contains one or more ballots. Every ballot in stratum $s$ is in exactly one batch.
- $n_p$: number of ballots in batch $p$. $N_s = \sum_{p=1}^{P} n_p$.
- $v_{pi} \in \{0, 1\}$: the reported votes for candidate $i$ in batch $p$
- $a_{pi} \in \{0, 1\}$: actual votes for candidate $i$ in batch $p$. If the contest does not appear on any ballot in batch $p$, then $a_{pi} = 0$.
- $V_{w\ell,s} \equiv \sum_{p=1}^{P}(v_{pw} - v_{p\ell})$: Reported margin in stratum $s$ of reported winner $w \in \mathcal{W}$ over reported loser $\ell \in \mathcal{L}$, in votes.

[13] So are some forms of preferential and approval voting, such as Borda count, and proportional representation contests, such as D'Hondt [12]. See https://github.com/pbstark/S157F17/blob/master/audit.ipynb for a derivation of ballot-level comparison risk-limiting audits for super-majority contests. (Last visited 14 May 2018.) Changes for IRV/STV are more complicated.

- $V_{w\ell}$: Overall reported margin of reported winner $w \in \mathcal{W}$ over reported loser $\ell \in \mathcal{L}$, in votes, for the entire contest (not just stratum $s$)
- $V$: smallest reported overall margin between any reported winner and reported loser: $V \equiv \min_{w \in \mathcal{W}, \ell \in \mathcal{L}} V_{w\ell}$
- $A_{w\ell,s} \equiv \sum_{p=1}^{P} (a_{pw} - a_{p\ell})$: actual margin in the stratum of reported winner $w \in \mathcal{W}$ over reported loser $\ell \in \mathcal{L}$, in votes
- $A_{w\ell}$: actual margin of reported winner $w \in \mathcal{W}$ over reported loser $\ell \in \mathcal{L}$, in votes, for the entire contest (not just in stratum $s$)

### 6.2 Reduction to maximum relative overstatement

If the contest is entirely contained in stratum $s$, then the reported winners of the contest are the actual winners if

$$\min_{w \in \mathcal{W}, \ell \in \mathcal{L}} A_{w\ell,s} > 0.$$

Here, we address the case that the contest may include a portion outside the stratum. To combine independent samples in different strata, it is convenient to be able to test whether the net overstatement error in a stratum exceeds a given threshold.

Instead of testing that condition directly, we will test a condition that is sufficient but not necessary for the inequality to hold, to get a computationally simple test that is still conservative (i.e., the risk is not larger than its nominal value).

For every winner, loser pair $(w, \ell)$, we want to test whether the overstatement error exceeds some threshold, generally one tied to the reported margin between $w$ and $\ell$. For instance, for a stratified hybrid audit, we set the threshold to be $\lambda_s V_{w\ell}$.

We want to test whether

$$\sum_{p=1}^{P} (v_{pw} - a_{pw} - v_{p\ell} + a_{p\ell})/V_{w\ell} \geq \lambda_s.$$

The maximum of sums is not larger than the sum of the maxima; that is,

$$\max_{w \in \mathcal{W}, \ell \in \mathcal{L}} \sum_{p=1}^{P} (v_{pw} - a_{pw} - v_{p\ell} + a_{p\ell})/V_{w\ell} \leq \sum_{p=1}^{P} \max_{w \in \mathcal{W}, \ell \in \mathcal{L}} (v_{pw} - a_{pw} - v_{p\ell} + a_{p\ell})/V_{w\ell}.$$

Define

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

Then no reported margin is overstated by a fraction $\lambda_s$ or more if

$$E \equiv \sum_{p=1}^{P} e_p < \lambda_s.$$

Thus if we can reject the hypothesis $E \geq \lambda_s$, we can conclude that no pairwise margin was overstated by as much as a fraction $\lambda_s$.

Testing whether $E \geq \lambda_s$ would require a very large sample if we knew nothing at all about $e_p$ without auditing batch $p$: a single large value of $e_p$ could make $E$ arbitrarily large. But there is an *a priori* upper bound for $e_p$. Whatever the reported votes $v_{pi}$ are in batch $p$, we can find the potential values of the actual votes $a_{pi}$ that would make the error $e_p$ largest, because $a_{pi}$ must be between 0 and $n_p$, the number of ballots in batch $p$:

$$\frac{v_{pw} - a_{pw} - v_{p\ell} + a_{p\ell}}{V_{w\ell}} \leq \frac{v_{pw} - 0 - v_{p\ell} + n_p}{V_{w\ell}}.$$

Hence,

$$e_p \leq \max_{w \in \mathcal{W}, \ell \in \mathcal{L}} \frac{v_{pw} - v_{p\ell} + n_p}{V_{w\ell}} \equiv u_p. \tag{4}$$

Knowing that $e_p \leq u_p$ might let us conclude reliably that $E < \lambda_s$ by examining only a small number of batches—depending on the values $\{u_p\}_{p=1}^P$ and on the values of $\{e_p\}$ for the audited batches.

To make inferences about $E$, it is helpful to work with the *taint* $t_p \equiv \frac{e_p}{u_p} \leq 1$. Define $U \equiv \sum_{p=1}^P u_p$. Suppose we draw batches at random with replacement, with probability $u_p/U$ of drawing batch $p$ in each draw, $p = 1, \ldots, P$. (Since $u_p \geq 0$, these are all positive numbers, and they sum to 1, so they define a probability distribution on the $P$ batches.)

Let $T_j$ be the value of $t_p$ for the batch $p$ selected in the $j$th draw. Then $\{T_j\}_{j=1}^n$ are IID, $\mathbb{P}\{T_j \leq 1\} = 1$, and

$$\mathbb{E}T_1 = \sum_{p=1}^P \frac{u_p}{U} t_p = \frac{1}{U} \sum_{p=1}^P u_p \frac{e_p}{u_p} = \frac{1}{U} \sum_{p=1}^P e_p = E/U.$$

Thus $E = U\mathbb{E}T_1$. So, if we have strong evidence that $\mathbb{E}T_1 < \lambda_s/U$, we have strong evidence that $E < \lambda_s$.

This approach can be simplified even further by noting that $u_p$ has a simple upper bound that does not depend on $v_{pi}$. At worst, the reported result for batch $p$ shows $n_p$ votes for the "least-winning" apparent winner of the contest with the smallest margin, but a hand interpretation would show that all $n_p$ ballots in the batch had votes for the runner-up in that contest. Since $V_{w\ell} \geq V \equiv \min_{w \in \mathcal{W}, \ell \in \mathcal{L}} V_{w\ell}$ and $0 \leq v_{pi} \leq n_p$,

$$u_p = \max_{w \in \mathcal{W}, \ell \in \mathcal{L}} \frac{v_{pw} - v_{p\ell} + n_p}{V_{w\ell}} \leq \max_{w \in \mathcal{W}, \ell \in \mathcal{L}} \frac{n_p - 0 + n_p}{V_{w\ell}} \leq \frac{2n_p}{V}.$$

Thus if we use $2n_p/V$ in lieu of $u_p$, we still get conservative results. (We also need to re-define $U$ to be the sum of those upper bounds.) An intermediate, still conservative approach would be to use this upper bound for batches that consist of a single ballot, but use the sharper bound (4) when $n_p > 1$. Regardless, for

the new definition of $u_p$ and $U$, $\{T_j\}_{j=1}^n$ are IID, $\mathbb{P}\{T_j \leq 1\} = 1$, and

$$\mathbb{E}T_1 = \sum_{p=1}^{P} \frac{u_p}{U} t_p = \frac{1}{U} \sum_{p=1}^{P} u_p \frac{e_p}{u_p} = \frac{1}{U} \sum_{p=1}^{P} e_p = E/U.$$

So, if we have evidence that $\mathbb{E}T_1 < \lambda_s/U$, we have evidence that $E < \lambda_s$.

### 6.3 Testing $\mathbb{E}T_1 \geq \lambda_s/U$

A variety of methods are available to test whether $\mathbb{E}T_1 < \lambda_s/U$. One particularly "clean" sequential method is based on Wald's Sequential Probability Ratio Test (SPRT) ([14]). Harold Kaplan pointed out this method on a website that no longer exists. A derivation of this "Kaplan-Wald" method is given in Appendix A of [12]; to apply the method here, take $t = \lambda_s$ in their equation 18.

A different sequential method, the Kaplan-Markov method (also due to Harold Kaplan), is given in [10].

## 7 Ballot-polling audits of a tolerable overstatement in votes

In this section we develop a new method for ballot-polling audits that can test numerical margins, rather than just test whether a candidate won. This requires a different approach than that taken by [5].

Existing ballot-polling methods consider only the fraction of ballots with a vote for either $w$ or $\ell$ that contain a vote for $w$, making the statistical test one for a proportion. To allow the error to be partitioned across the strata via $\lambda_s$, the necessary inference is about the *difference* between the number of votes for $w$ and the number of votes for $\ell$. This introduces a nuisance parameter, the number of ballots with votes for either $w$ or $\ell$. We deal with the nuisance parameter by maximizing the $P$-value over all possible values of the nuisance parameter, which ensures that the test is conservative.

### 7.1 Conditional tri-hypergeometric test

We consider a single stratum $s$, containing $N_s$ ballots. Of the $N_s$ ballots, $A_{w,s}$ have a vote for $w$ but not for $\ell$, $A_{\ell,s}$ have a vote for $\ell$ but not for $w$, and $A_{u,s} = N_s - N_{w,s} - N_{\ell,s}$ have votes for both $w$ and $\ell$ or neither $w$ nor $\ell$, including undervotes and invalid ballots. We might draw a simple random sample of $n$ ballots ($n$ fixed ahead of time), or we might draw sequentially without replacement, so the sample size $B$ could be random. For instance, the rule for determining $B$ could depend on the data.[14]

Regardless, we assume that, conditional on the attained sample size $n$, the ballots are a simple random sample of size $n$ from the $N_s$ ballots in the population. In the sample, $B_w$ ballots contain a vote for $w$ but not $\ell$, with $B_\ell$ and

[14] Sampling with replacement leads to simpler arithmetic, but is not as efficient.

$B_u$ defined analogously. The conditional joint distribution of $(B_w, B_\ell, B_u)$ is tri-hypergeometric:

$$\mathbb{P}_{A_{w,s},A_{\ell,s}}\{B_w = i, B_\ell = j | B = n\} = \frac{\binom{A_{w,s}}{i}\binom{A_{\ell,s}}{j}\binom{N_s - A_{w,s} - A_{\ell,s}}{n-i-j}}{\binom{N_s}{n}}. \quad (5)$$

Define the diluted sample margin, $D \equiv (B_w - B_\ell)/B$. We want to test the compound hypothesis $A_{w,s} - A_{\ell,s} \leq c$. The value of $c$ is inferred from the definition $\omega_{w\ell,s} \equiv V_{w\ell,s} - A_{w\ell,s} = V_{w,s} - V_{\ell,s} - (A_{w,s} - A_{\ell,s})$. Thus,

$$c = V_{w,s} - V_{\ell,s} - \omega_{w\ell,s} = V_{w\ell,s} - \lambda_s V_{w\ell}. \quad (6)$$

The alternative is the compound hypothesis $A_{w,s} - A_{\ell,s} > c$.[15] Hence, we will reject for large values of $D$. Conditional on $B = n$, the event $D = (B_w - B_\ell)/B = d$ is the same as $B_w - B_\ell = nd$.[16]

The $P$-value of the simple hypothesis that there are $A_{w,s}$ ballots with a vote for $w$ but not for $\ell$, $A_{\ell,s}$ ballots with a vote for $\ell$ but not for $w$, and $N - A_{w,s} - A_{\ell,s}$ ballots with votes for both $w$ and $\ell$ or neither $w$ nor $\ell$ (including undervotes and invalid ballots) is the probability that $B_w - B_\ell \geq nd$. Therefore,

$$\mathbb{P}_{A_{w,s},A_{\ell,s},N_s}\{D \geq d \mid B = n\} = \sum_{\substack{(i,j):i,j\geq 0 \\ i-j\geq nd \\ i+j\leq n}} \frac{\binom{A_{w,s}}{i}\binom{A_{\ell,s}}{j}\binom{N_s - A_{w,s} - A_{\ell,s}}{n-i-j}}{\binom{N_s}{n}}. \quad (7)$$

### 7.2 Maximizing the *P*-value over the nuisance parameter

The composite null hypothesis does not specify $A_{w,s}$ or $A_{\ell,s}$ separately, only that $A_{w,s} - A_{\ell,s} \leq c$ for some fixed, known $c$. Define $\mathcal{S}$ to be the set of pairs $(i, j)$ such that $i, j \geq 0, i - j \geq nd$, and $i + j \leq n$. The (conditional) $P$-value of this composite hypothesis for $D = d$ is the maximum $P$-value for all values $(A_{w,s}, A_{\ell,s})$ that are possible under the null hypothesis,

$$\max_{A_{w,s},A_{\ell,s}\in\{0,1,\ldots,N\}: A_{w,s}-A_{\ell,s}\leq c, A_{w,s}+A_{\ell,s}\leq N_s} \sum_{(i,j)\in\mathcal{S}} \frac{\binom{A_{w,s}}{i}\binom{A_{\ell,s}}{j}\binom{N_s - A_{w,s} - A_{\ell,s}}{n-i-j}}{\binom{N_s}{n}}, \quad (8)$$

wherever the summand is defined. (Equivalently, define $\binom{m}{k} \equiv 0$ if $k > m$, $k < 0$, or $m \leq 0$.)

[15] To use Wald's Sequential Probability Ratio Test, we might pick a simple alternative instead, e.g., $A_{w,s} = V_{w,s}$ and $A_{\ell,s} = V_{\ell,s}$, the reported values, provided $V_{w,s} - V_{\ell,s} > c$.

[16] In contrast, the BRAVO ballot-polling method [5] conditions only on $B_w + B_\ell = m$.

**Characterizing the optimal solution** The following result enables us to only test hypotheses along the boundary of the null set.

**Theorem 1** *Assume that $n < A_{w,s}+A_{\ell,s}$. Suppose the composite null hypothesis is $N_w - N_\ell \leq c$. The $P$-value is maximized on the boundary of the null region, i.e. when $N_w - N_\ell = c$.*

*Proof.* Without loss of generality, let $c = 0$ and assume that $A_{u,s} = N_s - A_{w,s} - A_{\ell,s}$ is fixed. Let $N_{w\ell,s} \equiv A_{w,s} + A_{\ell,s}$ be the fixed, unknown number of ballots for $w$ or for $\ell$ in stratum $s$. The $P$-value $p_0$ for the simple hypothesis that $c = 0$ is

$$p_0 = \sum_{(i,j)\in\mathcal{S}} \frac{\binom{N_{w\ell,s}/2}{i}\binom{N_{w\ell,s}/2}{j}\binom{A_{u,s}}{n-i-j}}{\binom{N_s}{n}} = \sum_{(i,j)\in\mathcal{S}} T_{ij}, \tag{9}$$

where $T_{ij}$ is defined as the $(i, j)$ term in the summand and $T_{ij} \equiv 0$ for pairs $(i, j)$ that don't appear in the summation.

Assume that $c > 0$ is given. The $P$-value $p_c$ for this simple hypothesis is

$$\begin{aligned} p_c &= \sum_{(i,j)\in\mathcal{S}} \frac{\binom{(N_{w\ell,s}+c)/2}{i}\binom{(N_{w\ell,s}-c)/2}{j}\binom{A_{u,s}}{n-i-j}}{\binom{N_s}{n}} \\ &= \sum_{(i,j)\in\mathcal{S}} T_{ij} \frac{\frac{N_{w\ell,s}+c}{2}(\frac{N_{w\ell,s}+c}{2}-1)\cdots(\frac{N_{w\ell,s}}{2}+1)(\frac{N_{w\ell,s}-c}{2}-j)\cdots(\frac{N_{w\ell,s}}{2}-1-j)}{(\frac{N_{w\ell,s}+c}{2}-i)\cdots(\frac{N_{w\ell,s}}{2}+1-i)(\frac{N_{w\ell,s}-c}{2})\cdots(\frac{N_{w\ell,s}}{2}-1)}. \end{aligned}$$

Terms in the fraction can be simplified: choose the corresponding pairs in the numerator and denominator. Fractions of the form $\frac{\frac{N_{w\ell,s}}{2}+a}{\frac{N_{w\ell,s}}{2}+a-i}$ can be expressed as $1+\frac{i}{\frac{N_{w\ell,s}}{2}+a-i}$. Fractions of the form $\frac{\frac{N_{w\ell,s}}{2}-a-j}{\frac{N_{w\ell,s}}{2}-a}$ can be expressed as $1-\frac{j}{\frac{N_{w\ell,s}}{2}-a}$. Thus, the $P$-value can be written as

$$\begin{aligned} p_c &= \sum_{(i,j)\in\mathcal{S}} T_{ij} \prod_{a=1}^{c/2} \left(1+\frac{i}{\frac{N_{w\ell,s}}{2}+a-i}\right)\left(1-\frac{j}{\frac{N_{w\ell,s}}{2}-a}\right) \\ &> \sum_{(i,j)\in\mathcal{S}} T_{ij} \left[\left(1+\frac{i}{\frac{N_{w\ell,s}+c}{2}-i}\right)\left(1-\frac{j}{\frac{N_{w\ell,s}}{2}+1}\right)\right]^{c/2} \\ &= \sum_{(i,j)\in\mathcal{S}} T_{ij} \left[1+\frac{\frac{N_{w\ell,s}+c}{2}j+\frac{N_{w\ell,s}}{2}i+i}{(\frac{N_{w\ell,s}+c}{2}-i)(\frac{N_{w\ell,s}}{2}+1)}\right]^{c/2} \\ &> \sum_{(i,j)\in\mathcal{S}} T_{ij} \\ &= p_0 \end{aligned}$$

The last inequality follows from the fact that $i$ and $j$ are nonnegative, and that $i < \frac{N_{w\ell,s}+c}{2}$ (it is a possible outcome under the null hypothesis).

**Solving the optimization problem** We have found empirically (but have not proven) that given $N$, $c$, and the observed sample values $B_w$ and $B_\ell$, the tail probability $p_c$, as a function of $A_{w,s}$, has a unique maximum at one of the endpoints, where $A_{w,s}$ is either as small or as large as possible. If this empirical result is true in general, then finding the maximum is trivial; otherwise, computing the unconditional $P$-value is a simple 1-dimensional optimization problem on a bounded interval.

### 7.3 Conditional testing

If the conditional tests are always conducted at significance level $\alpha$ or less, so that $\mathbb{P}\{\text{Type I error}|B = n\} \leq \alpha$, then the overall procedure has significance level $\alpha$ or less:

$$\begin{aligned}\mathbb{P}\{\text{Type I error}\} &= \sum_{n=0}^{N} \mathbb{P}\{\text{Type I error}|B = n\}\mathbb{P}\{B = n\} \\ &\leq \sum_{n=0}^{N} \alpha\mathbb{P}\{B = n\} = \alpha. \end{aligned} \tag{10}$$

In particular, this implies that our conditional hypergeometric test will have a conservative $P$-value unconditionally.

## 8 Recommendations

Of the methods Colorado might use to audit cross-jurisdictional contests that include CVR counties and no-CVR counties, stratified "hybrid" audits seem the most palatable, given the constraints on time for software development and the logistics of the audit itself. The workflow for counties would be the same as it was in November, 2017. Simulations suggest that this approach is relatively efficient.

What needs to change is the risk calculation, including the algorithms that determine when the audit can stop. Those algorithms could be implemented in software external to RLATool. The minimal modification to RLATool required to conduct a hybrid audit is to allow the sample size from each county to be controlled externally, *e.g.*, by providing a parameter file for each round, rather than using the current formula built into RLATool, which is based on contest margins within each county alone. The parameter file could be generated by external software using data about the audit exported via the RLATool `rla_export` command.

### 8.1 Stratum sample sizes

The statistical constraints on the two sample sizes are weak: increasing the sample size in one stratum generally allows the other sample size to be decreased. Allocating the sample across the two strata is therefore largely a political decision, as discussed in section 3. In general, when the contest outcome is correct, the total workload will be minimized by assigning a disproportionately large (compared to the number of ballots cast) amount of the work to the CVR stratum. The software described below can test the implications of different sampling allocations on the workload in each stratum and the total workload.

### 8.2 Software and examples

Examples of stratified hybrid audits are in Jupyter notebooks available at https://www.github.com/pbstark/CORLA18.

The first example, in `hybrid-audit-example-1`, is a hypothetical medium-sized election with 110,000 ballots cast, of which 9.1% were cast in no-CVR counties. The diluted margin is 1.8%. In 95 of 100 simulations, a stratified "hybrid" audit at risk limit 10% with sample sizes of 500 ballots in the CVR stratum and 700 ballots in the no-CVR stratum (1,200 ballots in all) would have sufficed to confirm the outcome, if the reported results were correct.

In contrast, an unstratified ballot-level comparison audit with risk limit 10% could have terminated after examining 263 ballots if it found no errors, and a ballot-polling audit of the entire contest would have been expected to examine about 14,000 ballots, more than 10% of ballots cast. The hybrid audit is thus not as efficient as a ballot-level comparison audit, but far more efficient than a ballot-polling audit.

Another conservative method, discussed in Section 3.2, involves conducting a ballot-level comparison audit statewide, treating any ballot selected from the no-CVR county as if it had a two-vote overstatement. In this numerical example, that method would lead to a full hand count.

The second example, also in `hybrid-audit-example-1`, is a hypothetical large statewide election with 2 million ballots cast, of which 5% were cast in no-CVR counties. The contest has a diluted margin of nearly 20% and the risk limit is 5%. The workload for a hybrid stratified audit is quite low: In 98% of 10,000 simulations, auditing 43 ballots from the CVR stratum and 20 ballots from the no-CVR stratum would have sufficed to confirm the outcome at a 5% risk limit.

If it were possible to conduct a ballot-level comparison audit for the entire contest, an audit at risk limit 5% could terminate after examining 31 ballots if it found no errors. The additional work needed to do the hybrid stratified audit falls mainly in the no-CVR stratum.

A second notebook, `hybrid-audit-example-2`, illustrates the workflow for conducting a hybrid stratified audit of an election with 2 million ballots cast. The reported margin is just over 1%, but the reported winner and reported loser are actually tied in both strata. The risk limit is 5%. We use Fisher's method to

combine the audits in the CVR stratum (sample size 500) and no-CVR stratum (sample size 1000). The maximum Fisher's combined $P$-value is over 20%, so the audit cannot stop at that point.

These notebooks can be modified and run with different contest sizes, margins, and risk limits to estimate the workload in different scenarios.

## References


1. Bañuelos, J., Stark, P.: Limiting risk by turning manifest phantoms into evil zombies. Tech. rep., arXiv.org (2012), http://arxiv.org/abs/1207.3413, retrieved 17 July 2012
2. California Secretary of State: California Secretary of State Post-Election Risk-Limiting Audit Pilot Program 2011-2013: Final Report to the United States Election Assistance Commission. http://votingsystems.cdn.sos.ca.gov/oversight/risk-pilot/final-report-073014.pdf Retrieved 6 May 2018 (2014)
3. Grimmett, G.R., Stirzaker, D.R.: Probability and Random Processes. Oxford University Press (August 2001), http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/0198572220
4. Higgins, M., Rivest, R., Stark, P.: Sharper p-values for stratified post-election audits. Statistics, Politics, and Policy **2**(1) (2011), http://www.bepress.com/spp/vol2/iss1/7
5. Lindeman, M., Stark, P., Yates, V.: BRAVO: Ballot-polling risk-limiting audits to verify outcomes. In: Proceedings of the 2011 Electronic Voting Technology Workshop / Workshop on Trustworthy Elections (EVT/WOTE '11). USENIX (2012)
6. Lindeman, M., Stark, P.B.: A gentle introduction to risk-limiting audits. IEEE Security and Privacy **10**, 42–49 (2012)
7. Rivest, R.L.: Bayesian tabulation audits: Explained and extended (January 1, 2018), https://arxiv.org/abs/1801.00528
8. Stark, P.: Conservative statistical post-election audits. Ann. Appl. Stat. **2**, 550–581 (2008), http://arxiv.org/abs/0807.4005
9. Stark, P.: Auditing a collection of races simultaneously. Tech. rep., arXiv.org (2009), http://arxiv.org/abs/0905.1422v1
10. Stark, P.: Risk-limiting post-election audits: $P$-values from common probability inequalities. IEEE Transactions on Information Forensics and Security **4**, 1005–1014 (2009)
11. Stark, P.: Super-simple simultaneous single-ballot risk-limiting audits. In: Proceedings of the 2010 Electronic Voting Technology Workshop / Workshop on Trustworthy Elections (EVT/WOTE '10). USENIX (2010), http://www.usenix.org/events/evtwote10/tech/full_papers/Stark.pdf
12. Stark, P.B., Teague, V.: Verifiable european elections: Risk-limiting audits for d'hondt and its relatives. JETS: USENIX Journal of Election Technology and Systems **3.1** (2014), https://www.usenix.org/jets/issues/0301/stark
13. Stark, P.B., Wagner, D.A.: Evidence-based elections. IEEE Security and Privacy **10**, 33–41 (2012)
14. Wald, A.: Sequential tests of statistical hypotheses. Ann. Math. Stat. **16**, 117–186 (1945)