A New Method for Stratified Risk-Limiting Audits

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] Verified Voting Foundation

Abstract. Risk-limiting audits (RLAs) offer a statistical guarantee: if a full manual tally of the paper ballots would show that the reported election outcome is wrong, an RLA has a known minimum chance of leading to a full manual tally. The risk limit is the maximum chance the audit will not in that case lead to a full manual tally. RLAs generally rely on random samples. Audit risk calculations are simplest for random samples of individual ballots drawn with replacement from all validly cast ballots. However, stratified sampling—partitioning the population of ballots into disjoint strata and sampling independently from the strata—may simplify logistics or increase efficiency. For example, some counties in Colorado (comprising 98.2% of voters) have new voting systems that allow auditors to check how the system interpreted each ballot, which allows efficient "ballot-level comparison" audits. Previous approaches to combining information from all counties into a single RLA of a statewide contest would require major procedural changes, or would sacrifice the efficiency of ballot-level comparison audits. We provide a simpler, more efficient approach using stratified sampling. Cast ballots are divided into two strata: those ballots cast in counties with newer systems, and the rest. Samples of individual ballots are drawn from those strata independently. A generalization of ballot-level comparison auditing in the first stratum and a generalization of ballot-polling auditing in the second are to find P-values for the hypotheses that the "overstatement error" in each stratum exceeds a threshold. The stratum-level P-values are combined using Fisher's combining function. The combined P-value is maximized P-value over all combinations of error in the two strata that for which the reported outcome could be incorrect. The audit stops when the maximum combined P-value is less than the risk limit. This approach to election audits is new, and immediately applicable in Colorado. We provide an open-source reference implementation and exemplar calculations in Jupyter notebooks.

Keywords: stratified sampling, nonparametric testing, Fisher's combining function, sequential hypothesis tests, Colorado risk-limiting audits (CORLA), maximizing P-values over nuisance parameters

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 contest is a procedure that has a known minimum chance of leading to a full manual tally of the ballots if the electoral outcome according to that tally would differ from the reported outcome. *Outcome* means the winner or winners (or, for instance, whether there is a runoff)—not the numerical vote totals. 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* [14] is required to check whether the paper trail is trustworthy.

Current methods for risk-limiting audits are generally *sequential hypothesis testing procedures*: 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 conducted in California, Colorado, Ohio, and Denmark, and are required by law in Colorado (CRS 1-7-515) and Rhode Island (SB 413A and HB 5704A).

The most efficient and transparent sampling design for risk-limiting audits selects individual ballots uniformly at random, with or without replacement [11]. Risk calculations for such samples can be made quite simple without sacrificing rigor [12,5]. However, to audit contests that cross jurisdictional boundaries then requires coordinating sampling in different counties, and may require different counties to use the lowest-common denominator method for assessing risk from the sample, which might not take full advantage of the capabilities of some voting systems. For instance, any system that uses paper ballots as the official record can conduct *ballot-polling* audits, while *ballot-level comparison audits* require systems to generate *cast-vote records* that can be checked manually against a human reading of the paper [4,5]. (Ballot-polling and ballot-level comparison audits are described below.) We show it then can be advantageous to use a stratified sample, selecting ballots independently from counties that can conduct ballot-level comparison audits and from those that cannot.

Stratified RLAs have been considered previously, primarily to conform with existing audit laws under which counties draw audit samples independently of each other, but also to allow auditors to start the audit before all vote-by-mail or provisional ballots have been tallied, by sampling independently from ballots cast in person, by mail, and provisionally, as soon as subtotals for each group are available [8,3]. However, extant methods address only a single approach to auditing, batch-level comparisons, and only a particular test statistic.

Here, we provide a more general approach to using stratified samples in RLAs. The approach involves finding the maximum P-value (over a vector of nuisance

parameters) of all allocations of tabulation error across strata for which a full count would find a different electoral outcome than was reported. (A *nuisance parameter* is a property of the population that is not of direct interest, but that affects the probability distribution of the data. The total *overstatement* across strata determines whether the reported outcome is correct; the overstatements in individual strata are nuisance parameters that affect the distribution of the audit sample. *Overstatement* is error that made the margin of one or more winners over one or more losers appear larger than it really was.)

The core of the method is to test whether the overstatement error in a single stratum exceeds a quota. Fisher's combining function is used to merge the P-values for those tests into a single P-value for the hypothesis that the overstatement in every stratum exceeds its quota. If that hypothesis can be rejected for *all* stratum-level quotas that could change the outcome—that is, if the maximum combined P-value is sufficiently small—the audit can stop; otherwise, the audit must inspect more ballots.

In fact, it is not necessary to consider all possible quotas: the P-value involves a difference of monotonic functions, which allows us to find upper and lower bounds everywhere from the values on a discrete grid. We present a numerical procedure, implemented in Python implementation, to find bounds on the maximum P-value when there are two strata. The procedure can be generalized to more than two strata.

1.1 Voting systems and audit strategies

Voting systems that export cast vote records (CVRs) in a way that the paper ballot corresponding to each CVR can be identified uniquely and retrieved, and for which the CVR corresponding to any particular paper ballot can be found, can be audited using *ballot-level comparison audits* [5], which involve comparing CVRs to the auditors' interpretation of voter intent directly from paper ballots. We call counties with such voting systems *CVR counties*. Ballot-level comparison audits are currently the most efficient approach to risk-limiting audits in that they require examining fewer 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) can be audited using *ballot-polling audits* [4,5], if 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. Ballot-polling audits generally require examining more ballots than ballot-level comparison audits to attain the same risk limit.

There is no literature on how to combine ballot-polling with ballot-level comparisons to audit (*cross-jurisdictional contests*), such as gubernatorial contests and statewide ballot measures, that include voters in CVR counties and voters in no-CVR counties. Existing methods either would require all counties to use ballot-polling, or would require no-CVR counties to perform *batch-level comparisons* (comparing manual tallies of physical groups of ballots to reported results for those groups). Batch-level comparison audits were found in California to be

less efficient than ballot-polling audits [1].[3] Moreover, to conduct batch-level comparison audits would require different pre-audit data export from voting systems, different random sampling procedures, different audit logistics, and different audit data uploads—large changes for no-CVR counties and large changes to Colorado's audit software, RLATool, including its data structures and user interface for counties.

Our new method solves the problem of auditing cross-jurisdictional contests, with little change to county procedures and to RLATool.

Section 2 presents the new approach to stratified auditing. Section 3 illustrates the method by combining ballot-polling in one stratum with ballot-level comparisons auditing in another. This requires straightforward modifications to the mathematics behind ballot-polling and ballot-level comparison to allow the overstatement to be compared to specified thresholds other than the overall contest margin; those modifications are derived in sections 3.1 and 3.2. Section 4 gives numerical examples for simulated audits, using parameters intended to reflect how the procedure would work in Colorado. We provide example software implementing the risk calculations for our recommended approach in a Python Jupyter notebook.[4] Section 5 gives recommendations and considerations for implementation.

2 Stratified audits

Stratified sampling involves partitioning a population into non-overlapping groups, and drawing independent random samples from those groups. [8,3] developed RLAs based on comparing stratified samples of batches of ballots to hand counts of the votes in those batches: batch-level comparison RLAs, using a particular test statistic. The method we develop here is more general and more flexible: it can be used with any test statistic, and test statistics in different strata need not be the same—which is key to combining ballot-polling with ballot-level comparisons.

Here and below, we consider 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 [9,12]. We use terminology drawn from a number of papers, notably [5].

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. Overstatements cast doubt on outcomes; understatements do not, even though they are tabulation errors.

[3] 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.

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

We use w to denote a reported winner and ℓ to denote a reported loser. The total number of reported votes for candidate w is V_w and the total for candidate ℓ is V_ℓ. Thus $V_w > V_\ell$, since w is reported to have gotten more votes than ℓ.

Let $V_{w\ell} \equiv V_w - V_\ell > 0$ denote the contest-wide margin (in votes) of w over ℓ. We have S strata. Let $V_{w\ell,s}$ denote the margin (in votes) of reported winner w over reported loser ℓ in stratum s. Note that $V_{w\ell,s}$ might be negative in one stratum, but $\sum_{s=1}^{S} V_{w\ell,s} = V_{w\ell} > 0$. Let $A_{w\ell}$ denote the margin (in votes) of reported winner w over reported loser ℓ that a full hand count would show: the *actual* margin, in contrast to the *reported* margin $V_{w\ell}$. Reported winner w really beat reported loser ℓ if and only if $A_{w\ell} > 0$. Define $A_{w\ell,s}$ to be the actual margin (in votes) of w over ℓ in stratum s.

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

An RLA is a test of the hypothesis that the outcome is wrong, that is, that w did not really beat ℓ: $\sum_s \omega_{w\ell,s} \geq V_{w\ell}$. The null is true if and only if there exist *some* S-tuple of real numbers $(\lambda_s)_{s=1}^s$ such that $\sum_s \lambda_s = 1$ and $\omega_{w\ell,s} \geq \lambda_s V_{w\ell}$ for all s.[5] Thus if we can reject the composite hypothesis $\cap_s \{\omega_{w\ell,s} \geq \lambda_s V_{w\ell}\}$ at significance level α for all (λ_s) such that $\sum_s \lambda_s = 1$, we can stop the audit, and the risk limit will be α.

2.1 Fisher's combination method

Fix $\lambda \equiv (\lambda_s)_{s=1}^s$, with $\sum_s \lambda_s = 1$. To test the conjunction hypothesis that stratum null hypotheses are true, that is, that $\omega_{w\ell,s} \geq \lambda_s V_{w\ell}$ for all s, we use Fisher's combining function. Let $p_s(\lambda_s)$ be the P-value of the hypothesis $\omega_{w\ell,s} \geq \lambda_s V_{w\ell}$. If the null hypothesis is true, then

$$\chi(\lambda) = -2 \sum_{s=1}^{S} \ln p_s(\lambda_s) \tag{1}$$

has a probability distribution that is dominated by the chi-square distribution with $2S$ degrees of freedom.[6] Fisher's combined statistic will tend to be small when all stratum-level null hypotheses are true. If any is false, then as the sample size increases, Fisher's combined statistic will tend to grow.

If, for all λ with $\sum_s \lambda_s = 1$, we can reject the conjunction hypothesis at level α, the audit can stop, i.e., if the minimum value of Fisher's combined statistic over all λ is larger than the $1 - \alpha$ quantile of the chi-square distribution with $2S$ degrees of freedom, the audit can stop.

[5] "If" is straightforward. For "only if," suppose $\omega_{w\ell} \geq V_{w\ell}$. Set $\lambda_s = \frac{\omega_{w\ell,s}}{\sum_t \omega_{w\ell,t}}$. Then $\sum_s \lambda_s = 1$, and $\omega_{w\ell,s} = \lambda_s \omega_{w\ell} \geq \lambda_s V_{w\ell}$ for all s.

[6] If the stratum-level tests had continuously distributed P-values, the distribution would be exactly chi-square with $2S$ degrees of freedom, but if any of the P-values 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 [2].

If the audit is allowed to "escalate" in steps, increasing the sample size sequentially, then either the tests used in the separate strata have to be sequential tests, or multiplicity needs to be taken into account, for instance by adjusting the risk limit at each step. Otherwise, the overall procedure can have a risk limit that is much larger than α. For examples of controlling for multiplicity when using non-sequential testing procedures in an RLA, see [8,**?**].

$p_s(\lambda)$ could be a P-value for the hypothesis $\omega_{w\ell,s} \geq \lambda_s V_{w\ell}$ from any test procedure. We assume, however, that p_s is based on a one-sided test, and that the tests for different values of λ "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 3.1 and 3.2.

One could a function other than Fisher's to combine the stratum-level P-values into a P-value for the compound hypothesis, provided it satisfies these properties (see [6]):

- the function is non-increasing in each argument and symmetric with respect to rearrangements of the arguments
- the combining function attains its supremum when one of the arguments approaches zero
- for every level α, the critical value of the combining function is finite and strictly smaller than the function's supremum

For instance, one could use Liptak's function, $T = \sum_i \Phi^{-1}(1 - p_i)$, or Tippett's function, $T = \max_i(1 - p_i)$.

Fisher's function is convenient for this application because the tests in different strata are independent, so the chi-squared distribution dominates the distribution of $\chi(\cdot)$ when the null hypothesis is true. If tests in different strata were correlated, the null distribution of the combination function would need to be calibrated by simulation; some other combining function might have better properties than Fisher's [6].

2.2 Maximizing Fisher's combined P-value for $S = 2$

We now specialize to $S = 2$ strata. In Colorado, one would comprise all ballots cast in CVR counties and the other all ballots cast in non-CVR counties. The set of $\lambda = (\lambda_1, \lambda_2)$ such that $\sum_s \lambda_s = 1$ is then a one-dimensional family: if $\lambda_1 = \lambda$, then $\lambda_2 = 1 - \lambda$. For a given set of data, finding the maximum P-value over all λ is thus a one-dimensional optimization problem.

The software we provide approximates the maximum via 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, since the objective function is continuous.

Here is a rigorous approach using bounds on Fisher's combining function χ for all λ. (A lower bound on χ implies an upper bound on the P-value: if,

for all λ, the lower bound is larger than the $1-\alpha$ quantile of the chi-squared distribution with 4 degrees of freedom, the maximum P-value is no larger than α.)

Some values of λ 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

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

Let λ_- and λ_+ be lower and upper bounds on λ.

Recall that $p_s(\cdot)$ are monotonically increasing functions, so, as a function of λ, $p_1(\lambda)$ increases monotonically and $p_2(1-\lambda)$ decreases monotonically. 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 lower bound for χ 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 piecewise-constant lower and upper bounds for $\chi(\lambda)$.

If, for all $\lambda \in [\lambda_-, \lambda_+]$, the lower bound on χ is larger than the $1-\alpha$ quantile of the chi-square distribution with 4 degrees of freedom, the audit can stop. On the other hand, if for some $\lambda \in [\lambda_-, \lambda_+]$, the upper bound is less than the $1-\alpha$ quantile of the chi-square distribution with 4 degrees of freedom, or if $\chi(a_k)$ is less than this quantile at any grid point, the sample size in one or both strata needs to increase. If the lower bound is less than the $1-\alpha$ quantile on some interval, but $\chi(a_k)$ is above this quantile at every grid point $\{a_k\}$, then one should improve the lower bound by refining the grid and/or by increasing the sample size in one or both strata.

3 Strategies for auditing cross-jurisdictional contests

The stratified auditing method in Section 2 makes it possible to conduct RLAs of cross-jurisdictional contests and still take advantage of the efficiency of ballot-level comparison methods when not all jurisdictions have voting systems that support them—as is the case in Colorado, where most ballots are cast in precincts with equipment that produces CVRs, but some are not.

CRS 1-7-515 requires Colorado to conduct risk-limiting audits beginning in 2017. The first risk-limiting election audits under this statute were conducted in November, 2017; the second were conducted in July, 2018.[7] Counties cannot audit cross-jurisdictional contests on their own: margins and risk limits apply to entire contests, not to the portion of a contest included in a county. Colorado has not yet conducted a RLA of a cross-jurisdictional contest (some counties audited their portion of a cross-jurisdictional contest as if the contest were entirely

[7] See https://www.sos.state.co.us/pubs/elections/RLA/2017RLABackground.html

contained in the county, but the result is not an RLA of the contest). To audit statewide elections and contests that cross county lines, Colorado will need to implement new approaches and make some changes to RLATool.

As of this writing, about 98% of active Colorado voters are in CVR counties. As discussed above, Colorado could simply revert to ballot-polling audits for cross-jurisdictional contests that include votes in no-CVR counties, but that would entail a loss of efficiency in CVR counties.

Alternatively, they could use batch-level comparison audits, with single-ballot batches in CVR counties and larger batches in non-CVR counties.[8] The statistical theory for such audits has been worked out (see, e.g., [8,9,10,12] and Section A, below); indeed, this is the method that was used in several of California's pilot audits, including the audit in Orange County.

However, to use the method in Colorado would require major changes to RLATool, for reporting batch-level contest results prior to the audit, for drawing the sample, for reporting audit findings, and for determining when the audit can stop. The changes would include modifying data structures and the county user interface. No-CVR counties would also have to revise their audit procedures. Among other things, they would need to report vote subtotals for physically identifiable groups of ballots before the audit starts. And no-CVR counties with voting systems that can only report subtotals by precinct might have to make major changes to how they handle ballots, for instance, sorting all ballots by precinct.

Fortunately, the stratified audit approach of Section 2 makes possible a "hybrid" RLA that keeps the advantages of ballot-level comparison audits in CVR counties but does not require major changes to how no-CVR counties audit, nor major changes to RLATool.

In order to use Equation 1, we must develop tests for the overstatement error that are appropriate for the corresponding voting system. Sections 3.1 and 3.2 describe these tests for overstatement in the CVR and no-CVR strata, respectively.

3.1 Comparison audits of overstatement quotas

To use comparison auditing in the approach to stratification described above requires extending previous work to test whether the overstatement error exceeds $\lambda_s V_{w\ell}$, rather than simply $V_{w\ell}$. Appendix A derives this generalization for arbitrary batch sizes, including batches consisting of one ballot. The derivation considers only a single contest, but the MACRO test statistic [9,12] automatically extends the result to auditing any number of contests simultaneously. The derivation is for plurality contests, including "vote-for-k" plurality contests. Majority and super-majority contests are a minor modification [8].[9]

[8] Since so few ballots are cast in no-CVR counties, cruder approaches might work, for instance, pretending that no-CVR counties had CVRs, but treating any ballot sampled from a no-CVR county as if it had a 2-vote overstatement error. See [**?**].

[9] So are some forms of preferential and approval voting, such as Borda count, and proportional representation contests, such as D'Hondt [13]. For a derivation of

3.2 Ballot-polling audits of a tolerable overstatement in votes

To use the new stratification method with ballot-polling requires a different approach than [4] took: their approach tests whether w got a larger *share* of the votes than ℓ, but we need to test whether the margin in votes in the stratum in exceeds a threshold (namely, $\lambda_s V_{w\ell}$). This introduces a nuisance parameter, the number of ballots with votes for either w or ℓ. We address this by maximizing the P-value over all possible values of the nuisance parameter. Appendix B develops the test.

Unlike the test in [4], which is based on Wald's Sequential Probability Ratio Test [15], this test is not inherently sequential. Hence, to use it in an audit that can "escalate" in steps, multiplicity must be accounted for. See [8,**?**] for an approach based on Bonferroni's inequality.

4 Numerical examples

Jupyter notebooks containing calculations for stratified hybrid audits intended to be relevant for Colorado are available at https://www.github.com/pbstark/CORLA18.

`hybrid-audit-example-1` contains two examples. The first 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.

The second 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 about 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.

If it were possible to conduct a ballot-level comparison audit for the entire contest, a RLA at risk limit 5% could terminate after examining 31 ballots if it found no errors. The additional work for the hybrid stratified audit is disproportionately in the no-CVR counties, as arguably it should be.

A second notebook, `hybrid-audit-example-2`, illustrates the workflow for a hybrid stratified RLA of an election with 2 million ballots cast. The reported

ballot-level comparison risk-limiting audits for super-majority contests, see https://github.com/pbstark/S157F17/blob/master/audit.ipynb. (Last visited 14 May 2018.) Changes for IRV/STV are more complicated.

margin is just over 1%, but the reported winner and reported loser are actually tied in both strata. The risk limit is 5%. For a sample of 500 ballots from the CVR stratum and 1000 ballots from the no-CVR stratum, the maximum Fisher's combined P-value is over 20%, so the audit cannot stop there.

A third notebook, `fisher_combined_pvalue`, illustrates the numerical methods used to check whether the maximum Fisher's combined P-value of a stratified hybrid audit is below the risk limit. It includes code to set up the audits in the CVR stratum and in the no-CVR stratum, to find the lower and upper bounds λ_- and λ_+ for λ, evaluate Fisher's combining function along a grid of possible values, and to compute bounds on the P-value via Equation 3.

5 Discussion

We give a new class of procedures for RLAs based on stratified random samples. The method is agnostic about the capability of voting equipment in different strata, unlike previous methods, which worked only for comparison audits in every stratum.

Among other things, the new approach solves a problem in Colorado by allowing ballot-polling in some counties to be combined with ballot-level comparisons in others to conduct RLAs of contests that cross jurisdictional lines, such as statewide contests and many federal contests.

We give numerical examples relevant to Colorado, implemented in Jupyter notebooks can be modified to estimate the workload for different contest sizes, margins, and risk limits. Generally, increasing the sample size in one stratum allows the other sample size to be decreased. 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. In our numerical experiments, the new method requires auditing far fewer ballots than previous approaches would.

A Comparison audits for an overstatement quota

A.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.
- $V_{w\ell}$: 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)

A.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 P-value is not larger than its nominal value).

For every winner, loser pair (w, ℓ), we want to test whether the overstatement error exceeds some threshold, generally one tied to the reported margin between w and ℓ. 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 λ_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 λ_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 jth 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, $\mathbb{P}\{T_j \le 1\} = 1$, 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 $E < \lambda_s$.

A.3 Testing $\mathbb{E}T_1 \ge \lambda_s/U$

A variety of methods are available to test whether $\mathbb{E}T_1 < \lambda_s/U$. One particularly elegant sequential method is based on Wald's Sequential Probability Ratio Test (SPRT) ([15]). Harold Kaplan pointed out this method on a website that no longer exists. A derivation of this *Kaplan-Wald* method is in Appendix A of [13]; 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].

B Ballot-polling derivation

In this section, we derive a ballot-polling test of the hypothesis that the margin (in votes) in a single stratum exceeds a threshold c. The test is not sequential, so if it is to be used in a RLA that escalates in steps from an initial sample size to larger sample sizes before leading to a full hand count, multiplicity must be taken into account, for instance, using Bonferroni's inequality. (See, e.g., [8,**?**].)

B.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 ℓ, $A_{\ell,s}$ have a vote for ℓ but not for w, and $A_{u,s} = N_s - N_{w,s} - N_{\ell,s}$ have votes for both w and ℓ or neither w nor ℓ, 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.[10]

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 ℓ, with B_ℓ and B_u defined analogously. The conditional joint distribution of (B_w, B_ℓ, B_u) is tri-hypergeometric:

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

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

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,

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

The alternative is the compound hypothesis $A_{w,s} - A_{\ell,s} > c$.[11] 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 $B_w - B_\ell = nd$.[12]

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

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

B.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 a fixed 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,

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

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

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 $N_w - N_\ell = c$.*

[11] 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 $V_{w,s} - V_{\ell,s} > c$.

[12] In contrast, the BRAVO ballot-polling method [4] conditions only on $B_w + B_\ell = m$.

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 ℓ 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_ℓ, 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 is true always, then finding the maximum is trivial; otherwise, computing the unconditional P-value is a simple 1-dimensional optimization problem on a bounded interval.

B.3 Conditional testing

If the conditional tests are always conducted at significance level α or less, so that $\mathbb{P}\{\text{Type I error}|B = n\} \leq \alpha$, then the overall procedure has significance level α 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 the conditional hypergeometric test will have a conservative P-value unconditionally.

References

1. 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)
2. 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
3. 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
4. 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)
5. Lindeman, M., Stark, P.B.: A gentle introduction to risk-limiting audits. IEEE Security and Privacy **10**, 42–49 (2012)
6. Pesarin, F., Salmaso, L.: Permutation tests for complex data: Theory, applications, and software. John Wiley and Sons, Ltd., West Sussex, UK (2010)
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.: Risk-limiting vote-tabulation audits: The importance of cluster size. Chance **23**(3), 9–12 (2010)
12. 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
13. 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
14. Stark, P.B., Wagner, D.A.: Evidence-based elections. IEEE Security and Privacy **10**, 33–41 (2012)
15. Wald, A.: Sequential tests of statistical hypotheses. Ann. Math. Stat. **16**, 117–186 (1945)