Draft: 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] Department of Political Science, Columbia University, NY, USA and Verified Voting Foundation

Abstract. Risk-limiting audits (RLAs) offer a statistical guarantee that if a full manual tally would show that the reported election outcome is wrong, an RLA has a known minimum chance of leading to a full manual tally. RLAs generally rely on random samples; 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 make the audit more efficient. For example, some Colorado counties (comprising 98.2% of voters) have new voting systems that allow auditors to check how the system interpreted each ballot; the rest do not. Previous approaches to combining information from all counties into a single RLA of a statewide contest would require counties with new voting systems to use a less efficient method than their equipment permits, or would require major procedural changes. We provide a simpler, more efficient approach: stratify cast ballots into ballots cast in counties with newer systems and those cast in counties with legacy systems; sample individual ballots from those strata independently; apply a generalization of ballot-level comparison auditing in the first stratum and of ballot-polling auditing in the second to test the hypothesis that the "overstatement error" in each stratum exceeds a threshold; combine the stratum-level results using Fisher's nonparametric combination of tests; and find the maximum combined P-value over all partitions of the error. The audit can stop when the maximum is less than the risk limit. The method for combining information from different strata and for combining different audit strategies (ballot-polling and comparison) is new, and immediately applicable in Colorado. We provide an open-source reference implementation of the method 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 is a procedure that has a known minimum chance of leading to a full manual tally of the ballots if the electoral outcome of that tally would differ from the reported outcome. *Outcome* means the winner or winners, 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 election accounting and security procedures were followed.

Risk-limiting audits are generally (but not necessarily) incremental, relying on sequential hypothesis tests: 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 is to sample individual ballots uniformly at random from the ballots cast in the contest, with or without replacement [11]. Risk calculations for samples of individual ballots can also 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 risk calculations, e.g., ballot-polling rather than ballot-level comparisons [4,5], both of which are described below.

It may be more tractable logistically to use a stratified sample, selecting ballots independently from different counties or groups of counties. Stratification may also be useful within counties, for instance, to allow auditors to start auditing before all vote-by-mail or provisional ballots have been tallied, by sampling independently from ballots cast in person, by mail, and provisionally.

While stratified RLAs have been considered previously [8,3], the methods addressed only a single approach to auditing, batch-level comparisons, and only a particular test statistic.

This paper introduces a more general approach to using stratified samples in RLAs. The approach involves finding a “worst case” P-value (over a vector of nuisance parameters) for the null hypothesis that a full count would find different winners than were 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. Here, only the total *overstatement* determines whether the reported outcome is correct, but how that error is partitioned across strata affects 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.)

For each fixed value of the nuisance parameters, Fisher's combining function is used to combine independent P-values from the strata into an overall P-value that the error in each stratum exceeds a threshold. Conceptually, the procedure considers every possible partition of outcome-changing error across the strata, and calculates the largest P-value across all such partitions, using Fisher's combining function. If the maximum P-value is less than the risk limit, the audit can stop; otherwise, the audit must inspect more ballots.

In practice, it is not necessary to consider all possible partitions the allowable error across strata: the P-value is 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 (with a 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 compare 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) do not allow auditors to check how the system interpreted voter intent for individual ballots. Election results involving those counties can still be audited using *ballot-polling audits* [4,5], 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. Ballot-polling audits generally require examining more ballots than ballot-level comparison audits to attain the same risk limit.

There is currently no literature on how to combine ballot polling in non-CVR counties with ballot-level comparisons in CVR counties to audit contests that include voters in both. Existing methods either would require all counties to use the lowest common denominator (ballot-polling, which does not take advantage of the CVRs), or would require no-CVR counties to perform *batch-level comparisons* (manually tallying physical groups of ballots for which the voting system reported subtotals). Batch-level comparison audits were found in California to be less efficient than ballot-polling audits [1].[3]

Our new approach to stratification solves this problem, facilitating RLAs of contests that cross jurisdictional boundaries (*cross-jurisdictional contests*), such

[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.

as gubernatorial contests and statewide ballot measures, which in Colorado include CVR and non-CVR counties. As of July, 2018, Colorado has not performed a risk-limiting audit of a cross-jurisdictional contest.

Section 2 presents the new method for stratified audits. Section 3 illustrates the method by combining ballot-polling in one stratum with ballot-level comparisons auditing in another. This requires straightforward modifications of both ballot-polling and ballot-level comparison to allow the overstatement to be compared to 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 the cast ballots into non-overlapping groups, and drawing independent random samples from those groups. [8,3] discuss stratified sampling in batch-level comparison audits, using a particular test statistic. The method we develop here is more general and more flexible: in principle it can be used with any test statistic, and the 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; the key reference is [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.

We use w to denore 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, so that $V_w > V_\ell$, 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 two strata, $s = 1, 2$. 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 $V_{w\ell,1} + V_{w\ell,2} = V_{w\ell} > 0$. Let $A_{w\ell}$ denote the margin (in votes) of reported winner w over reported loser ℓ that a full hand count of all validly cast ballots would show, that is, the *actual* margin rather than the *reported* margin. 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; this too may be negative.

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

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 $\omega_{w\ell} \equiv \omega_{w\ell,1} + \omega_{w\ell,2} < V_{w\ell}$.

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 $\omega_{w\ell,2} \geq (1-\lambda)V_{w\ell}$.[5] If, for all λ, 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.)

2.1 Fisher's combination method

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.[6] 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 λ_1 and $\lambda_2 = 1-\lambda_1$, we can reject the conjunction hypothesis at level α, 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 λ 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.

$p_s(\lambda)$ 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 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.

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

[6] 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 [2].

One could use another function besides Equation 1 to combine P-values. Combining functions must satisfy the following properties [6]:

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

Other combining functions include Liptak's combining function, $T = \sum_i \Phi^{-1}(1-p_i)$, and Tippett's combining function, $T = \max_i(1-p_i)$.

Fisher's combining function is convenient for this application because the tests in different strata are statistically independent, so analytic formulas based on the chi-squared distribution can be used to calibrate the distribution of $\chi(\lambda_1, \lambda_2)$. If the tests across strata were correlated, the distribution of the combination function would need to be calibrated by simulation and another combining function might have better properties [6].

2.2 Maximizing Fisher's combined P-value

The audit can stop if the maximum of Fisher's combined P-value over all λ is not larger than α, the risk limit. For a given set of audit data, finding the maximum P-value over all λ 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 λ.

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 λ. (Lower bounds translate directly into an upper bound on the P-value as a function of λ: 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 α.) Let λ_- be the smallest possible value of λ and λ_+ be the largest possible value of λ. 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

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

Recall that $p_s(\cdot)$ increases monotonically in its argument, so $p_1(\lambda)$ is monotonically increasing in λ and $p_2(1-\lambda)$ is monotonically decreasing in λ. 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 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 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 all $\lambda \in [\lambda_-, \lambda_+]$, the upper bound is anywhere 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 of the grid points $\{a_k\}$, the sample size in one or both strata needs to be larger.

If the lower bound is less than the chi-square $1-\alpha$ quantile on some interval, but $\chi(a_k)$ is above this quantile at each of the grid points $\{a_k\}$, then it is indeterminate whether the audit can be stopped. One should construct a tighter lower bound by refining the grid, dividing each interval $[a_k, a_{k+1})$ into smaller intervals and recomputing the lower bound. If parts of the refined lower bound are still less than the chi-square $1-\alpha$ quantile, then we cannot conclude that the risk limit is attained. The sample sizes in one or both strata must be increased.

3 Strategies for auditing heterogeneous voting systems

The stratified auditing method in Section 2 solves the problem of auditing contests for which ballots were cast using heterogeneous voting equipment. This is the case in Colorado, where most ballots are cast in precincts with equipment that produces CVRs, but a small fraction are not.

CRS 1-7-515 requires Colorado to conduct risk-limiting audits beginning in 2017. The first set of coordinated risk-limiting election audits across the state took place in Colorado in November, 2017.[7] Those audits only covered contests restricted to a single county: counties could audit independently. Colorado's "uniform voting system" program[8] led many Colorado counties to purchase (or to plan to purchase) voting systems that export CVRs, and thus can be audited using ballot-level comparisons. It is estimated that by June, 2018, 98.2% of active Colorado voters will be in CVR counties. These counties cannot audit ballots independently: margins and risk limits apply to entire contests, not to the portion of a contest included in a county. To audit statewide elections and contests that cross county lines, Colorado will need to implement new approaches that account for heterogeneous voting equipment.

Several auditing strategies could be used in Colorado. The first approach is to use ballot-polling audits for all cross-jurisdictional contests. This would not take advantage information provided by CVRs and would increase the workload disproportionately in the CVR counties.

Another approach is to perform a comparison audit across all counties, but to use batches consisting of more than one ballot (and to perform batch-level

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

[8] https://www.sos.state.co.us/pubs/elections/VotingSystems/UniformVotingSystem.html

comparisons) in legacy counties and batches consisting of a single ballot (and to perform ballot-level comparisons) in CVR counties.[9]

However, Colorado's software would need to be modified in order to allow sampling batches with unequal probabilities and to calculate risk appropriately. The mathematical details for calculating batch-level error bounds, drawing the samples with probability proportional to an error bound, and calculating the attained risk from the sample results are worked out in published papers [9,10,12] and in Section A. Indeed, this is the method that was used in several of California's pilot audits, including the audit in Orange County.

Moreover, this approach 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.

Finally, the stratified audit approach from Section 2 can solve the problem of auditing in CVR and no-CVR counties. Every ballot cast in the contest is in exactly one of two strata, CVR counties and no-CVR counties. The CVR counties would conduct a ballot-level comparison audit while the legacy counties independently conduct a ballot-polling audit, to then be combined by a central organization. This approach does not unduly increase the workload in CVR counties to compensate for legacy equipment, nor does it force the legacy counties to audit entire batches of ballots.

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.

Of these methods, 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.

3.1 Batch comparison audits of a tolerable overstatement in votes

Previous methods for comparison auditing must be modified to handle two new requirements. The first relates to partitioning the permissible overstatement through the parameters λ_s, as discussed in Section 2. The second handles batch-level comparison audits.

For this framework, we'd like to test whether the overstatement of any margin (in votes) exceeds some fraction λ of the overall margin $V_{w\ell}$ between reported winner w and reported loser ℓ. 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 must test other values of λ, as described in Section 2.

[9] 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.

We address the second requirement by deriving a method for batches of arbitrary size, which might be useful to audit contests that include CVR counties and legacy counties. This modification comes from changing the assumed upper bound on the possible overstatement in batches. We keep the *a priori* error bounds tighter than the "super-simple" method [12]. To keep the notation simpler, we consider only a single contest, but the MACRO test statistic [9,12] 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].[10]

The mathematical derivations are in Appendix A.

3.2 Ballot-polling audits of a tolerable overstatement in votes

In in Appendix B, 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 [4].

Existing ballot-polling methods consider only the fraction of ballots with a vote for either w or ℓ 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 λ_s, the necessary inference is about the *difference* between the number of votes for w and the number of votes for ℓ. This introduces a nuisance parameter, the number of ballots with votes for either w or ℓ. 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.

4 Numerical examples

Examples of stratified hybrid audits, like what could be used in Colorado, 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

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

not as efficient as a ballot-level comparison audit, but far more efficient than a ballot-polling audit.

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.

5 Discussion

We have developed a general procedure for auditing stratified random samples. Previous methods for auditing stratified random samples of ballots were tailored to comparison audits. Our proposed method is more flexible, in that it allows for different hypothesis tests in each stratum. Thus, it can be applied to a variety of situations where stratified random sampling is appropriate: sampling ballots independently from several counties, or even from different groups of ballots within a county. To our knowledge, Fisher's combination method has not been used to combine independent audits before.

Furthermore, we have presented a case study in using the proposed stratified method to audit cross-jurisdictional contests in which some jurisdictions use voting equipment that can export a CVR and some jurisdictions use legacy equipment. The accompanying Jupyter notebooks can be modified and run with different contest sizes, margins, and risk limits to estimate the workload in different scenarios. 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. 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. In our numerical experiments, the proposed stratification method greatly reduced the number of ballots needed to confirm the reported contest outcome compared to other auditing approaches.

A Ballot comparison derivation

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 risk 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, $\mathbb{P}\{T_j \leq 1\} = 1$, 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 $E < \lambda_s$.

A.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) ([15]). 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 [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

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.[11]

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

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:

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

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,

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

The alternative is the compound hypothesis $A_{w,s} - A_{\ell,s} > c$.[12] 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 $B_w - B_\ell = nd$.[13]

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 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,

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

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

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

[13] In contrast, the BRAVO ballot-polling method [4] 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* $N_w - N_\ell \le c$*. 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 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 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.

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 our 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)