Preparing to Audit Colorado's 2018 Primaries

Mark Lindeman
Neal McBurnett
Kellie Ottoboni
Ronald L. Rivest
Philip B. Stark

Draft February 7, 2018

Abstract

Colorado's current audit software (RLATool) needs to be improved to audit partisan primaries in Colorado in 2018, even to draw the sample for the audit: the current version of RLATool does not allow the user to select the sample size, nor does it directly allow an unstratified random sample to be drawn across counties. Similarly, RLATool needs to be modified to recognize that contests can cross jurisdictional boundaries; currently, it treats every contest as if it were entirely contained in a single county. Margins and risk limits apply to entire contests, not to the portion of a contest included in a county. Second, to audit a contest that includes voters in "legacy" counties (counties with voting systems that cannot export cast vote records) and voters in counties with newer systems requires new statistics, if one wants to keep the efficiency of ballot-level comparison audits that the newer systems afford. Third, auditing contests that appear only on a subset of ballots can be made much more efficient if the sample can be drawn from just those ballots that contain the contest. While allowing samples to be restricted to ballots reported to contain a particular is not essential for the June, 2018 primaries, it will be necessary eventually to make it feasible to audit smaller contests.

1 Introduction

A risk-limiting audit (RLA) of an election is a procedure that has a known, pre-specified minimum chance of correcting the electoral outcome if the outcome is incorrect—that is, if the reported outcome differs from the outcome that a full manual tabulation of the votes would find. RLAs require a durable, voter-verifiable record of voter intent, such as paper ballots, and they assume that this audit trail is sufficiently complete and accurate that a full hand tally would show the true electoral outcome. That assumption is not automatically satisfied: a *compliance audit* is required.

Risk-limiting audits are generally incremental: they examine more ballots, or batches of ballots, until either (i) there is strong statistical evidence that a full hand tabulation would confirm the outcome, or (ii) the audit has led to a full hand tabulation, the result of which becomes the official result.

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

CRS 1-7-515 required Colorado to implement risk-limiting audits beginning in 2017. (There are provisions to allow the Secretary of State to exempt some counties.) The first statewide risk-limiting election audits took place in Colorado in November, 2017.

Colorado's "uniform voting system" program has led many Colorado counties to purchase (or to plan to purchase) voting systems that are auditable at the ballot level: those systems export cast vote records (CVRs) for individual ballots in a manner that allows the corresponding paper ballot to be identified, and conversely, make it possible to find the CVR corresponding to any particular paper ballot. We call counties that have such systems "CVR" counties. It is estimated that by June, 2018, 98.2% of active Colorado voters will be in CVR counties. CVR counties can perform "ballot-level comparisons," which 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.

Other counties ("legacy" or "non-CVR" counties) have systems that do not allow auditors to check how the system interpreted voter intent for in-

dividual ballots. Their election results can still be audited, provided their voting systems create a voter-verifiable paper trail (*e.g.*, voter-marked paper ballots) that is conserved to ensure that it remains accurate and intact, and organized well enough to permit ballots to be selected at random. Pilot audits in California suggest that the most efficient way to audit such systems is by "ballot-polling" (in contrast to "batch-level comparisons," for example).

There is currently no literature on how to perform risk-limiting audits of contests that include CVR counties and non-CVR counties by combining ballot polling and ballot-level comparisons. Existing methods would either require all counties to use the lowest common denominator, ballot-polling (which does not take advantage of the CVRs, and thus is expected to require more auditing than a method that does take advantage of the CVRs), or would require non-CVR counties to perform batch-level comparisons, which were found in California to be (generally) less efficient than ballot-polling audits.[1]

This document focuses on near-term requirements for risk-limiting audits in Colorado: June and November 2018.

1.1 Colorado in June, 2018

We understand that for June, 2018, Colorado Secretary of State Wayne Williams intends to require a risk-limiting audit of at least one statewide contest in addition to a countywide contest in each county.

Auditing efficiency is controlled in part by how well the audit can limit the sample to ballots that contain the contests under audit. Some contests are on (essentially) every ballot, for instance the governor's race. Others, such as mayoral contests, may appear on only a small fraction of ballots cast in a county. Partisan primaries—even for statewide office—are somewhere in between, because in general no single party's primary appears on every ballot cast in the state. Thus, either we accept a cut to efficiency and sample ballots from counties (or collections of counties) but keep the simplicity of being able to sample uniformly, or we develop a way to focus the auditing on the ballots that contain the contest. The latter requires external information, e.g., from SCORE, as discussed below.

Moreover, party primaries for statewide offices (and perhaps other con-

[1] See Rivest (2018) for a different (Bayesian) approach to auditing contests that include both CVR counties and non-CVR counties.

tests) will include CVR counties and non-CVR counties, so we need a method to audit across mixed jurisdictional voting technology.

This report addresses both issues, providing a handful of ways of dealing with heterogeneous voting technology, varying in efficiency, complexity, and on whom any additional audit burden falls.

2 Crude (and unpleasant) approaches

2.1 Hand count the legacy counties

The simplest approach to combining legacy counties with CVR counties is to require every legacy county to do a full hand count of the primaries, and to conduct a ballot-level comparison audit in CVR counties, based on contest margins adjusted for the results of the manual tallies in the CVR counties. For instance, imagine a contest with two candidates, reported winner w and reported loser ℓ. Suppose 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 the reported winner. Suppose that a full manual tally of the votes in the legacy counties shows V'_w votes for w and V'_ℓ votes for ℓ. Suppose that a total of N ballots were cast in the CVR counties. Then the diluted margin for the comparison audit in the CVR counties is $[(V_w - V'_w) - (V_\ell - V'_\ell)]/N$. This approach is presumably unacceptable because it would require every legacy county to do a full hand count. (But it would provide an incentive for those counties to upgrade their systems sooner rather than later, and it does not penalize CVR counties for the fact that their legacy siblings have not yet upgraded.)

2.2 Subtract error bounds for the legacy counties from vote totals

If ballot accounting and SCORE can give good upper bounds on the number of ballots cast in each contest in legacy counties, there are simple upper bounds on the total possible overstatement error each legacy county could contribute to the overall contest results; those can be subtracted from the overall margin (as in the previous subsection) and the remainder of the contests can be audited in CVR counties against the adjusted margins. For instance, consider a primary that appears on N ballots in a legacy counties.

Suppose that in legacy counties, the overall, statewide contest winner, w, is reported to have received V'_w votes, and some loser, ℓ, is reported to have received V'_ℓ votes. (Note that V'_ℓ could be greater than V'_w: w is not necessarily the reported winner in the legacy counties.) Then the most overstatement error that the county could possibly have in determining whether w in fact beat ℓ is if every reported undervote, invalid vote, or vote for a different candidate, t, had in fact been a vote for ℓ (producing a 1-vote overstatement), and every vote reported for w was in fact a vote for ℓ (producing a 2-vote overstatement). The reduction in the margin that would produce is $N - V'_w - V'_\ell + 2V'_w = N + V'_w - V'_\ell$ votes.

This approach may be unacceptable for at least two reasons: first, if the margin is small, it could easily lead to a full hand count in every county. Second, even if it doesn't lead to a full hand count, it penalizes CVR counties for the fact that non-CVR counties have not upgraded their systems, because it reduces the margin in every contest that includes a legacy county.

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

A third simple-but-pessimistic approach is to sample uniformly from all counties as if one were performing a ballot-level comparison audit everywhere, but to treat any ballot selected from a legacy county as a two-vote overstatement. This approach is probably unacceptable for at least two reasons: first, if the margin is small, it could easily lead to a full hand count in every county. Second, even if it doesn't lead to a full hand count, it penalizes CVR counties for the fact that non-CVR counties have not upgraded their systems, because it will require expanding the sample (across all counties) every time a ballot is selected from a legacy county.

3 Variable batch sizes

A third approach is to perform a comparison audit across all counties, but to use batches consisting of more than one ballot (batch-level comparisons) in legacy counties and batches of a single ballot (ballot-level comparisons) in CVR counties. The constraint here is that the non-CVR counties need to be able to report vote subtotals for physically identifiable batches. If a county's voting system can only report subtotals by precinct but the county

does not sort paper ballots by precinct, this approach might require revising how the county handles its paper; we understand that this is the case in many Colorado counties.

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

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

While those two approaches are laborious, they would provide a viable short-term solution, especially combined with information from SCORE to check that the reported batch-level results contain the correct number of ballots for each contest under audit. Moreover, it does not unduly increase the workload in CVR counties to compensate for the fact that some other counties have not upgraded their voting systems.

This kind of variable-batch-size comparison audit approach would require modifying or augmenting RLATool in several ways:

1. the CVR reporting tool would need to be modified to allow non-CVR counties to report batch-level results in a manner analogous to how CVR counties report ballot-level results, or an external tool would need to be provided.

2. the sampling algorithm would have to allow sampling batches—and sampling them with unequal probability, because efficient batch-level audits involve sampling batches with probability proportional to a bound on the possible overstatement error in the batch. It would also need to calculate the appropriate sampling probability for each batch (of whatever size). Again, this could be accommodated using an external tool to draw the sample from legacy counties.

3. the risk calculations would need to be modified. This, too, could be done with external software, with suitable provisions for capturing audit data from RLATool or directly from legacy counties.

None of these changes is enormous; the mathematics and statistics are already worked out in published papers, and there is exemplar code for calculating the batch-level error bounds, drawing the samples with probability proportional to an error bound, and calculating the attained risk from the sample results. Indeed, this is the method that was used in several of California's pilot audits, including the audit in Orange County. A derivation of a method for comparison audits with variable batch sizes is given below in section 6.

4 Stratified "hybrid" audits

Other approaches involve *stratification*: partitioning the cast ballots into non-overlapping groups and sampling independently from those groups. One could stratify by county, but in general it is simpler and more efficient statistically (i.e., results in auditing fewer ballots) to minimize the number of strata. We consider methods that use two strata: CVR counties and non-CVR counties. Collectively, the ballots cast in CVR counties comprise one stratum and the ballots cast in legacy counties comprise a second stratum; every ballot cast in the contest is in exactly one of the two strata. We assume that the samples are drawn from the two strata independently.

4.1 Partitioning the total permissible overstatement into strata

The simplest approach to stratification involves partitioning the risk limit and the tolerable overstatement error of the tabulation into two pieces, one for the (pooled) CVR counties and one for the (pooled) non-CVR counties. Let $V_{w\ell} > 0$ denote the contest-wide margin (in votes) of reported winner w over reported loser ℓ. 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. Let $A_{w\ell}$ denote the margin (in votes) of reported winner w over reported loser ℓ that a full hand count of the entire contest would show, that is, the *actual* margin rather than the *reported* margin. Reported winner w really beat reported loser 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.

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 iff $\omega_{w\ell} \equiv$

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

Pick $\lambda_1 \in \Re$ and define $\lambda_2 = 1 - \lambda_1$. If $\omega_{w\ell,1} < \lambda_1 V_{w\ell}$ *and* $\omega_{w\ell,2} < \lambda_2 V_{w\ell}$, candidate w really received more votes than candidate ℓ. Some (λ_1, λ_2) pairs can be ruled out *a priori*, because (for instance) $\omega_{w\ell,s} \in [-2N_s, 2N_s]$, where N_s is the number of ballots cast in stratum s. There are other simple, sharper bounds, sketched below.

The choice of λ_1, the strata risk limits $\{\alpha_s\}$, and details of the audit procedures affect the workload and the overall risk limit. (See section 4.1.1.)

For ballot-level comparison audits, auditing to ensure that $\omega_{w\ell,s} < \lambda_s V_{w\ell}$ is discussed in section 6; it is a minor modification of the method embodied in RLATool.

For ballot-polling audits, auditing to ensure that $\omega_{w\ell,s} < \lambda_s V_{w\ell}$ is discussed in section 7. Note that this requires a more substantial modification of the standard ballot-polling calculations, because the standard calculations consider only the fraction of ballots with a vote for either w or ℓ that contain a vote for w, while we need to make an inference about the difference between the number of votes for w and the number of votes for ℓ. This introduces an additional nuisance parameter, the number of ballots with votes for either w or ℓ.

4.1.1 Combining stratum-level risk limits

We audit to test the two hypotheses $\{\omega_{w\ell,s} \geq \lambda_s V_{w\ell}\}_{s=1}^2$, independently for the two strata. If we reject *both* hypotheses, we conclude that the contest outcome is correct; otherwise, we manually re-tabulate the contest in one or both strata, depending on the audit rules. Those rules matter: generally, the two audits will need to be conducted to smaller risk limits individually than the desired risk limit for the contest as a whole, unless proceeding to a full hand tabulation in one stratum automatically triggers a full hand tabulation in the other stratum.

Recall that the samples are drawn independently from the two strata. Pick $\alpha_1, \alpha_2 \in (0, \alpha)$. (Below we discuss the choice further.) Also pick λ_1. Then if $\omega_{w\ell,1} < \lambda_1 V_{w\ell}$ and $\omega_{w\ell,2} < \lambda_2 V_{w\ell}$, the outcome is correct. We audit each stratum s to test the hypothesis $\omega_{w\ell,s} \geq \lambda_s V_{w\ell}$ with risk limit α_s, as if it were its own election. We want to know the relationship between those two stratum-level "risks" and the overall risk that the audit will not correct the outcome if the outcome is wrong. The overall risk depends on what we do if the audit in one stratum leads to a full manual tally of that stratum.

Here are the possibilities. The outcome is certainly correct if, in both strata, the net overstatement in the stratum is less than its threshold. For the outcome to be wrong, one or both strata need to have net overstatement $\omega_{w\ell,s}$ greater than its corresponding threshold $\lambda_s V_{w\ell}$. That is, if $\omega_{w\ell,1}+\omega_{w\ell,2} \geq V_{w\ell}$, then $\omega_{w\ell,1} \geq \lambda_1 V_{w\ell}$ or $\omega_{w\ell,2} \geq \lambda_2 V_{w\ell}$, or both. If the allocated overstatement is exceeded in only one stratum, h, then the chance that the stratum will be fully hand counted is at least $1 - \alpha_h \geq 1 - \alpha$.

If both $\omega_{w\ell,1} \geq \lambda_1 V_{w\ell}$ and $\omega_{w\ell,2} \geq \lambda_2 V_{w\ell}$, then the chance both are fully tabulated is $(1 - \alpha_1)(1 - \alpha_2)$, since the audit samples in the two strata are independent.

What should we do if the audit leads to a full tally in one stratum, h? We consider two options. The simpler is to automatically require a full hand count of the other stratum, to set the record straight. If the audit uses this rule, then we can take $\alpha_1 = \alpha_2 = \alpha$, and the procedure will have risk limit α.

A second approach is to adjust the contest margin for the results of the manual tally in the fully counted stratum, h, and continue to audit in the other, but against the overall margin, adjusted for the "known" tally in the stratum that had been counted: we test against the share $V_{w\ell} - A_{w\ell,h} \equiv \lambda'_t V_{w\ell}$, rather than against its (originally allocated) share $\lambda_t V_{w\ell}$. Then to reject the new null hypothesis in stratum t is to conclude that the overall outcome is correct.

The statistical wrinkle is that adjusting for the manual tally in the hand-counted stratum h changes the hypothesis being tested in the other stratum t in a way that is itself random: whether the original null $\omega_{w\ell,s} \geq \lambda_t V_{w\ell}$ is tested or the new null $\omega_{w\ell,s} \geq \lambda'_t V_{w\ell}$ is tested depends on what the sample reveals in stratum h. If the hypothesis does change, there is only one value possible for λ'_t—which depends on the reported margin $V_{w\ell}$ and the count $A_{w\ell,h}$ in stratum h—but it is unknown until $A_{w\ell,h}$ is known.

We assume that before any data are collected, the audit specifies two families of tests: for each stratum s, a family of level-α_s tests of the null hypothesis that the overstatement in the stratum is greater than or equal to c, for all feasible values of c. That is,

$$\Pr\{\text{reject hypothesis that} \omega_{w\ell,s} \geq c_s || \omega_{w\ell,s} \geq c_s\} \leq \alpha_s, \tag{1}$$

for $s = 1, 2$, and all feasible c_s. Moreover, we insist that the test depend on data only from ballots selected from its stratum. Because the samples in the

two strata are independent, for all feasible pairs c_1, c_2,

$$\Pr\{\text{reject neither hypothesis} \omega_{w\ell,s} \geq c_s, \;\; s = 1,2 || \omega_{w\ell,s} \geq c_s \} \text{for both} s = 1,2\} \geq (1-\alpha_1)(1-\alpha_2). \tag{2}$$

What is the chance that the audit leads to a full hand tabulation if the outcome is incorrect? One way the audit can lead to a full hand tally is if it leads to a full count in one stratum, the null hypothesis in the other stratum is changed, and the audit in the second stratum then proceeds to a full manual tally. (There are other ways the audit can lead to a full hand tally, for instance, if neither null hypothesis is rejected, but this is one way.)

If the outcome is wrong, there is at least one stratum in which the overstatement $\omega_{w\ell,s}$ exceeds the threshold $\lambda_s V_{w\ell}$. Let h be one such stratum. Then the chance the audit in stratum h leads to a full manual tally in that stratum is at least $(1 - \alpha_h)$. If the audit leads to a full manual tally in stratum h and the overall outcome is wrong, then the (new) null hypothesis in the other stratum, t must be true. If we started to audit that new hypothesis *ab initio*, the chance that we would reject it would be at most α_t, so the chance the audit would lead to a full hand count of stratum t is at least $1 - \alpha_t$. The question is whether "changing hypotheses" could make that chance smaller. The inequality 2 shows that it cannot: for any feasible pair of overstatements, $c = (c_1, c_2)$, if $\omega_{w\ell,1} \geq c_1$ and $\omega_{w\ell,2} \geq c_2$, the chance that neither the hypothesis $\omega_{w\ell,1} \geq c_1$ nor the hypothesis $\omega_{w\ell,2} \geq c_2$ will be rejected is at least $(1 - \alpha_1)(1 - \alpha_2)$.

And therefore, for this procedure, the chance that there will be a full hand count in both strata is at least $(1 - \alpha_1)(1 - \alpha_2)$ if the outcome is incorrect, even if the probability were zero that both of the original audits would proceed to a full hand count. The overall risk limit is thus not larger than $1 - (1 - \alpha_s)(1 - \alpha_t)$.

As an example, suppose we want the overall risk limit to be 5%. If we use a risk limit of 4% in the no-CVR stratum and a risk limit of 1.04% in the CVR stratum, the risk limit will be $1 - 0.96 \times 0.9896 < 0.05$.

4.2 Constraining the total overstatement across strata

A more statistically efficient approach to ensuring that the overstatement error in the two strata does not exceed the margin is to try to constrain the *sum* of the overstatement errors in the two strata, rather than constrain the pieces separately: there are many ways that the total overstatement could be

less than $V_{w\ell}$ without having the overstatement $\omega_{w\ell,s}$ in stratum s less than $\lambda_s V_{w\ell}$, $s = 1, 2$. To that end, imagine *all* values λ_1. If, for all such pairs, we can reject the hypothesis that the overstatement error in stratum 1 is greater than or equal to $\lambda_1 V_{w\ell}$ *and* the overstatement error in stratum 2 is greater than or equal to $\lambda_2 V_{w\ell}$, then we can conclude that the outcome is correct.

To test the conjunction hypothesis (i.e., that both of those null hypotheses are false), we use Fisher's combining function. Let $p_s(\lambda)$ be the p-value of the hypothesis $\omega_{w\ell,s} \geq \lambda V_{w\ell}$. 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 the combination

$$\chi(\lambda_1, \lambda_2) = -2 \sum_{s=1}^{2} \ln p_s(\lambda_s) \tag{3}$$

has a probability distribution that is dominated by the chi-square distribution with 4 degrees of freedom. (If the two tests had continuously distributed p-values, the distribution would be exactly chi-square with four 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 results in (**?**).)

Hence, if, for all λ_1 and $\lambda_2 = 1 - \lambda_1$, the combined statistic $\chi(\lambda_1, \lambda_2)$ is greater than the $1 - \alpha$ quantile of the chi-square distribution with 4 degrees of freedom, the audit can stop.

The calculation of $p_s(\lambda)$ uses the procedures discussed in sections 6 and 7.

5 Sampling from subcollections

To audit contests that are contained on only a fraction of the ballots cast in one or more counties efficiently requires the ability to sample from just those ballots (or, at least, from a subset of all ballots that contains every such ballot). Because the CVRs cannot be entirely trusted (otherwise, the audit would be superfluous), we cannot rely on them to determine which ballots contain a given contest. However, if we have independent knowledge of the number of ballots that contain a given contest (e.g., from the SCORE system), then there are methods that allow the sample to be drawn from ballots whose CVRs contain the contest and still limit the risk rigorously. See Benaloh et al. (2011) and Bañuelos and Stark (2012) for details.

6 Comparison audits of a tolerable overstatement in votes

We consider auditing in a single stratum 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 might want to test other values of λ, as described above.

In Colorado, comparison audits have been ballot-level (i.e., batches consisting of a single ballot). In this section, we derive a method for batches of arbitrary size, which might be useful for Colorado to audit contests that include CVR counties and legacy counties. We keep the *a priori* error bounds tighter than the "super-simple" method (Stark, 2010). To keep the notation simpler, we consider only a single contest, but the MACRO test statistic (Stark, 2009b, 2010) 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 such as bond measures are a minor modification (Stark, 2008).[2]

6.1 Notation

- $\mathcal{W}$: the set of reported winners of the contest
- $\mathcal{L}$: the set of reported losers of the contest
- N_s ballots were cast in all in the stratum. (The contest might not appear on all N_s ballots.)
- P "batches" of ballots are in stratum s. A batch contains one or more ballots. Every ballot in stratum s is in exactly one batch.
- n_p: number of ballots in batch p. $N_s = \sum_{p=1}^{P} n_p$.
- $v_{pi} \in \{0, 1\}$: the reported votes for candidate i in batch p

[2]So are some forms of preferential and approval voting, such as Borda count, and proportional representation contests, such as D'Hondt (Stark and Teague, 2014). Changes for IRV/STV are more complicated.

- $a_{pi} \in \{0, 1\}$: 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_s: smallest reported margin in the stratum among all C contests audited using the same sample: $V_s \equiv \min_{w\in\mathcal{W},\ell\in\mathcal{L}} V_{w\ell,s}$
- V: smallest reported overall margin among all C contests audited using the same sample: $V \equiv \min_{w\in\mathcal{W},\ell\in\mathcal{L}} V_{w\ell}$
- $A_{w\ell,s} \equiv \sum_{p=1}^{P}(a_{pw} - a_{p\ell})$: actual margin in the stratum of reported winner $w \in \mathcal{W}$ over reported loser $\ell \in \mathcal{L}$, in votes
- $A_{w\ell}$: actual margin of reported winner $w \in \mathcal{W}$ over reported loser $\ell \in \mathcal{L}$, in votes, for the entire contest (not just in stratum s)

6.2 Reduction to maximum relative overstatement

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

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

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

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

For every winner, loser pair (w, 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 simple stratified audit, we might take 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

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

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 $V_{w\ell} \geq V$ and $0 \leq v_{pi} \leq n_p$,

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

Thus if we use $2n_p/V$ in lieu of u_p, we still get conservative results. (We also need to re-define U to be the sum of those upper bounds.) An intermediate, still conservative approach would be to use this upper bound for batches that consist of a single ballot, but use the sharper bound 4 when $n_p > 1$. Regardless, for the new definition of u_p and U, $\{T_j\}_{j=1}^n$ are IID, $\mathbb{P}\{T_j \leq 1\} = 1$, and

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

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

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

To test whether $\mathbb{E}T_1 < \lambda_s/U$, there are a variety of methods available. One particularly "clean" sequential method is based on Wald's Wald (1945) Sequential Probability Ratio Test (SPRT). Harold Kaplan pointed out this

method on a website that no longer exists. A derivation of this "Kaplan-Wald" method is given in Stark and Teague (2014, Appendix A); 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 Stark (2009a).

7 Ballot-polling audits of a tolerable overstatement in votes

7.1 Conditional hypergeometric test

We consider a single stratum s, containing N_s ballots. We will sample individual ballots without replacement from stratum s. 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.[3]

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. Then, conditional on $B = n$, the joint distribution of (B_w, B_ℓ, B_u) is tri-hypergeometric:

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

The test statistic will be the diluted sample margin, $D \equiv (B_w - B_\ell)/B$. This is the sample difference in the number of ballots for the winner and for the loser, divided by the total number of ballots in the sample. We want to test the compound hypothesis $A_{w,s} - A_{\ell,s} \le c$. 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,

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

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

The alternative is the compound hypothesis $A_{w,s} - A_{\ell,s} > c$.[4] Hence, we will reject for large values of D. Conditional on $B = n$, the event $D = (B_w - B_\ell)/B = d$ is the event $B_w - B_\ell = nd$.

Suppose we observe $D = d$. The test will condition on the event $B = n$ and $B_w + B_\ell = m$. (In contrast, the BRAVO ballot-polling method (Lindeman et al., 2012) conditions only on $B_w + B_\ell = m$.)

Given $B = n$, all samples of size n from the ballots are equally likely, by hypothesis. Hence, in particular, all samples of size n for which $B_w + B_\ell = m$ are equally likely. There are $\binom{A_{w,s}+A_{\ell,s}}{m}\binom{N_s-A_{w,s}-A_{\ell,s}}{n-m}$ such samples. Among these samples, B_w may take values $i = 0, 1, \ldots, m$. For a fixed i, there are $\binom{A_{w,s}}{i}\binom{A_{\ell,s}}{m-i}\binom{N_s-A_{w,s}-A_{\ell,s}}{n-m}$ samples with $B_w = i$ and $B_\ell = m - i$.

The factor $\binom{N_s-A_{w,s}-A_{\ell,s}}{n-m}$ counts the number of ways to sample $n - m$ of the remaining ballots. If we divide out this factor, we simply count the number of ways to sample ballots from the group of ballots for w or for ℓ. There are $\binom{A_{w,s}+A_{\ell,s}}{m}$ equally likely samples of size m from the ballots with either a vote for w or for ℓ, but not both, and of these samples, $\binom{A_{w,s}}{i}\binom{A_{\ell,s}}{m-i}$ contain i ballots with a vote for w but not ℓ. Therefore, conditional on $B = n$ and $B_w + B_\ell = m$, the probability that $B_w = i$ is

$$\frac{\binom{A_{w,s}}{i}\binom{A_{\ell,s}}{m-i}}{\binom{A_{w,s}+A_{\ell,s}}{m}}.$$

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 sum of these probabilities for events when $B_w - B_\ell \geq nd$. This event occurs for $B_w \geq \frac{m+nd}{2}$. Therefore,

$$\mathbb{P}_{A_{w,s},A_{\ell,s},N_s}\{D \geq d \mid B = n, B_w + B_\ell = m\} = \sum_{i=(m+nd)/2}^{\min\{m,A_{w,s}\}} \frac{\binom{A_{w,s}}{i}\binom{A_{\ell,s}}{m-i}}{\binom{A_{w,s}+A_{\ell,s}}{m}}. \quad (6)$$

This conditional p-value is thus the tail probability of the hypergeometric distribution with parameters $A_{w,s}$ "good" items, $A_{\ell,s}$ "bad" items, and a

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

sample of size m. This calculation is numerically stable and fast; tail probabilities of the hypergeometric distribution are available and well-tested in all standard statistics software.

In simulations, we compared the conditional hypergeometric test to testing using the tri-hypergeometric distribution directly. We used a population size of 1000 and varied the sampling rate from 1% to 20%, the fraction of ballots for w or ℓ from 50% to 90%, and the actual margin from 1% to 10%. The conditional hypergeometric test performed better than the tri-hypergeometric test: it tended to correctly reject the null hypothesis more frequently and in one example, its p-value was smaller than the corresponding tri-hypergeometric p-value about 80% of the time. Furthermore, the tri-hypergeometric distribution is not widely available in statistical packages and its p-value calculation involves summing over two variables (as opposed to just one in the hypergeometric case), making calculations much slower. For these reasons, we propose conditioning on both $B = n$ and $B_w + B_\ell = m$ to yield the conditional hypergeometric test.

7.2 Maximizing the p-value over the null set

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. 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,

$$\max_{A_{w,s},A_{\ell,s}\in\{0,1,\ldots,N\}:A_{w,s}-A_{\ell,s}\leq c,A_{w,s}+A_{\ell,s}\leq N_s} \sum_{i=(m+nd)/2}^{\min\{m,A_{w,s}\}} \frac{\binom{A_{w,s}}{i}\binom{A_{\ell,s}}{m-i}}{\binom{A_{w,s}+A_{\ell,s}}{m}}, \qquad (7)$$

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

7.2.1 Optimizing over the parameter c

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

Theorem 1. *Assume that $A_{w,s} > cm/2$, $B_w > B_\ell$, and $c > 0$. 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 $N_w - N_\ell = c$.*

THIS SEEMS TO HOLD FOR $c < 0$ TOO. THE PROOF DOESN'T SO MUCH MATTER ON $c = 0$ AND $c > 0$, BUT THE DIFFERENCE BETWEEN THE TWO VALUES. SHOULD THINK ABOUT WHETHER THIS IS WORTH ADDING, WHETHER THE ASSUMPTIONS THERE ARE USEFUL.

Proof. Without loss of generality, let $c = 0$ and assume $A_{w,s}$ is fixed.

The p-value p_0 for the simple hypothesis that $c = 0$ is

$$p_0 = \sum_{i=(m+nd)/2}^{\min\{m,A_{w,s}\}} \frac{\binom{A_{w,s}}{i}\binom{A_{w,s}}{m-i}}{\binom{2A_{w,s}}{m}} = \sum_{i=(m+nd)/2}^{\min\{m,A_{w,s}\}} T_i, \tag{8}$$

where T_i is defined as the ith term in the summand and $T_i \equiv 0$ for $i > \min\{m, A_{w,s}\}$.

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

$$\begin{aligned} p_c &= \sum_{i=(m+nd)/2}^{\min\{m,A_{w,s}+c\}} \frac{\binom{A_{w,s}+c}{i}\binom{A_{w,s}}{m-i}}{\binom{2A_{w,s}+c}{m}} \\ &= \sum_{i=(m+nd)/2}^{\min\{m,A_{w,s}\}} T_i \frac{(A_{w,s}+c)(A_{w,s}+c-1)\cdots(A_{w,s}+1)(2A_{w,s}+c-m)\cdots(2A_{w,s}+1)}{(A_{w,s}+c-i)\cdots(A_{w,s}+1-i)(2A_{w,s}+c)\cdots(2A_{w,s}+1)} + \Delta, \end{aligned}$$

where $\Delta \equiv \sum_{i=A_{w,s}+1}^{A_{w,s}+c} \frac{\binom{A_{w,s}+c}{i}\binom{A_{w,s}}{m-i}}{\binom{2A_{w,s}+c}{m}}$ is nonzero whenever $A_{w,s} > m - c$.

Terms in the fraction can be simplified: choose the corresponding pairs in the numerator and denominator. Fractions of the form $(A_{w,s}+c+j)/(A_{w,s}+j-i)$ can be expressed as $1+i/(A_{w,s}+j-i)$. Fractions of the form $(2A_{w,s}+k-m)/(2A_{w,s}+k)$ can be expressed as $1-m/(2A_{w,s}+k)$. Thus, the p-value can be written as

$$
\begin{aligned}
p_c &= \sum_{i=(m+nd)/2}^{\min\{m,A_{w,s}\}} T_i \prod_{j=1}^{c} \left(1 + \frac{i}{A_{w,s}+j-i}\right)\left(1 - \frac{m}{2A_{w,s}+j}\right) + \Delta \\
&> \sum_{i=(m+nd)/2}^{\min\{m,A_{w,s}\}} T_i \prod_{j=1}^{c} \left(1 + \frac{i}{A_{w,s}+c-i}\right)\left(1 - \frac{m}{2A_{w,s}+c}\right) + \Delta \\
&= \sum_{i=(m+nd)/2}^{\min\{m,A_{w,s}\}} T_i \left[\left(1 + \frac{i}{A_{w,s}+c-i}\right)\left(1 - \frac{m}{2A_{w,s}+c}\right)\right]^c + \Delta \\
&= \sum_{i=(m+nd)/2}^{\min\{m,A_{w,s}\}} T_i \left[1 + \frac{i(2A_{w,s}+c) - m(A_{w,s}-i+c) - im}{(A_{w,s}+c-i)(2A_{w,s}+c)}\right]^c + \Delta \\
&= \sum_{i=(m+nd)/2}^{\min\{m,A_{w,s}\}} T_i \left[1 + \frac{A_{w,s}(2i-m) - c(m-i)}{(A_{w,s}+c-i)(2A_{w,s}+c)}\right]^c + \Delta
\end{aligned}
$$

This can be bounded. The assumption that $B_w > B_\ell$ implies that $m = B_w + B_\ell \leq 2i \leq 2B_w$ for all $i \leq B_w$. Thus, $2i - m \geq 1$ and $m - i \geq m - B_w \geq m/2$. So,

$$
p_c > \sum_{i=(m+nd)/2}^{\min\{m,A_{w,s}\}} T_i \left[1 + \frac{A_{w,s} - cm/2}{(A_{w,s}+c-i)(2A_{w,s}+c)}\right]^c + \Delta > \sum_{i=(m+nd)/2}^{\min\{m,A_{w,s}\}} T_i = p_0
$$

□

7.2.2 Optimizing over the parameter $A_{w,s}$

We have shown empirically (but do not prove) that this tail probability, as a function of $A_{w,s}$, has a unique maximum at one of the endpoints. If the empirical result is true, then finding the maximum is trivial; otherwise, it is a trivial one-dimensional optimization problem to compute the unconditional p-value.

7.3 Conditional testing

If the conditional tests are always conducted at significance level α or less, i.e., so that $\mathbb{P}\{\text{Type I error} | B = n, B_w + B_\ell = m\} \leq \alpha$, then the overall

procedure has significance level α or less:

$$\begin{aligned}\mathbb{P}\{\text{Type I error}\} &= \sum_{n=0}^{N}\sum_{m=0}^{N}\mathbb{P}\{\text{Type I error}|B=n, B_w+B_\ell=m\}\mathbb{P}\{B=n, B_w+B_\ell=m\} \\ &\leq \sum_{n=0}^{N}\sum_{m=0}^{N}\alpha\mathbb{P}\{B=n, B_w+B_\ell=m\} = \alpha. \end{aligned} \tag{9}$$

In particular, this implies that our conditional hypergeometric test will have the correct risk limit unconditionally.

References

J.H. Bañuelos and P.B. Stark. Limiting risk by turning manifest phantoms into evil zombies. Technical report, arXiv.org, 2012. URL `http://arxiv.org/abs/1207.3413`. Retrieved 17 July 2012.

J. Benaloh, D. Jones, E. Lazarus, M. Lindeman, and P.B. Stark. SOBA: Secrecy-preserving observable ballot-level audits. In *Proceedings of the 2011 Electronic Voting Technology Workshop / Workshop on Trustworthy Elections (EVT/WOTE '11)*. USENIX, 2011. URL `http://statistics.berkeley.edu/~stark/Preprints/soba11.pdf`.

M. Lindeman, P.B. Stark, and V. Yates. 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, to appear 2012.

Ronald L. Rivest. Bayesian tabulation audits: Explained and extended, January 1, 2018. URL `https://arxiv.org/abs/1801.00528`.

P.B. Stark. Conservative statistical post-election audits. *Ann. Appl. Stat.*, 2:550–581, 2008. URL `http://arxiv.org/abs/0807.4005`.

P.B. Stark. Risk-limiting post-election audits: P-values from common probability inequalities. *IEEE Transactions on Information Forensics and Security*, 4:1005–1014, 2009a.

P.B. Stark. Auditing a collection of races simultaneously. Technical report, arXiv.org, 2009b. URL `http://arxiv.org/abs/0905.1422v1`.

P.B. Stark. 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. URL `http://www.usenix.org/events/evtwote10/tech/full_papers/Stark.pdf`.

Philip B. Stark and Vanessa Teague. Verifiable european elections: Risk-limiting audits for d'hondt and its relatives. *JETS: USENIX Journal of Election Technology and Systems*, 3.1, 2014. URL `https://www.usenix.org/jets/issues/0301/stark`.

A. Wald. Sequential tests of statistical hypotheses. *Ann. Math. Stat.*, 16: 117–186, 1945.