Preparing to Audit Colorado's 2018 Primaries

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

Draft January 11, 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.

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 sample can be focused on 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 contests) 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 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 that $\omega_{w\ell,s} \geq \lambda_s V_{w\ell}$, 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.

Typically, the two audits need to be conducted to smaller risk limits individually than the target overall 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 $\alpha_1, \alpha_2 \in (0, \alpha)$. (More discussion of the choice appears below.) 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 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. That depends in part on what we do if the audit in a given stratum leads to a full manual tally of that stratum.

Here are some scenarios. The outcome is certainly correct if both net overstatements are less than their respective thresholds. For the outcome to be wrong, one or both strata needs to have net overstatement $\omega_{w\ell,s}$ greater than its corresponding threshold $\lambda_s V_{w\ell}$. 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 it's only stratum s, then the chance that

stratum s will be fully hand counted is at least $1 - \alpha_s \geq 1 - \alpha$.

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 $1 - (1 - \alpha_1)(1 - \alpha_2)$, 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? The simplest solution is to require a full hand count of the other stratum, to set the record straight. If this is the rule, then we can take $\alpha_1 = \alpha_2 = \alpha$, and the procedure will have risk limit α.

Alternatively, we might adjust the contest margin for the "known" vote tally in the fully counted stratum (call the stratum t), and continue to audit in the other, but against the entire adjusted margin $V_{w\ell} - A_{w\ell,t} \equiv \lambda'_s V_{w\ell}$, rather than against the share $\lambda_s V_{w\ell}$. Then to reject the null hypothesis in that stratum means that the overall outcome is still correct.

The wrinkle is that treating the votes in the fully counted stratum as known changes the hypothesis being tested in a way that is itself random: whether the original null or a new null is tested depends on what the sample in the other stratum shows. (However, if the hypothesis changes, there's only one value possible for the new λ_s—which depends on the reported margin and the count in the other stratum—but it's unknown until the other stratum count is known.)

The solution is through conditioning. The samples in the two strata are independent. Think of the overall procedure as concluding that the outcome (without a full hand count in both strata) if:

- the original hypotheses are rejected in both strata (neither stratum is fully hand tabulated)
- conditionally on escalating to a full count in stratum s, the threshold λ_t is adjusted in the other stratum, and the hypothesis that the overstatement error in that stratum is greater than the new limit, $\lambda'_t V_{w\ell}$ is rejected.

These two possibilities are mutually exclusive, so the chance that either occurs is the sum of their chances. If the outcome is incorrect, the chance that the audit ends with the first possibility is at most $\max(\alpha_1, \alpha_2)$. (It is the maximum because the overstatement could be concentrated in the stratum audited to the higher risk limit.)

The chance it ends with the second possibility when the outcome is incorrect is more complicated to compute. If the outcome is incorrect, at least one

of the strata has more error than its allocated tolerance. The chance that the audit proceeds to a full count in that stratum is at least $1 - \alpha_s$. What happens in the other stratum? The value of λ'_t is fixed, but unknown before the audit starts. Consider the conditional probability that a sequential test would reject the hypothesis that the margin is less than $\lambda'_s V_{w\ell}$, given that the other stratum goes to a full count. Because the tests in the two strata are independent, that's the same as the unconditional probability. If we are using a sequential test in the remaining stratum t, the chance of ever rejecting the hypothesis is at most α_t if the (new) null is true. The conditioning just "delays" looking at the value of the test statistic for the new null hypothesis; waiting does not increase the overall chance of incorrectly rejecting the null, because the test is legitimately sequential.

Thus an upper bound on the overall risk limit is

$$\alpha_{\text{bound}} = \max(\alpha_1, \alpha_2) + \max((1 - \alpha_1)\alpha_2, \alpha_1(1 - \alpha_2)). \tag{1}$$

This can be simplified. Consider the second term:

$$\max((1 - \alpha_1)\alpha_2, \alpha_1(1 - \alpha_2)) = \max(\alpha_2, \alpha_1) - \alpha_1\alpha_2. \tag{2}$$

Hence $\alpha_{\text{bound}} = 2\max(\alpha_1, \alpha_2) - \alpha_1\alpha_2$. If we set the risk limit to be the same in both strata ($\alpha_1 = \alpha_2 = \alpha_r$), then the overall risk limit is not larger than $2\alpha_r - \alpha_r^2$. To have a risk limit of 0.05, then, we could take

$$\alpha_r = \frac{2 \pm \sqrt{4 - 4 \times 0.05}}{2} = 1 \pm \sqrt{.95}.$$

The relevant root is $1 - \sqrt{.95} = 0.0253$. Thus if we audit each stratum to a risk limit of 0.0253, the overall risk limit for the audit is not larger than 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, 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 $\omega_{w\ell,1} \geq \lambda_1 V_{w\ell}$ and $\omega_{w\ell,2} \geq \lambda_2 V_{w\ell}$, 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 $p_s(\lambda)$ using sequential procedures is 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 approach (Stark, 2009, 2010) trivially 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).[1]

6.1 Notation

- N ballots were cast in all. (The contest might not appear on all N ballots)
- $\mathcal{W}$: the set of reported winners of the contest
- $\mathcal{L}$: the set of reported losers of the contest
- n_p: number of ballots in batch 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 in batch p, then $a_{pi} = 0$.

[1] 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.

- $V_{w\ell} \equiv \sum_{p=1}^{N}(v_{pw} - v_{p\ell})$: Reported margin in the stratum of reported winner $w \in \mathcal{W}$ over reported loser $\ell \in \mathcal{L}$.
- V: smallest reported margin in the stratum among all C contests audited using the same sample: $V \equiv \min_{w\in\mathcal{W},\ell\in\mathcal{L}} V_{w\ell}$
- $\mu = V/N$: the "diluted stratum margin," the margin in the stratum in votes divided by the total number of ballots in the stratum
- $A_{w\ell} \equiv \sum_{p=1}^{N}(a_{pw} - a_{p\ell})$: actual margin in the stratum of reported winner $w \in \mathcal{W}$ over reported loser $\ell \in \mathcal{L}$

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

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

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

We won't test that inequality directly. Instead, 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 (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 a fraction λ of the overall margin $V_{w\ell}$, that is, we want to establish

$$\sum_{p=1}^{N}(v_{pw} - a_{pw} - v_{p\ell} + a_{p\ell})/V_{w\ell} < \lambda.$$

Now the maximum (over all winner, loser pairs) of sums is not larger than the sum of maxima; that is,

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

Hence, if

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

all the reported outcomes must be correct. 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 the reported outcomes of all the contests must be correct if

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

To see that a different way, suppose that the outcome of one or more contests is wrong. Then there is some contest c and some reported (winner, loser) pair $w \in \mathcal{W}, \ell \in \mathcal{L}$ for which

$$0 \geq A_{w\ell} = V_{w\ell} - (V_{w\ell} - A_{w\ell}) = V_{w\ell} - \sum_{p=1}^{N} (v_{pw} - a_{pw} - v_{p\ell} + a_{p\ell}),$$

i.e.,

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

Diving both sides by $V_{w\ell}$ gives

$$\sum_{p=1}^{N} \frac{v_{pw} - a_{pw} - v_{p\ell} + a_{p\ell}}{V_{w\ell}} \geq 1.$$

But

$$\begin{aligned} \frac{v_{pw} - a_{pw} - v_{p\ell} + a_{p\ell}}{V_{w\ell}} &\leq \max_{w \in \mathcal{W}, \ell \in \mathcal{L}} \frac{v_{pw} - a_{pw} - v_{p\ell} + a_{p\ell}}{V_{w\ell}} \\ &\leq \max_{w \in \mathcal{W}, \ell \in \mathcal{L}} \frac{v_{pw} - a_{pw} - v_{p\ell} + a_{p\ell}}{V_{w\ell}} = e_p, \end{aligned}$$

so if the outcome is wrong, $E = \sum_p e_p \geq 1$. Thus a risk-limiting audit can rely on testing whether $E \geq 1$. If the hypothesis $E \geq 1$ can be rejected at significance level α, we can conclude that all the reported outcomes are correct.

Testing whether $E \geq 1$ would require a very large sample if we knew nothing at all about e_p without auditing ballot p: a single large value of e_p could make E arbitrarily large. Fortunately, there is an *a priori* upper

bound for e_p. Whatever the reported votes v_{pi} are on ballot p, we can find the potential values of the actual votes a_{pi} that would make the error e_p largest, because a_{pi} can only be zero or one:

$$\frac{v_{pw} - a_{pw} - v_{p\ell} + a_{p\ell}}{V_{w\ell}} \leq \frac{v_{pw} - 0 - v_{p\ell} + 1}{V_{w\ell}}.$$

Hence,

$$e_p \leq \max_{w \in \mathcal{W}, \ell \in \mathcal{L}} \frac{v_{pw} - v_{p\ell} + 1}{V_{w\ell}} \equiv \tilde{u}_p.$$

Knowing that $e_p \leq \tilde{u}_p$ might let us conclude reliably that $E < 1$ by examining only a small fraction of the ballots—depending on the values $\{\tilde{u}_p\}_{p=1}^N$ and on the values of $\{e_p\}$ for the audited ballots.

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

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

$$\mathbb{E}T_1 = \sum_{p=1}^N \tilde{u}_p/\tilde{U} t_p = \frac{1}{\tilde{U}} \sum_{p=1}^N \tilde{u}_p \frac{e_p}{\tilde{u}_p} = \frac{1}{\tilde{U}} \sum_{p=1}^N e_p = E/\tilde{U}.$$

Thus $E = \tilde{U}\mathbb{E}T_1$.

So, if we have strong evidence that $\mathbb{E}T_1 < 1/\tilde{U}$, we have strong evidence that $E < 1$.

This approach can be simplified even further by noting that $\tilde{u}_p$ has a simple upper bound that does not depend on any v_{pi}. At worst, the CVR for ballot p shows a vote for the "least-winning" apparent winner of the contest with the smallest margin, but a hand interpretation shows a vote for the runner-up in that contest. Since $V_{w\ell} \geq V$ and $0 \leq v_{pi} \leq 1$,

$$\tilde{u}_p = \max_{w \in \mathcal{W}, \ell \in \mathcal{L}} \frac{v_{pw} - v_{p\ell} + 1}{V_{w\ell}} \leq \max_{w \in \mathcal{W}, \ell \in \mathcal{L}} \frac{1 - 0 + 1}{V_{w\ell}} \leq \frac{2}{V}.$$

Thus, if we define $u_p \equiv 2/V$ and sample ballots at random with probability proportional to u_p, in fact we will sample ballots with *equal* probability.

Define

$$U \equiv \sum_{p=1}^{N} \frac{2}{V} = \frac{2N}{V} = 2/\mu$$

and re-define $t_p \equiv e_p/u_p$ (rather than $e_p/\tilde{u}_p$); let T_j be the value of t_p for the ballot selected at random in the jth draw, as before. Then still $\{T_j\}_{j=1}^n$ are IID, $\mathbb{P}\{T_j \leq 1\} = 1$, and

$$\mathbb{E}T_1 = \sum_{p=1}^{N} \frac{u_p}{U} t_p = \frac{1}{U} \sum_{p=1}^{N} u_p \frac{e_p}{u_p} = \frac{1}{U} \sum_{p=1}^{N} e_p = E/U = \frac{\mu}{2} E,$$

i.e.,

$$E = \frac{2}{\mu} \mathbb{E}T_1.$$

So, if we have evidence that $\mathbb{E}T_1 < \mu/2 = 1/U$, we have evidence that $E < 1$.

7 Ballot-polling audits of a tolerable overstatement in votes

We propose a conservative test as follows. Suppose that the outcome is incorrect (this is the null hypothesis). Then there is error in the tabulation of votes in one or both strata such that, net, the tabulation errors in the two strata combine to favor the reported winner(s) by at least the reported margin(s).

We consider a single contest and assume that the social choice function is plurality (first-past-the-post) or vote for k of C candidates (e.g., a city council election). Those restrictions can be relaxed somewhat; in particular, slight modifications allow the approach to work with majority, super-majority, approval, and Borda.

There are two collections of ballots, $\mathcal{B}_1$, which contains N_1 ballots, and $\mathcal{B}_2$, which contains N_2 ballots. The total number of ballots is $N \equiv N_1 + N_2$. We also refer to the two collections as *strata*.

There are C candidates. The set $\mathcal{W}$ are the candidates reported to have won; the set $\mathcal{L}$ are those reported to have lost. For the social choice function under consideration, candidate i belongs in $\mathcal{W}$ if and only if she is reported to have received more votes than every candidate in $\mathcal{L}$. We audit by looking at (winner, loser) pairs. The audit stops when there is sufficiently strong

statistical evidence that each winner $w \in \mathcal{W}$ in fact received more votes than each loser $\ell \in \mathcal{L}$.

Let π_s^c denote the fraction of ballots in stratum s that show a vote for candidate c. The total number of votes for candidate c is $N_1\pi_1^c + N_2\pi_2^c$. Candidate $w \in \mathcal{W}$ really beat candidate $\ell \in \mathcal{L}$ if and only if

$$\sum_{s=1}^{2} N_s \pi_s^w > \sum_{s=1}^{2} N_s \pi_s^\ell, \tag{4}$$

or, equivalently, iff

$$\sum_{s=1}^{2} N_s (\pi_s^w - \pi_s^\ell) > 0. \tag{5}$$

To simplify the notation, for any $\pi \equiv (\pi_1, \pi_2)$, define

$$\bar{\pi} \equiv \frac{1}{N} \sum_{s=1}^{2} N_s \pi_s, \tag{6}$$

and for $p \equiv (p_1, p_2)$, define $\bar{p}$ analogously. Then candidate $w \in \mathcal{W}$ really beat candidate $\ell \in \mathcal{L}$ if and only if $\bar{\pi}^w > \bar{\pi}^\ell$.

There are a variety of ways to proceed; we sketch two here and visit others below.

7.1 Ballot polling for the weighted sum of proportions

The first approach is to test the hypothesis $\bar{\pi}^w \leq \bar{\pi}^\ell$ for every (w, ℓ) pair. Unstratified ballot-polling audits are based on the conditional probability that a randomly selected ballot shows a vote for w, given that it shows a vote for w or for ℓ (**??**). If that conditional probability is larger than 1/2, w really received more votes than ℓ.

Because the fraction of ballots that do not show a vote for either w or ℓ can differ between strata, the conditional probabilities in the two strata separately are not enough to tell whether w really beat ℓ: one also needs to know the total number of ballots with votes for either w or for ℓ in each stratum. Those numbers could be estimated from the same samples used to estimate the conditional probabilities, but the uncertainty of those estimates would need to be taken into account, and because of the dependence between the estimated fraction of ballots for w among those for w or ℓ and the estimated number of ballots for either w or ℓ, that is difficult to do in a sharp way.

In this section, to illustrate the germ of the approach, we assume for simplicity that the contest is a plurality contest with $C = 2$ candidates and that every ballot has a valid vote either for w or for ℓ, but not for both, so there are N valid votes in all. This assumption rules out undervotes and invalid ballots, which need to be taken into account in real elections. We show below in section **??** how that can be done.

The reported winner purportedly received a fraction p_s^w of the votes in stratum s. Thus $\bar{p}^w > \bar{p}^\ell$ and $N_1 p_1^w + N_2 p_2^w > N/2$. In stratum s, we use Wald's sequential probability ratio test (SPRT) to test the hypothesis that $\pi_s^w = p_s$ against the alternative hypothesis that the reported results are correct, that is, that $\pi_s^w = p_s^w$.

7.2 Ballot polling without replacement, no undervotes or invalid ballots

The mathematics in this section essentially recapitulates **?**, pp43–44 in different notation, with some simplifications.

There is a population of N items. Item j has "value" $a_j \in \{0, 1\}$.

We want to test the hypothesis H_0 that $\frac{1}{N}\sum_j a_j = 1/2$ against the alternative hypothesis H_1 that $\frac{1}{N}\sum_j a_j = \gamma$, for some fixed $\gamma > 1/2$.

We will draw items sequentially, without replacement, such that the chance that item i is selected in draw j, assuming it has not been selected already, is $1/(N - j + 1)$. Let $\mathcal{B}_{j-1}$ be the indices of the items selected up to and including the $j-1$st draw, and $\mathcal{B}_0 \equiv \emptyset$.

Let $\mathbb{B}_j$ denote the index of the item selected at random in the jth draw.

The chance that the first draw $\mathbb{B}_1$ gives an item with value 1, i.e., $\Pr\{a_{\mathbb{B}_1} = 1\}$, is $\frac{1}{N}\sum_b a_b$. Under H_0, this chance is $p_{01} = 1/2$; under H_1, this chance is $p_{11} = \gamma$.

Given the values of $\{a_{\mathbb{B}_k}\}_{k=1}^i$, the conditional probability that the ith draw gives an item with value 1 is

$$\Pr\{a_{\mathbb{B}_i} = 1 | \mathcal{B}_{i-1}\} = \frac{\sum_{b \notin \mathcal{B}_{i-1}} a_b}{N - i + 1}.$$

Under H_0, this chance is

$$p_{0i} = \frac{N/2 - \sum_{b \in \mathcal{B}_{i-1}} a_b}{N - i + 1}.$$

Under H_1, this chance is

$$p_{1i} = \frac{N\gamma - \sum_{b \in \mathcal{B}_{i-1}} a_b}{N - i + 1}.$$

Let X_i be the indicator of the event that the ith draw gives an item with value 1, i.e., the indicator of the event $a_{\mathbb{B}_i} = 1$. The likelihood ratio for a given sequence $\{X_k\}_{k=1}^i$ is

$$\text{LR} = \frac{\prod_{k=1}^i p_{1k}^{X_k}(1-p_{1k})^{1-X_k}}{\prod_{k=1}^i p_{0k}^{X_k}(1-p_{0k})^{1-X_k}}.$$

This can be simplified. Note that p_{0k} and p_{1k} have the same denominator, $N-k+1$, and that the numerators share a term. Define $A(k) \equiv \sum_{b \in \mathcal{B}_{i-1}} a_b$. Then

$$\text{LR} = \prod_{k=1}^{i} \left(\frac{N/2 - A(k)}{N\gamma - A(k)} \right)^{X_k} \left(\frac{N - N/2 - (k-1-A(k))}{N - N\gamma - (k-1-A(k))} \right)^{1-X_k}$$

$$= \prod_{k=1}^{i} \left(\frac{N/2 - A(k)}{N\gamma - A(k)} \right)^{X_k} \left(\frac{N/2 - k + 1 + A(k)}{N(1-\gamma) - k + 1 + A(k)} \right)^{1-X_k}$$

If H_0 is true, the chance that LR is ever greater than $1/\alpha$ is at most α.

7.3 Ballot polling without replacement, accounting for undervotes and invalid votes

Suppose that the population contains N_w ballots with a valid vote for w but not ℓ, N_ℓ ballots with a valid vote for ℓ but not w, and N_u ballots with votes for both w and ℓ or for neither w nor ℓ. Then $N = N_w + N_\ell + N_u$. Let $N_{w\ell} \equiv N_w + N_\ell$ be the number of ballots in the population with a valid vote for w or ℓ but not both.

Suppose we have made k draws from the pool of ballots and have observed $n_w(k)$ votes for the reported winner, $n_\ell(k)$ votes for the reported loser, and $n_U(k)$ votes for other candidates or invalid votes. $n_w(k) + n_\ell(k) + n_U(k) = k$.

The probability that the $k+1$st ballot is for the reported winner, conditional on all the previously observed ballots, is

$$\mathbb{P}(X_{k+1} = w \mid n_w(k), n_\ell(k), n_U(k)) = \frac{\#\text{ ballots for w not yet drawn}}{N-k}$$

Under the null, the number of votes for candidate w remaining is $\frac{N_{w\ell}}{2} - n_w(k)$. Under the alternative, the number is $\frac{N_{w\ell}}{2} + \mu - n_w(k)$ for $\mu > 0$. Thus, the likelihood ratio for observing a vote for candidate w, conditional on the previous ballots, is

$$LR_{k+1}(w) = \frac{\frac{N_{w\ell}}{2} - n_w(k)}{\frac{N_{w\ell}}{2} + \mu - n_w(k)}.$$

By the same reasoning, the likelihood ratio for observing a vote for the reported loser, conditional on the previous ballots, is

$$LR_{k+1}(\ell) = \frac{\frac{N_{w\ell}}{2} - n_\ell(k)}{\frac{N_{w\ell}}{2} - \mu - n_\ell(k)}.$$

The fraction of invalid ballots in the population is the same under the null and alternative, therefore the likelihood ratio for drawing an invalid ballot is always equal to 1. Therefore, at step $k+1$, the overall likelihood ratio is

$$LR_{k+1} = \prod_{i=1}^{k+1} LR_i(w)^{\mathbb{I}(X_i = w)} LR_i(\ell)^{\mathbb{I}(X_i = \ell)}.$$

More compactly, the likelihood ratio of the null to the alternative (assuming $N_{w\ell}$ is even) is

$$\text{LR} = \frac{\binom{N_{w\ell}/2}{n_w}\binom{N_{w\ell}/2}{n_\ell}}{\binom{N_{w\ell}/2+\mu}{n_w}\binom{N_{w\ell}/2-\mu}{n_\ell}}. \tag{7}$$

The likelihood ratio is a function of the nuisance parameter $N_{w\ell}$. The test is conservative when the likelihood ratio is maximized, which occurs when $N_{w\ell}$ is as large as possible. Given $n_U(k)$, we know that $N_{w\ell} \leq N - n_U(k)$. This maximal value changes each time we observe a ballot that is not for the winner or loser.

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`.

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. Auditing a collection of races simultaneously. Technical report, arXiv.org, 2009. 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`.