Comments re statistics of auditing the 2018 Colorado elections

by Mark Lindeman, Ronald L. Rivest, Philip B. Stark and Neal McBurnett
(with comments, edits, and suggestions by others)
January 3, 2018

Overview

This document provides comments related to the statistical aspects of auditing the June 2018 primary election and the November 2018 general election in Colorado, based on experience gained in the audit of the November 2017 Colorado general election, and the expected differences in election characteristics and audit objectives between the November 2017 election and the forthcoming elections.

This document is for submission to the CO Secretary of State's office (before noon MST January 3, 2018).

This document provides comments on the statistical aspects only, and not on the many other aspects of auditing, such as anonymity, transparency, or the management of ballot review (among many issues).

We do not pretend that this note covers all of the relevant statistical issues; we just try to hit what seem to be the important issues. (We'd of course be happy to advise on any other issues that seem problematic.)

Context: Audit of November 2017 CO general election

As required by CO state law, some of the November 2017 election results in Colorado were audited. We make the following few remarks for context.

1. Overall, the audit went well and achieved its stated objectives, which were limited. (For example, not all contests were audited, and some multi-jurisdiction contests were audited as if they were single-jurisdiction contests.) The published results so far are at CO Audit Center.
2. Software (RLATool) supporting the audit was provided by Free and Fair. (F&F web site for Colorado RLA)
3. The audit was efficient in that it involved examining relatively few ballots manually; at least one local election official found it to be less burdensome than the previous statutory audit.
4. The November 2017 audit should be viewed as a "first step" towards more complete audits as envisaged by CO law.

Outstanding issues from current implementation:

1. **Verifying that CVRs sum to the reported contest results**: Our understanding is that at this time, the Secretary of State has the ability to compare reported contest results with the CVRs it receives, but there is no publicly observable means to verify that this comparison has been done correctly.
2. **Checking hash values.** We understand that the state checks the hash values provided by county election officials against the CVRs, but that presently there is no procedure for county election officials or the public to confirm that the hash of the CVRs being used by the state RLA tool indeed matches the original hash value.

General remarks:

It seems best to "take the long view" and not expend too much effort on issues that are known to be short-term. In particular, issues relating to the fact that not all counties provide ballot-level CVRs perhaps can be handled via some expedient means this year, as they should be resolved once counties finishing upgrading their equipment.

To audit the primaries, there are several potential strategies. We recommend strategies that build functionality that will be required eventually, even after all counties can produce CVRs. For instance, building tools that make it possible to audit smaller contests seems best to us.[1]

Planning issues:

1. It needs to be decided whether the software should treat primaries for various parties as "separate elections" or just as separate contests within a larger single election. (Recommendation: the latter, unless the ballots can be physically sorted by party as part of the tabulation, and the state wants to run two instances of the RLA tool. But there is no reason the various contests in the several primaries cannot just be considered as separate contests in one overall election. Care needs to be taken when computing diluted margins to use the right denominator, of course. The denominator is the number of ballots in the collection from which the sample is drawn.)
2. It needs to be decided which contests are eligible to be audited. (We offer no specific recommendation, but it would be good if every contest, independent of scope or margin, had some chance of being selected as an audited contest. Moreover, to avoid the possibility that the selection of contests for audit would be perceived as partisan, it would be best if the selection of contests were random)

[1] An alternative for June would be to audit both major parties' primaries statewide, computing the sample sizes from the overall diluted margins. Although one party's contest may be audited more heavily than the risk limit would require, the workload impact should be manageable when distributed statewide, and will contribute to opportunistic auditing of other contests.

3. To audit contests that cross jurisdictional boundaries, contests and candidates need to be **named uniformly** across the state. (Recommendation: the SoS should develop and endorse naming conventions for contests and candidates.)

Issues relating to input provided to RLAtool and structuring of RLAtool:

1. RLATool needs to be restructured to have contests (not votes within counties) be a first-rate entity.
    a. For that to work, SoS needs to establish **uniform naming conventions** for contests and candidates (as noted above).
    b. RLATool has to have the notion of **sampling from contests**. Risk limits, diluted margins, and other quantities need to be defined as per-contest quantities, not per-county quantities. The risk needs to be computed for entire contests, not for the portion of a contest in a county.
    c. RLATool will also have to aggregate contest CVRs to get contest totals, then check that those contest totals agree with the sums of the subtotals reported separately by counties.
    d. RLATool will have to calculate diluted margins for contests, taking into account the reported results and the total number of ballots from which the sample will be drawn.
2. Get information from SCORE on the number of returned/voted ballots that contain each contest.
    a. RLATool has to be modified to have a UI to ingest those data from SCORE; ideally, SCORE would export things in a machine-readable format that RLATool can read.
    b. RLATool has to know what CVRs contain what contests; that can be inferred from the non-null fields in the CVRs. (Note, however, that the fields present in a given CVR may affect the anonymity of the corresponding ballot.)
3. RLATool has to be able to do sanity checks that there aren't more CVRs with any given contest than there "should" be, according to SCORE. (We need to think about what to do if there are more CVRs than there should be, which may happen.)
4. RLATool needs to be able to sample from CVRs that are supposed to contain a contest (across counties), augmented by "zombie" CVRs if the number of CVRs is less than SCORE says should be there. Note that sampling in a non-uniform manner increases the complexity of calculating risk levels, so care should be taken in interpreting the evidence for other contests. Risk calculations in RLATool should automatically substitute overstatements (in the case of comparison audits) or the appropriate equivalent (e.g., a vote for every reported loser, in the case of ballot-polling audits) whenever a "zombie" is selected. (This last item is already part of "ballot not found'' reports from the Audit Board.)

Issue: contests with small diluted margins in a county

If a contest under audit has a small diluted margin in a county, it is hard to efficiently sample ballots containing that contest by merely sampling ballots at random from those cast in that county. Although this works as a matter of statistics, it is not efficient. Are there better approaches?

One approach is to use CVR data to identify the location of ballots having the relevant contest. Provided appropriate cross-checks are in place to ensure that there are not more CVRs containing a given contest than there should be, a procedure that relies on the CVRs can work, **even though the CVR data can be trusted**. (If the CVRs were entirely trustworthy, there would be no need ro an audit!)

Can we somehow validate the CVR information regarding the ballot styles of individual ballots?

We suggest that **using SCORE data** may suffice for providing such validation. That is, we assume that SCORE can provide trustworthy information regarding the number of ballots in each county that contain a given contest (including ballots that are undervoted in those contests). If the number provided by SCORE turns out to be greater than or equal to the number of CVRs that have the contest, then the CVR data can be used as a perhaps-reliable guide to the location of all the ballots having the given contest. The following caveats apply:

1. The SCORE data is likely to vary slightly from the CVR counts, for a variety of reasons. The statistics needs to handle this.
2. This approach **does not work** in "ballot-polling counties," where there are no ballot-level CVRs but only aggregate tallies. (This is a serious, if short-term, problem. It may be particularly relevant for minor party primary elections.)
3. If the CVR count is less than the SCORE count, then "zombies" may be used to make up the difference. (A "zombie" is a fictitious ballot that is interpreted as containing a valid vote for every non-winning candidate in the contest.) See Limiting Risk by Turning Manifest Phantoms into Evil Zombies
4. If the CVR count is more than the SCORE count, some procedure is needed to correct this situation. (Details omitted here, and not so clear.)
5. It is unclear what level of trust SCORE data deserves. While it is arguably independent of the CVR data, it is nonetheless software-produced and subject to adversarial attack.

We observe that using SCORE data also has potential anonymity issues (even though that is not the focus for this note).

Issue: multi-county contests "of mixed type"

How should we perform risk measurement for a multi-county contest (including statewide contests) in which some one or more counties have CVRs for each ballot, and one or more

counties do not have such CVRs? (We call such counties “ballot-polling counties” and “comparison counties” respectively.)

1. There are a number of **”frequentist approaches” that** may be used. Some are based on having two strata: one for all the ballot-polling counties and one for the comparison counties. Some are based on ballot-level comparison audits, with special treatment of ballots selected from counties with legacy systems (which are incapable of producing CVRs that can be matched to physical ballots). The simplest frequentist approach is to use a comparison audit, but to perform the comparison at the level of individual ballots in CVR counties and at the level of reporting batches in counties with legacy systems. This would require modifying RLATool in several ways (to ingest batch-level totals, to calculate risk in a slightly more general way, and to allow unequal sampling probabilities for batches of different sizes), but the approach has been implemented and tested in practice. Philip Stark is working on a writeup of such approaches.

2. A ``**Bayesian approach”** may be used.
   See Bayesian Tabulation Audits: Explained and Extended or
   (arXiv posting of Bayesian Tabulation Audits)
   The Bayesian approach has been implemented and tested in the github repository
   Audit-Lab

Issue: Frequentist or Bayesian methods

While frequentist approaches fit the standard definition of a risk-limiting audit, Bayesian approaches (which some view as more “heuristic” in character, and others view as more mainstream, given their success in other fields) offer a contrasting set of opportunities and considerations. They can be used concurrently with frequentist approaches. The technical debate over their use hinges on whether they are risk-limiting, and if so, what their risk limit is (They are certainly “risk-limiting” in a Bayesian sense.)

Issue: Sampling

Sampling is at the heart of any statistical audit (e.g. a risk-limiting audit).

One dimension to consider is whether the sampling will be stratified or not. The statistical calculations are simpler (and sample sizes are generally smaller) if sampling is not stratified, but stratification offers some logistical advantages. For instance, it can “de-couple” the escalation of the audit in different counties.
Multi-county contests have natural strata, since each county has its own collection of paper ballots.. While it is possible to consider all of the relevant counties together as a single unstratified collection, one may gather logistical efficiency by not doing so, generally at the cost of requiring larger samples. Methods for combining ballot polling with comparison audits

currently require stratification, although it is possible to use as few as two strata by combining all of the ballot-polling counties into one stratum and all of the comparison-audit counties into another stratum.

There are other ways in which stratification may be employed. For example, it is possible to stratify a sample according to the reported vote in the CVR. That is, just sample ballots which appear to be votes for Jones (one stratum), or not for Jones (another stratum). Such refinements are possible but probably not of interest for the forthcoming CO audits, since they make opportunistic auditing very ineffective.

One must be careful to understand how a proposed stratification interacts with multi-contest auditing. For instance, suppose contest X is on all the ballots in counties A and B, and contest Y is on all the ballots in counties B and C. Suppose further that contest X has a small margin, while contest Y has a large margin. Then one would expect to sample ballots from counties A and B at a higher rate than the ballots in county C. This takes care of contest X well, but leaves contest Y in an interesting state, where county B ballots are sampled heavily and county C ballots are lightly sampled. The risk-measurement procedure may need to take into account differing sampling rates in different counties. (There are frequentist and Bayesian approaches for doing this, with varying levels of efficiency; details omitted here.)

A convenient way of organizing sampling across the state is to give each ballot a unique “ballot key” derived in a pseudo-random manner from the random seed and the ballot location information. We might think of the ballot key as a large unique integer. Then Audit Central can specify that auditing targets for each contest by specifying which collections of paper ballots are relevant, and for each relevant such collection, the number of ballots desired. The understanding is that if a collection must report on N ballots, then it is the set of N ballots in that collection with the lowest ballot keys that are desired. Such a process maximizes the “overlap” between audits of different contests, as the ballot-keys are the same (independent of contest), which can yield substantial efficiencies.