Delayed Stratification for Timely Risk-Limiting Audits

Philip B. Stark

2 April 2019 (revised 22 October 2022)

Abstract

Some jurisdictions, including California, continue to accept ballots after election day until shortly before the results are certified. This compresses the amount of time available to conduct a risk-limiting audit, if the audit cannot start before all validly cast ballots have been tabulated. One way to allow audits to start sooner is to use stratified sampling, drawing independent samples from groups of ballots tabulated in different time periods. Stratified sampling generally increases the number of ballots that must be inspected to attain a given risk limit. This paper shows that if there is an upper bound on the number of validly cast ballots, audits can start anytime, without stratifying the sample, using existing sequential auditing methods. The approach gives local election officials flexibility to sample tabulated ballots before every ballot has been received, had its provisional status resolved and/or signature cured, and been tabulated (if it is determined to have been validly cast).

Assumptions

We assume that the audit will use a *sequential* testing method, so that the test gives a valid P-value for all sample sizes. At time t, N_t ballots have been tabulated, and there is an upper bound $\bar{N}_u$ on the total number of validly cast ballots that have not yet been tabulated.

Procedure

1. Pick an initial sample size n. (Methods to select the sample size are discussed below.)

2. Select n indices from 1 to $N_t + \bar{N}_u$. Let n_t denote the number of sampled indices corresponding to already tabulated ballots (sample indices between 1 and N_t).

3. Audit the sample of n_t already-tabulated ballots.

4. Tabulate the remaining ballots.

5. Suppose there were $N_u \leq \bar{N}_u$ such ballots. Let n_u denote the number of indices in the original sample that are greater than N_t but less than or equal to N_u.

   - Create "pseudo-data" for those n_u ballots, treating each in the worst-case way (so as to maximize the P-value).[1]

   - combine the n_t actual data from the audited ballots with the n_u pessimistic pseudo-data for the as-yet unaudited ballots.

     - If the risk for that combination data and pseudo-data is not larger than the risk limit, the audit can stop.

     - otherwise:

       * create a ballot manifest for the (newly tabulated) N_u ballots

       * draw a (new) random sample of size n_u from the manifest.[2] Those n_u ballots, together with the n_t, comprise a simple random sample of size $n_t + n_u$ from the total population of $N_t + N_u$ validly cast ballots.

       * audit the new sample of n_u ballots in any order, any number at a time.

         · after a ballot is audited, replace one of the pessimistic pseudo-data with the audit data for that ballot

         · if the risk becomes less than the risk limit at any point, the audit can

---

[1]For instance, in a ballot comparison audit, assume pessimistically that there was a 2-vote overstatement for each of those ballots; for a ballot polling audit, assume that every ballot had a valid vote for every reported loser.

[2]This is important for security: if the original selections had been used, an adversary could contrive to produce a ballot manifest that "forced" pre-determined ballots to be sampled.

stop

* if all n_u ballots have been audited and the risk is still greater than the risk limit, augment the overall sample of $n_t + n_u$ ballots by sampling uniformly at random from the $N_t + N_u$ tabulated ballots (that is, continue to audit as if the ballots had never been divided into two groups).

Theorem. The risk limit for the resulting audit is conservative.

Proof. The measured risk at each stage is never smaller than it would be if the ballots for which there were "pseudo-data" had instead actually been inspected.

Setting the initial sample size n

There are a number of strategies for picking the initial sample size, depending on how much the LEO wants to avoid having to sample from the $N_u \leq \bar{N}_u$ ballots that had not been tabulated when the audit commenced (at the cost of auditing more ballots in all). For instance, one could assume that the reported results for the N_t tabulated ballots is representative of what the remaining N_u ballots will show, to estimate what the overall reported vote shares will be. If the LEO wants to completely avoid sampling from the as-yet-untabulated ballots, set the initial sample size so that the expected risk will be below the risk limit even when pessimistic pseudo-data are used in place of the $\leq n - n_t$ ballots drawn from the untabulated ballots. That might not be possible (i.e., no sample size would be large enough), depending on the relative sizes of N_t and $\bar{N}_u$, and the hypothesized vote shares. If the LEO simply wants to get started auditing before all votes have been tabulated, s/he could make less pessimistic assumptions in setting the initial value of n.

Numerical Example

We imagine auditing a single plurality contest with candidates Alice, Bob, and Carol. We will use a ballot comparison audit.

Suppose that 95,000 ballots have been tabulated and at most 5,000 remain untabulated. (The untabulated might include unresolved provisional ballots, ballots with signature problems awaiting possible cure, late-arriving vote-by-mail ballots, etc.) Among the tabulated ballots, Alice is reported to have received 50,000 votes, Bob is reported to have received 40,000 votes, Carol is reported to have received 3,000 votes, and there are reported to be 2,000 undervotes and invalid ballots: Alice is the reported winner.

The smallest *diluted margin* is $(45000 - 40000)/95000 = 0.0526$. The 95,000 tabulated ballots represent 95% of the 100,000 possible ballots. If we sample at random from the 100,000, we should expect about 5% of the ballots to be among the not-yet-tabulated. If 5% of the ballots cast in the election showed a 2-vote overstatement, that could alter the outcome. Hence, it would not surprising if the audit has to examine some ballots selected from the group that hadn't been tabulated, no matter how large the initial sample size is.