White Paper: Election Audit Strategy

PART 2: Toward a better understanding of Statistical Election Audits

Ray Lutz
Citizens Oversight, Inc.

V0.1 2018-10-22
V0.2 2018-12-05 -- Community Draft for Comment
V0.3 2018-12-31 -- Revised Workload Section
V0.4 2019-01-28 -- Consistency edits

Table of Contents

1.0 Introduction

This is the second part regarding Election Audit Strategy. Please see Part 1 for background information.

Risk Limiting Audits (RLAs) have been proposed as a way to limit the risk that the election tabulation could be wrong by reviewing a limited statistical sample of ballots and either comparing them to the reported result (referred to as a "Polling RLA" or PRLA) or comparing ballot by ballot to the original reported results, which must also be broken down to the ballot level, and sufficient tracking must be available to be able to compare the physical ballot to the official computer report (referred to as a "Ballot Comparison RLA" or BCRLA). This analysis will focus on getting a better understanding of how such audits would perform, and then proposing an approach that balances all the concerns.

Indeed, we have a number of competing proposals, each with very respected authorities in the subject area promoting the viability of that approach. At the same time, people are implementing these proposals, and almost no one really understanding the true nature of the results nor how these competing proposals can be compared. (The situation reminds me of that faced by Galileo in the mid 1500s when the authorities of the day were adhering to Aristotelian physics, which was developed at a time when they believed that the true nature of the world could be determined simply through human thought, while no one wanted to actually stoop to conducting experiments (as that sort of thing was considered below the ruling class). Galileo finally broke through this tradition and used simple experiments to provide proof that the laws of physics for 1800 years were wrong. Perhaps one of the oft-noted examples is when he proved by supposedly dropping objects from the tower of Pisa that objects that were larger fell just as fast as a small object, as long as the small object was not so light that air resistance would slow it. This corrected the model as promoted by Aristotle from about 350 BCE, and introduced the tradition of using experimentation as a basis for determining the truth of assertions made about the world. Then, Galileo went on to question the earth-centered model of the universe for which he nearly died and spent the rest of his life in house arrest.)

Luckily enough for us, election audits are actually quite easy to model in software, and so we can run thousands of audits on various elections and compare the results (and hopefully without fearing house arrest).

Hopefully, this will serve as a means of discussion, so that any misconceptions or misunderstandings I may have can be resolved as well as providing what I believe will be a much better understanding of what we are faced with.

Leading proposals for audits are compared, including "Super-Simple Simultaneous Single-Ballot Risk-Limiting Audits," by Philip B. Stark[1] (S4RLA), (also described in "A Gentle Introduction to Risk-limiting Audits," by Mark Lindeman and Philip B. Stark[2] (Gentle)), "Bayesian Tabulation Audits Explained and Extended," by Ron Rivest[3], and for polling audits: "BRAVO: Ballot-polling Risk-

1 https://www.usenix.org/legacy/events/evtwote10/tech/full_papers/Stark.pdf
2 https://www.stat.berkeley.edu/~stark/Preprints/gentle12.pdf
3 http://people.csail.mit.edu/rivest/pubs/Riv18a.pdf

limiting Audits to Verify Outcomes", by Lindeman, Stark and Yates[4] (BRAVO), and "Clip Audit," by Ron Rivest[5].

The approach taken in this paper is to try each of the audits thousands of times on modeled elections and compare how each performs. As a result, we will propose an alternative and simplified conservative approach which can be easily understood and will provide a chart for each election so the officials and the public can see how risky the audit is.

The proposed method is called "Balanced Risk Audit with Workload Limitation" or BRAWL, for reference, and a method based on this approach is proposed for both ballot comparison and ballot polling audits. Although a system of simple tables can be used to conduct a single audit, we find that frequently, races are audited in parallel and the sample sizes are based on a closer race which is on the same ballot. In those cases, producing the chart showing the location of the audit with respect to thousands of simulated audits provides a fantastic visual aid. This paper will show how this can be used in an actual pilot risk limiting audit in Orange County from June, 2018. The BRAWL approach promotes the use of much lower risk limits for most elections of 0.1% or lower.

Some issues to be aware of:

Limited Scope

As generally considered, the "risk limiting audits" and similar approaches are concerned with a very narrow step in the entire election process, but it is a very important step to get right. These audits are concerned only with the actual tabulation after the voters have been authenticated, ballots cast, transported, scanned, and tabulated. Each one of those steps has its own risk factors. Limiting the risk at each step is important because the confidence of each step (the inverse of the risk) is multiplied with the confidence at each other step to get the total confidence. The total risk cannot be less than the risk of any single step, so that is why it is important that the risk is minimized in each step of the process.

Inadvertent "fix up"

There is also risk that during the audit process itself, mistakes will be made. Or perhaps more likely, election officials will innocently and inadvertently (or perhaps intentionally) "fix up" any inconsistencies they see along the way, just as they have during the rest of the election. But the audit process is different. Fixing up errors is generally not allowed at this stage because it eliminates precious data from the analysis. We are working with only a very small sample of the ballots that were processed and "fix up" of inconsistencies is tempting to gain an all-clear from the audit. If the audit process is at all involved (as these definitely are), requiring many manual steps, then it will not be clear which errors are part of the audit process rather than the election, and that can be a problem. Simpler is better.

4 https://www.usenix.org/system/files/conference/evtwote12/evtwote12-final27.pdf
5 http://people.csail.mit.edu/rivest/pubs/Riv17b.pdf

For example, in Los Angeles, they audit using a 1% batch comparison audit. They have systematically rescanned any batches of ballots that vary significantly (say over three votes) from the computer report in an effort to diagnose the problem. That does not seem too bad at first glance, but the problem is that they then consider only how the scanner did on the batch in the rescan of the ballots, and, once made, they only compare with that report (and not the original report). If it compares correctly, they say all is well. When in fact, the rescan of the ballots simply confirmed that the original scan was faulty, and the rescan is like modifying the original results. A case of inadvertent fix-up just described and almost no one realizes the magnitude of the error.

In the ballot comparison audits being proposed, a similar act would be to say "Oh, I see that we probably got the wrong ballot here. Here is one that does match the computer report, let's use that instead." With statistical audits that rely on a very small sample, any variances are very important and if they are disregarded through this sort of innocent correction process, then they will not work at all.

Ballot Manifest

There is a great reliance on what is called the "Ballot Manifest" which may be something that is generated by the computer to list all the ballots included in the tally and where to find them. The audits under consideration assume this document is pristine and already checked. In reality, we probably also need some sort of sampling inspection to verify that the ballot manifest also reflects accurately the ballots which are included in the audit.

Most Elections Have Wide Margins of Victory

The margin of victory is typically defined in general practice as the reported margin between the ballot option that has the most votes and the immediate loser, including all the other options in the calculation. So if there were three ballot options where A got 45%, B got 40% and C got 15% of the vote, the margin of victory is defined as 5%. The usual analysis considers only the ballots in question between A and B, which is 85% of the total, and then the percentages are 53% for A and 47% for B, or a margin of 6%, i.e. slightly wider when taken alone.

With that in mind, consider Figure 1, actual data of margins of victory (using the traditional, tighter definition) for all congressional elections for the years 2012-2016.

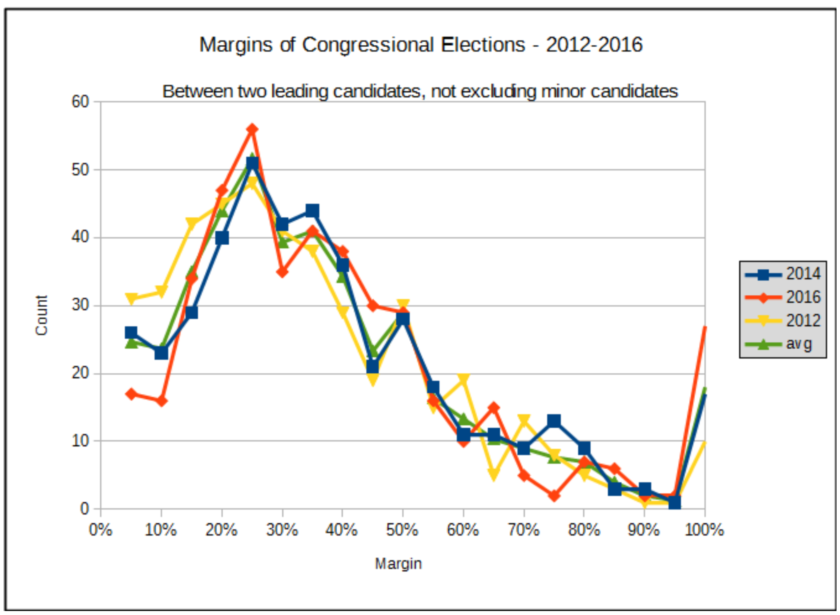


Figure 1: Margins of Victory of Congressional Elections
Source: Ballotpedia.org

We can see that the most common margin is about 25%, and when analyzed, about 90% of elections have margins greater than 10%, and 80% of elections have margins greater than 20%.

We will find that statistical audits are great for larger margins and start to get costly when the margins get very tight. Yet, since 90% of elections have margins of victory greater than 10%, it will be clear that these methods work very well indeed for nearly all elections, and thus are a very good approach (most of the time).

2. The Model

How such audits would preform in practice is actually fairly easy to model in software for simple races and voting types. For this purpose, this analysis will make use of "R," which is a programming language and free software environment for statistical computing supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. (The modeling algorithm has also been translated to Python for those who prefer to use that language.)

The model we generate here actually relies on very simple concepts that can be easily understood and proven to be true regardless of the platform used. We will consider a single race between two options "A", and "B", where the reported result has A as the winner and B as the loser by a reported percentage margin of victory, but due to either inaccuracies in the counting equipment, mistakes, or fraud, "B" is the actual winner by at least one vote. If the method can accomplish this, the it can also detect any race that has been modified by any greater extent at that same margin, of course subject to the caveats already mentioned, in that there are other steps that also have a risk associated with them.

Ballots

This model will concern itself with only a single race between two options, and it assumes that all ballots in the election contain this race. Our model generally considers 100,000 ballots, as that is a common size of election districts that are interested in conducting such audits, although any number of ballots can be utilized (the code actually models the entirety of the ballots with 20 integers).

For ballot comparison methods, we are in fact interested in not just the actual votes on those ballots, but also the difference in the actual votes and those that were reported by the election system software. The official report is generally called the "Cast Vote Record" (CVR) and (for ballot comparison methods) there must be a single CVR record that corresponds to each ballot. Any Ballot Comparison Audit requires that the CVR be available so it can be compared to the physical ballot. Ballot polling audits do not require that a CVR be paired up with each ballot, and that it is common with existing election management software today because they do not track every ballot so it can be paired up.

The table below shows the various types of votes simulated in the election. To save space, instead of having say 100,000 integers, each with a value, the election is modeled as 9 bins, resulting in a total of

20 integers for any number of ballots, although there is a slight and insignificant processing time penalty.

Overstatements	Reported	Actual	Bin	Comment
0	A	A	6	CVR matches the actual ballot exactly
0	B	B	7	
0	N	N	8	
2	A	B	0	Two Vote Overstatement: This "flip" of the vote on a single ballot is unlikely to occur by equipment failure or due to improper ballot interpretation but is a likely hack because it requires a minimum number of ballots to be affected. However, it can also occur due to mis-feeds where a different ballot is scanned instead of the appropriate ballot.
1	N	B	1	One Vote Overstatement: CVR shows an extra vote for the winner or is missing a vote for the loser. This is expected by recurring equipment errors or by a hacker changing CVR, if the hacker wanted to conceal the hack as an equipment failure.
1	A	N	2	
-1	N	A	3	One Vote Understatement: This type of error is likely only due to sporadic equipment errors and it is not of concern because it does not put the race into question, but it does add "noise" to the problem.
-1	B	N	4	
-2	B	A	5	Two Vote Understatement is highly unlikely but can occur due to mis-feeds where a different ballot is scanned instead of the appropriate ballot. These do not put the race into question but does add "noise" to the problem.

If the ballot has an undervote (no vote at all in a race with two options) or an overvote (both options are marked), then this is a Non-vote ("N"), which is equivalent regardless of which case it is. With these three possibilities, we wind up with nine possible cases.

For ballot comparison audits that only require the raw number of overstatements, we can just look at the number of overstatements associated with that bin. One overstatement occurs when the reported results overstated the margin for the reported winner A, i.e. the actual vote was for B or a non-vote. Understatements are the opposite and we will process those as negative overstatements. For polling audits, we look only at the actual vote. Bayesian audits will use both reported and actual values on each ballot.

Parameters

To create the model election, we have the following input parameters

Parameter	**Description**
nTotalBallots	Total number of voted ballots. 100000 is used in our simulations but to create an accurate model of any real election, the simulator allows any number of votes to be entered.
marginpct	Margin of victory, as a % of the voted ballots. 5 is 5%. This the "pairwise" margin of

	victory only between the two ballot options and does not include any other ballot options, and does not include unvoted ballots, which may exist in the H_0 case. This margin also determines the number of 1-vote overstatements that are required to "flip" the election, which is simply the total_ballots * marginpct / 100. Please note that there may be a number of non-voted ballots which are modeled, which means that A + B + N = nTotalBallots, where N is the number of non-votes.
noise1pct	The percent of nTotalBallots of offsetting 1-vote overstatements and 1-vote understatements. This can be estimated prior to the start of the audit and then modified after the first sample of ballots is processed. (The general subject of noise will be treated with more clarity in the text.)
noise2pct	The percent of total_ballots which is noise of offsetting 2-vote overstatements and 2-vote understatements. Since these are regarded as rare, we only model this as 0.
hack2pct	The percent of non-noise overstatements (as determined by marginpct) which are two-vote overstatements. Typical values are 0 and 100. For example, for nTotalBallots = 100,000 and marginpct=3, hack2pct=100, then there are 1500 two vote overstatements in the election and no 1-vote overstatements.
nSamples	The number of samples in the modeled audit. This is not the minimum or recommended limit of sampling, but the number of ballots in the modeled audit if there is no stopping point. For this first case, we will use 2,000 samples.
nTrials	The number of trials of each hypothesis. Generally 1,000 trials is used.

We will focus on the case with 100,000 ballots and a 3% margin in the race in question. We will have two cases to consider, H_0: the null hypothesis (that the reported results are correct, A wins) and H_1: that B wins. For H_1, the minimum change required to flip this race is either 3,000 1-vote overstatements or 1,500 2-vote overstatements. Since there are two kinds of 1-vote overstatements, we will arbitrarily split these between the two types evenly. For algorithms that only consider overstatements and understatements, this modeling simplification does not affect the result.

Those parameters determine how the array is constructed which models the nTotalBallots in the election. We construct

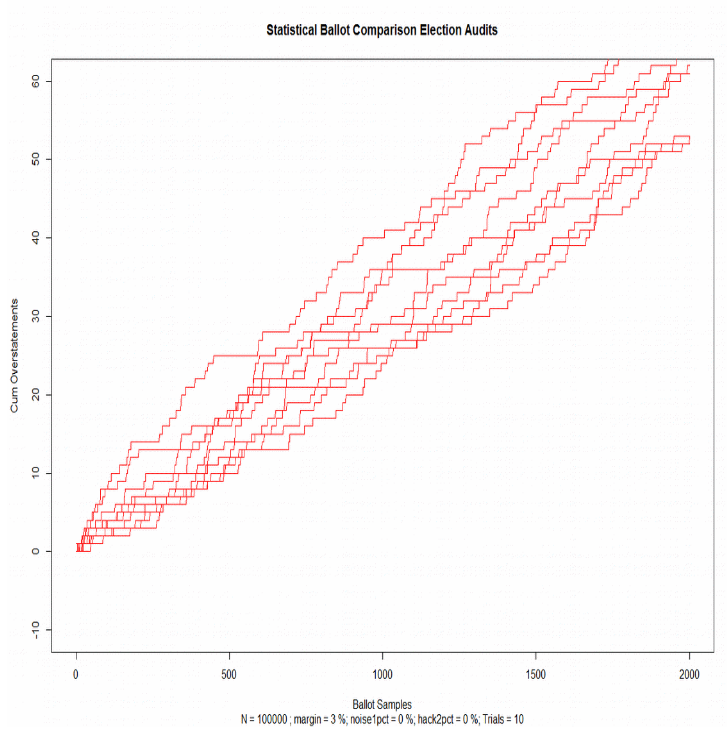


Figure 2: 10 trials each with 2000 samples from 100,000 ballots with 3000 overstatements

our array as bins, with each bin being the appropriate number of one of the nine cases described above. Because we are sampling randomly, it matters not that the ballots are actually grouped into bins.

The use of bins is a means to reduce the amount of memory required for the simulations, but it is equivalent to having nTotalBallots values, and sampling from those.

Hypotheses

We have the two cases, H_0 -- the null hypothesis -- that must average out to 0 overstatements even if there is "noise" that does not effectively change the result, and where A wins, and H_1, which has the same noise, but also enough overstatement votes to barely flip the election in favor of B.

If we randomly sample this array nSamples times, then we have the sample vector with length n in the order that we pulled the samples from the ballot array. The models will (generally) pull 2,000 samples in each audit trial from the 100,000 ballots without replacement, because that is really how ballots should be drawn, since there is substantial overhead to drawing a sample, and in a simulation, it is easy to model sampling without replacement (while using mathematical equations, it can complicate matters quite a lot).

On the average, for a 3% margin, the sample should contain

$$2000 * 0.03 = 60 \text{ overstatements,}$$

and on average, one overstatement is expected every 33 ballots.

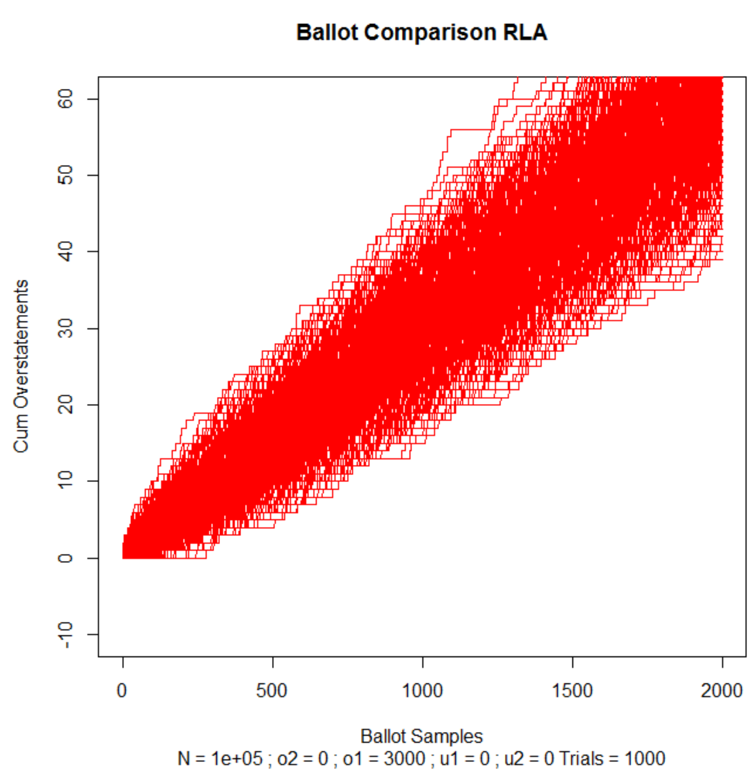


Figure 3: H_1 with 1000 trials of 2000 samples each

Figure 2 shows 10 audits with samples drawn at random with no noise added. The x-axis represents the number of ballots sampled in sequence, while the y-axis is the net cumulative number of overstatements. There will be some variation each time samples are drawn at random, and this accounts for the various different tracks of each set of samples. Each line represents the samples of an audit drawn sequentially.

Figure 2 does not include additional "noise" due to any understatements that do not contribute to the total. Thus, all tracks only go up, and they never go down at all, in this case.

The same case run with 1,000 audit trials is shown in Figure 3. In essence, this is the election with enough machine errors or fraudulent manipulation to just barely flip the election is is what the audit is actually looking for.

Dilution

In practice, the race may not be on all the ballots in the district. If the race is only contained in some of the ballots, then the number of ballots to be sampled will increase by the inverse of that faction. Thus, if only 25% of the ballots contain the race of interest, then we will need to sample 1/0.25 = 4x as many ballots. The dilution fraction can be estimated based on the number of ballots cast in the district for that race vs. all the ballots cast, or it can be estimated as the ballots begin to be sampled. For our purposes here, we will only have a single race on each ballot, and all ballots contain that race, so the dilution fraction is 1, i.e. no dilution.

Noise

It is realistic to assume that for real elections, there will be spurious errors where perhaps certain voters do not fill the bubble right or the ballot is askew and the scanner misses the mark. If these errors happen on a sporadic or even periodic basis, then it will likely affect the different ballot options equally both as understatements and as overstatements. Any time they do not balance, then it is not considered noise. If they balance, then it fits this definition of noise. If we assume a 0.2% noise rate, it will affect both H_0 and H_1. For the former, in 100,000 ballots for this race, there are 200 1-vote overstatements and 200 1-vote understatements. For H_1, there are 200 1-vote understatements and 3200 1-vote overstatements, therefore netting to 3000 overstatements, to flip the race. In this case, on the average, every 500 ballots will have a single understatement which is balanced by one overstatement. The amount of noise is more a function of the process being used and over time, we will likely get to know what the expected noise is for given approaches to conducting elections. For now, the 0.2% level seems about right (if not a bit high) based on the author's experience with the audits witnessed in CA and FL. It is important to note that the 200 ballots in error

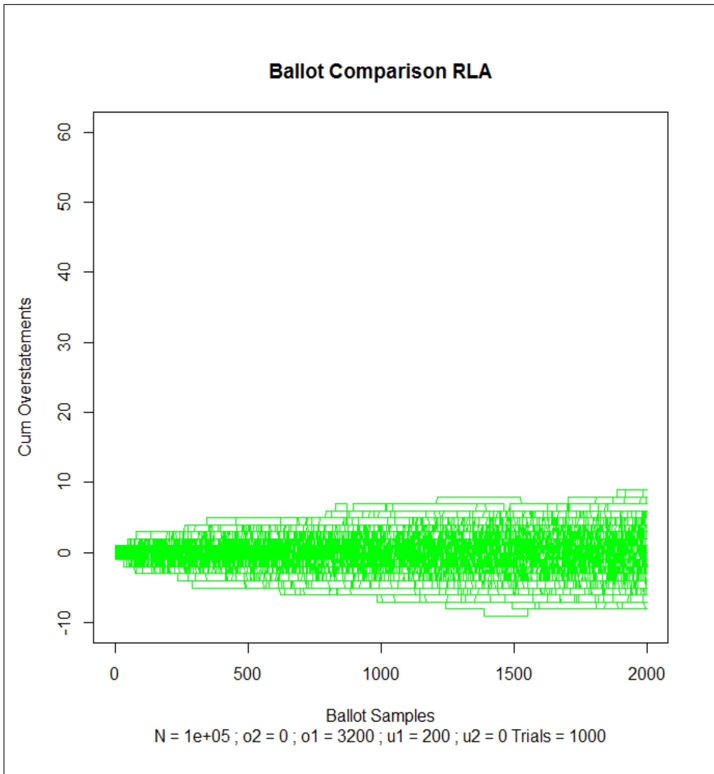


Figure 4: Null Hypothesis with 0.2% noise

out of 100K and other 200 the other way, is for the given race only (not summed for all races on the ballot).

To estimate the noise, we can evaluate the number of understatements that are seen in the initial set of sample ballots and we we see more than expected, then this implies more noise and will cause a required increase the minimum sample size. It may also be possible to do a better job of estimating the noise if we look at all races on every ballot sampled to see how many variances there are, and then divide that by the number of races on each ballot. Otherwise, the number of samples we take is really not enough to get a very trustworthy estimate.

When noise is added, the effect is not too dramatic but it adds some dispersion to both the cases, H_0 and H_1.

If we had no noise, then the H_0 case, i.e. null hypothesis, is simply a horizontal line at 0. This is not realistic.

Figure 4 illustrates the dispersion of the null hypothesis samples with 0.2% noise added. Total of the election is exactly as stated, but there are just as many overstatements as understatements, and they cancel each other out. Any that do not cancel out are considered part of the H_1 hypothesis.

Please note that when we say the overstatements cancel out the understatements, this is with regard to the actual count only. Any given algorithm may consider overstatements differently from understatements, and the simulator includes this information when the algorithms are used.

Dispersion

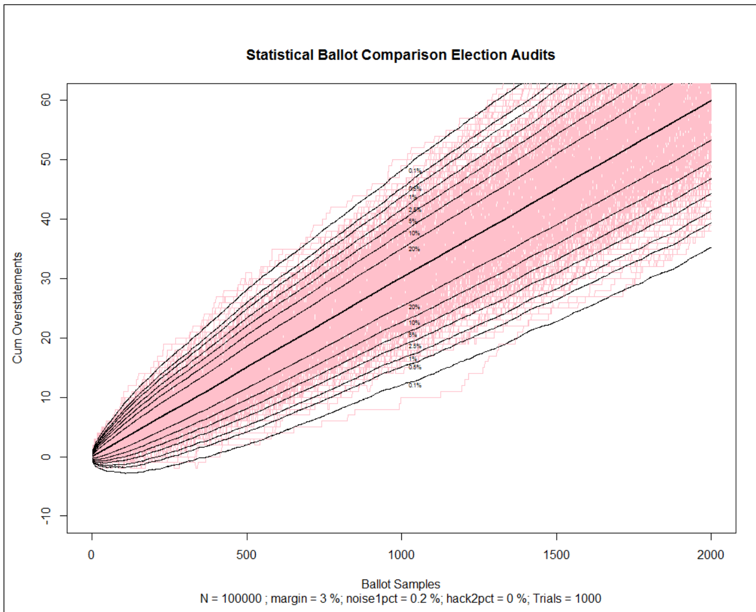


Figure 5: H_1 with dispersion curves

For each hypothesis, we can evaluate the mean number of overstatements among all the trials, and draw some curves which estimate the dispersion of the samples. Figure 5 shows those curves. Please note that in these plots, the mean is not a just line drawn where it probably should be, but instead it is a series of dots, one for each sample which is the mean of the 1,000 audit trials at that sample point. In contrast, the dispersion curves are drawn based on the mean and standard deviation at each point, and the % numbers estimate the fraction of audits that will extend beyond that point, away from the mean. These assume the normal distribution. We can see that the 0.1% level estimates that one audit out of the 1,000 in the trial will extend beyond that line, and in the lower extent, we can see than indeed, one audit does extend beyond that point.[6] (In later revisions of the simulator, we used the actual distribution rather than using the standard deviation assumption and found the results to be approximately the same.)

Combining the two hypotheses

At this point, we can combine the two cases in Figure 6. As the margin increases, the top distribution will tilt up, and as the reported margin of victory tightens, the top distribution will tilt down, like a pair of scissors. As you can imagine, once the top distribution gets very close to the bottom one, it is very difficult to clearly tell them apart using a statistical sampling method. But since most elections have wide margins, the two distributions will normally be nicely separated.

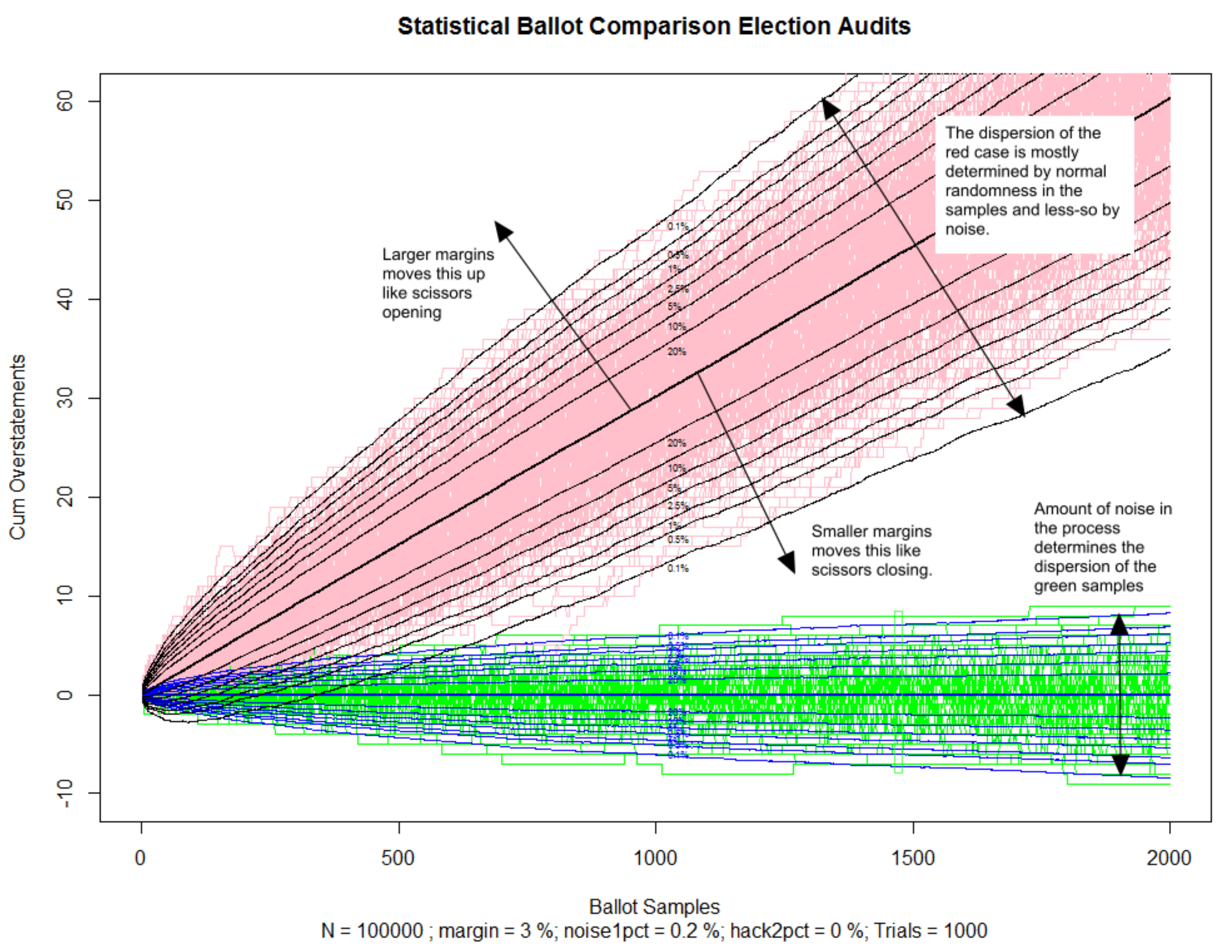


Figure 6: H_0 and H_1 distributions at 3% margin and 0.2% noise

6 A refinement of this analysis might be to abandon the use of the normal assumption at each point and instead determine the dispersion by averaging many more trials to get smooth curves for analysis.

Clearly, at this margin, the two distributions can be easily discriminated to any level of confidence desired and within a reasonable number of ballot samples. The traditional methodology for discrimination is to choose a risk level, and then split the risk between the two types of failures, either (a) calling for a hand-count (or other deterministic approach, which we will just call "hand-count" to comprise all such methods) when the election is not flipped, or (b) inappropriately confirming the election when it is indeed flipped. The conservative and safe approach is to avoid confirming an election unless you are very sure it is correct, while more easily calling for a hand count inappropriately, as that does not have any risk that the election will be improperly confirmed (but does result in additional counting).

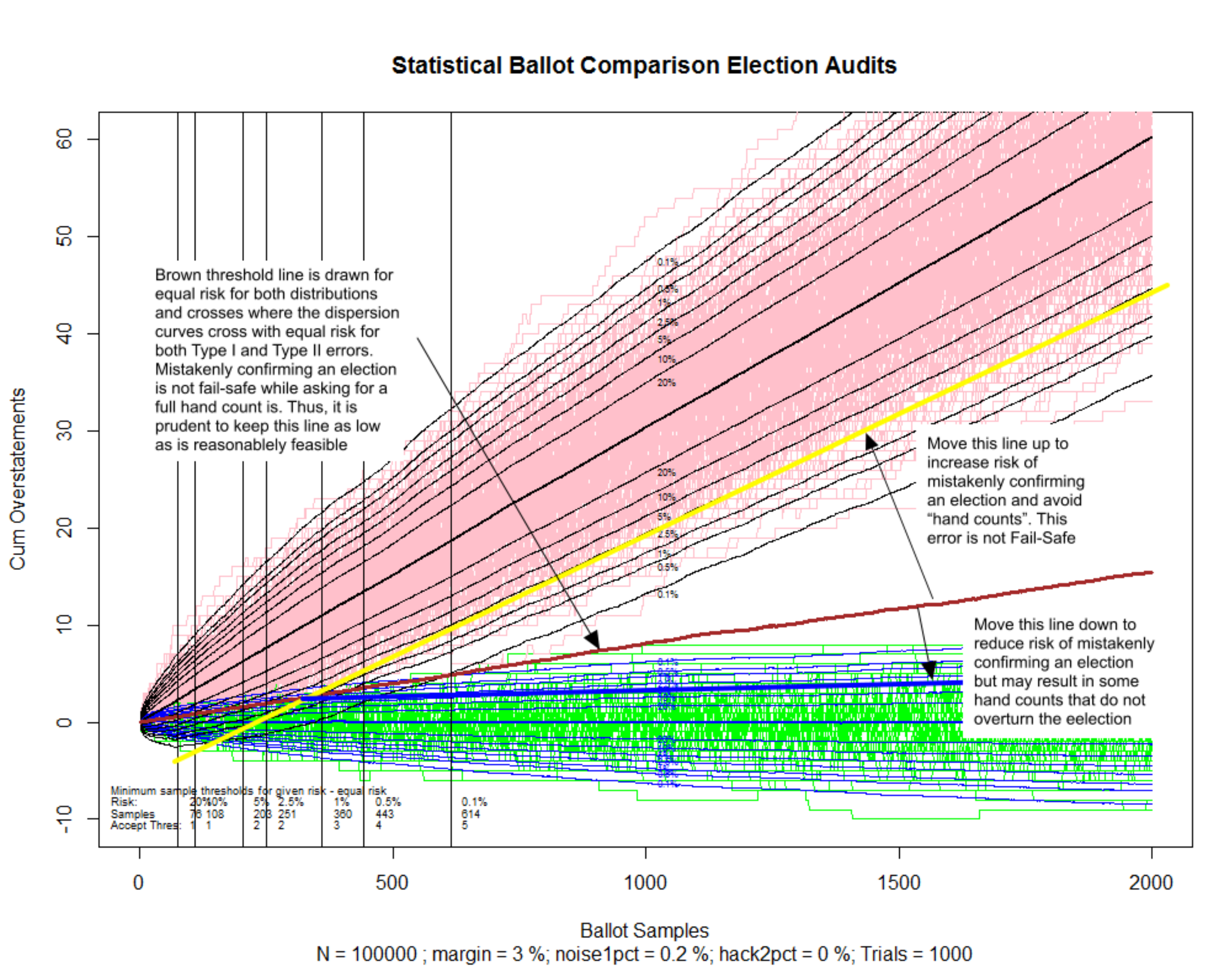


Figure 7: Dispersion Threshold Analysis

Discrimination

We can determine the minimum sample thresholds when the two distributions cross over the threshold curves. The question is, where should the threshold line be drawn? The brown line in Figure 7 is set to be proportionally between the means of the two distributions, based on the dispersion of each. (It is tempting to follow the top distribution by following a line perpendicular to the mean but the proper way is to always stay in along a vertical line.)

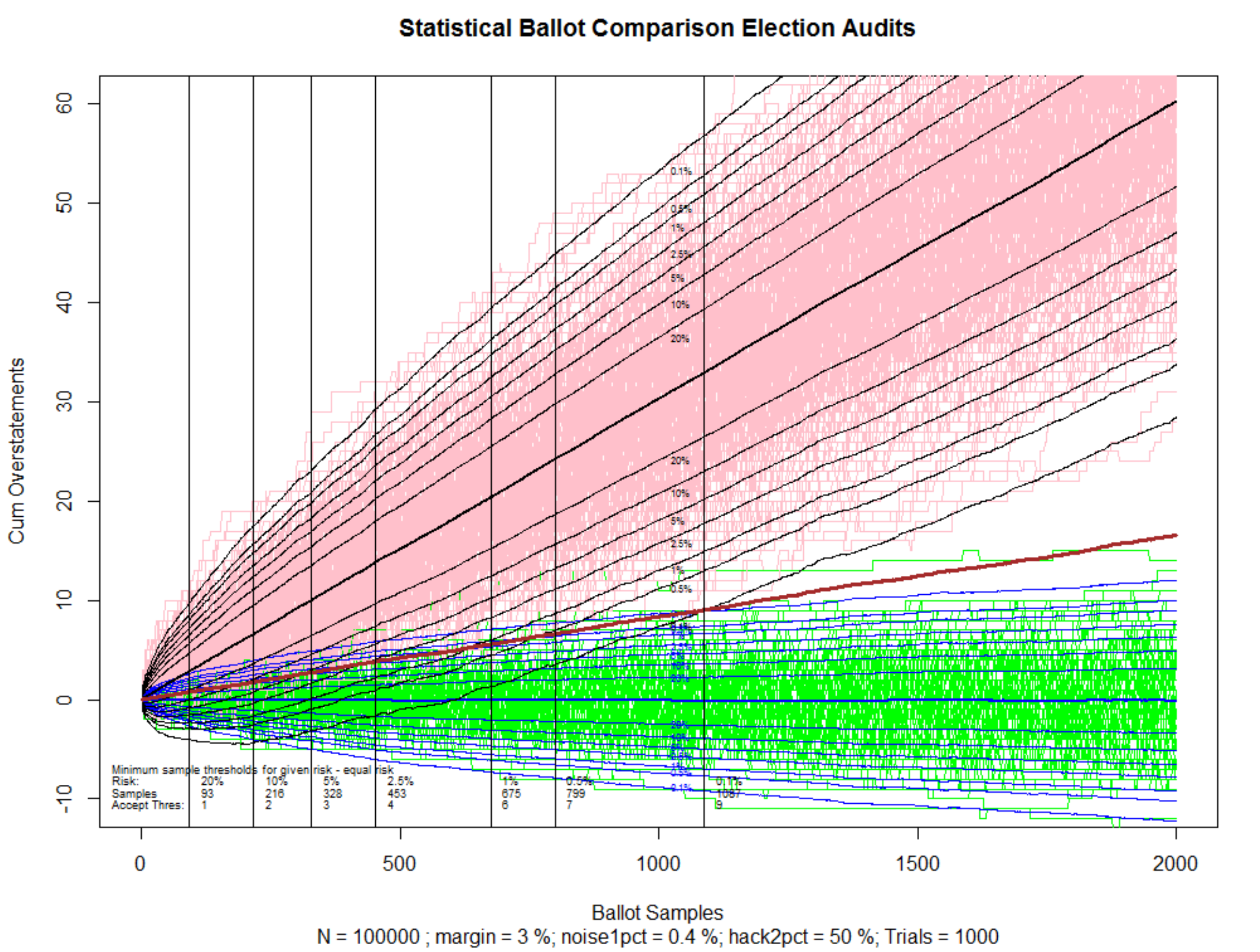


Figure 8: Traditional Thresholds with 0.4% noise and 50% 2-vote overstatements in H1

The vertical lines are drawn where the dispersion curves cross each other so the risk is the same of making the two types of error. For example, at 614 ballots, there is an equal chance of 0.1% that we will confirm an election in error and a 0.1% that we will in appropriately call for a full hand count.

Because the number of samples is really quite low, and to vastly improve the confidence in the election, there is no good reason not to use a much lower risk level, such as using the 0.1% curves and sampling 614 ballots. Given the amount of noise assumed (and can be confirmed by the number of understatements in the sample) we confirm the election if we have 5 or fewer 1-vote overstatements and call for a hand count (deterministic count) at anything over that number. The average of the hacked election is nearly 20 overstatements at that point, but if we have 6 overstatements in 614 ballots, there is something severely wrong with the election anyway. (We will refine this method later in this document.)The big question is where to put this threshold line. If we want to push that elections should be run with a very small amount of noise, then the line should be closer to the green distribution, and if any election seems strange, look into it further. Putting the line up to incorrectly confirm a flipped election is allowing extremely sloppy elections with a large amount of noise.

To be even a bit more safe, we may want to assume twice that amount of noise and half of the hack is 2-vote overstatements. This situation is shown in Figure 8.

If we know the noise, we can set a limit on the number of ballots that need to be included in the sampling process. We see here that with twice as much noise, the number of samples required goes from 614 to 1087, going up only 75%. But noise is a function of the process and we want to encourage clean and error free processes. (It will be prudent to characterize the noise in various processes used for tabulation.)

3. Compare with S4RLA

We will now compare this situation with some popular proposals for Risk Limiting Audits. The first is that which is described by Stark in the paper we are calling S4RLA. Refer to Figure 9.

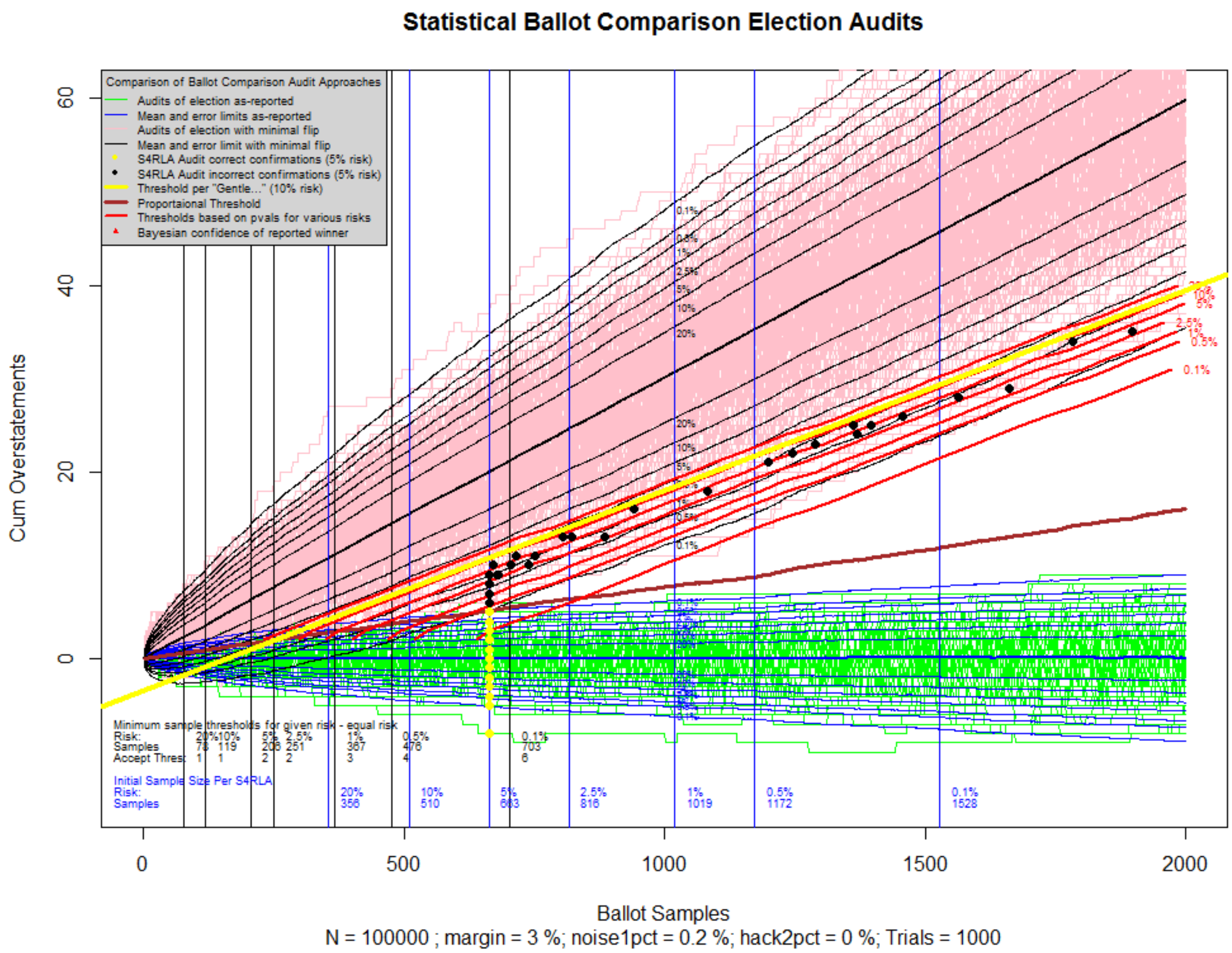


Figure 9: Addition of S4RLA thresholds and correct and incorrect confirmations

We will go back to 0.2% noise. and we will plot the pVal thresholds for the risk levels we are considering, 20%, 10%, 5%, 2.5%, 1%, 0.5% and 0.1%, shown with the dark red lines and annotated at the right end in Figure 9. The vertical blue threshold lines are the nMin values (starting sample sizes) as calculated according to the equations in S4RLA. The Yellow dots are correct confirmations of H_0 case while the black dots are incorrect confirmations of the H_1 case, at 5% risk level. The yellow line is the threshold for 10% risk as provided in "Gentle" assuming only 1-vote overstatements and no "noise." Even with an incredible amount of noise, these thresholds have the "wrong" slope, in the author's view, which is unfortunately defined to fail to produce the correct result a given % of the time, given a flipped race.

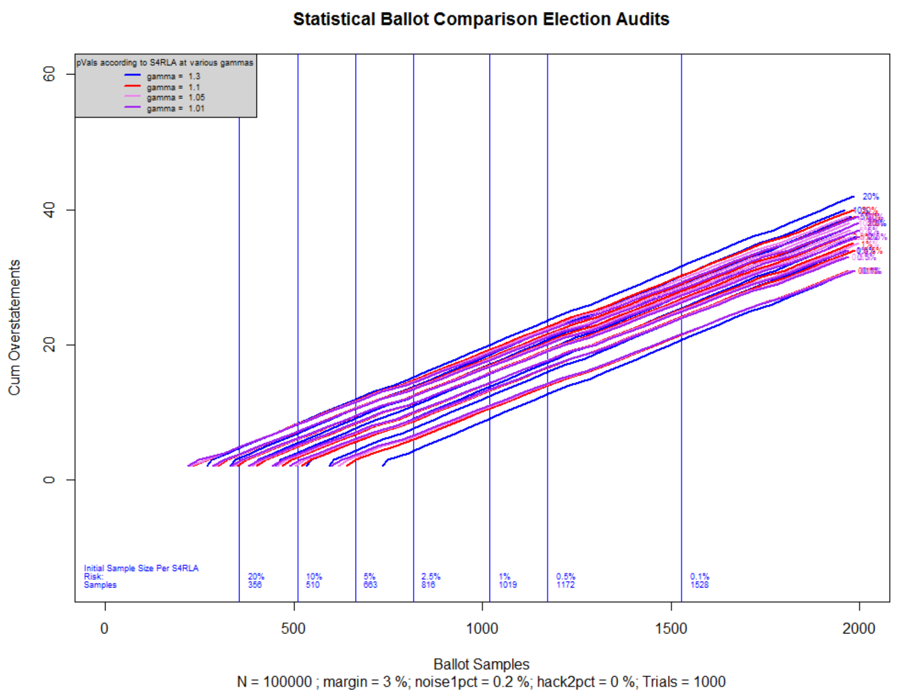


Figure 10: S4RLA pVal thresholds at various gammas

What is striking with these thresholds is that they seem much too high and are not at all conservative, in that they allow improper confirmation of a hacked election rather easily and allow very sloppy operations by elections officials.

As part of this modeling effort, the two adjustable parameters to the S4RLA method were investigated. Gamma is largely used to model how much noise there is in the process, but it hardly moves the thresholds much at all. And what is strange about this "inflator" term, is you can't get rid of it by setting it to 1 because then the log of 1 is zero and the equations are undefined with a division by zero. Thus, we do not recommend this "Super Simple" method because it is actually sort of hard to use.

Next, we looked at the variation in the acceptance threshold in the S4RLA approach by holding gamma constant at an inactive value of 1.0001 and adjusting lambda to several values. Lambda must be between 0 and 1 but you can't go much below 0.4. we note that the slope does get closer to what we view as the optimal (proportional risk) threshold, shown as the dashed line in Figure 11. But they are still located in the wrong place vertically and not as long to the left a desired. With all that said, it may be possible to find an optimal combination of lambda and gamma settings that will work to match the proportional threshold. Yet, this is not taught by the S4RLA document.

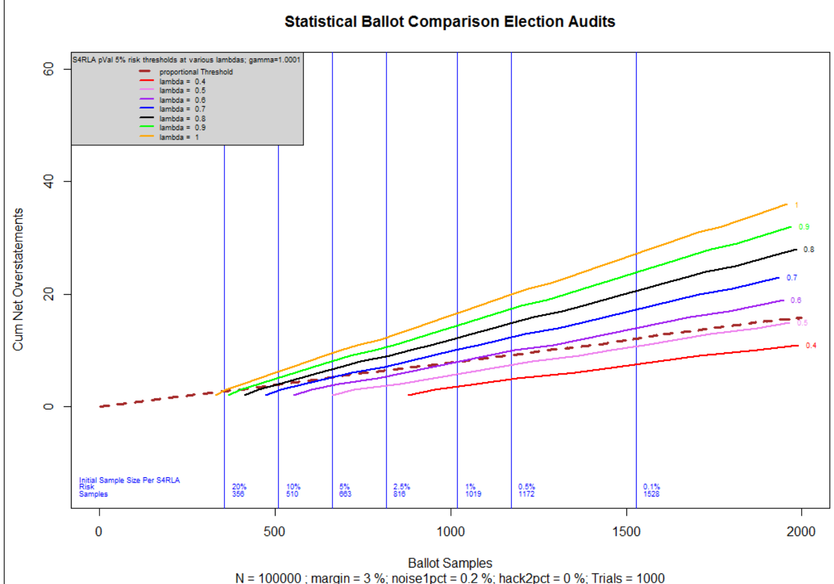


Figure 11: S4RLA risk thresholds at various Lamda values at Gamma=1.0001

4. Compare with Bayesian Method

The Bayesian method as described by Ronald Rivest was also tested in comparison with the model. To allow these results to be easily seen, we will remove the S4RLA curves from the plot. Since the Bayesian method is somewhat time consuming compared to the calculations involved with the other methods, we will provide the probability that A will win if a full hand count was performed at a few key thresholds rather than at every sample in every trial. The results can be seen in Figure 12.

It seems that some sample sizes, it may be feasible to set a threshold of perhaps 88% for the equivalent of a 5% risk limit under the S4RLA method. We will note that although not provided here, all tests of the H_0 case resulted in 100% for A or nearly so.

This result is inconsistent and non-monotonic and there does not seem to be an easy linkage between the real risk level and the win probability, as seen in Figure 13.

The performance of the Bayesian approach improved substantially using pseudocounts of 0 (and is the one shown here -- using pseudocounts of 1 resulted in pretty much everything being 100% for A even in the "B wins" case) and so there may be other ways to tweak this approach so it is useful. The Bayesian approach is actually very good in some respects, as it can handle many different district sizes, multi-option races, and unusual voting schemes, so I am hoping that perhaps it can be tuned up to be a viable method. At present, however, it is not ready for deployment.

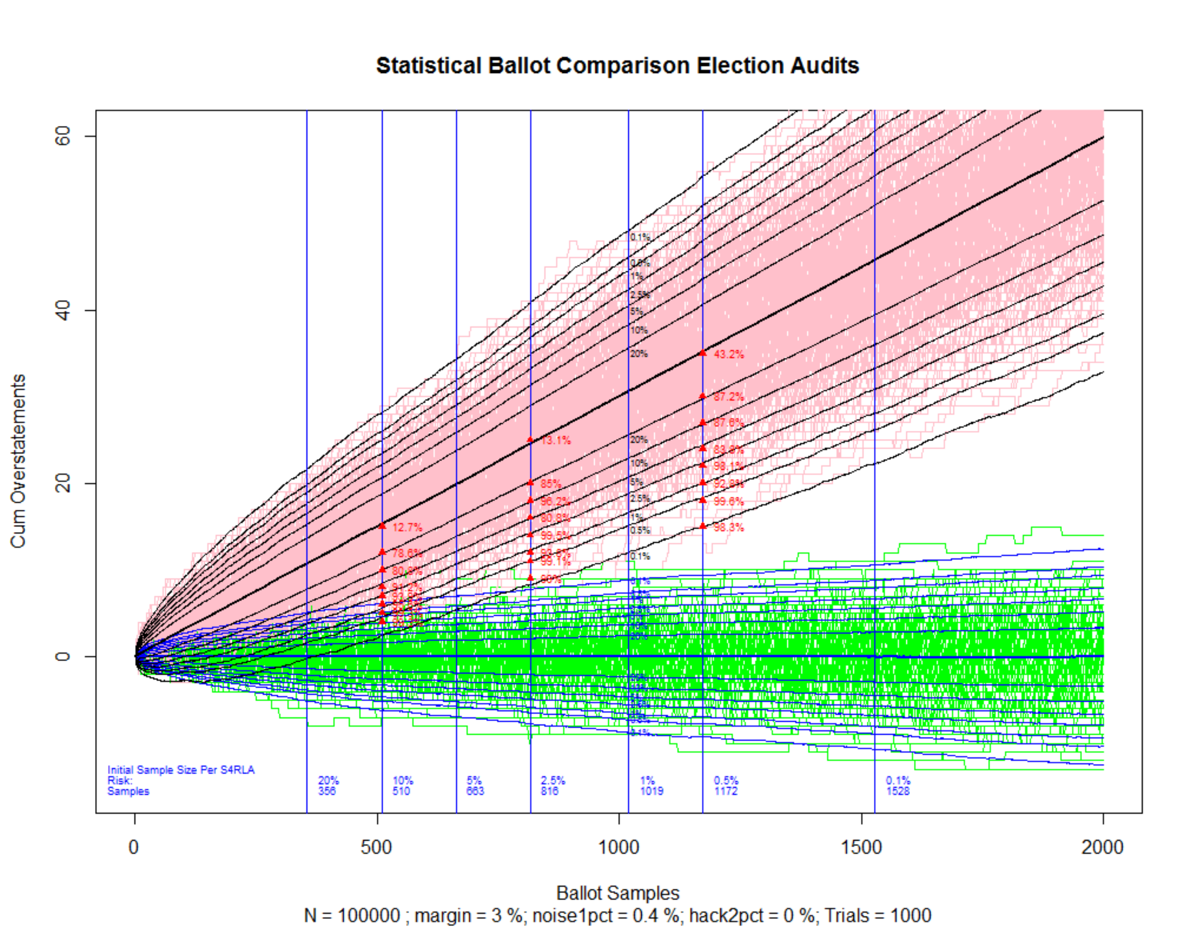


Figure 12: Example Bayesian audits

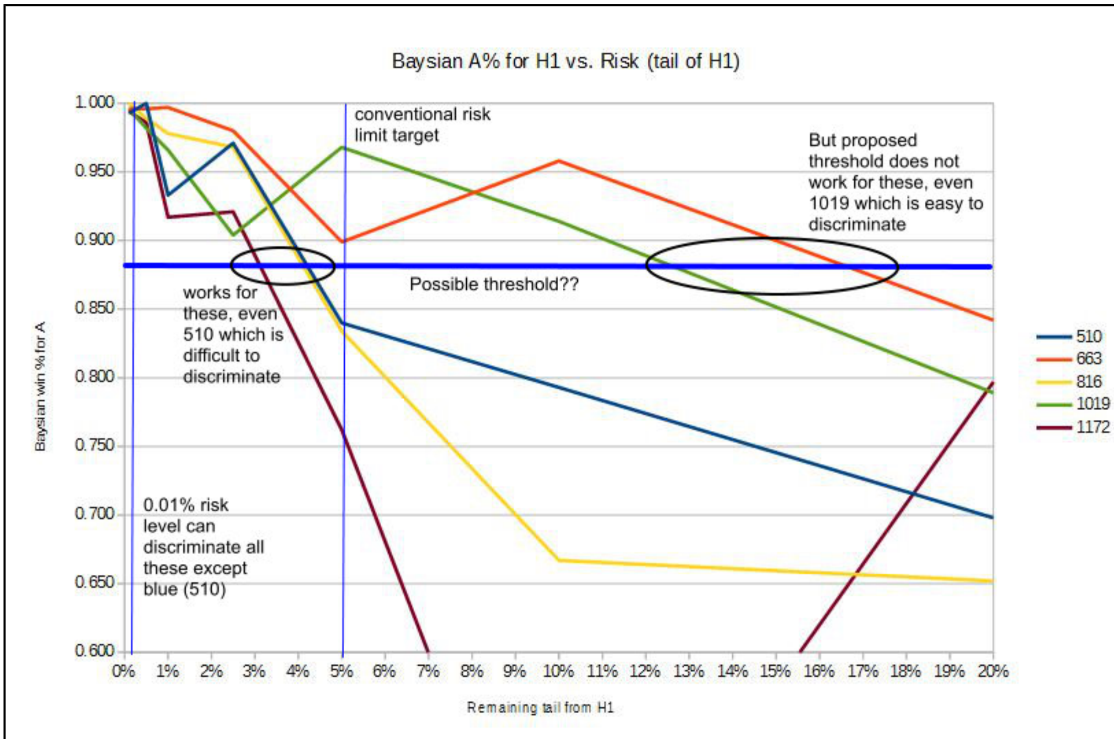


Figure 13: Bayesian results at various sample levels showing issues with setting the threshold

5. The BRAWL method

We now suggest an audit process based on the audit simulations done here. We are looking for a minimum risk, that is to reduce the risk level for most races to less than 0.1%, i.e. we are 99.9% sure that the confirmation is correct, and we can easily bound the work required. For tighter races, we will need to compromise a bit more, and so it will be a balanced approach. The work load can be clearly determined up front, so we can perhaps call this the "Balanced Risk Audit with Workload Limitation" (BRAWL) as a practical (and simpler) way to deploy risk limited audits.

For the purposes of this proposal, we will assume for the moment that 0.2% noise is a valid limit for the amount of noise to be encountered. This will need to be validated based on the process being used, and the initial samples gathered during the inspection process. As mentioned, it will be helpful to gather as much information as possible from the ballots that are sampled until the noise is more fully understood for that process. By gathering the rate of all discrepancies, even in races that are not being considered for the audit, the noise in due to the process can be determined.

Samples should be processed in blocks

There is one other factor contributing to the overall design of this approach. It is important to separate the steps of 1) of reviewing the paper ballots and inputting that information into essentially an "Actual" cast vote record set (ACVRs), and 2) the comparison with reported values -- the Reported Cast Vote Records. RCVRs. The ACVR file should be secured using a computed secure hash value and both published before the subsequent steps in the process, and the RCVR must be similarly secured and published. This reduces the risk that an election worker will "fix up" any discrepancies. Because of all this overhead for each time new ballots are added to the set of sampled ballots, it will be best if we have very few stages in the process. We will want an initial set of samples which are evaluated, and if those do not meet the initial requirements, then we will expand the sample one time (or optionally divide the group in half and do two more groups).

The goal and methodology of BRAWL

The goal will be to reduce the risk of confirming "A" as the winner when "B" actually won to only 0.1% for most elections, and we will include also a "conservatism factor" of 20% because we just are not quite sure that everything will go our way. This is a standard approach in engineering designs of systems to avoid failures for reasons we can't quite predict up front. For purposes of explanation, we will zoom into the key areas of this plot so you can see clearly, in Figure 14.

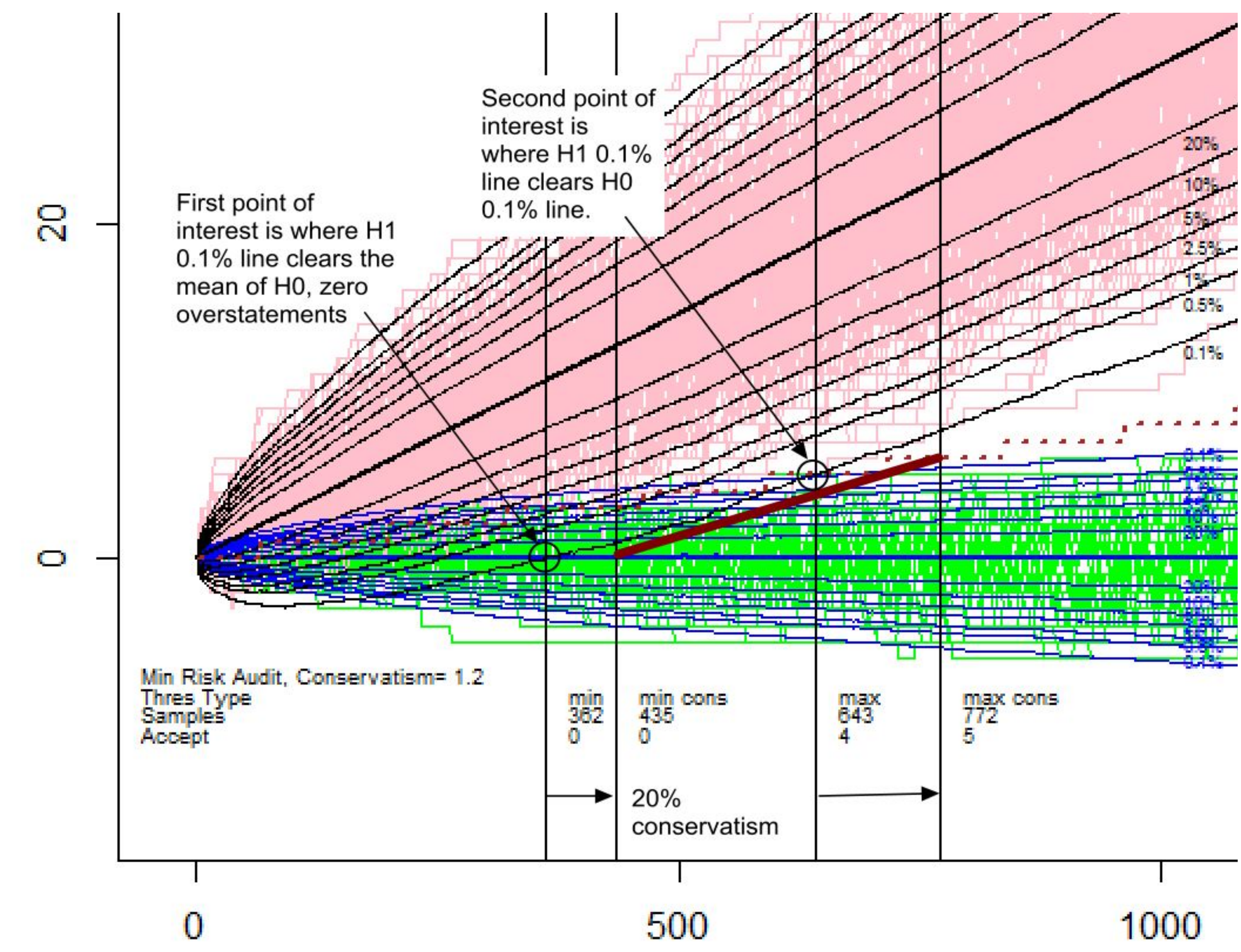


Figure 14: BRAWL at 3% margin with 0.2% noise

According to the model and the statistics of necessary number of overstatements for the election to flip, the first threshold is calculated at the point where the 0.1% curve of the red H_1 distribution crosses the 0 axis. You can see that it occurs at 362 ballots. We add a 20% conservatism factor and define the initial sample size to be 435 samples. At this point, if the H_0 distribution is exceptionally free of "intrusion" by the H_1 distribution at the x axis where the net cumulative overstatements (determined by adding overstatements and subtracting understatements) is less than or equal to zero, then the audit can stop. However, if the net cumulative overstatements is not <= 0, the audit should then escalate. We propose that it not escalate incrementally, but all in one (or perhaps two, half-sized) stage(s). The next size is calculated first by the crossing of the two 0.1% dispersion limits for both distributions, and that happens at 643 samples. Then, we will inflate that number by the conservatism factor of 20%, giving us the ultimate sample size of 772 samples. When we have a situation, such as this case, where we go beyond the maximum threshold (643 in this case) then we will go with the number of cumulative overstatements provided by the proportional threshold described earlier. This is provided on this plot by the dotted red line. At that point in the audit, we have a clear go/no-go threshold that is far outside the limits of the two distributions, as shown in the histogram view in Figure 15. If we have 5 cumulative overstatements or less, then we consider the election confirmed. More than 5, and the audit stops, and a full hand count (or other deterministic approach) is used.

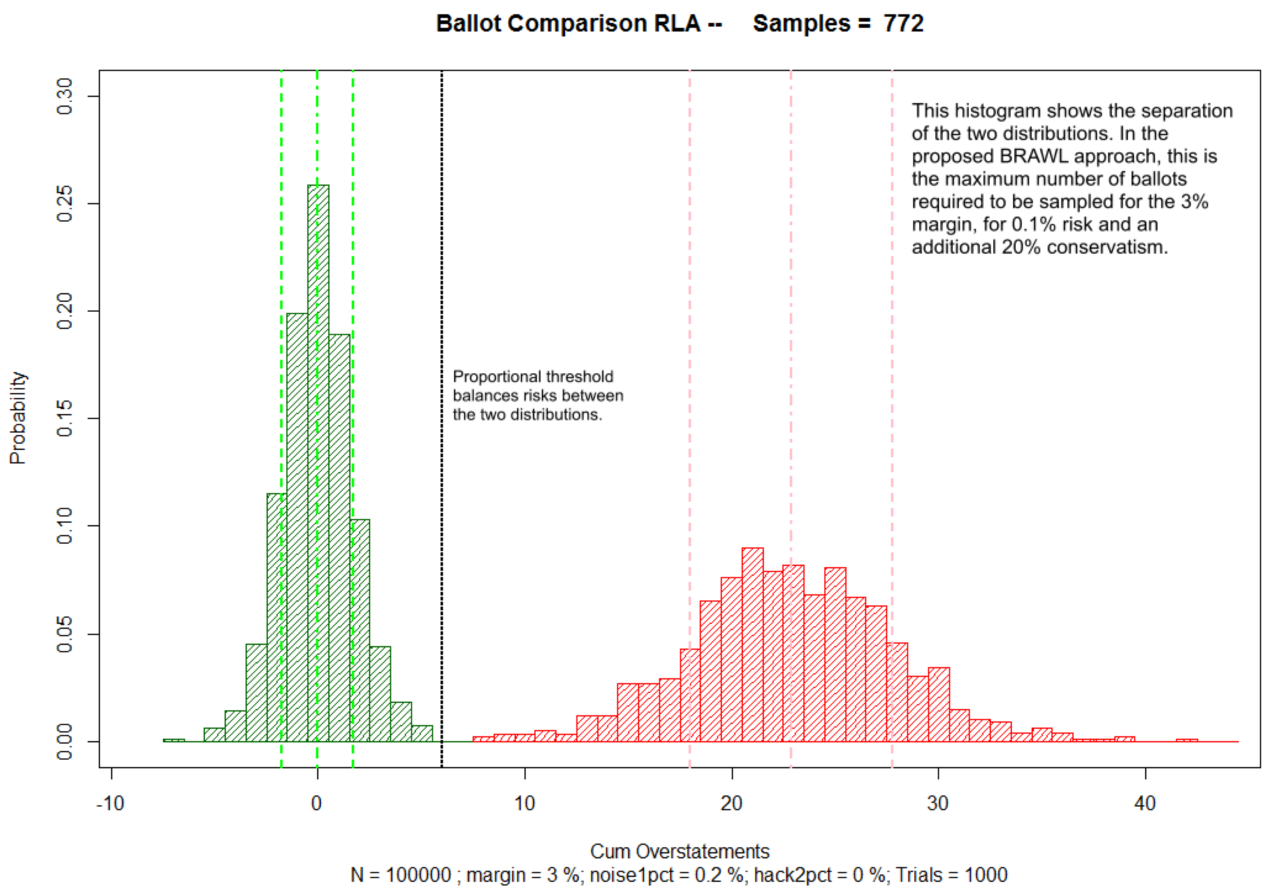


Figure 15: Separation of the two distributions after 772 samples at 3% margin.

Summary of BRAWL

In summary, the audit in this case proceeds as follows:

1. randomly pull 435 ballots from secure storage, preferably half by referring to a computer-generated ballot manifest and the other half strictly by physical location.
2. Evaluate the actual marking on the ballots and enter those into a computer file we will call the Actual CVR, ACVR. Secure that file with a secure hash and publish both so outsiders can make copies so as to detect any changes.
3. Compare ACVR with the official Reported Cast Vote Record set (RCVR) to determine the net cumulative overstatements.

4. If less than or equal to zero, stop and confirm the election.
5. If not, then randomly pull an additional 337 ballot from secure storage as mentioned, for a total of 772 ballots.
6. Evaluate the actual marking on the ballots and enter those into a second computer file ACVR2. Secure that file with a secure hash and publish both so outsiders can make copies so as to detect any changes.
7. Compare ACVR2 with RCVR to determine the net cumulative overstatements, including those from the initial sample.
8. If the net cumulative overstatements is less than or equal to 5, stop and confirm the election.
9. Otherwise, move to a full-hand count or other deterministic process.

The second part of BRAWL

That covers the first part of the BRAWL method, which we could call the minimal risk method, since it reduces the risk of improperly calling the election for the wrong winner, in the case when a full hand count would actually reveal it to only 0.1%, and actually a bit less because of the 20% conservatism.

It turns out that this method will work for most elections until the margin of victory gets down to less than 2% or so. To figure this out, we configured the modeling program to be run over approximately 40,000 simulated audits at various margins, to determine the thresholds for initial sample and maximum conservative sample. The result is the plot shown in Figure 16.

This plot provides the results of this analysis and provides some approximate numbers based on the assumptions already described. First, you will note that the curves follow the 1/x relation form, and as the margins shrink, the number of ballots in the sample grows very large indeed. The top four curves are the points just described for the BRAWL method. We will be interested in the Conservative Min (cmin) and Conservative Max (cmax), the annotation provides the sample sizes required and in parenthesis, the overstatement threshold. The other two curves are the even more risky 5% min and max curves which will come into play shortly.

Conceptually the 1/x form of these curves is because (in Figure 6) the red distribution is moving down toward the green one, and like a pair of scissors, the intersection of the two moves faster and faster (to the right) as they move together. Depending on how significant the noise is, it may take a very large sample size for the two distributions to have enough separation so they can be discriminated, if indeed that ever happens. Staying with the 0.1% risk level for 2% margin requires 1382 samples with the 20% conservatism included. An election with a reported 1% margin requires 3736 ballot samples (worst case), literally off the chart in this plot. At this point, and only for these tight elections, we believe it is okay to drop down to the 5% (for example) risk level. If we do that, we move to the 5max curve. At

1500 samples, we can detect an election down to just under 1% margin, taking the risk that we may be wrong at 5%. It is only in these tight races that we need to entertain such a risk.

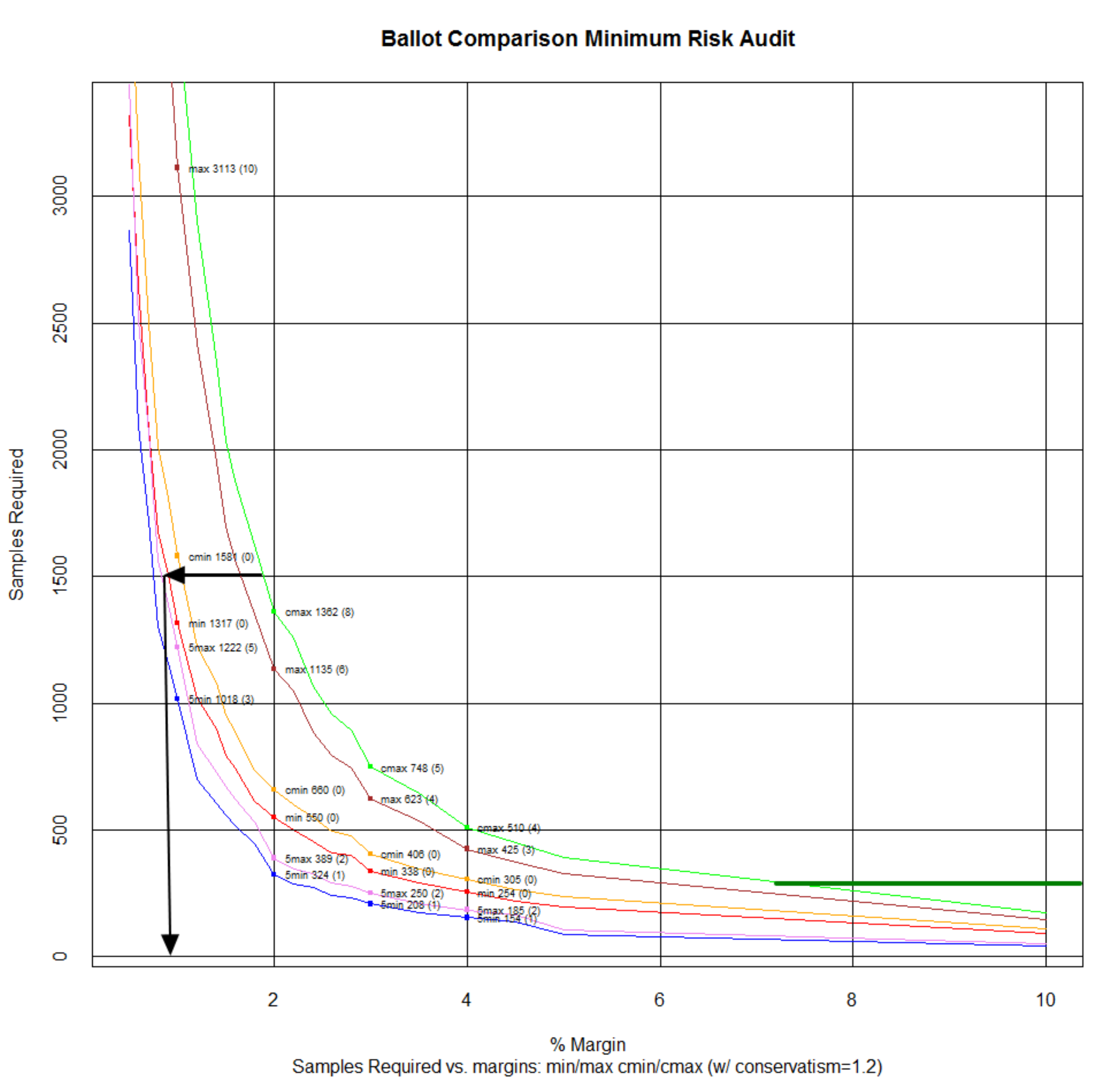


Figure 16: Compromises in BRAWL for small and large margins

With margins any closer than this, we do not recommend using ballot comparison risk limit audits at all, that is for any margins less than 1% (at 0.2% noise). The amount of work is just too great, and there are other ways to skin the cat.

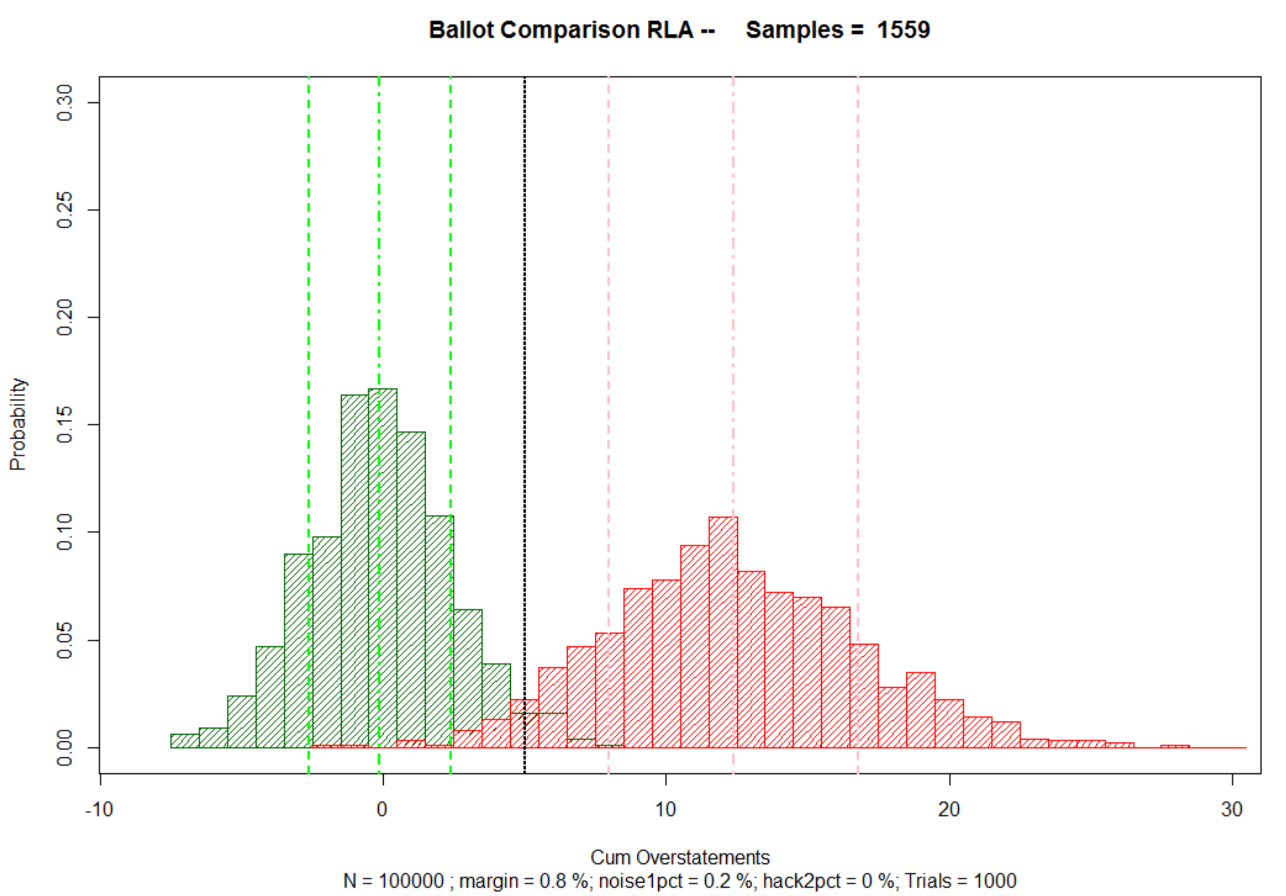


Figure 17: Histogram of 0.8% margin at 1559 samples and threshold for 5% balanced risk.

Now at the other end, the number of samples even for the conservative minimal risk of 0.1% begins to shrink to very small values and a commensurate small number of overstatements. Given the structure of elections, we believe that any sample size less than say 200 or 300 is imprudent. Therefore, we recommend that the minimum is between 200 and 300 samples, and the number of net cumulative overstatements will grow according to the margin and the location of the proportional threshold described earlier respective to the margin of the race.

This is shown in Figure 16 as the straight green line extending the top line to the right.

With these changes in place, we have an audit procedure that can be described and completely (for simple races and voting schemes) using a set of tables. Even if the tables are extensive, simple lookups are feasible which will ease the complexity of proving that correct procedures are followed, and it greatly will simplify the instructions for conducting the audits. Without the refinements described to balance the workload, both to decrease it slightly as the races get tight and also to increase it slightly when the number of samples get too small, we have a table which describes the minimum number of samples, an inflated number, and the number of net cumulative overstatements which are the threshold for each.

Margin	Min	cMin	Max	CMax	5Min	5Max	os Max	os CMax	os 5Min	os 5Max
0.5	3376	4052	8936	10724	2840	3408	17	21	5	7
0.6	2506	3008	6791	8150	2057	2469	15	19	4	6
0.7	2066	2480	5260	6312	1720	2064	13	16	4	6
0.8	1714	2057	4335	5202	1299	1559	12	15	3	5
0.9	1496	1796	3619	4343	1152	1383	11	13	3	5
1	1281	1538	3050	3660	1012	1215	10	12	3	5
1.2	1061	1274	2395	2874	722	867	9	11	2	4
1.4	835	1002	1920	2304	597	717	8	10	2	4
1.5	789	947	1684	2021	545	654	7	9	2	4
1.6	754	905	1563	1876	532	639	7	9	2	4
1.8	633	760	1368	1642	450	540	7	9	2	4
2	583	700	1164	1397	340	408	6	8	1	2
2.2	473	568	925	1110	283	340	5	7	1	2
2.4	456	548	900	1080	275	330	5	6	1	3
2.6	417	501	805	966	248	298	5	6	1	2
2.8	392	471	746	896	234	281	5	6	1	2
3	349	419	646	776	216	260	4	5	1	2
3.5	290	348	533	640	181	218	4	5	1	2
4	256	308	488	586	156	188	4	5	1	2
4.5	229	275	393	472	139	167	3	4	1	2
5	199	239	338	406	94	113	3	4	0	1
10	85	102	142	171	43	52	2	2	0	1

It is the case, however, that audits will be concerned with not just the closest race, but also providing a risk estimation for other races that are opportunistically audited in the same sample of ballots pulled for the closest races that will drive the number of ballots audited. In those cases, it is nice to provide a plot showing the actual audit and how it compares with simulated audits, so as to provide a visual description which can be easily understood without any mathematics background. An example of such a plot is shown in Figure XX regarding the results of the Orange County June 2018 and the simulation approach.

6. Ballot Polling Audits

We now turn to polling audits. Since the reader likely now has a bit better idea of what the charts mean, this topic can be covered a bit more rapidly using a similar simulation model. Consider Figure 18. The plot provides audits of the election as-reported in green, and if the election has enough errors or hacks to flip, in pink.. This example considers a 10% margin race.

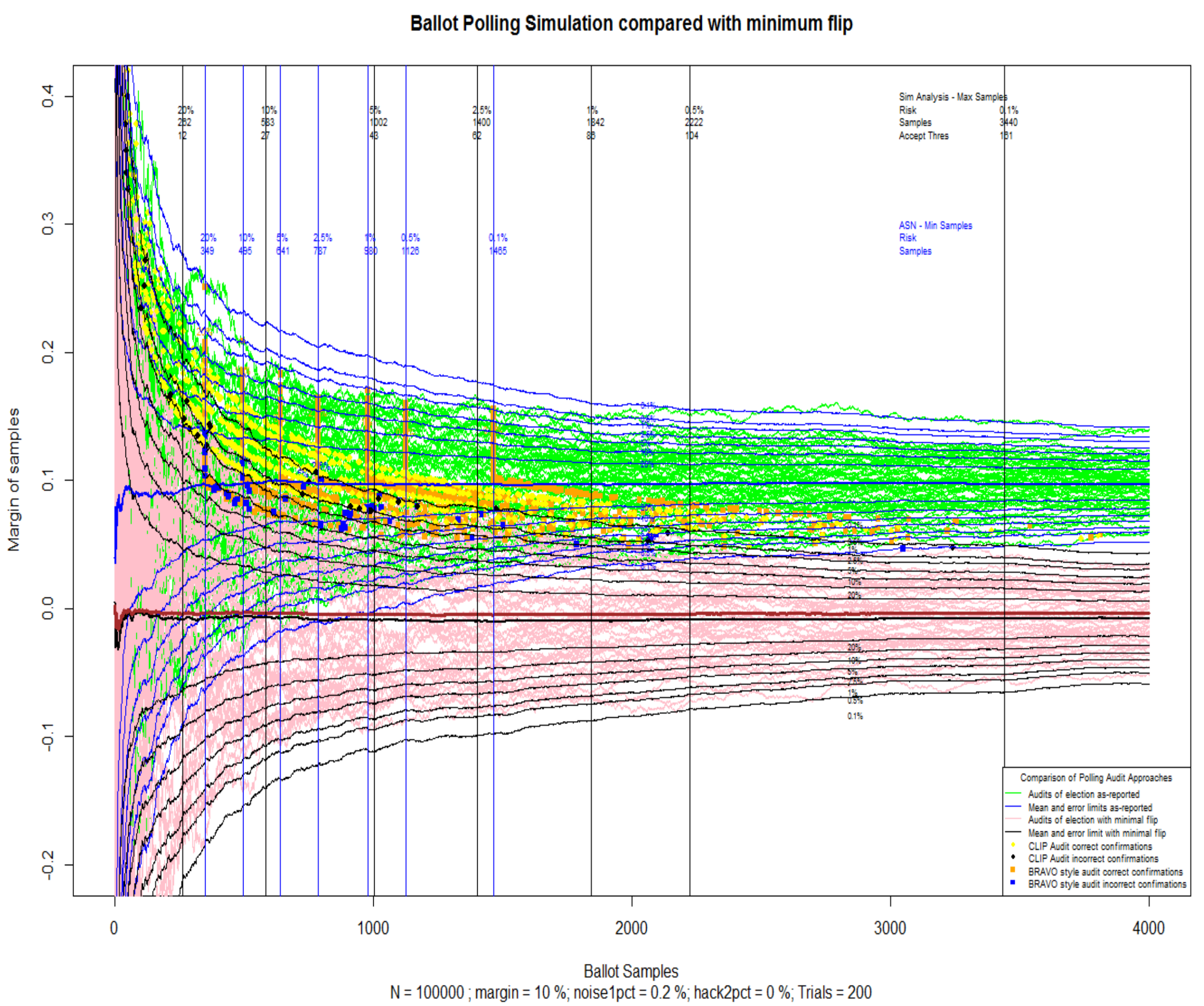


Figure 18: Polling audit simulation showing BRAVO and CLIP methods.

Here, to add to our understanding of how each of the audits progress, we have excluded all audits (pink and green) except those that are either at a maximum or minimum among all the audits at at least one point in the progress of the audit when compared to all other audits in 200 trials.

If you turn your head to the right, the distributions look like two redwood trees that have very large bases and overlap until they get tall enough and thin enough. These will continue to thin until at the very end, all audits would converge on a single margin, the result.

As the first few ballots are added to the total average, the average can swing wildly. Even the average of all the audits wiggles quite bit at first. But over time, the green election shows that there is a 10% margin, with A winning over B. The pink curves show that with the H_1 hypothesis, the margin averages to just under 0%, and the election should be awarded to B.

Also drawn on these curves are the same dispersion lines as were drawn in the ballot comparison case. According to this analysis, there should be no reason for any audit procedure to continue to audit the election after about 3440 samples, because at that point, the distributions have separated so that there is only 0.1% risk in either direction.

Also shown on this plot are the number of samples for the distributions to meet the criteria of 0.5%, 1%, 2.5%, 5%, 10% and 20% risk, that is, where the black dispersion lines of the pink error case cross the blue dispersion lines of the green distribution, the as-reported win of 10% in A's favor. These lines would be valid if the auditors first sample all the ballots required to that line and then conducted the test. If the audit is broken up into blocks or done ballot-by-ballot, then there is some risk at each step that you will misinterpret the result, and thus, these vertical lines may be substantially off by the portion of audits misinterpreted along the way.

The two approaches considered in this paper to deal with a polling audit described by BRAVO and CLIP. The progress of these audits is shown with the vertical starting positions (Average Sample Number -- ASN) shown in blue for BRAVO for each risk level, Then, when that threshold is met, you will see many audits being correctly confirmed as shown with the orange dots. When BRAVO incorrectly confirms the flipped election hypothesis (pink) then that is shown with a blue dot.

The CLIP audit does not have minimums, and the intent is to have it proceed on a ballot-by-ballot basis, starting right away. Correct confirmations by CLIP are shown with yellow dots and incorrect confirmations by black dots.

According to this analysis, there is no reason for any polling audit at this margin to continue past about 3440 samples drawn, because then the two 0.1% dispersion lines have crossed and the two distributions have parted to allow discrimination. Unfortunately, the BRAVO and CLIP do not use any hard-stop and provide vary little guidance as to how long these can go on. And we note that two of the audits (within this range) continue after this point in the BRAVO audit.

What these audits do is to slightly optimize so that for many cases it can stop early. They (generally) confirm audits that are clearly beyond the limits of the pink distribution.

The key interest in these plots is the area where the two "trees" overlap. It is certainly quite safe to confirm any election in an audit that has the green tree, for example beyond the 0.1% extent of the pink "tree." The overlapping part will look like a steeple.

What BRAVO and CLIP neglect to do is to consider the pink tree beyond the extent of the green tree, i.e. they do not call the audit off when it is clearly a flipped election -- the lower pink tree when the green tree is not overlapping. This is considered a poor algorithm because it is far harder to continue to pull ballots at random than it is to conduct a full hand tally of the election once it is known that the election is that far out of whack.

Using this type of audit, the concept of noise is not meaningful, as there are no overstatements nor understatements, only the actual votes accumulated compared with the reported results. Also, there is no reliance on a detailed set of Cast Vote Records nor (potentially) on any manifest. For these reasons, this type of audit is simpler and to that extent, better. But the severe downside is the sheer number of ballots to count.

7. BRAWL Polling Method

If we utilize the general concepts of the Balanced Risk approach, then a proposed alternative to the popular methods is proposed, which we will call the BRAWL ballot polling audit. This method will be described conceptually with Figure 19.

First, auditors should progress through the audit in batches of about 500 ballots at a time. Batches are recommended because it is best to separate the evaluation process from the audit process, but it is less severe of a requirement in this type of audit because there is no detailed CVR to compare with, and thus reduced danger of any fix-up as the process develops.

Second, auditors develop the margin based on the ballots sampled. If it is beyond the 0.1% dispersion line in the positive direction of the error case (marked by the 0.1% black line and shown with the dark-green line), confirm the election. On the other hand, if it is below the 0.1% (blue) dispersion line, then reject the reported results and move to a sequential hand count (or other deterministic method).

If not, then get another batch of 500 ballots and do it again. Worst case, this audit will continue to require, at most, about 3660 ballots, if you consider the 0.1% dispersion lines.

We must point out that the even though we are evaluating each audit based on the 0.1% dispersion lines, this does not mean we only include 0.1% of the "other"case. That would only happen if you gathered up all the samples and evaluated it only one time after processing 3660 samples. If we evaluate the 0.1% tail for each batch, then 0.1% will be mistakenly discriminated at each batch. By the time we get to the end of the tail and have processed about 8 batches, about 0.1% x 8 = 0.8% have "leaked out" meaning we are off by nearly 1% in the raw risk estimation. The benefit of conducting the audit process in blocks is therefore clear, since the number of errors is limited by the number of blocks.

So the threshold shown by evaluating the strict thresholds (shown along the top and with the light vertical lines) will need to be adjusted by the samples required to adjust it by that (approximately) 1% just mentioned. In any case, the "tip of the steeple" can be lopped off at an appropriate location to limit the workload required.

Once the location is determined where the audit will be terminated, the threshold will be set at one-half the reported margin. If the calculated margin by the actual ballots sampled by the audit is above that threshold, the election is confirmed, otherwise, the reported results are rejected and the audit moves to a full sequential hand count (or other deterministic method).

The early termination point can be determined for any desired risk level, and the go/no go margins can be determined in advance, and again, a simple and standardized table-lookup approach can be utilized so any code developed to assist in the process can be easily confirmed as correct.

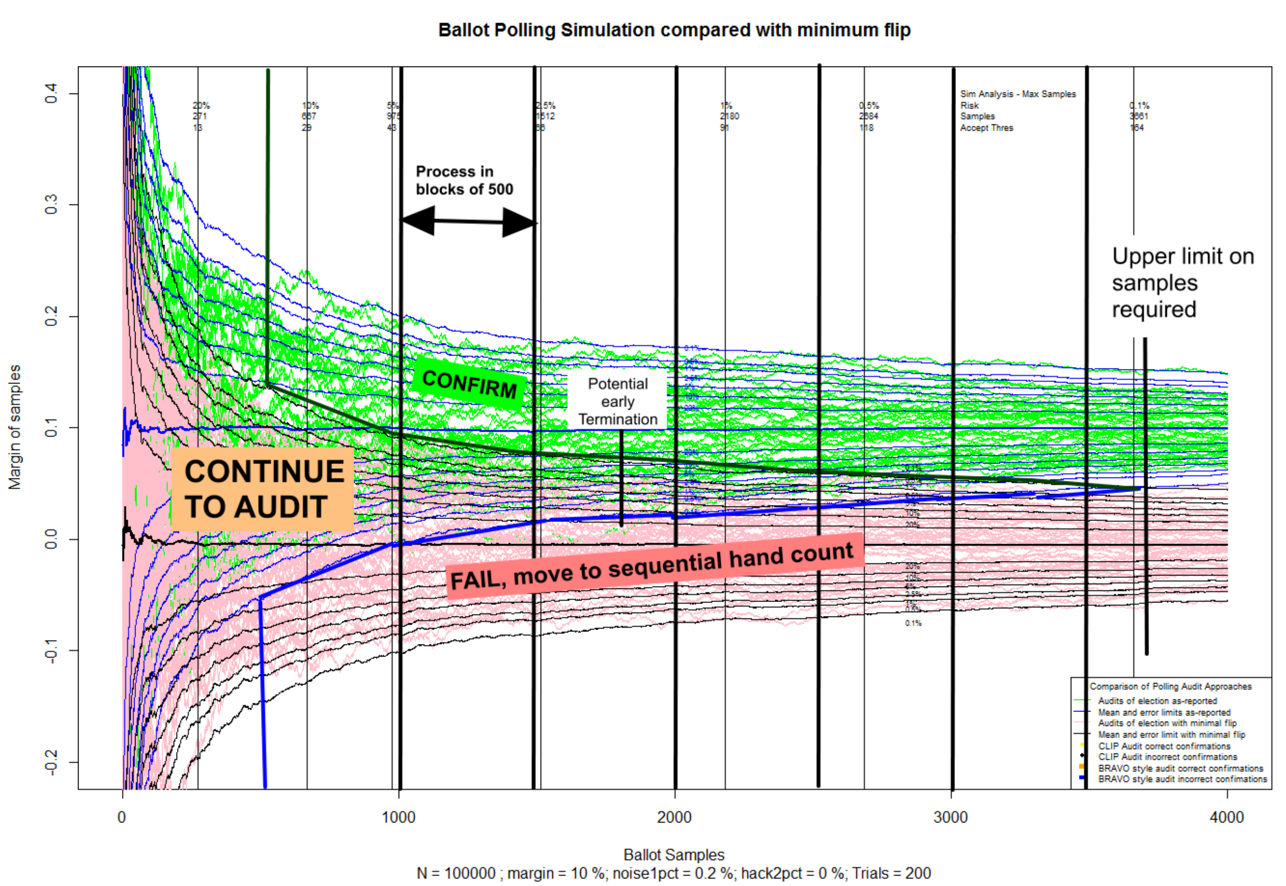


Figure 19: BRAWL ballot polling audit method.

We would like at this stage to get an idea of the required maximum number of samples required. To do this, we simply run the same simulation multiple times while varying the reported margin. Figure 20 provides a sense for how the maximum ballots required for a given risk limit will vary vs. the margin of the race. It is clear that the knee of this curve is at about 5% margin. Any races closer than that and this approach, and any other polling audit approach will begin to get very costly in terms of work.

Please note that these are the maximum counts. Most polling audits will complete much sooner, but it is all just a matter of chance in how the ballots are pulled. Election officials that are planning on conducting polling audits should bear in mind that the workload may extend to the values shown, and on top of that, a full sequential "hand count" may still be called for. Based on workload criteria, officials can always opt for a "full hand count" at any point.

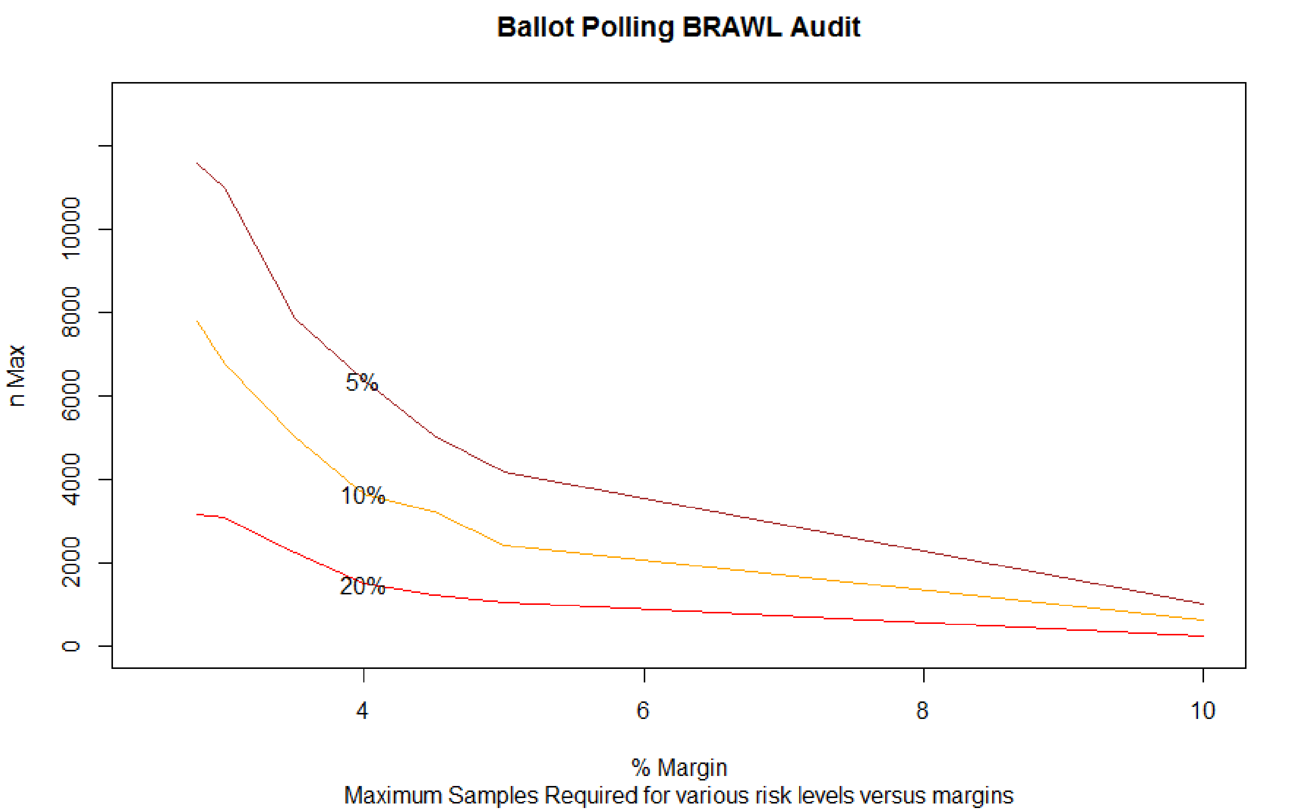


Figure 20: Maximum ballots required for polling audits (BRAWL)

8. Batch Comparison Audits

The most popular type of audit today is the batch-comparison audit, where each batch of ballots selected for the audit is hand-tallied, and compared with the computer report for that batch. Although we do not attempt to model these, some comments are warranted.

The "1% Manual Tally" in California[7] and the "Voting System Audit" in Florida,[8] among others, are batch-comparison audits. The benefits of this type of audit is that it is simple to implement, is easy to understand, provides a check for some types of fairly extensive hacks of the election, provides a direct check that the equipment is working correctly, and validates the organizational structure of the ballots (precincts, batches, etc) to some extent, equivalent to what a manifest audit would require.

7 California Election Code, §15360 -- https://codes.findlaw.com/ca/elections-code/elec-sect-15360.html
8 Florida Statute §101.591 "Voting System Audit" -- http://www.leg.state.fl.us/statutes/index.cfm?App_mode=Display_Statute&Search_String=&URL=0100-0199/0101/Sections/0101.591.html

Key guidelines for batch comparison audits:

(a) all batches in the election should be subjected to the random selection process, with about the same likelihood of being selected,

(b) the random selection process should insure that the batches selected are a complete surprise,

(c) the batches of ballots should be kept in the secure chain of custody and not handled prior to the manual tally,

(d) the full set of computer results, broken down by batch, should be frozen prior to the random selection and manual tally, and

(d) the number of batches should be greater than the number of scanning machines so as to check them all. In some counties (San Diego, Santa Barbara) the number of scanners is far greater (10x) than the number of batches sampled so there is no way to cover all the potential vulnerabilities.[9]

This audit as implemented in CA and FL is not strong enough to guarantee detection of even some very extensive hacks to any significant degree from a statistical standpoint.

Strength of Batch Comparison Audits

The probability of catching a hack depends on how many precincts are affected by the hack. Assuming a contest with two options, that contest would need to have a close margin for any hack to be feasible, with nearly 50% of the votes already cast for the desired option. The votes that could be affected in any precinct would be the other (just over) 50% of the votes for the undesired option. Any hacker would probably modify only about 10% (i.e. 20% of the remaining margin) or it would be considered too obvious to attempt. In any case, the hack must be spread over a number of precincts.

If we consider the case of a district with only 100 precincts (so the result can be extrapolated to larger districts), and assuming a 1% batch comparison audit, then one precinct will be chosen for the audit, and if the precincts are about the same size, then each precinct is about 1% of the total number of votes. As mentioned, it is our opinion that at most 10% of the total ballots could be affected in any one precinct by a hack that would not be obvious. If those votes are flipped in the precinct (add one vote to the desired candidate while subtracting one from the other), then it doubles the effect in that one precinct, and it is like moving the election by 0.2% for each precinct included in the hack. Considering a total move of 5%, then 25 precincts (5/0.2) would be required to implement the hack. Choosing one precinct out of the 100 means there is a 25% chance that one of the 25 hacked precincts would be chosen, and the hack therefore detected. This is far lower than the 95% confidence level normally

9 In San Diego and Santa Barbara, they have used the Diebold precinct scanner but not at the precinct. These are used in the central office and the ballots from precincts are transported to the central office by precinct workers on election night. In San Diego in the 2016 primary election, they set up 160 scanner machines and each one scanned 10 precincts. But the 1% manual tally only requires that 16 precincts be chosen. Thus, only about 10% of the machines will be subject to the audit if the random selection happens to choose precincts processed by different machines. They do not attempt to optimize the selection process to insure that different machines are tested.

considered a goal in other sampling scenarios. To reach that level of confidence would require that more than 10% of the precincts included instead of 1%. Some of these assumptions could change but no matter what, sampling 1% of the precincts provides a low confidence level to detect this type of hack.

Game Theory Issues

Unlike truly random processes, however, the attack vectors are not truly random, as the behavior of the hacker is dependent on whether he thinks he might be caught. If the audit is conducted correctly, it would present a risk that the hack would be detected. With this fact known by the fraudster, he would likely not attempt the hack at all. For this reason, an audit process may be successful in thwarting attacks even if the risk of being caught is not as significant as would otherwise be necessary if the process that produces the errors was purely stochastic in nature (and does not have any understanding that it might be caught). Most manufacturing production strategies include the assumption that the cause of the error is either random or periodic in nature, and is not an employee that is trying to cause failures.

This is a topic in the mathematical field of game theory, and to date, I have not seen the election process analyzed using this framework. The typical analysis is strictly statistical without any consideration of how even a weak audit will affect the likelihood of a hacker attempting to modify the election.

As defined, this type of audit (as implemented in CA) does not escalate even if discrepancies are found, and the actions required of the election officials if any discrepancies are found is not clearly defined. Intense public scrutiny can help keep the officials honest. These drawbacks could be partially rectified by implementing procedures which would require escalation to include more batches to be tallied if the result was very close and/or if a significant number of discrepancies were detected. And if the subject is properly analyzed with game theory included, we may find that these are better than originally thought.

There are other variations of these audits. In Florida, each county chooses only one race, and performs a batch-comparison audit on 1% of the precincts related to that race. Unfortunately, they do not perform any audit at all if they also perform a "manual recount of overvotes and undervotes" of <u>any</u> race in the county. This audit process is very weak and is essentially nonexistent if an "automatic recount" occurs.

9. Ballot Image Audits (BIAs)

The other major approach to auditing is based on using secured Ballot Images. This option does not exist unless ballot images are produced, but this is becoming more common all the time. As we will show, the creation, securing and saving ballot images is a reasonable and prudent goal. Images should be created without lossy compression, in simple formats to avoid hidden "metadata" fields that can contain any additional information.

Once you have validated the images as described below, they can be used for just about any audit approach, including a ballot comparison RLA, or a 100% independent recount audit. Ballot images should be secured using block-chain style security i.e. secure hash message digests which are published and signed.

There is a hazard that an insider could modify the CVR and ballot image, but may not have access to the ballots themselves. The only way this can be done is to modify the ballot images before they are secured, and then the CVR is generated from the modified ballot images. This possibility can be reduced by comparing the ballot images with the paper ballots to confirm they are an accurate representation. This should occur after the ballot images are secured per the Technical Brief – "Block-Chain Style Cybersecurity For Digital Ballot Images" And as it turns out, validating ballot images is slightly better if done after any deterministic process (such as a 100% retabulation) that would alter the margin of victory, so that the correct sample size is used.

Image Validation

To validate images, images are compared with the chain-of-custody secured paper ballots. The number of ballots that must be sampled is related to the narrowest margin of victory in any race, and the level of certainty we wish to have. The certainty is (1.0 - risk) that we would improperly accept as valid the images when in fact they contain a hack of sufficient size to flip the narrowest race. We assume the most efficient hack, i.e. a 2-vote overstatement (flip of the race by moving the vote from the undesired (winning) option to the desired (losing) option). This is most efficient hack because the fraudster can modify the outcome by modifying the least number of ballots. Any other modification that alters the result will be less efficient, will require altering more ballots, so those will also be detected if we can detect the most efficient hack.

The question then is: How many ballots must be sampled before at least one of the modified images in the most efficient hack is selected? To calculate this, we must know that the diluted margin is the (race specific margin) * (faction of ballots that include that race).

This is best answered by considering the probability of continuously not selecting a modified image.

Assuming that x% of the ballot images have been modified so they do not match the original ballots, then the probability of not selecting one of those is 1-x. If we do that over and over, then we multiply that probability each time, (1-x)*(1-x)*(1-x)..and so on or (1-x)**n. With every sample, the chance that we will continue to not hit one of the modified images will be reduced accordingly. When (1-x)**n = risk, then n is the minimum number of samples needed.[10]

10 To explain this for an intuitive understanding, consider a fair coin. How many flips will it take so that 99% of the time, you will see both heads AND tails? The first flip, you will get head or tails with 100% probability, let's assume we get heads. Now, we consider how probable it will be to continue to flip the coin and get heads over and over, until the probability of getting that series is less than 1%. The first flip, we have a 50% chance of getting heads again. Flip again, and the chance is 50% each time, but to get them in series, we multiply the probabilities, 0.50*0.50 =0.25 chance of flipping two more times, then 12.5% (3), 6.25% (4), 3.125% (5), 1.5625% (6), 0.78125% (7). So it took seven more flips. Using Equation [1], n = CEIL (log (0.01) / log (1-.5)) = CEIL (-2/-0.3) = CEIL (6.64) = 7.

If one ballot is modified, it could affect the margin by twice that amount, because both candidates could be modified (add one to the lower and subtract one from the winner). Thus, for this application, x = margin/2. First solving for n, and rounding any fraction up.

risk = (1-(margin/2))**n	
LOG(risk) = LOG((1-(margin/2))**n)	take log of both sides.
LOG(risk) = n*LOG(1-(margin/2))	pull out exponent
n = CEILING (LOG(risk) / LOG(1-margin/2))	solve for n and round fractions up. [1]

This produces the curves shown in Figure 21 provides the number of samples required vs. the margin using linear scales (deemed appropriate because sampling the ballots does require a given amount of work.)

The best way to perform the validation sampling is to divide the number of ballots to be reviewed by two and select the ballots in two different ways. For the first set, randomly select a ballot image, and then using the manifest, look for the matching ballot among the physical ballots and pair it up. For the second set, randomly select physical ballots without using any manifest or computer report, and then attempt to locate the corresponding image using any information available. This approach covers the cases when a ballot is scanned twice (and two records would exist in the image set but not in the physical set) or not scanned at all (not in the image set but found in the physical ballot set.) Each of those two possibilities can affect the total by only one vote, and so we need to sample them half as often as the flipped-vote case to discover if they are prevalent enough to modify the outcome, because there would need to be twice as many ballots affected.

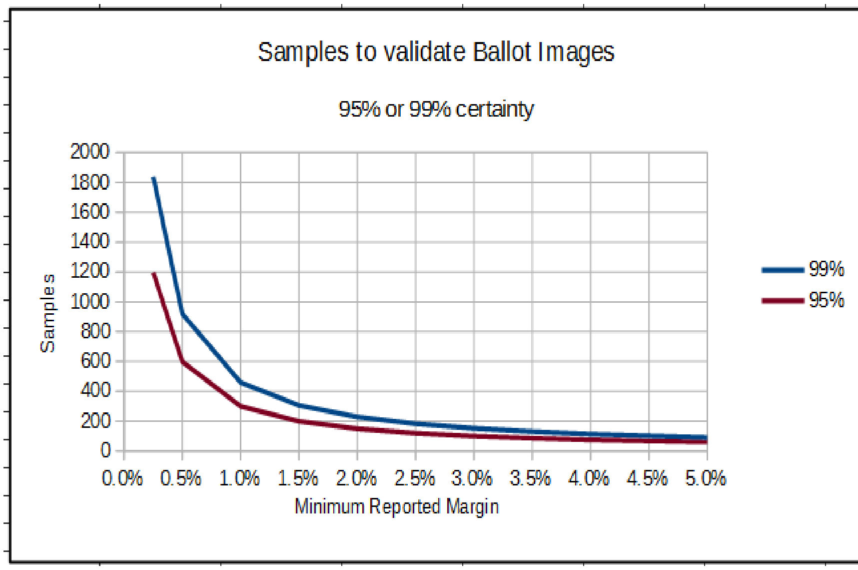


Figure 21: Samples to validate ballot images

The process of selecting and comparing ballot images with physical ballots should be public and documented with recorded side-by-side comparisons, so it will not be feasible for compromised officials to cover up ballots hacked with at least the minimal hack.

It is considered mandatory to have a unique identifier on the physical ballot that can be compared with the image to confirm that the correct image is being matched with the correct ballot. In a perfect world, if even a single case is encountered where the ballot image does not match the ballot and the image has been found to be modified, inserted or deleted, then indeed we do have a problem as that should never occur, and may be the sign of a hack.

Yet, there is a chance that the image is of the correct ballot but is not a fair representation and then this would indicate likely a scanner or compression issue, or that some ballots are missing from the ballot set (misfed) or a ballot accidentally scanned twice. As the ballot images are compared with the ballots, if the image does not match the ballot, an evaluation is required to determine if the ballot was modified or if the image was modified. If it does appear that the image was modified, this would cause a root-cause analysis to determine how the images were modified and by whom.

For election systems that base their interpretation of the ballots on digital image processing of ballot images, an altered image is an extremely horrendous event to occur, and it means the entire tabulation must be questioned. The policy likely should be that the entire tabulation should be redone from scratch, of course after making corrections so the fraud or severe mistakes could not recur.

Image Validation Requires less work and may be automated

The number of ballots that must be examined to simply validate the ballot set is similar to that in the various ballot comparison RLA approaches, but in those cases, there is an expectation that there is some variation in how the voter intent is interpreted. Just because a system misinterprets voter intent does not mean there is sufficient cause to be immediately suspicious that a hacker or fraudster has compromised the election. We expect a certain amount of "noise" where the Reported CVR does not match the Actual CVR after human interpretation.

This is not the case with the ballot images. There is no tolerance for any variation, save equipment faults (such as misfeeds or double feeds). The images should be identical and match for every ballot.

Therefore, because of this clearcut go/no-go criteria, the number of ballots required to validate a ballot image set will always be less than a ballot comparison audit which relies on a cumulative number of errors (overstatements) before there is any concern.

Auditing the tabulation using Secured Images

Validating ballot images can be done before or after the images are used to conduct a risk-limiting audit based on the images, or a 100% re-tabulation of the election results -- a deterministic process -- rather than a statistical audit. When margins become very tight, statistical audits become too unwieldy as they require many manual steps and many ballot samples. Ballot image audits can be automated and replicated by several competing outsiders. The most we can do with statistical audits is to watch them very carefully and hope election officials are not "fixing up" as they go (and carefully designed procedures can indeed limit the risk of such innocent fix-up.

Ballot image audits (BIAs) are less subject to inadvertent and innocent "fix-up" by election officials and employees. This, in itself, may be sufficient justification to warrant a serious look at BIAs versus the statistical audits that have many manual steps involved, and are very complex procedures unto themselves. It is hard to know if you are correctly correcting the audit process or improperly correcting the data that is being audited.

On the other hand, the downside is that validating images does mean that you have to be able to access both the physical ballot and image for that ballot, similar to the S4RLA audit.

The ballots could be reviewed "by hand" by looking over the images, which is comparable with the hand count mentioned, but can be done with many people on the internet in a crowd-sourcing arrangement that can be very satisfying to the public. Or the images could be compared with an auditing program by an independent auditor.

It is most likely, however, that the secured ballot images will be regarded as much more reliable than the paper, because the paper can be easily modified by someone with a pencil, lost or destroyed. Properly secured ballot images cannot be modified without detection, and once published, impossible to delete.

BIA supports "Divide and Conquer" and "Test Early, Test Often"

The two key tactics to help test just about anything are encapsulated in the phrases, "Divide and Conquer" and "Test Early, Test Often."

One key shortcoming to the ballot-sampled Polling RLA or Ballot Comparison RLA is that they do not attempt to follow these tactics. The entire ballot set is treated as one random pool, and there are no steps that divide and conquer. And as a result, they do not assist in determining what exactly is wrong -- no diagnostic hints are provided.

The ballot pool is actually not one large vat of marbles. The ballots are quite structured in their organization, into precincts, batches, districts, counties, etc. This structure is largely lost or suppressed by fully statistical procedures.

The approach which validates ballots images, and then applies another audit technique to the validated samples divides the testing into two phases. When such a split occurs, it provides two benefits. First, each step is simpler and is easier to understand. Secondly, the split allows diagnosis of the problem.

Once the images have been validated, attack vectors based on modifying, adding or subtracting the images have been excluded. After that point, the images can be relied upon (within the risk parameter specified). If the image does not match the CVR, then we know the CVR has been modified, rather than the image modified. Thus, a split in testing provides more information that is useful for diagnosis.

Non-validated Ballot Images are Still Useful

If secured ballot images are available but they have not been validated by comparing with the paper ballots, can they be used to audit the election? Of course they can, but there is an increased risk. And because doing a full 100% ballot image audit is not a statistical process, the increased risk is probably less than the risk already tolerated by statistical procedures.

As mentioned, using ballot images splits the problem into two parts. If the ballot images are not validated, that means they have not been compared with the paper, and there is some (small) chance that a hacker may have altered ballot images prior to the creation of the CVR. It will not catch a hack consisting of modifying ballot images after they were created and before they were secured. The CVR will be created based on the altered ballot images, so there will be no difference between the CVR and the Ballot Images in this case. The hack could be detected with any check that compares paper with CVR or Images.

Auditing ballot images means we are comparing them with the reported CVR set. This will detect any hack that modifies the CVR after the secured ballot images are created, which is largely the entire set of central-tabulator hacks which are possible today. The additional hazards created by creating secured ballot images is far fewer than the hazards which existed without secured ballot images, and this is a net benefit.

Indeed, it would be quite difficult and hazardous for any compromised insider to modify the ballot images in the tiny window of time between image creation and image security. With administrative controls, such as improved procedures, this window can be minimized to a point where ballot images can be relied upon for the range of margins where hand counts are not called for, and image validation may be skipped without a huge increase of risk. And, as mentioned, the increase in risk is probably far less than the risk already tolerated by statistical procedures. BIAs do not have the large increase in manual overhead as do statistical audits, and so they are the only kind reviewed so far, that can easily deal with sub-2% margins.

10 The Open Ballot Initiative (Formerly, Open Canvass)

One auditing approach that relies on the existence of secured and validated ballot images is based on the notion that the full set of ballot images can be distributed to several competing groups that will generate their own independent tabulation, creating a full set of cast vote records (CVRs) which can then be easily compared to discover where the disparate groups disagree on the results. This we have called the Open Ballot Initiative (TOBI) because it is based on the concept that once validated, the ballot image evidence can then be easily available for review by anyone.

TOBI suggests that the CVR sets generated by the various third parties are compared ballot by ballot (likely by machine) with the official CVR result (if available) to create a result in a standardized format. This is simply comparing a two CVR sets, in essence comparing two large tables, and it is something computers are especially good at (and can also be easily spot-checked manually). (If the CVR is not available, then the competing third parties would compare with each other.)

Any ballots where the competing tabulations disagree can be flagged so those can be reviewed in much more detail. Indeed, at some point, the corresponding physical ballots may be consulted and compared with the images as each can provide a secondary check to each other. We don't need to review all the ballots, only the ballots where the different parties disagree in terms of voter intent or how the vote was extracted from the image. And, each may wish to review paper as well not only to confirm the images that are in question, but also to confirm the set of images as a whole.

This sort of comparison is not a statistical process, it is a rigorous, 100% review of all the ballot images by different parties using their own software and algorithms for discerning voter intent. So in the end, the only risk factor is the underlying risk that the images are compromised, and that can be minimized by image validation with a minimum of ballots compared.

It is likely that the various competing parties will come up with differing interpretations for some of the ballots. The ballots where voter intent was interpreted differently by the different parties and their different software versions will be some relatively small number of ballots X. If X < half the smallest vote margin, then even if those were all interpreted correctly by exhaustive review, there is no way the that interpretation can change the outcome.

Unlike RLAs, TOBI provides a mechanism for improvements to the voter-intent interpretation heuristics. Over time, the various competing parties -- including the election officials -- can improve their voter-intent algorithms and will have fewer discrepancies. Statistical approaches like PRLA or CRLA do not have this beneficial characteristic.

The fact that TOBI provides the ballot data to other parties provides outsider review of the election so that no insider can cheat. Traditional self-audits which rely on the election officials to also honestly report mistakes require intense scrutiny to insure honest reporting. Independent processing is, in essence, extremely intense scrutiny of the election.

Such audits avoid the tendency for insiders to innocently "fix-up" discrepancies in the audit. Such fix-up can be a killer to statistical audits, as a very few discrepancies will tilt the scale one way or the other. This hazard does not exist in BIAs.

It is interesting to note that if this methodology is used, there is no need to worry about whether the source code of the code providing interpretation of the ballot images is "open-source" or proprietary. The multiple-party comparison process eliminates this from concern, as long as the comparison actually is done, and image validation is done properly and openly.

Although really any audit process can be performed on ballot images, these audits can be done automatically. The most important substitute is for the "full hand count," which is frequently mentioned by other audit procedures as the way to deal with elections that are too close to easily discriminate statistically.

Automating audits based on ballot images will require that the various (sometimes 100s) of different ballot styles are correctly dealt with. A given race may be located in different places on the ballot and in different languages. Finding these and then recognizing voter intent properly can be a fairly large challenge. But today, given that computers are driving cars, we are confident that these issues are not show stoppers.

11. Comprehensive Risk Estimation

Most of the "Risk Limiting Audit" calculations that estimate risk focus on a very specific and limited scope while relying on low or zero risk for other steps. These methods sometimes suggest that they will limit the risk that the election outcome is incorrect to some specific value, sometimes shrouded in law. Unfortunately, the calculations presented frequently do not provide a comprehensive calculation and underestimate the actual risk.

In any of these audits, the risk is based on a calculation of samples drawn at random from the paper ballot evidence of the election. The ballot evidence must be reliable -- and the risk calculations assume that the ballot evidence is pristine and that the sampling of the ballots is complete and trustworthy.

Of course, those notions are hardly ever real outside the ideal notions used by mathematicians. Instead, we must assign a risk to each step of the process. Or better, a confidence factor, which is (1 - risk) for that stage. The total confidence is the product of all the confidence factors. We obtain an equation of probability similar in form to the famous "Drake Equation" which is used to estimate the likelihood that we might interact with extraterrestrial life.

<table>
<tr><th>Risk Step</th><th>Ballot Polling Audit</th><th>Ballot Comparison Audit</th><th>Batch Comparison Audit</th><th>Ballot Image Audit</th></tr>
<tr><td>Risk that ballots are modified / added / deleted prior to scanning</td><td colspan="4">All processes rely on a robust chain of custody prior to scanning the ballots. Audit procedures should review the number of ballots cast at polling places and by mail to ensure all ballots are included and invalidated ballots are minimized.</td></tr>
<tr><td>Risk that images are modified / added / deleted after scanning ballots and prior to securing images.</td><td colspan="3">Images are not used directly but may be used to reduce risk that ballots are modified prior to sampling or full-hand count is used.</td><td>Non-zero risk can be reduced by comparing images to paper, and reduced by minimizing window of opportunity</td></tr>
<tr><td>Risk that ballots will be modified prior to audit sampling, sample</td><td>Non-zero -- sample could be preselected and precounted to</td><td>non-zero - if unmodified ballots are provided as samples, hack will</td><td>non-zero - may modify only some batches and provide</td><td>0% - does not rely on paper ballots or sampling</td></tr>
</table>

Risk Step	Ballot Polling Audit	Ballot Comparison Audit	Batch Comparison Audit	Ballot Image Audit
manipulated to provide desired results	cover up hack.	not be detected.	unmodified batches for audit	
Requires frozen CVR set	No. 0% risk but does rely on reported margin	Yes	Yes	Does not need a CVR but may want to compare with the official CVR. 0% risk
Random Selection	Requires oversight of random number procedure			0% risk
Drawing samples	Requires robust observation, may have 100% risk if done without observation		Pulling of audited batches is simple as they can remain in sealed boxes until audited	0% risk
Reliant on a ballot manifest	Usually, and if so, non-zero risk	Yes - non-zero risk as manifest may not include all ballots	no - batches are audited as stored	no - 0% risk
Data entry of ballots may be manipulated by custom software	non-zero if custom DRE-like software is used	non-zero if custom DRE-like software is used	non-zero if custom DRE-like software is used	0% as no data entry is required as images are used directly.
Compatible with secured ballots due to possible judicial contest	Non-zero risk - Audit may not be able to be performed by court order		Yes - batches need not be compromised	Yes -- Image audit can be performed without obtaining access to ballots
Compatible with third party and competitive auditing	No	No	No	Yes
Relies on expensive full hand count if audit finds problems	Yes - 1% to 2% risk in full hand count		No -- generally there is no escalation.	No. Audit is deterministic and has lower risk than full hand count.
May confirm a hacked election	Yes. This is the normal risk cited in Risk-Limiting Audit procedures		Yes, may miss hacked election about 75% of the time.	0%. Deterministic procedure does not include risk at this stage.

All risk factors should be included in the overall risk for any procedure.

12. Economic Comparison

One thing that is largely unaddressed is the comparative costs of performing these various types of audits. This section will attempt to understand this.

Manual Tally Cost - Batch Comparison Audit

We have good data on the time it takes to perform the 1% Manual Tally as performed in California. Generally, these times do not include the overhead of accessing the batches, as those are pulled prior to the start of the manual tally process. It is very important for the ballots to be sorted by precinct to implement the 1% manual tally and also to implement any recounts[11] which may be appropriate. If you don't have the ballots (at least partially) sorted by precinct, then you can't easily find the ballots that include the race of interest.

The manual tally itself can be conducted using the read-and-tally method, with teams of three people, one reader and two talliers. The result of the tally is provided to the supervisor who compares it with the computer result. If the tally result matches the computer, then they stop. Otherwise, it is re-tallied by other teams to determine if the computer result is incorrect. The total time elapsed shown in Figure 22 and on average, is just under 15 seconds per ballot-contest, and nearly 4 minutes per ballot on the average to tally all contests on the ballot. These times do not include overhead such as breaks, meal periods, etc. This table only reflects the times for tallying the Polls Ballots[12], and they do not include any VBM ballots nor provisional ballots.

Studies have shown that for ballots with many races, the read-and-tally method is faster and more accurate than the sort-and-stack method[13]. But if you are doing only one race, the sort and stack method could be faster. To use that, the ballots are sorted into "n" stacks, each representing the vote for each of the "n" ballot options. Then, the stacks are counted. The count should total to the total number of ballots.

11 To be clear, the term "recount" used here refers to the official recount process as defined by the Secretary of State and not part of the audit process. Usually, even if the ballots are not perfectly sorted by precinct, they are sorted down to perhaps 32 larger districts so that a given race to be recounted does not have to involve all the batches.

12 Polls Ballots are those that are completed in-person at polling places and are naturally sorted by precinct. The other three groups of ballots are the "Early VBM", "Later VBM", and "Accepted Provisional" ballots. The Early VBM ballots are those that are received and processed prior to election night, while Later VBM ballots are completed after election night. In CA, there is a significant (up to about 40% of all ballots) still to be processed after election night.

13 Goggin, Bryne, Gilbert, "Post-Election Auditing Effects of Procedure and Ballot Type on Manual Counting Accuracy, Efficiency, and Auditor Satisfaction and Confidence" http://www.copswiki.org/Common/M1725 (2012)

seq	precinct cons	contests per ballot	ballots	time (hrs)	time minutes	mins per ballot	total contests	mins per contest-ballot
32	110150	26	150	9.58	575	3.83	3900	0.15
363	237200	25	182	18.16	1,090	5.99	4550	0.24
368	240000	26	203	13.33	800	3.94	5278	0.15
418	270510	27	281	37.50	2,250	8.01	7587	0.30
597	376700	27	226	12.42	745	3.30	6102	0.12
637	403500	12	138	6.50	390	2.83	1656	0.24
670	404230	11	70	4.25	255	3.64	770	0.33
686	405400	13	148	7.67	460	3.11	1924	0.24
857	420520	12	150	11.42	685	4.57	1800	0.38
877	423900	11	47	2.17	130	2.77	517	0.25
991	442800	16	159	5.75	345	2.17	2544	0.14
1229	487000	15	191	15.35	921	4.82	2865	0.32
1332	528200	12	137	8.83	530	3.87	1644	0.32
1418	538500	13	128	7.5	450	3.52	1664	0.27
1431	546600	13	92	3.75	225	2.45	1196	0.19
1454	549280	14	123	5.67	340	2.76	1722	0.20
					AVERAGE:	**3.85**		**0.24**

Figure 22: Actual data of time cost of 1% manual tally for Polls ballots in San Diego 2016 primary

Practiced teams can do a better job than new workers. We note that one precinct of 281 ballots took 37 hours to count because it had to be passed to other teams. For those people who are proponents of hand counting all ballots at the precinct after the election is over, we remind you that this is a very tedious process and actually very difficult. If we can use machines to help us perform this work, and still be confident of the results, this is the goal of the audit procedures.

Also, there is another factor. The 1% Manual Tally in CA is a batch-comparison audit. It compares the result with the computer report. If you do not have the computer report to rely on, then it is about twice as costly to do the tally, because you need to tally each precinct at least twice by separate teams to reduce the likelihood of error.

What is surprising about this process is how difficult it is for humans to perform because of the sheer boredom of the work. Yet, these data provide a very good starting point. Unfortunately, we do not have solid data on the other procedures so we will make good-faith estimates.

Polling RLA Cost

The PRLA must randomly select physical ballots and tally the result. Thus, the cost of pulling the ballot will be on top of the 4 minute cost to tally any individual ballot. For sake of comparison, we will assume that the ballots can each be accessed within ten minutes[14], and then each ballot tabulated in 4 minutes, for 14 minutes total[15].

Ballot Comparison Audit Cost

A ballot comparison audit requires that the physical ballot be paired up with the CVR record for that ballot and then tallied. So we must find not just a random physical ballot but a specific ballot, and then tally it. Total time probably 20 minutes to get both the physical and CVR ballot and 4 minutes to tally and compare. There is an additional cost to determine if the process must continue or if it can stop. We will include another minute for that, for 25 minutes total.

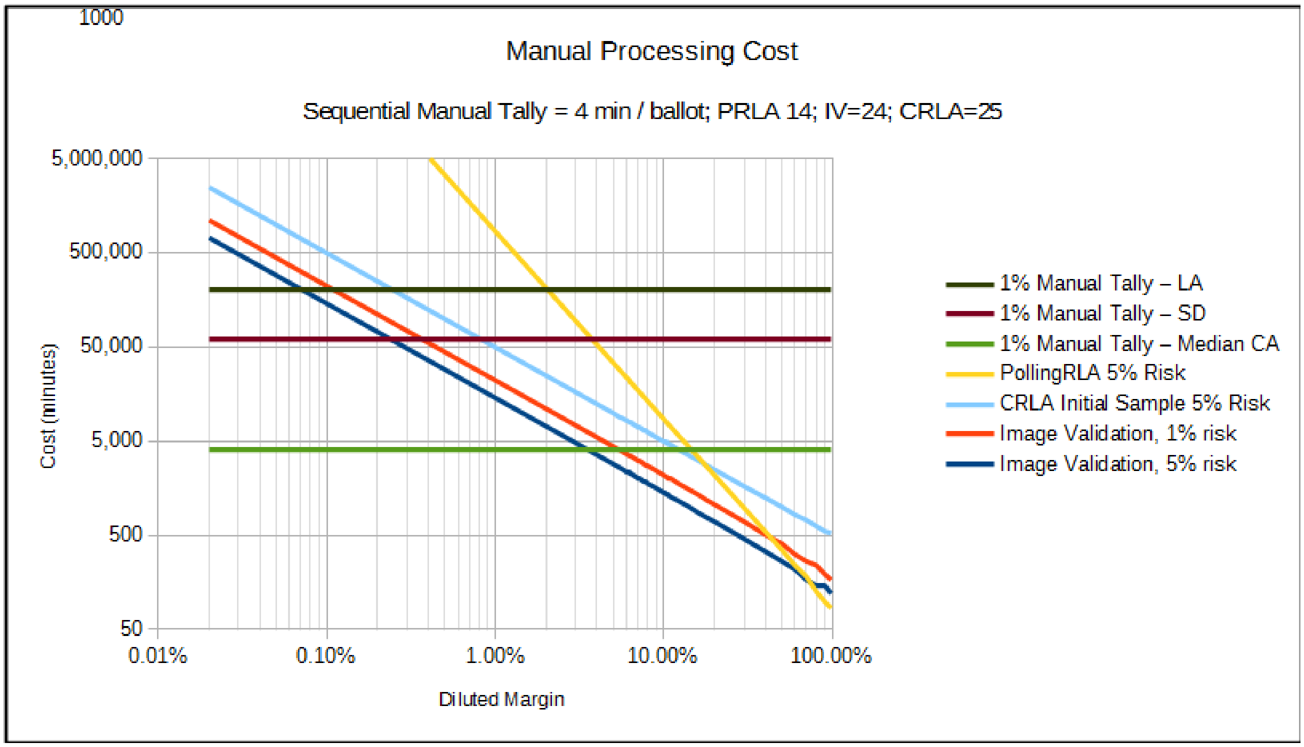


Figure 23: Economic Analysis of Auditing Methods - THIS NEEDS TO BE UPDATED

Image Verification Cost

The cost to validate images will be similar to the cost included in the Ballot Comparison Audit, but it will not take 4 minutes to tally the ballot as it need only be compared to determine that they are the exact same ballot. Usually the style of marking in hand-marked ballots can assist in this comparison. Comparing the vote on the ballot need not be part of the comparison process. Thus, we will estimate

14 Actual data regarding the time required to access physical ballots is not available as of this publication. Any source of such real data is fondly appreciated

15 The report of the Orange County pilot audit conducted in June, 2018, reported the entire cost of auditing five races on 160 ballots pulled at random was estimated to be $4,000 ($25 per ballot). The 14 minute estimate is within the range of credible values given this single data point.

that IV will take the same time as the BCRLA except for the 1 extra minute added for the complexity of the BCRLA process.

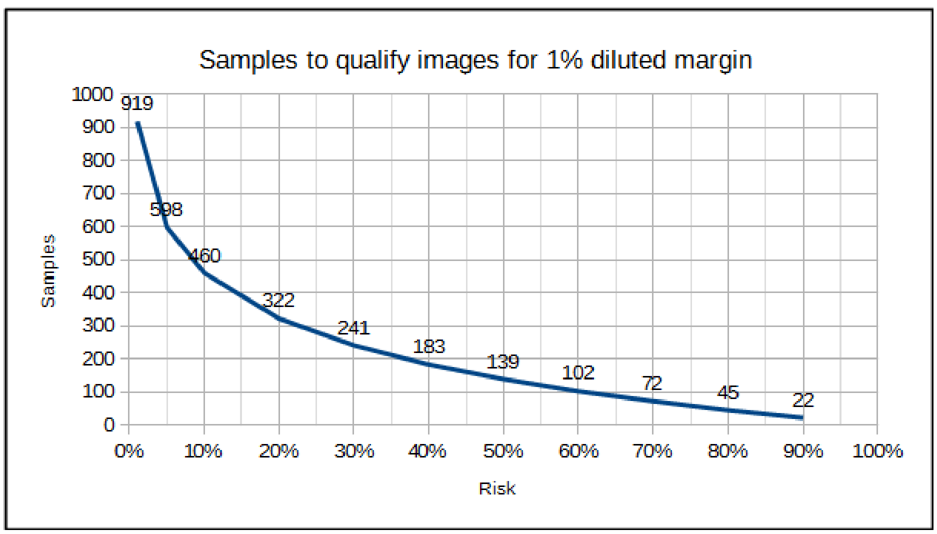
One method for validating ballot images is to randomly pull them, then re-scan them using an off-the-shelf scanner, and use a comparison program to compare the ballots. They will not compare bit-for-bit, but we can compare them on an image basis and likely do a very good job of finding serious problems. Remember, we are looking for image that have the votes flipped, so they are substantially different, but for humans, this difference may be difficult to find unless the votes are extracted and compared.

Figure 23 provides a chart to allow visual comparison of these methods. The cost to perform the manual tally is based on the actual data for a sequential manual tally. However, if we had just one race that was so close that it needed to be tallied, it might roughly take about the same amount of time as the 1% manual tally for all races. So roughly speaking, we will consider that these levels also represent the cost to tally one race.

We show the median CA county, which has about 100,000 voters, and San Diego, which is the #2 county in the state (and #6 in the nation) with 1.5 million voters, and Los Angeles County, the largest district in the country with about 5 million voters[16].

13 Combined Strategy

As mentioned earlier, the Ballot Comparison and Image Verification (IV) curves regarding the samples required are related only to the margin of victory and not the actual size of the district. Essentially, if those curves are above the line corresponding to the manual count cost for that county (for one race) then, a hand count will be less expensive. Considering Ballot Comparison Audits, half the counties in the state should trigger a full hand count of any race with diluted margin under 10%, because the overall cost to perform a Ballot Comparison Audit is higher than just doing the hand count for those races. If multiple races exist that are close, then a statistical audit may be economical for slightly tighter margins. Importantly, about 90% of contests have margins of victory over 10%, so these audits are cost effective for the vast majority of contests.



16 The largest districts in CA do not faithfully perform the 1% manual tally on all strata as they tend to omit the Later VBM and Provisional Ballots to reduce the cost.

Then, on races that have margins closer than 10% (polling audits) or closer than 3% (ballot comparison audits) should use ballot image audits.

Ray Lutz, MSEE, has worked in the document imaging industry, contributed to national and international facsimile standards, and is the founder of Citizens' Oversight Projects, a public interest group that has conducted oversight and reviewed election procedures.

More information: http://citizensoversight.org

Contact information:
Ray Lutz; raylutz@citizensoversight.org
619-820-5321