Risk-Limiting Vote-Tabulation Audits: The Importance of Cluster Size

Philip B. Stark

Department of Statistics

University of California, Berkeley

9 June 2010

Revised 7 December August 2010

Chance, *23(3)*, 9–12, 2010

1 Vote-Tabulation Audits

Post-election vote-tabulation audits compare hand counts of votes in a collection of groups (clusters) of paper records ("voter-verified paper audit trail," VVPAT) to reported machine counts of the votes in the same clusters. Vote-tabulation audits can serve a variety of roles, including process monitoring, quality improvement, fraud deterrence, and bolstering public confidence. All of these raise statistical issues.

This note focuses on audits that check whether the machine-count outcome

is correct. The outcome is the set of winners, not the numerical vote totals. The machine-count outcome is correct if it agrees with the outcome that a full hand count of the paper audit trail would show. Hand counts can have errors, but many jurisdictions define the correct outcome to be the outcome a hand count shows. Moreover, when the hand count of a cluster of ballots disagrees with the machine count, jurisdictions typically repeat the hand count until they are satisfied that the problem is with the machine count, not the hand count.

Generally the only legally acceptable way to prove that a machine count outcome is wrong—and to repair it—is to count the entire audit trail by hand. An audit that has a pre-specified chance of requiring a full hand count if the machine-count outcome is wrong—no matter what caused the outcome to be wrong—is called a *risk-limiting audit*. The *risk* is the maximum chance that there won't be a full hand count when the machine-count outcome is wrong.

2 The Role of Statistics in Risk-Limiting Audits

Statistics lets us reduce the amount of counting when the machine-count outcome is right, while ensuring that there is still a big chance of counting the entire audit trail if that outcome is wrong. Risk-limiting audits can be couched as hypothesis tests. The null hypothesis is that the machine-count outcome is incorrect. To reject the null is to conclude that the machine-count outcome is correct. A type I error occurs if we conclude that the machine-count outcome is correct when a full hand count would show that it is wrong. The significance

level is the risk.

It is natural and convenient to test the null hypothesis sequentially: Draw a random sample of clusters and audit them. If the sample gives strong evidence that the null hypothesis is false, stop auditing. Otherwise, expand the sample and evaluate the evidence again. Eventually, either we have counted all the clusters by hand and thus know the correct outcome, or we stopped auditing without a full hand count. We can limit the risk to level α by designing the audit so that the chance it stops short of a full hand count is at most α in every scenario in which the machine-count outcome is wrong.

The amount of hand counting needed to confirm that a correct outcome is indeed correct depends on the sampling design, the margin, the number of ballots cast, the number and nature of the differences the audit finds, and on the number of votes for each candidate in each of the clusters from which the sample is drawn. Clusters typically correspond to precincts or to precincts divided by mode of voting (e.g., votes cast in person versus votes cast by mail). We shall see that using smaller clusters can dramatically reduce the amount of hand counting when the machine-count outcome is right.

3 Heuristic Examples

3.1 Jelly Beans

We have 100 4-ounce bags of various flavors of jelly beans. Some bags have assorted flavors, some only a single flavor. Each 4oz bag contains 100 jelly beans,

so there are 10,000 in all. I love coconut jelly beans, and want to estimate how many there are in the 100 bags. The canonical flavor assay for jellybeans is *destructive tasting*, so the more we test, the fewer are left to share.[1] Consider two approaches.

1. Pour the 100 bags into a large pot and stir well. Then draw 100 beans without looking. Estimate the total number of coconut jelly beans to be the number of coconut jelly beans in the sample, times 100.

2. Select one of the 4oz bags at random. Estimate the total number of coconut jelly beans to be the number of coconut jelly beans in that bag of 100, times 100.

Both estimates are statistically unbiased, but the first has much lower variability. Mixing disperses the coconut jelly beans pretty evenly. The sample is likely to contain coconut jelly beans in roughly the same proportion as the 100 bags do overall, so multiplying the number in the sample by 100 gives a reasonably reliable estimate of the total.

In contrast, a single bag of 100 selected at random could contain only coconut jelly beans (if any of the bags has only coconut) or no coconut jelly beans (if any of the bags has none). Since the bags can have quite different proportions of coconut jelly beans, 100 beans selected the second way can be quite likely to contain coconut jelly beans in a proportion rather different from the overall proportion, and multiplying the number of coconut beans in that bag by 100

[1] Yes, I will share my jelly beans with you. After I pick out the coconut ones. And the banana. And ....

could have a large chance of being far from the total number of coconut jelly beans among the 10,000.

To get a reliable estimate by counting the coconut jelly beans in randomly selected bags, we would need to test quite a few bags (i.e., quite a few *clusters*), not just one. It's more efficient to mix the beans before selecting 4oz. Then 4oz suffices to get a reasonably reliable estimate.

Conversely, suppose that a sample of 100 beans drawn the first way contains no coconut jelly beans. We would then have 95% confidence that there are no more than 293 coconut beans among the 10,000. In contrast, if a sample drawn the second way contains no coconut jelly beans, we would only have 95% confidence that there are no more than 9,500 coconut jelly beans among the 10,000. To have 95% confidence that there are no more than 293 we would have to test at least 63 of the 100 bags, not just one bag: 63 times as many jelly beans as a simple random sample requires.

3.2 How salty is the stock?

We have 100 12-ounce cans of stock, of a variety of brands, styles, and types: chicken, beef, vegetable, low-sodium, regular, etc. We want to know how much salt there is in all 1,200 ounces of stock as a whole. The salt assay ruins the portion of the stock that is tested: The more we test, the less there is to eat. Consider two approaches:

1. Open all the cans, pour the contents into a large pot, stir well, and remove a tablespoon of the mix. Determine the amount of salt in that tablespoon,

multiply by the total number of tablespoons in the 100 cans (1T = 0.5oz, so the total number of tablespoons in the 100 cans is $12 \times 100 \times 2 = 2{,}400$T).

2. Select a can at random, determine the amount of salt in that can, and multiply by 100.

Both estimates are statistically unbiased, but the first estimate has much lower variability: That single tablespoon is extremely likely to contain salt in roughly the same concentration the 100 cans have on the whole.

In contrast, a can selected the second way can be quite likely to contain salt in a concentration rather different from the 1,200 ounces of stock as a whole, unless all the cans have nearly identical concentrations of salt.

For the first approach, we can get a reliable estimate of the total salt from a single tablespoon (0.5oz) of stock. But for the second approach, even 12 ounces of stock is not enough to get a reliable estimate. The first approach gives a more reliable result at lower cost: It spoils less stock.

To get a reliable estimate by sampling cans, we would need to assay quite a few cans selected at random. A single can is not enough, even though it contains 24 tablespoons of stock—far more than we need in the first approach. It's more efficient and cheaper to mix the stock before selecting the sample.

4 Connection to election auditing

A vote-tabulation error that causes the machine-count margin to appear larger than the true margin is like a coconut jelly bean or a fixed quantity of salt. A

precinct or other cluster of ballots is like a bag of jelly beans or a can of stock. Drawing the audit sample is like selecting a scoop or a bag of jelly beans or a tablespoon or can of stock.

Counting ballots by hand has a cost: The more we have to count, the greater the cost. Hence, we want to count as few ballots as possible as long as we can still determine whether the electoral outcome is correct—whether the number of errors is insufficient to account for the margin of victory. Similarly, testing the flavor of jelly beans or assaying the salt in the soup also has a cost. (Although I'd volunteer to determine the flavor of jelly beans, gratis.)

There are also costs for reporting votes in small clusters and organizing ballots so that those clusters can be retrieved, just as there are costs involved in opening all the bags of jelly beans and mixing them together, and in opening all the cans of soup and mixing them together. Reporting votes for small clusters of ballots can also reduce voter privacy.

In the food examples, the first approach is like auditing individual ballots or small clusters. All the ballots are mixed together well. A relatively small sample can give a reliable estimate of the difference between the machine counts and what a full hand count would show for the entire contest.

In the food examples, the second approach is like auditing using precincts or other large clusters of ballots. Many errors that increased the apparent margin could be concentrated in a small number of clusters, because there is no mixing across clusters. A single cluster drawn using the second approach doesn't tell us much about the overall rate of vote-tabulation errors, no matter how large

the cluster is (within reason). To compensate for the lack of mixing across clusters of ballots, we need to audit many clusters, just like we need to count the coconut jelly beans in many bags or assay many cans of soup if we don't mix their contents across clusters before drawing the sample.

5 Numerical Examples

Suppose we have 50,000 ballots in all, 500 ballots cast in each of 100 precincts. We will draw a random sample of 500 ballots to tally by hand to check against machine subtotals. Consider the three ways of selecting 500 ballots: (i) drawing a precinct at random, (ii) drawing 10 clusters of 50 ballots at random without replacement, and (iii) drawing 500 individual ballots at random without replacement (a simple random sample). Method (i) gives the least information about the whole contest; method (iii) gives the most, as we shall see. The smaller the clusters are, the harder it is to hide error from the random sample.

Suppose that for 1,000 (i.e., 2%) of the ballots, the machine interpreted the vote to be for the machine-count winner but a manual count would show a vote for the apparent loser. What is the chance that the hand count of the votes in the sample finds any of those 1,000 ballots? For method (iii), the chance does not depend on how the misinterpreted ballots are spread across precincts: It is about 99.996%, no matter what. But for methods (i) and (ii), the chance does depend on how many incorrectly interpreted ballots there are in each cluster. For simplicity, assume that when a precinct is divided into 10 clusters, the

misinterpreted ballots by precinct	randomly selected precinct of 500	10 randomly selected clusters of 50	simple random sample of 500
10 in every precinct	100%	100%	99.996%
10 in 98 precincts, 20 in 1 precinct	99%	∼100%	99.996%
20 in 50 precincts	50%	99.9%	99.996%
250 in 4 precincts	4%	33.6%	99.996%
500 in 2 precincts	2%	18.4%	99.996%

Table 1: Chance that a sample of 500 ballots contains at least one misinterpreted ballot in various scenarios.
There are 100 precincts containing 500 ballots each, and 1,000 of the 50,000 ballots (2%) are misinterpreted. Column 1: the way in which the 1,000 misinterpreted ballots are spread across precincts. Columns 2–4: the way in which the sample is drawn. Column 2: 1 precinct of 500 ballots drawn at random. Column 3: 10 clusters of 50 ballots drawn at random without replacement. Column 4: a simple random sample of 500 ballots. When a precinct is subdivided into 10 clusters, the number of misinterpreted ballots in those clusters is assumed to be equal.

number of misinterpreted ballots in each of those 10 clusters is the same. For instance, if the precinct has 20 misinterpreted ballots, each of the 10 clusters has 2 misinterpreted ballots.

Table 1 gives the resulting probabilities. They vary widely. In the case most favorable to precinct-based sampling, hand counting a single randomly selected precinct is guaranteed to find a misinterpreted ballot (10, in fact). But the chance falls quickly as the misinterpreted ballots are concentrated into fewer precincts. In the scenario least favorable to precinct-based sampling, the chance is only 2% for a randomly selected precinct and 18.4% for 10 randomly selected clusters of 50—but remains 99.996% for simple random sampling.

If misinterpretations are caused by equipment failures in precincts, that might concentrate errors in only a few precincts. If misinterpretations occur because pollworkers accidentally provided voters pens with the wrong color or

type of ink, that might concentrate errors in only a few precincts. If a fraudster tries to manipulate the outcome, he or she might target the ballots in only a few precincts, either to avoid detection or for logistical simplicity. In these three hypotheticals, if the sample is drawn by selecting an entire precinct it could easily be squeaky clean. But with the same counting effort, the chance of finding at least one error if the 500 ballots are drawn as a simple random sample remains extremely high, 99.996%, whether the misinterpreted ballots are concentrated in only a few precincts or spread throughout all 100.

Even when the sample does find some misinterpreted ballots, the percentage of such ballots in the sample can be much lower than the percentage in the contest as a whole. As before, suppose that for 1,000 (i.e., 2%) of the ballots, the machine interpreted the vote to be for the machine-count winner but a manual count would show them to be for the apparent loser. What is the chance that the percentage of misinterpreted ballots in the sample is at least 1%?

Table 2 gives the answers for the same set of scenarios. In the situation most favorable to precinct-based sampling, hand counting a single randomly selected precinct is guaranteed to reveal that at least 1% of the ballots were misinterpreted (in fact, it will show that 2% were). But the chance falls quickly as the misinterpreted ballots are concentrated into fewer precincts. In the case least favorable to precinct-based sampling, the chance is only 2% for a randomly selected precinct and 18.4% for 10 randomly selected clusters of 50—but remains 97.2% for simple random sampling. Using smaller clusters increases the chance that the percentage of misinterpreted ballots in the sample will be close to the

misinterpreted ballots by precinct	randomly selected precinct of 500	10 randomly selected clusters of 50	simple random sample of 500
10 in every precinct	100%	100%	97.2%
10 in 98 precincts, 20 in 1 precinct	99%	∼100%	97.2%
20 in 50 precincts	50%	62.4%	97.2%
250 in 4 precincts	4%	5.7%	97.2%
500 in 2 precincts	2%	18.4%	97.2%

Table 2: Chance that the percentage of misinterpreted ballots in a sample of 500 is at least 1% in various scenarios
There are 100 precincts containing 500 ballots each, and 1,000 of the 50,000 ballots (2%) are misinterpreted. Column 1: the way in which the 1,000 misinterpreted ballots are spread across precincts. Columns 2–4: the way in which the sample is drawn. Column 2: 1 precinct of 500 ballots drawn at random. Column 3: 10 clusters of 50 ballots drawn at random without replacement. Column 4: a simple random sample of 500 ballots. When a precinct is subdivided into 10 clusters, the number of misinterpreted ballots in those clusters is assumed to be equal.

percentage of misinterpreted ballots in the contest as a whole. Smaller clusters yield more reliable estimates.

Suppose the hand counts and machine counts match perfectly for a sample drawn in one of the three ways—no errors are observed. What could we conclude about the percentage of misinterpreted ballots in the contest as a whole, at 95% confidence? Table 3 gives the answers. For the same counting effort, the simple random sample tells us far more about the rate of misinterpreted ballots in the contest as a whole.

6 Discussion

Audits that have a guaranteed minimum chance of leading to a full hand count whenever the machine-count outcome is incorrect—thereby repairing the

randomly selected precinct of 500	10 randomly selected clusters of 50	simple random sample of 500
95.0%	25.7%	0.58%

Table 3: Upper 95% confidence bounds for the number of misinterpreted ballots for three ways of drawing 500 ballots at random, when the sample contains no misinterpreted ballots.
There are 100 precincts containing 500 ballots each. Columns 1–3: the way in which the sample is drawn. Column 1: 1 precinct of 500 ballots is drawn at random. Column 2: 10 clusters of 50 ballots drawn at random without replacement. Column 3: a simple random sample of 500 ballots is drawn. The bounds are obtained by inverting hypergeometric tests.

outcome—are called risk-limiting audits. The risk is the largest chance that the audit will not proceed to a full hand count when the machine-count outcome is incorrect. Risk-limiting audits can be implemented as sequential tests of the null hypothesis that the machine-count outcome is incorrect. The significance level is the risk.

I introduced risk-limiting audits in 2007 and conducted six field-pilots of risk-limiting audits in 2008 and 2009. In April 2010, the Board of Directors of the ASA endorsed risk-limiting audits and called for risk-limiting audits to be conducted for all Federal and statewide contests and a sample of smaller contests [10]. While 22 states have laws that require some kind of post-election vote-tabulation audit[2], none currently requires risk-limiting audits. California bill AB 2023, which recently passed the California State Assembly by a unanimous bipartisan vote, calls for an official pilot of risk-limiting audits in 2011. The ASA, California Common Cause, Verified Voting Foundation, and Citizens for Election Integrity Minnesota have endorsed AB 2023. AB 2023 will be heard by the California State Senate Committee on Elections, Reapportionment and

[2]Verified Voting, `http://www.verifiedvoting.org/`, last accessed 30 May 2010.

Constitutional Amendments in June 2010.

Reducing cluster size can dramatically reduce the hand counting required for risk-limiting audits. For instance, a 2009 risk-limiting audit in Yolo County, CA, audited a cluster sample of 1,437 ballots to attain a risk limit of 10%. Clusters were precincts, split by mode of voting (in person versus by mail). A simple random sample of just 6 ballots—about 240 times fewer—would have sufficed instead, if no errors were found [8].

There are tradeoffs: Using smaller clusters requires vote tabulation systems and procedures that report subtotals for smaller clusters, and it requires elections officials to be able to locate and retrieve the paper trail for those clusters. There is also a tradeoff between cluster size and voter privacy. If a group of voters can be linked to a cluster of ballots with similar voting patterns, one can determine how those voters voted.

The biggest impediment to efficient risk-limiting audits is the inability of current commercial vote tabulation systems to report the machine interpretation of small clusters of ballots or individual ballots. The next generation of vote tabulation systems should be designed with auditing in mind.

Acknowledgement. I am grateful to Mark Lindeman for helpful comments.

References

[1] J. L. Hall, L. W. Miratrix, P. B. Stark, M. Briones, E. Ginnold, F. Oakley, M. Peaden, G. Pellerin, T. Stanionis, and T. Webber. Implementing risk-limiting post-election audits in California. In *Proc. 2009 Electronic Voting Technology Workshop/Workshop on Trustworthy Elections (EVT/WOTE '09)*, Montreal, Canada, August 2009. USENIX.

[2] M. Lindeman, M. Halvorson, P. Smith, L. Garland, V. Addona, and D. McCrea. Principles and best practices for post-election audits. `www.electionaudits.org/files/best%20practices%20final_0.pdf`, 2008.

[3] L. Saldaña. California assembly bill 2023. `www.leginfo.ca.gov/pub/09-10/bill/asm/ab_2001-2050/ab_2023_bill_20100325_amended_asm_v98.html`, 2010.

[4] P.B. Stark. Conservative statistical post-election audits. *Ann. Appl. Stat.*, 2:550–581, 2008.

[5] P.B. Stark. Election audits by sampling with probability proportional to an error bound: dealing with discrepancies. statistics.berkeley.edu/∼stark/Preprints/ppebwrwd08.pdf, 2008.

[6] P.B. Stark. A sharper discrepancy measure for post-election audits. *Ann. Appl. Stat.*, 2:982–985, 2008.

[7] P.B. Stark. CAST: Canvass audits by sampling and testing. *IEEE Transactions on Information Forensics and Security, Special Issue on Electronic Voting*, 4:708–717, 2009.

[8] P.B. Stark. Efficient post-election audits of multiple contests: 2009 California tests. Technical report, Social Science Research Network, 2009. 2009 Conference on Empirical Legal Studies.

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

[10] American Statistical Association. American Statistical Association statement on risk-limiting post-election audits. `www.amstat.org/outreach/pdfs/Risk-Limiting_Endorsement.pdf`, 2010.