



Percentage-based
vs.
SAFE
Vote Tabulation Auditing:
A Graphic Comparison

[this page is intentionally blank]

Percentage-based versus SAFE Vote Tabulation Auditing: A Graphic Comparison

John McCarthy*, Howard Stanislevic**, Mark Lindeman***, Arlene Ash****, Vittorio Addona*****, Mary Batcher*****

Abstract

Trustworthy elections require comprehensive auditing and corrective action to eliminate major errors in counting votes. In this paper, we address just one component of electoral audits: specifying how many randomly selected precincts should undergo hand-counted audits to decide whether the winner determined by an electronic tally should be confirmed. Several pending electoral-integrity bills specify hand audits of 2% to 10% of all precincts. However, percentage-based audits are usually inefficient, because they use large samples for large jurisdictions, even though the sample needed to achieve good accuracy is much more affected by the closeness of the race than the size of the population. Percentage-based audits can also be ineffective, since close races may require auditing a large fraction of the total – even a 100% hand recount – to provide confidence in the outcome. This paper presents the SAFE (Statistically Accurate, Fair and Efficient) alternative to percentage-based sampling, based on the same statistical principles that inform audits in business and finance. In recent federal elections, highly reliable SAFE audits would have required about the same total effort and resources as the percentage-based audits now being considered. However, SAFE audits ensure high confidence in all electoral outcomes by using auditing resources more efficiently and employing large samples only when necessary.

Introduction

To verify election winners, Congress and several states are considering laws to require comparing machine tabulations with hand counts of paper ballots for randomly chosen precincts.¹ Since hand counts cost time and money, just enough precincts should be recounted to rule out election-altering miscounts, which may arise for a variety of reasons, including hardware malfunctions, unintentional programming errors, malicious attempts to alter election outcomes, or “undervotes” caused by ballot marks that interfere with correct counting. The key question is: *how many is enough for an adequate random sample?*

Electoral audits, like financial or manufacturing audits, are undertaken to avoid bad outcomes – such as monetary fraud, faulty drug composition, or declaring someone to be the winner who did not get the most

*Verified Voting Foundation; **E-Voter Education Project; ***Political Studies Program, Bard College; ****Chair of the Subcommittee on Electoral Integrity of the American Statistical Association’s Scientific and Public Affairs Committee; *****Mathematics and Computer Science, Macalester College; *****Chair of the American Statistical Association Working Group on Accurate and Fair Elections. *Please send all comments and suggestions for revision to john@verifiedvoting.org*

¹ One bill, H.R.811, has been reported from committee in the House of Representatives:

<http://thomas.loc.gov/cgi-bin/bdquery/z?d110:h.r.00811:>

The Chair of the Senate Rules Committee has also introduced a bill, S.1487, for consideration:

<http://thomas.loc.gov/cgi-bin/bdquery/z?d110:s.1487:>

For a list of states that already some kind of election auditing, see

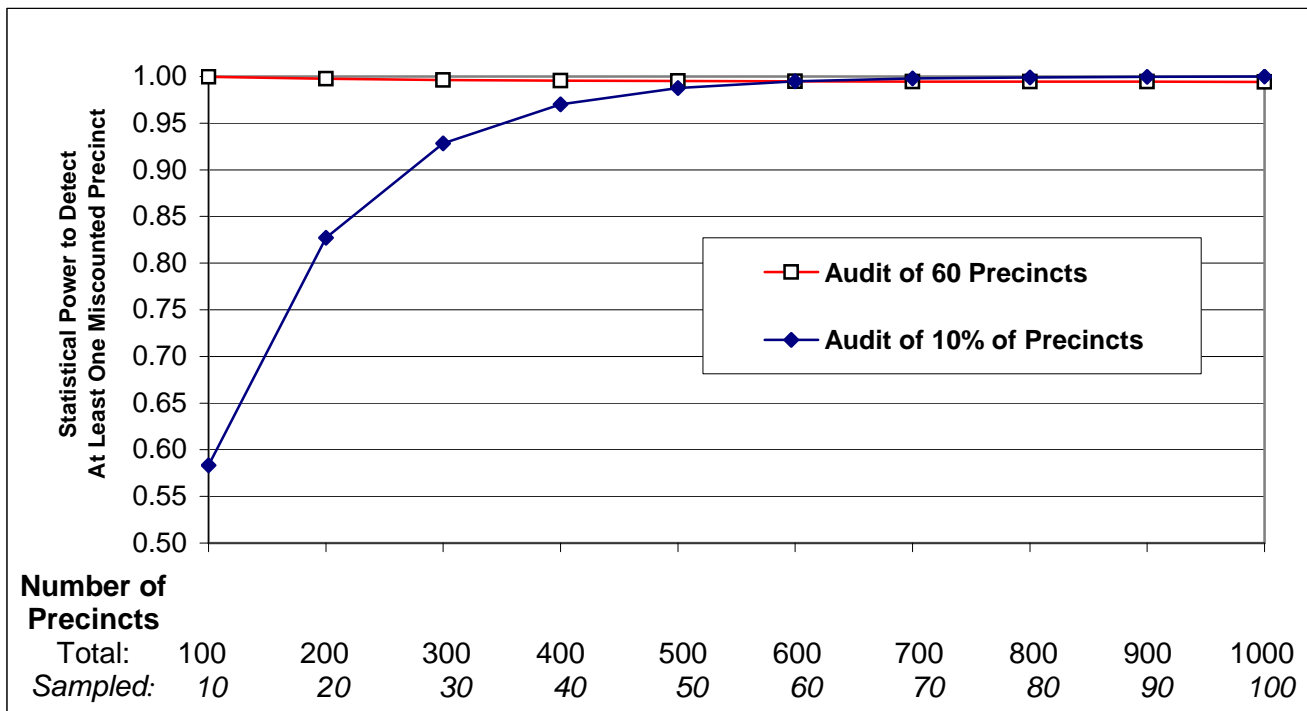
http://www.verifiedvoting.org/downloads/Manual_Audit_Provisions.pdf. New Jersey’s legislature is currently considering a bill modeled on the SAFE approach described at the end of this paper [S.507 as amended].

We use the term “precinct” throughout this paper because it is the most widespread election unit of analysis, but most of our points also apply to other possible sampling units, as discussed below on page 3 and in footnote 6

votes. As in finance or manufacturing, audits are equally able to detect both accidental and malicious errors. But financial audits and quality control tests set sample sizes with a quality goal in mind: specifically to be very likely to detect errors that are large enough to be harmful. In contrast, most proposed election laws and regulations specify auditing a fixed percentage of precincts, or, perhaps, auditing 3%, 5% or 10% of precincts, depending upon the closeness of the margin of victory.

Theory predicts and research has illustrated² that percentage-based audits are frequently inefficient (too large) or ineffective (too small). This is because *the statistical power³ of a method for detecting election-altering miscounts depends principally on the size of the sample – not the fraction of precincts sampled.* For example, suppose that unknown to the auditor, 8% of the total precincts are miscounted. The audit proceeds by randomly examining some number of precincts. If at least one sampled precinct has miscounts, additional investigations will occur; otherwise, the election will be incorrectly confirmed. Thus, the statistical power of the audit is defined as the probability (a number between 0 and 1) that the audit sample contains at least one precinct from the 8% of precincts with miscounts. Clearly, the more precincts that are audited the higher the power. In Figure 1, we compare the statistical power of an audit of 60 precincts to an audit of 10% of the total number of precincts for jurisdictions (e.g., Congressional District, state, state legislative district, etc.) that range in size from 100 to 1000 precincts (and thus the 10% audit samples range from 10 to 100 precincts).

Figure 1: Statistical Power of a 10% Audit vs. an Audit of 60 Precincts When 8% of Precincts Have Miscounts: By Jurisdiction Size



Note that the 60-precinct sample has 99% or greater statistical power *regardless of the total number of precincts*. As the graph suggests, if 8% of the precincts have miscounts, a 60-precinct sample is sufficient

² Saltman 1975; Theisen 2005; Dopp 2006; Stanislevic 2006; Dopp and Stenger 2006; Aslam, Popa and Rivest 2007.

³ The concept and mathematical definition of statistical power is discussed further below and in Appendix A.

for jurisdictions of unlimited size. In contrast, the power of the 10% audit is less than 60% when only 10 precincts (10% of 100) are sampled. Also, sampling 100 precincts (10% of 1,000) is not much more effective than sampling 60 precincts.⁴ This is the basic problem with fixed-percentage audit sizes – they are unnecessarily large for many races, yet too small in others.

Figure 1 suggests that we could specify a sample size of 60 precincts for every audit, but 60 can be too small. Errors in races with narrow margins of victory (where miscounts in fewer than 8% of all precincts could change the outcome) cannot be detected reliably with a sample size of 60. When elections are close, audits must examine more precincts to have a good chance of detecting levels of miscounts large enough to change the outcome.

In this paper, we show the advantages of vote tabulation audits using the SAFE (“Statistically Accurate, Fair, and Efficient”) approach. SAFE auditing, in contrast to percentage-based auditing, emulates statistical methods used for quality control in finance and manufacturing by specifying sample sizes that are *statistically accurate* for confirming election outcomes with a high level of confidence, and *efficient* in allocating auditing resources. SAFE auditing is *fairer* and more equitable than percentage-based auditing because SAFE audits have a pre-specified, high probability of detecting outcome-altering miscounts in all elections – from statewide races to those in a single Congressional district or county.

This paper uses the terms “audits” and “auditing” to mean supervised comparisons of hand-to-eye manual counts versus machine tabulations in a subset of precincts, selected at random shortly after an election and before the results are certified. Hand-to-eye manual audits of voter-verified paper ballots are needed as independent checks, because electronic recounts cannot verify the fidelity of electronic tallies. Many computer experts have stressed the need for software-independent ways to confirm voter intent.⁵

For convenience we use the word “precinct” throughout, although the appropriate audit unit is the smallest cluster that is separately tallied and reported in the unofficial election results released prior to the audit. Thus, if a precinct’s votes consist of tallies from two machines whose paper and electronic votes are readily distinguished,, then the two machines could be treated as if they were unrelated precincts.⁶ (Proposals to audit or to sample ballots, rather than entire precincts, are beyond the scope of this paper.)⁷

Manual counts should also be done in precincts with obvious problems (such as machine failure), and for routine quality improvement monitoring, even in the absence of doubt about who won. We will assume – and strongly recommend – that election officials (and perhaps candidates and political parties) can designate some precincts with apparently anomalous returns to be audited.⁸ Comprehensive auditing

⁴ The relationship between statistical power, sample size, and election margin is discussed further below and in Appendix B.

⁵ See "Requiring Software Independence in VVSG 2007: STS Recommendations for the TGDC" (Discussion draft posted Dec. 1, 2006), <http://vote.nist.gov/DraftWhitePaperOnSIinVVSG2007-20061120.pdf>, p. 2. A voting system is “software-independent” if a previously undetected change or error in its software cannot cause an undetectable change or error in an election outcome. The 2007 Voluntary Voting System Guidelines to be issued by the Election Assistance Commission are widely expected to call for software-independent voting systems.

⁶ Thus, a random sample of auditing units might include just one of the two machine tallies.

⁷ See, for instance, Neff 2003; Wand 2004; Simon and O’Dell 2006; unknown authors (“Titanium Standard”) 2006. Some proposals would use electronic identifiers to verify that specific individual votes were counted correctly; others would sample from all ballots to estimate the vote shares in the entire election.

⁸ Selection of high-interest precincts to be audited in addition to the randomly sampled precincts is consistent with methods used in financial and quality audits.

should also examine many other parts of the electoral process.⁹ Another important question is how to follow up when the initial audit casts doubt upon an electoral outcome. But we do not pursue these broader aspects of audit design here.¹⁰

If the initial sample size is inadequate, material discrepancies may never be found. Therefore, this paper focuses on a single question: how should we determine the number of randomly selected precincts to be selected for a routine audit to verify the apparent winner?

Goals and Assumptions

The primary goal of a vote tabulation audit is to confirm that the winner in the electronic count is the person or position favored by the most voters *when that is true* (i.e., that a complete hand count of the paper ballots would confirm the outcome) and to reveal miscount irregularities that could be sufficiently large to change the outcome. This paper addresses:

- (1) How many precincts should be randomly sampled for auditing to ensure that outcome-altering miscounts are detected with a high probability? Briefly, what should be the sample size?
- (2) How can we measure the effectiveness (for detecting outcome-altering miscounts) of different ways of calculating and specifying sample sizes?
- (3) What principles should we use to determine the smallest sample size that has an acceptably high chance of detecting a potentially outcome-altering miscount?

In this paper we make several simplifying assumptions:

- Every vote is cast in one and only one precinct;¹¹
- Precincts to be audited will be chosen at random *after* an election has taken place and after the unofficial vote counts for each auditable unit are publicly reported; and
- Every precinct has an equal chance to be included in an audit's random sample.¹²

The principles and methods described here apply reasonably well to more complex situations; however, defining optimal procedures in such settings exceeds the scope of this paper.

⁹ Many of these other kinds of issues are outlined in David Marker, John Gardenier, and Arlene Ash, "Statistics Can Help Ensure Accurate Elections" (President's Invited Column) *Amstat News*, June 2007 <http://www.amstat.org/publications/amstat/index.cfm?fuseaction=pres062007>. An upcoming Brennan Center report (Norden et al. 2007) provides a review of the literature, detailed discussion of different auditing methods, and an excellent set of recommendations for creating, improving, and using election audits, and discusses how audits and other procedures can address threats to election integrity.

¹⁰ For instance, we would require a manual count of at least one precinct within each county where an audited race appears on the ballot – added, if necessary, to the initial random sample – to detect county-specific problems. Further, discrepancies that do *not* necessarily cast doubt upon the outcome of a race might still trigger further actions to punish malfeasance and to reduce errors in future elections. We will discuss these and other related topics in future papers.

¹¹ We also assume that the precinct-level vote counts add up to the official totals. Often "early" and absentee votes are not allocated to particular precincts. These votes can be audited in various ways (such as grouping them into precinct-size "bundles" or "pseudo precincts"), but that is beyond the scope of this paper.

¹² It may be preferable to give larger precincts a higher probability of inclusion: see Rivest 2007.

Statistical Power and Election Audits: Two Key Principles

Elections differ widely in numbers of precincts: a typical Congressional District contains about 500 or fewer precincts, while a statewide election in California involves about 22,000 precincts. Regardless of how many precincts are involved, each election-specific audit must sample and examine enough precincts to achieve an acceptable level of statistical power. Two key factors largely determine whether a sample is large enough to verify an election outcome: sample size (i.e., the number of precincts selected at random) and margin of victory (i.e., the percentage of total votes cast that separates the winner from the runner-up in a given election).

First, as illustrated in Figure 1, for a constant margin of victory, the accuracy of an estimate from a sample depends primarily on the *absolute size* of the sample, not its percentage of the population.

Second, the audit sample size needed to confirm the outcome depends on the winning margin. For example, if a candidate appears to have won by a wide margin, a random audit of 10 precincts that shows no miscounts can provide high confidence in that outcome. But if the margin is razor-thin – as in the governor’s race in Washington State in 2004 – nothing less than a complete hand recount may suffice.

Because of these two principles, no percentage-based rule – calling for 2%, 3% or even 10% sampling – is suitable for determining audit sample sizes.

Statistical Power: The Chance that an Audit Will Detect an Outcome-Altering Miscount

Many people believe that only a 100% hand count can determine the winner. Generally, less than 100% is required. For example, suppose that a race involved 500 precincts and there were 500 voters in each precinct (250,000 voters in all). Suppose the winner got 60% of the electronic vote count and the loser, 40% (150,000 vs. 100,000).¹³

If 420 of the 500 precincts (84% of the precincts in this case) are verified (by hand count) to have given 60% of the vote to the putative winner, then no possible allocation of the remaining votes will change the outcome, since 60% of 84% of the votes is already a majority (50.4%). Thus, we can have 100% certainty that an election result will stand without hand counting all the ballots.

However, we also can have a very high level of confidence in the outcome of this 60-40% race after carefully examining the vote counts of far fewer than 420 randomly selected precincts. This is because there would have to be at least 25,000 “flipped votes” – that is, instances in which a vote for the apparent loser was miscounted as a vote for the apparent winner, or 50,000 more “lost votes” for the apparent loser than for the apparent winner, or some combination of these problems for the hand count to change who won. How could this happen?

If hand (re)counting would cause the apparent loser to gain 50 votes in each of 500 precincts, a very small audit sample would detect this. The hardest errors to find (in a random sample of just a few precincts) are those where miscounts are limited to just a few precincts. But, with a 60-40 split, it would be very hard to “cram” all the miscounts into a tiny number of precincts – especially without arousing suspicion. Even if

¹³ In practice, the number of votes counted in any one race is generally smaller than the number of ballots, due to deliberate abstentions and/or uncounted attempts to vote. However, this makes no important difference to the argument made here.

miscounts were spread out over half the precincts, to be outcome-altering, their impact in each precinct would have to be quite large. That is, an average of 60% of the votes in these precincts (rather than the 40% originally recorded) would have to go to the losing candidate.

Miscounts that shift vote totals by more than 20% of the total votes (i.e., 20 percentage points) in any single precinct should be sufficiently noticeable to trigger a suspicion-based “challenge audit.”¹⁴ If so, then our random audits need only look for shifts of at most 20% per precinct. This assumption allows us to determine how many precincts must be miscounted in order to reverse the outcome of an election with a given margin of victory. We call this parameter the *Within-Precinct Miscount* (WPM).¹⁵

In our example, either half the precincts would need to shift by exactly 20%, or more than half by somewhat less. If at least 250 of the 500 precincts have miscounts, a random audit of just 7 precincts will find at least one, with probability better than 0.99.¹⁶ Statisticians call the probability that an audit sample will reveal a particular miscount its *power* (to detect that miscount).¹⁷ Thus, an audit of 7 precincts has 99% power to detect a miscount that occurs in half the precincts.

Even if only 100 precincts out of 500 contain errors, an audit sample of size 20 has 99% power to discover that there are miscounts. This result may seem surprising, because the first pick has an 80% chance of *not* finding an error. However, the chance of missing all 100 bad precincts *20 times in a row* is only about 1%.¹⁸

Quantifying Audit Efficacy

We define the efficacy of an audit as its statistical power to detect an outcome-changing level of miscount. Practical considerations will help specify the level of power sought (the higher the level, the larger the samples required). Our purpose is to explain how to apply a well-understood statistical method for determining the sample size needed to achieve a specified power.

We make two more simplifying assumptions:

- Miscounts in any precinct represent *at most* a shift (i.e., Within-Precinct Miscount) of 20 percentage points (such as the difference between a 60% and a 40% share of the vote); and
- An audit of a race with outcome-altering miscounts is successful if it finds at least one miscounted precinct.¹⁹ That is, the power of an audit (to detect a wrongly-determined race) is the probability that the auditing sample contains at least one miscounted precinct.

¹⁴ As mentioned earlier, we assume that “challenge” selection can be used to audit precincts with apparently anomalous returns in addition to those that are randomly sampled for audit. Appel (2007) underscores that unfortunately, under existing laws, even blatantly anomalous results would not necessarily trigger recounts or other recourse.

¹⁵ Saltman (1975) referred to this as the “maximum level of undetectability by observation.” The value 20% is frequently used in other studies of vote-tabulation auditing.

¹⁶ The chance that no problems show up on 7 draws is essentially the same as the chance that a fair coin comes up heads 7 times in a row = $\frac{1}{2} \times \frac{1}{2} \times \dots \times \frac{1}{2}$, that is, $\frac{1}{2}$ times itself 7 times = $\frac{1}{128}$ of 1%.

¹⁷ For a good intuitive description of statistical power, see J.K. Lindsey, *Revealing Statistical Principles* (New York: Oxford University Press, 1999), p. 116.

¹⁸ The exact probability is computed as the product of 20 terms: $(400/500) \times (399/499) \times \dots \times (381/481) = 0.01045 \sim 1\%$, since (if no bad precincts are sampled) each successive draw must remove one good precinct from the remaining pool.

¹⁹ Miscounts of one or two votes in a precinct may not be important, if they are quite small (that is, within the bounds of expected discrepancies between hand and machine counts across whole precincts) and do not disproportionately favor the same candidate from one precinct to the next.

We do *not* assume that larger miscounts are inherently impossible, only that they would trigger suspicion-based audits. The second assumption requires that when a miscounted precinct is encountered, auditors will take further steps to determine the winner. Importantly, while finding a miscount in a sampled precinct does not necessarily overturn an election, finding no miscounts does confirm the original winner.

Since elections are the bedrock of our democratic republic, we might reasonably require at least 95% probability that *if* miscounts might have altered an outcome, the audit sample will find at least some discrepancy that would trigger further investigation.

As illustrated above, and using formulas provided in Appendices A and B, we can compute the statistical power of a sample of n precincts to detect a miscount in an election jurisdiction with N precincts, when B of the N precincts contain miscounts. In the previous section, we saw that when B is 100 and N is 500 (i.e., 20% of the precincts have miscounts), an audit of $n = 20$ precincts has 99% power. However, what if there are far fewer miscounted precincts? If only 4% of precincts have miscounts (that is, $B = 20$), the power to detect at least one of these miscounted precincts with a 20-precinct audit drops below 57%. Thus, we would be nearly as likely to falsely conclude that there are no miscounts as to find one. To obtain 95% power here would require auditing $n = 69$ precincts; 99% power requires $n = 101$.²⁰

Of course, in an actual election we do not know how many precincts (if any) are miscounted, but (given the assumption that no precinct has a miscount shift larger than 20%) we can calculate the *smallest* number of miscounted precincts necessary to change the outcome of a particular election. Just above, for a 60-40 winning margin with precincts of equal size, we saw that at least half the precincts would have to have 20% miscounts to change the outcome. When margins are narrower, fewer miscounted precincts can change the outcome, so larger samples are needed to make it highly likely that the audit will find at least one among that smaller number of miscounted precincts.

Statistical Power of Percentage-based Audits

We use the framework just described to estimate the efficacy (statistical power) of the currently popular requirement that a percentage of precincts be audited. For example, Connecticut has just adopted a law requiring random audits of 10% of voting districts (precincts) in many elections.²¹

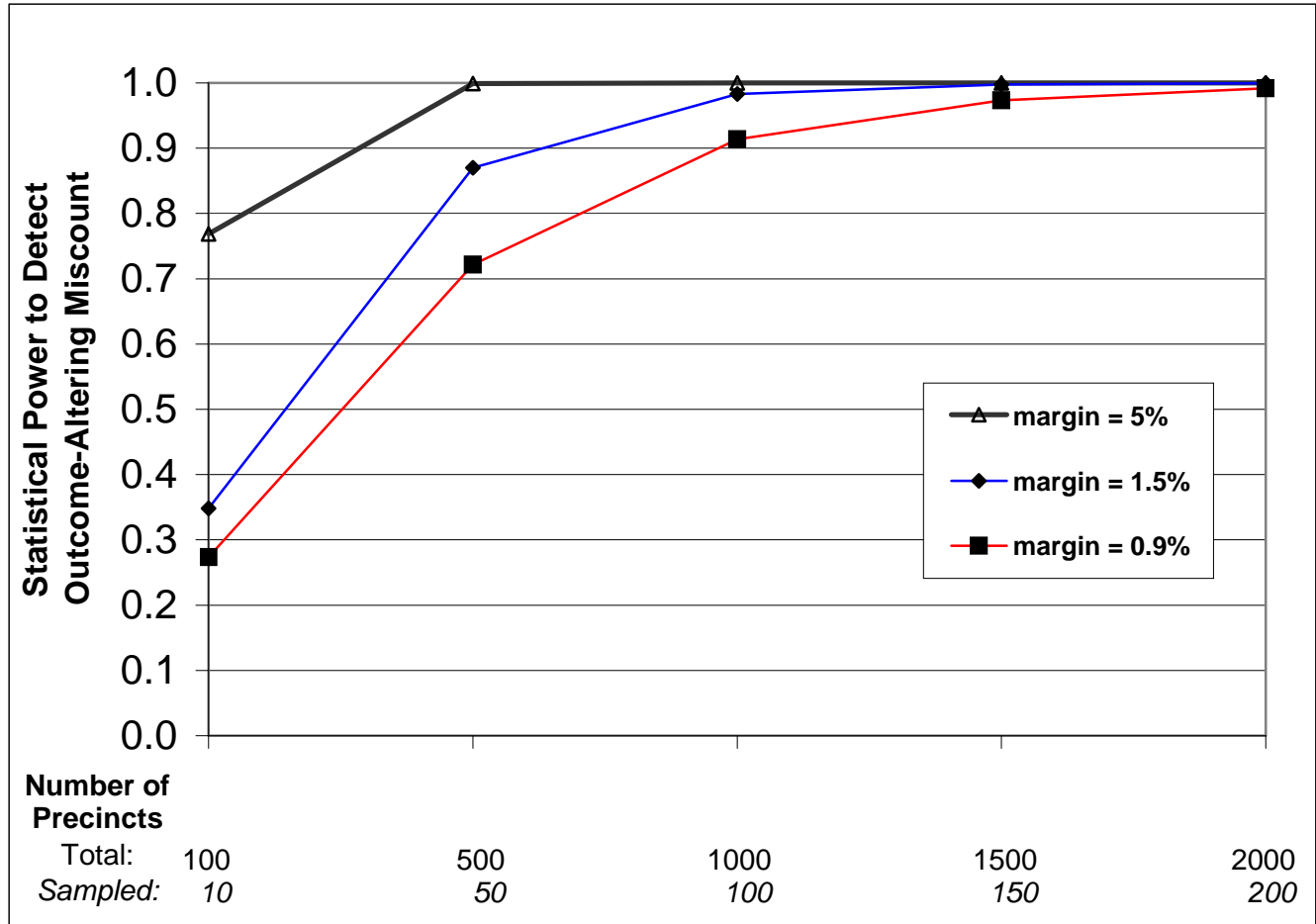
Will these audits be effective? The answer depends on the election: specifically on the winning margin and the total number of precincts in the jurisdiction. Figure 2 shows the power of a 10% audit for jurisdictions with five different numbers of precincts and three different winning margins. Here we assume that all precincts contain the same number of votes.²² To tie this figure closer to a real situation, note that Connecticut has 769 voting districts, so the power of a 10% statewide audit would fall about halfway between the second and the third values from the left in Figure 2.

²⁰ For the exact mathematical formula and further explanation of statistical power and the null hypothesis, see Appendix A.

²¹ The text of the law is available at <http://www.cga.ct.gov/2007/ACT/PA/2007PA-00194-R00SB-01311-PA.htm>.

²² Or, less stringently, that the average number of votes is the same for the precincts with problems as for all precincts. Technical details about the calculations for all figures are given in Appendix B.

**Figure 2: Statistical Power of 10% Audits for Districts:
By Number of Precincts Audited and Margin of Victory***



* Power here is calculated as the probability of finding at least one miscounted precinct when the number of miscounted precincts equals the number of average-sized precincts with 20% shifts needed to overturn the election.

Clearly, 10% audits can have limited power to detect outcome-altering errors when electoral margins are close or when the total number of precincts is small. For instance, if the winning margin is 0.9%, sampling 50 out of 500 precincts confers only about 72% power to detect an outcome-altering miscount. The fixed percentage approach also examines too many precincts when electoral margins are more than a few percentage points or the total number of precincts in the election is large (in particular, for statewide races in large states). With a margin of 5 percentage points, auditing 50 precincts yields better than 99% power, regardless of the number of precincts in the district. Thus, for a winning margin of 5 points or more, it is wasteful to audit much more than 50 precincts in order to verify the outcome. A 10% audit in California would sample about 2,200 precincts, while a sample of 200 already has 99% power to detect an

outcome-altering level of miscount, even with a winning margin of only 1%. If the goal is to verify election outcomes, a 10% audit will count too many precincts in some races and not enough in others.²³

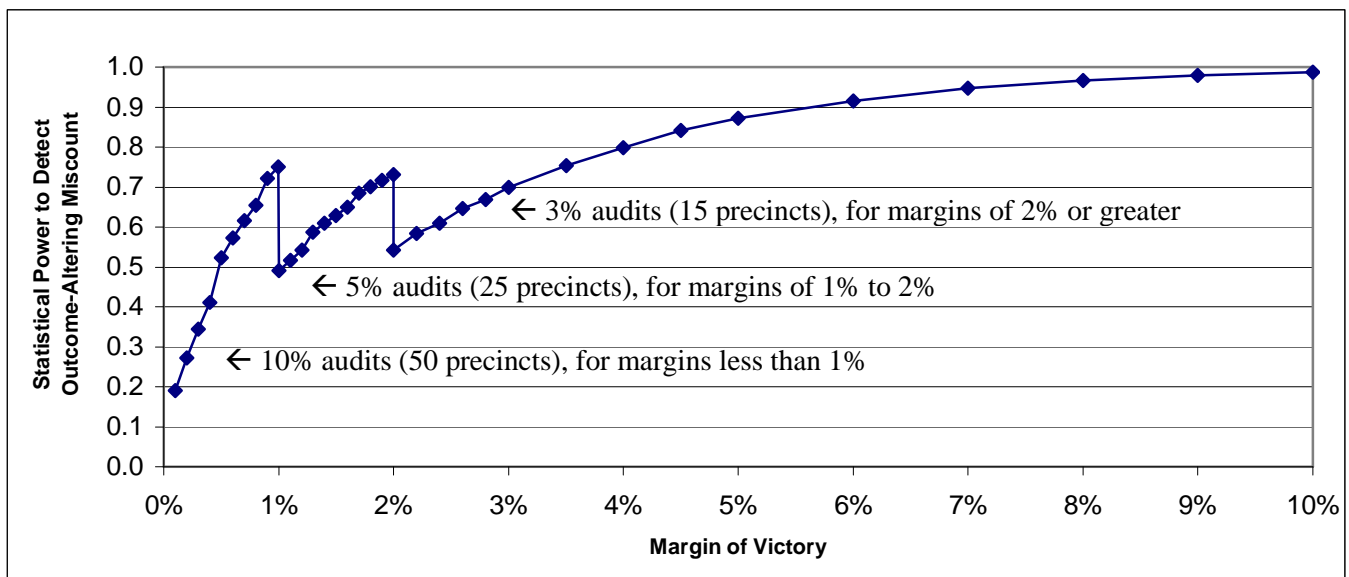
Statistical Power of Multi-Tier Percentage Audits

A multi-tier variant of percentage-based auditing specifies a percentage of precincts to be audited that depends roughly on the winning margin. For example:

- audit 3% of precincts if the winning margin is 2 percent or more of the total votes cast;
- audit 5% of precincts if the winning margin is at least 1 percent but less than 2 percent; and
- audit 10% of precincts if the winning margin is less than 1 percent of the total votes cast.

Since narrower margins require larger audits, this approach is better than specifying a single audit percentage. However, it still suffers from the basic problems of any percentage-based auditing requirement (see Figure 3). In modest-sized districts, sampling 5% or even 10% of the precincts does not achieve even 75% power for detecting election-altering miscounts when the winning margin is just 1 or 2 percent of the total votes cast.

Figure 3: Statistical Power of Three-Tiered 3-5-10% Audits in a 500-Precinct Jurisdiction: By Margin of Victory*



* Assumes the number of precincts with miscounts equals the minimum number needed to overturn the election if all miscounts are in average-sized precincts, each with a vote shift of 20 percentage points.

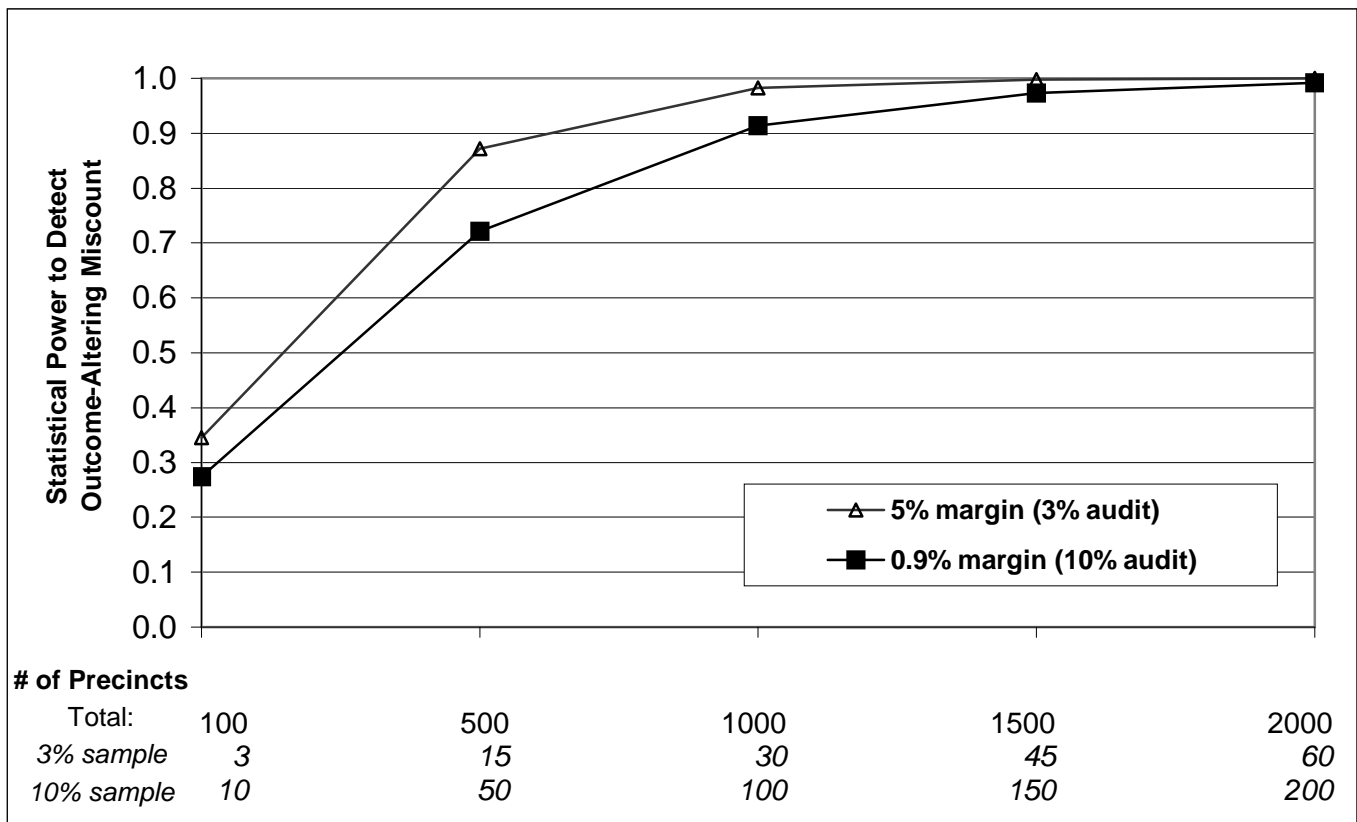
A tiered-audit requirement also is vulnerable to malicious shifting of vote margins into the part of each tier with the lowest power to detect miscounts. For example, Figure 3 shows that the power to detect an outcome-altering miscount is about 75% with a margin just under 1% (first tier), but drops to less than 50% with a margin of exactly 1%. Thus an attacker with some control of the voting system (or by stuffing

²³ There are other reasons to audit additional precincts, such as requiring a minimum percentage in each administrative jurisdiction (e.g., county) that could help catch jurisdiction-specific problems (such as ballot programming errors). See Theisen (2005), Norden et al. (2007).

a traditional ballot box), could add or remove only a few votes so that the reported margin of victory becomes 1% or slightly higher, thereby shifting the audit from the 10% sample tier to the 5% tier, cutting the sample size in half. The resulting decrease in statistical power, leveraged by changing only a few votes, makes it far less likely to detect larger miscounts that could change the outcome, especially those concentrated in relatively few precincts. A similar opportunity exists near the margin of 2%.

Most Congressional and State Legislative Districts in the United States have fewer than 500 precincts.²⁴ Figure 3 shows that for a jurisdiction with 500 precincts, when an election is decided by less than 3 percentage points, 3-5-10%-tiered audits have poor statistical power – almost always less than 70%, and sometimes less than 50%. That is, in a close race where the initial electronic counts declared the wrong winner, such audits have a 30 to 50% chance or more of confirming that incorrect outcome. Since most districts have *fewer* than 500 precincts, three-tiered audits will typically have even less power than shown here. Figure 4 shows just how large jurisdictions have to be before tiered audits achieve good power.

Figure 4: Statistical Power of Tiered 3% and 10% Auditing: By Jurisdiction Size*



* Assumes miscounts of at most 20% occur in precincts not larger than average.

Figure 4 depicts two possible margins (0.9% and 5%) from different tiers: the narrower margin triggers a 10% audit; the wider margin, a 3% audit. *For these margins*, a 3% audit is *uniformly more effective* (powerful) than a 10% audit because the sample size required for good power is much larger when the

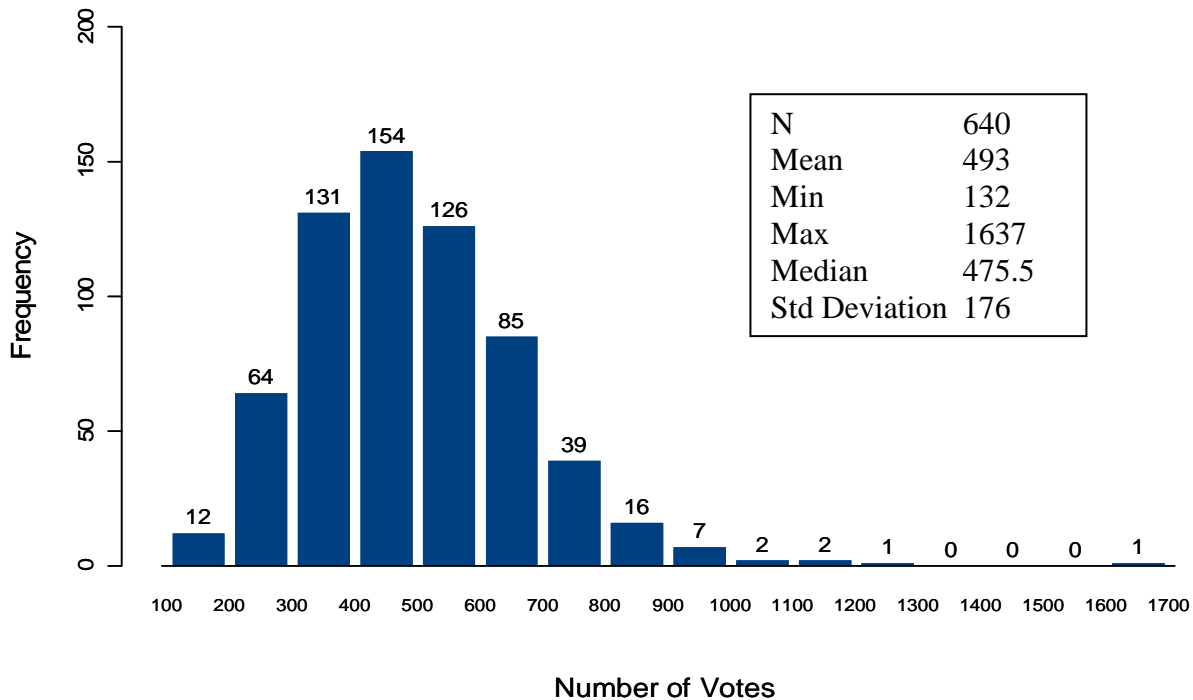
²⁴ Over thirty states have (on average) fewer than 500 precincts per Congressional district; only three average more than 800 precincts per district.

winning margin is less than 1% than with a winning margin of 5%. Most importantly, this figure shows that in House races (which usually involve far fewer than 1000 precincts), tiered percentage audits do not solve the problem they were intended to address; their power for auditing close races is predictably poor.

Implications of Variations in Precinct Size

Until now we have put aside the fact that election jurisdictions contain precincts with unequal numbers of voters. As Saltman noted in 1975, and Stanislevic (2006) has examined in detail more recently, variations in precinct size can further reduce the statistical power of vote tabulation audits.²⁵ For example, Figure 5 shows the variation in number of votes per precinct in the 640 precincts of the Fifth Congressional District of Ohio in the 2004 general election. CD-5’s smallest precinct vote total was only 132 (less than 1/20 of 1% of the 315,540 votes in the entire district), while its largest precinct vote total was 1637 (nearly ½ of 1% of the district total).

Figure 5: Distribution of Votes Counted in 2004 among the 640 Precincts of Ohio’s Fifth Congressional District



When vote miscounts are concentrated in larger precincts, fewer miscounted precincts are needed to alter an election. Consider an election with a winning margin of 6 points (for instance, 53% to 47%). This margin could be overturned by 20% shifts in precincts containing 15% of all votes.²⁶

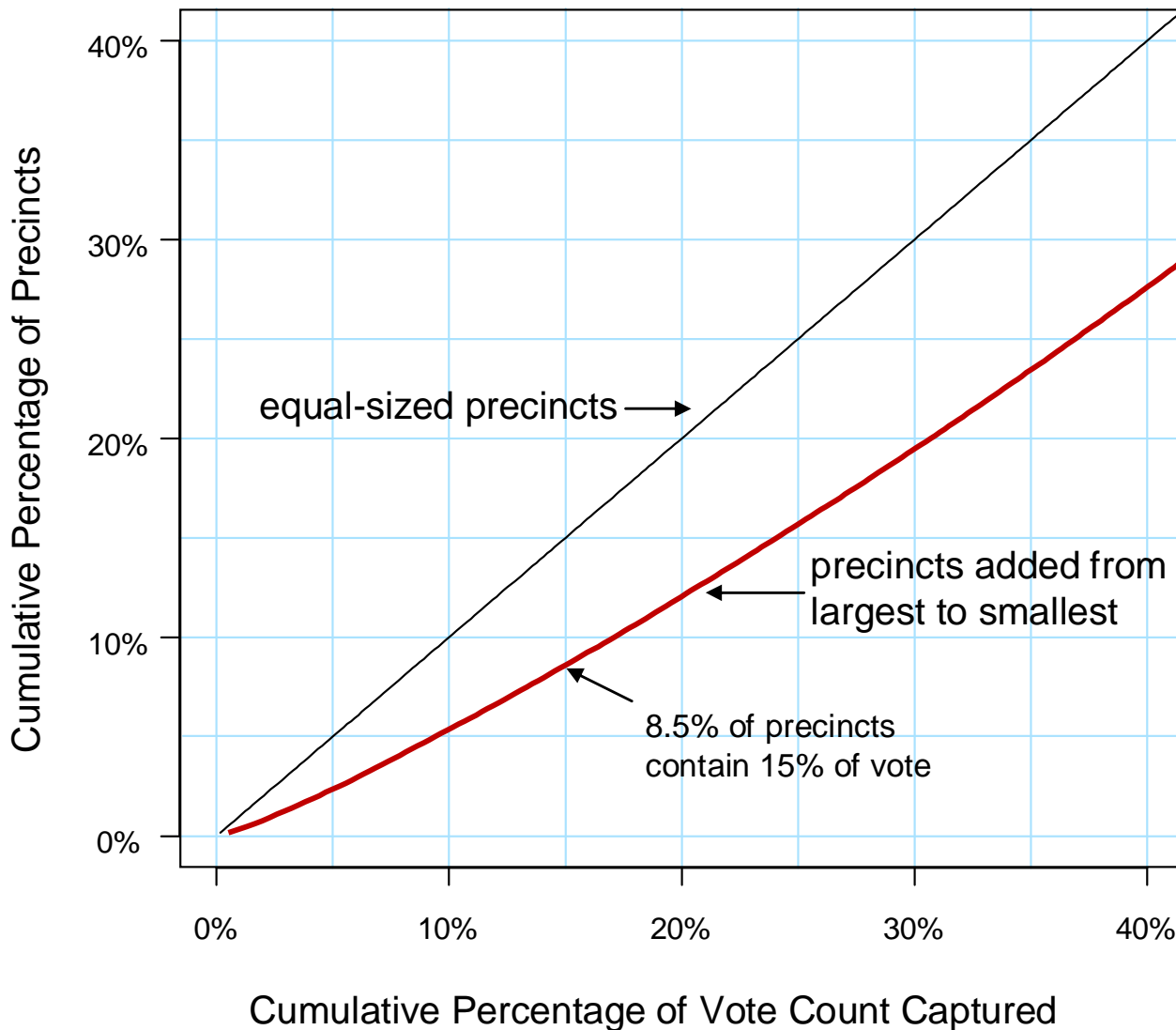
²⁵ Dopp and Stenger (2006) and Lobdill (2006) explore the same phenomenon.
²⁶ The latter percentage is calculated as half the margin (here 6% / 2 = 3%) divided by the Within-Precinct Miscount (here 20%): 3% / 20% = 15%. Equivalently, the relationship $V_m = (M / 2) / WPM$, where M is the margin in votes and WPM is the Within-Precinct Miscount percentage, can be used to calculate V_m , the number of votes there must be in the miscounted precincts to alter the outcome. (Our example expresses V_m and M as percentages instead of counts.)

Figure 6 shows that, *if the largest precincts are chosen first*, only about 8.5% (55) of Ohio CD-5’s precincts are needed to encompass 15% of the vote. An audit that has good power to detect a miscount when 15% of precincts (96) are miscounted may have poor power when only 55 precincts have miscounts.

If all precincts contribute the same number of votes, it takes the same percentage of the precincts to encompass a given percentage of the total vote. Thus, for example, any 15% of the precincts contain exactly 15% of the votes. This obvious fact is captured in the straight line in Figure 6. The curved line shows the extent to which – when the precinct-size distribution in a district varies as shown in Figure 5, and when the largest precincts are used first – many fewer precincts are needed to contain a given percentage of the total vote.

Figure 6: Fewer Precincts Can Affect More Votes When Precincts Vary in Size*

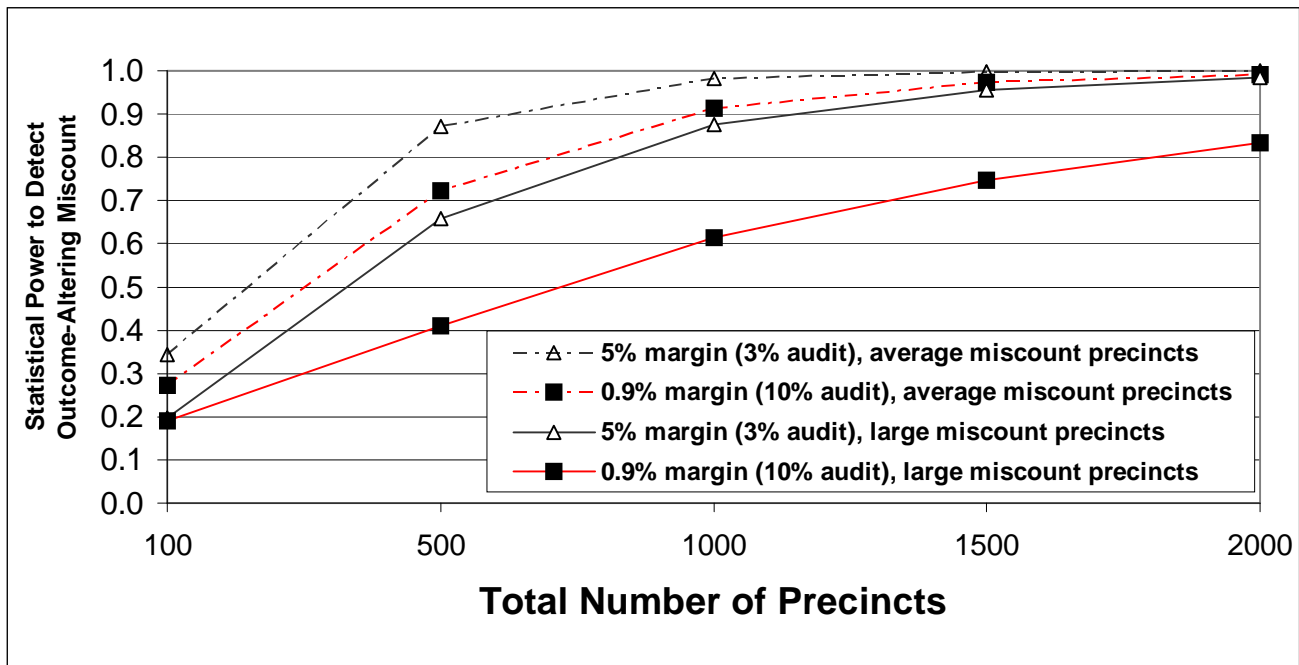
* Calculations assume that variable-sized precincts contain the same fractions of the total vote count as those in Ohio's Fifth Congressional District in 2004 (see Figure 5, above) and that precincts are added to the cumulative percentage from largest to smallest.



Applying the precinct size distribution of Ohio’s CD-5 to any total number of precincts, we can estimate the statistical power of audits when all the miscounts occur in the largest precincts. We do this for each election, by finding the smallest number of precincts that could possibly contain at least as many votes as needed to reverse the outcome.

Figure 7 compares the power of percentage-based audits when miscounts occur in average-size precincts (shown in Figure 4), to the power of the same audits under the worst-case assumption that miscounts occur in the largest precincts. Note that the statistical power to detect large-precinct miscounts (shown by the solid lines) is substantially lower than the statistical power to detect miscounts in average-sized precincts (dotted lines). For instance, in a Congressional District with 500 precincts, the estimated power of a 3% audit to confirm the outcome of a race with a 5-point margin (dotted lines with triangular markers) drops from about 87% to about 66% (solid line with triangle markers) under a “worst-case,” large-precinct assumption – increasing the risk of failing to detect outcome-altering miscounts by over 20 percentage points. For a 0.9% margin in the same district, the power to detect a miscount in the large-precinct scenario is over 30 percentage points less than the power to detect a miscount in the average-size precinct scenario.

Figure 7: Power Is Lower for Detecting Miscounts Concentrated in the Largest Precincts



* Power to detect outcome-altering miscounts when they reside in a minimum number of average-sized precincts (“average precincts”) versus when the same number of miscounted votes are in the “largest precincts.” Assumes precinct-size variation as in Figure 5. (Calculations are shown in Appendix B.)

The SAFE Alternative for Vote Tabulation Auditing

Statisticians and a growing number of election experts have urged replacing fixed percentage vote tabulation audits with the statistically grounded rules that have informed decades of financial auditing and quality control. We call this approach SAFE – for Statistically Accurate, Fair, and Efficient – vote

tabulation auditing. For each election contest, SAFE audits randomly select the number of precincts needed to achieve a specified level of statistical power to detect an outcome-altering miscount. For a full, technical description of the SAFE Method for auditing vote tabulations, see Appendix C.

However, the basic idea is that the sample size for each race depends upon the following four factors:

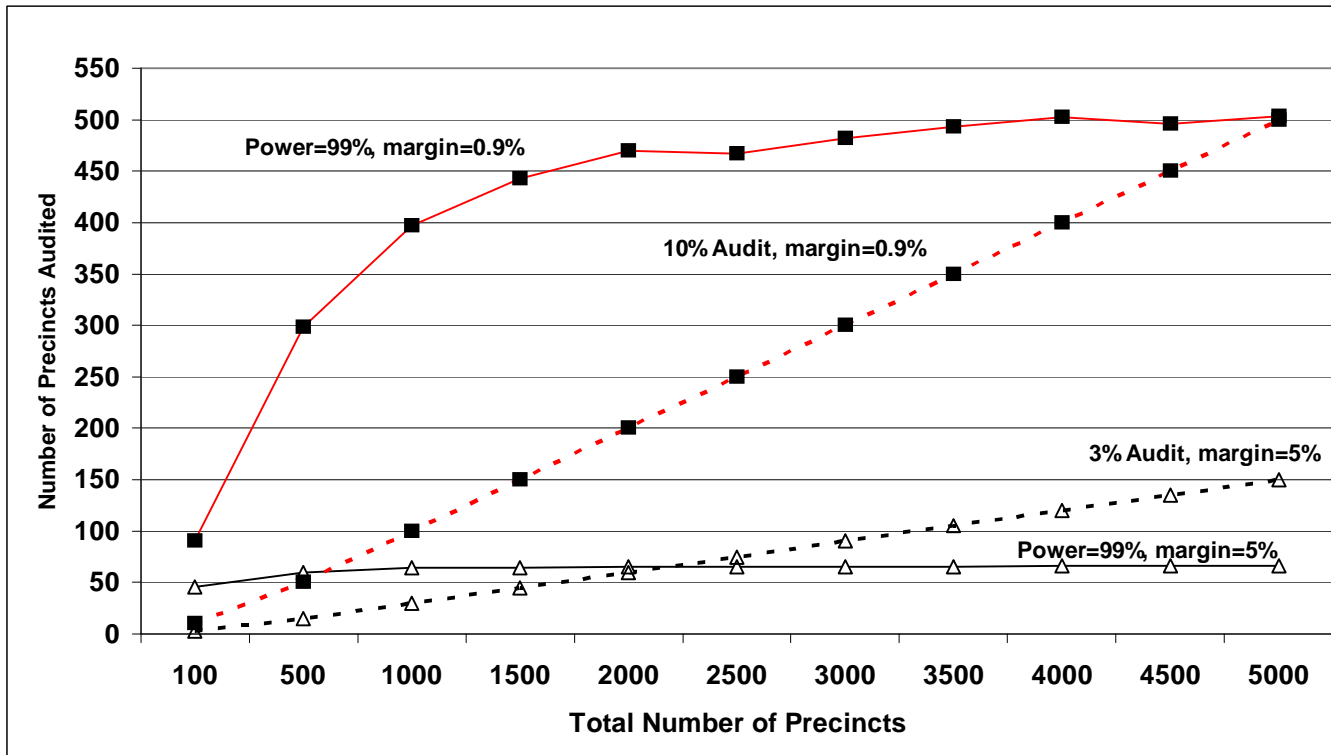
1. The level of statistical power we wish to achieve. For example, 99% power means that if an outcome-altering miscount occurred, we would fail to discover it only one time per 100 audits. When an election “passes” an audit that has high power, this is strong evidence supporting the initial reported outcome. Higher power requires larger samples, as is discussed in more detail in Appendix A.
2. The victory margin. This is the percentage of votes separating the winner from the runner-up. Narrower margins require larger audit samples.
3. The within-precinct miscount (WPM). This is a hypothesized maximum percent of all the votes that could be shifted in any single precinct without triggering a suspicion-based recount. All power calculations in this paper assume a WPM of 20%. The higher the WPM, the easier it is for large miscounts to “hide” in just a few precincts, requiring larger audits samples to retain a good chance of finding at least one precinct with a miscount.
4. The precinct-size distribution. When precincts in a jurisdiction vary widely in numbers of votes, it is possible to hide a given number of miscounted votes in fewer large precincts than if all the precincts were the same size. This is discussed in more detail in “Implications of Variations in Precinct Size,” above. In this paper, all subsequent calculations assume that precinct sizes are distributed as in Figure 5 above, and power is calculated as the probability of detecting at least one miscount when the miscounts all occur in the largest precincts. This conservative approach leads to larger samples than if the miscounts are assumed to occur in average- or constant-sized precincts.

The sample size for each SAFE audit is thus designed to have a desired statistical power to detect an outcome-altering miscount if one occurred. SAFE audits deploy auditing resources efficiently, effectively and fairly across all races. In the next section, we show how SAFE audits are more effective in small races, and more efficient in large races, than percentage-based audits.

Comparing Percentage-Based vs. SAFE Vote Tabulation Audits

To evaluate the relative merits of percentage-based and SAFE audits, Figure 8 shows the power of each method for total numbers of precincts ranging from 100 to 5,000, for two margins of victory (0.9% and 5%). All power calculations are computed under the same assumptions, including the assumption that miscounts occur in the largest precincts.

**Figure 8: Percentage-Based versus SAFE Vote Audit Sample Sizes:
By Jurisdiction Size***



*Slight “dips” in the SAFE lines are anomalies due to rounding the numbers of sampled precincts up to whole numbers in the formula used for calculating sample sizes (see Appendix B).

Figure 8 plots the number of precincts to be sampled (on the vertical) against the total number of precincts in the race. Dashed lines show audit sizes for two different margins of victory (0.9% and 5%) for the tiered percentage approach illustrated in Figure 3; solid lines show audit sizes for 99%-power SAFE audits for the same two margins of victory.

For either margin, regions of the graph where the dashed and solid lines are far apart signify areas of difficulty. When the dashed percentage line is substantially below the solid power-based line, the percentage-based audit examines *too few* precincts to be very likely to detect a miscount even when it is large enough to flip the election; when the lines are reversed, the percentage-based audit examines *more* precincts than necessary for a “clean” audit to provide confidence in the original outcome.

Percentage-based audits can be far too small. Consider, for example, a congressional election with a winning margin of 0.9 percentage points (square symbols in Figure 8) in a district with 500 precincts. The tiered percentage method (higher dashed line) requires a random audit sample of 50 precincts (10%). But good power (with a race this close) requires auditing almost 300 precincts (highest solid line). An audit of only 50 precincts in this situation has only about 41% power.

Figure 8 shows that the sample size demands of power-based methods level off as the total number of precincts increase. In contrast, percentage-based audits require larger samples in elections with many

precincts whether or not they are needed. For instance, for a 5-percentage-point margin, the audit size required for 99% power is always less than 70 precincts, no matter how many total precincts are involved. For a state with 5,000 precincts (about the size of Michigan), a 3% audit (150 precincts) is more than twice as large as needed to achieve 99% power here.

For larger states and larger margins, the efficiency advantage of power-based audits is even greater. For California, with almost 22,000 precincts, a 3% audit – over 650 precincts – is more than *nine* times as large as necessary to confirm, with 99% power, the outcome of an election with a 5 point margin. Yet many statewide races have even larger winning margins, and can be audited, with high power, by sampling 50 or fewer precincts.

Note that even for a race decided by only 0.9 points, an audit of about 500 precincts suffices even for large states. Therefore, contrary to many people's intuition, even a 3% audit in a California statewide election contest (over 650 precincts) would more than suffice to confirm the outcome of such a race with 99% power. A 10% audit confers little additional advantage, although it would require auditing *over 1,500 additional precincts*.

In statewide, landslide elections a fixed-percentage audit almost certainly imposes unnecessarily high auditing costs. Because auditing serves other functions, it may be wise to audit some minimum number of precincts (or precincts per county or per legislative district) even when the winning margin is large, as a check on the performance of voting machines and other aspects of the election process in each administrative jurisdiction.²⁷ However, no scenario, other than a very narrow victory margin, requires auditing hundreds of precincts.

Finally, to compare estimated costs for SAFE audits versus percentage-based audits, using the sample precinct size distribution of Figure 5, we analyzed the results of three complete sets of federal elections (2002, 2004 and 2006).²⁸ In Table 1, we examine power and resource requirements for the 1393 contested federal elections during that period, comparing a three-tiered audit requirement, to fixed-percentage and SAFE audits.²⁹

Clearly, 2% audits use the fewest resources, but fail to achieve power as great as 95% in 206 (almost 15% of) races. At the other extreme, 10% audits would have had at least 99% power to detect a worst-case scenario miscounts in 92.7% of races, but at the cost of auditing 57.6 million ballots. 99%-power SAFE audits would require examining only 23.0 million ballots.

None of the percentage-based audits would have adequately sampled all races, leading – even with 10% sampling – to 19 races in which an election-altering miscount would be more likely to be missed than detected.

²⁷ Norden et al. (2007) recommends auditing a minimum percentage of precincts in each administrative jurisdiction that runs elections (e.g., each county) to identify problems that may not necessarily alter election outcomes.

²⁸ Our analysis here updates the results of Lindeman and Stanislevic 2007.

²⁹ We further required that every audit (whether percentage-based or SAFE) include at least one precinct per county in the election, even if this requirement entailed a larger audit than would otherwise be necessary. For instance, Iowa has 99 counties and approximately 1966 precincts, so the smallest possible audit in a statewide election includes over 5% of precincts.

Table 1: Federal Elections (2002-2006) Achieving Various Levels of Power by Type of Audit*

Type of Audit	Percentage-based				SAFE	
	Tiered 3-5-10%	2%	3%	10%	99% power	95% power
<i>Power of the Audit</i>	<i>Number of elections (percent)</i>					
at least 99%	1152 (82.7%)	1089 (78.2%)	1152 (82.7%)	1292 (92.7%)	1393 (100%)	-
from 95% up to 99%	77 (5.5%)	98 (7.0%)	74 (5.3%)	31 (2.2%)	-	1393 (100%)
from 50% up to 95%	112 (8.0%)	137 (9.8%)	110 (7.9%)	51 (3.7%)	-	-
less than 50%	52 (3.7%)	69 (5.0%)	57 (4.1%)	19 (1.4%)	-	-
<i>Total hand-counted votes (in millions)</i>	20.5	15.3	19.4	57.6	23.0	19.0

* Results for 1393 contested elections for U.S. president, Senate, and House of Representatives. Vote margins were calculated from FEC data for 2002 and 2004; Dr. Adam Carr's Psephos archive for 2006. Numbers of precincts per election were estimated based on the 2004 EAC Election Day Survey³⁰; for House elections, the number of precincts in each state was divided by the number of Congressional Districts to estimate the precincts per District. If an audit size would otherwise be smaller than one precinct per county, one precinct per county is audited. Power is calculated to protect against miscounts residing in the largest precincts, as described in "Implications of Variations in Precinct Size" above.

Auditing fewer precincts in the races where statistical power exceeded 99% power could free up resources for effective sampling in the rather rare races where percentage-based audits have too little power.

SAFE audits designed to achieve 95 to 99% power require about the same resources (across all races) as 3-5-10% tiered audits. By definition and design, however, the power to detect outcome-reversing outcomes in SAFE-based audits is uniformly high across all races.

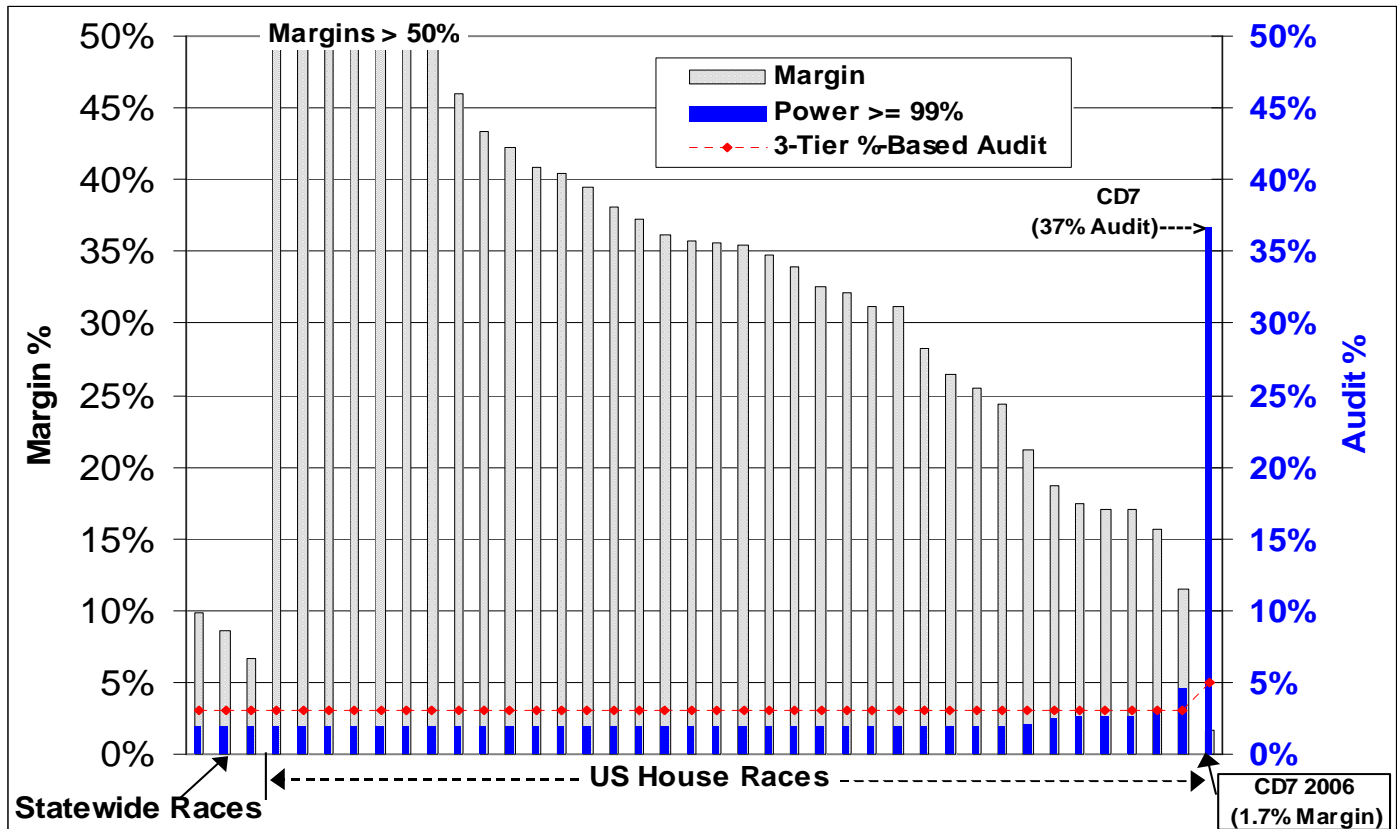
Clearly the opportunity to achieve very high assurance in *each* federal election, with little or no increase in the number of ballots that need to be hand-counted compared to the less powerful 3-5-10 percentage approach, should not be squandered.

A Single-State Example

Finally, abstracting from the above data, we examined the federal elections from 2002 to 2006 in the State of New Jersey, where an election auditing bill is pending in the legislature. Rather than the one-precinct-per-county minimum used in the calculations for the national study summarized in Table 1, the proposed New Jersey legislation calls for a minimum of 2% of precincts per Congressional District (CD). This allows counties to bear the burden of the minimum audit roughly in proportion to their population.

³⁰ The Election Assistance Commission's 2004 Election Day Survey can be found at http://www.eac.gov/election_survey_2004/intro.htm.

Figure 9: Winning Margins and What Audits Would Have Been Required for 3-Tier and SAFE Approaches in New Jersey General Federal Elections 2002-2006*



* Separately for 3 statewide and 37 contested House races, the races with the widest margins are shown at the left, with increasingly competitive races shown from left to right. Audit percentage scale is shown on right. The tiered approach requires audits of 3%, 5% or 10% for (respectively) margins 2% or higher; 1% to less than 2%; less than 1%. (No race had a margin of less than 1%.) SAFE audits are required to sample a minimum of 2% of precincts, regardless of power considerations. SAFE audits are designed with 99% power to detect an outcome-altering miscount, assuming 20% Within-Precinct Miscount in the largest precincts.

Figure 9 compares the size required by the tiered audit rule (marked with “diamonds”) to a SAFE audit with 99% statistical power (dark columns). The light gray columns indicate victory margins. The 99% SAFE audit requirement is more than satisfied by the 2% minimum in all but the 7 most competitive House races (at the far right); 3% sampling would be wasteful in the other 30 House races and all three statewide races. In 5 of the remaining races the SAFE requirement is met by sampling between 2% and 3% of precincts. One race (the 2nd closest House race) requires a sample of almost 5% (but would have gotten only 3% under the tiered audit).

On the far right of the chart is the only race (CD-7 in 2006, with a 1.7% margin) that would have required a really large audit (37% of the district’s estimated 476 precincts) to achieve 99% statistical power to detect an outcome-altering miscount. This race would have had a 5% audit under the tiered scheme. This was the only close federal election in New Jersey in six years. Dropping back to, say, 95% power here would require a 26% sample, rather than 37%. However, when a close race is such a rare event, we should

be able to afford the audit required to be fully confident in its outcome. In confirming the outcomes of these 40 New Jersey races, SAFE audits would have used minimum 2% sampling most of the time and only more than 5% in a single race to achieve at least 99% power. SAFE audits would have examined 3.0% of precincts per race overall, as would the 3-5-10 tiered audits. The tiered audit, however, would have been ineffective in the closest U.S. House race (1.7% margin), having less than 40% power to find a miscounted precinct, even if there had been enough miscounted precincts to change the outcome of that election.³¹ Clearly, SAFE auditing is more statistically accurate, fair and efficient.

Conclusions

Effective electoral oversight requires routine, comprehensive checks on the entire voting process. An important component is ensuring that the winner in an initial electronic tally is the same as would be identified in a 100% hand count of voter-verified paper ballots. The goal is to provide an independent check on electronically tabulated outcomes of races. In this paper, we describe SAFE (Statistically Accurate, Fair and Efficient) vote tabulation audits. Any electoral audit should be conducted within a larger framework of good practice, such as following well-specified, publicly supervised procedures; examining both randomly selected precincts and precincts with observed anomalies; and generating publicly accessible auditing data.³² However, one feature distinguishes SAFE audits: they randomly sample just enough precincts to make it very likely to detect a miscount if an election-changing miscount had occurred. SAFE audits are therefore both more efficient and more effective than, say, audits of 3%, 5% or 10% of precincts.

Election audits have been conducted in some state and local jurisdictions for years.³³ Saltman (1975) cites the 1% "manual tally" still used in California. Although professional auditors, statisticians and computer scientists should advise on standards and procedures, competent election officials and staff can implement the SAFE auditing sample size calculations, detailed in the Appendices below, without special assistance.

Whatever method the U.S. Congress adopts for establishing the size of audits, it should also allow States to adopt alternative SAFE audit procedures designed to achieve uniformly high levels of statistical power. Smaller audit samples should be allowed when they achieve at least 99% power to detect an outcome-altering miscount.

Acknowledgements

We want to thank a number of people gave us very useful comments and suggestions for improvement, particularly Gregory Bell, Judy Bertelsen, Kathy Dopp, Ed Gracely, Dave Hoaglin, Tom Piazza, and Judy Tanur.

³¹ Power =40% for a 5% audit (24 out of 476 total precincts) assuming 20% miscounts in the largest precincts.

³² Many of these points are also covered in recommendations by other groups concerned with election auditing. For example, see recommendations in Norden et al. (2007).

³³ For instance, the electionline.org briefing paper "Case Study: Auditing the Vote" (<http://www.electionline.org/Portals/1/Publications/EB17.pdf>) discusses the auditing experiences of several states including California and Minnesota.

References

- Appel, Andrew W. 2007. "Effective audit policy for voter-verified paper ballots in New Jersey." March 9, 2007. Published on the Internet at: <http://www.cs.princeton.edu/~appel/papers/appel-nj-audits.pdf>.
- Aslam, Javed A., Raluca A. Popa and Ronald L. Rivest. 2007. "On Estimating the Size and Confidence of a Statistical Audit." April 22, 2007. Published on the Internet at: <http://theory.csail.mit.edu/~rivest/AslamPopaRivest-OnEstimatingTheSizeAndConfidenceOfAStatisticalAudit.pdf>.
- Dopp, Kathy. 2006. "How Can Independent Paper Audits Ensure Election Integrity?" Updated July 25, 2006. Published on the Internet at: http://electionarchive.org/ucvAnalysis/US/paper-audits/Paper_Audits.pdf.
- Dopp, Kathy and Frank Stenger. 2006. "The Election Integrity Audit." September 25, 2006. Published on the Internet at: <http://electionarchive.org/ucvAnalysis/US/paper-audits/ElectionIntegrityAudit.pdf>.
- Lindeman, Mark, and Howard Stanislevic. 2007. "H.R.811: Fact & Friction – Part III." March 28, 2007. Published on the Internet at: <http://e-voter.blogspot.com/2007/03/hr811-fact-friction-part-iii.html>.
- Lobdill, Jerry. 2006. "Considering Vote Count Distribution in Designing Election Audits." Revision 2, November 26, 2006. Published on the Internet at: <http://vote.nist.gov/Considering-Vote-Count-Distribution-in-Designing-Election-Audits-Rev-2-11-26-06.pdf>.
- Neff, C. Andrew. 2003. "Election Confidence---A Comparison of Methodologies and Their Relative Effectiveness at Achieving It." Revision 6, December 17, 2003. Published on the Internet at: <http://www.votehere.net/papers/ElectionConfidence.pdf>.
- Norden, Lawrence, Aaron Burstein, Joseph Hall and Margaret Chen. 2007. "Post-Election Audits: Restoring Trust in Elections" (forthcoming at http://brennancenter.org/subpage.asp?key=38&proj_key=76).
- Poisson, Siméon Denis. 1837, *Recherches sur la probabilité des jugements en matière criminelle et en matière civile*. Available on Google Book Search at <http://books.google.com/books?id=vqnKMvTuXo4C&pg=PA1&dq=Recherches+sur+la+probabilit%C3%A9+des+jugements+en+mati%C3%A8re+criminelle+et+en+mati%C3%A8re+civile>.
- Rivest, Ronald L. 2007. "On Auditing Elections When Precincts Have Different Sizes." Draft of April 29, 2007. Published on the Internet at: <http://theory.csail.mit.edu/~rivest/Rivest-OnAuditingElectionsWhenPrecinctsHaveDifferentSizes.pdf>.
- Saltman, Roy G. 1975. "Effective Use of Computing Technology in Vote-Tallying." National Bureau of Standards, Final Project Report, March 1975, prepared for the General Accounting Office. [See particularly Appendix B, "Mathematical Considerations and Implications in Selection of Recount Quantities."] Available on the Internet at: http://csrc.nist.gov/publications/nistpubs/NBS_SP_500-30.pdf.

Simon, Jonathan D., and Bruce O'Dell. 2006. "An End to 'Faith-Based' Voting: Universal Precinct-based Handcount Sampling To Check Computerized Vote Counts In Federal and Statewide Elections."

September 8, 2006. Published on the Internet at:

<http://electiondefensealliance.org/files/UPSEndFaithBasedVoting.pdf>.

Stanislevic, Howard. 2006. "Random Auditing of E-Voting Systems: How Much is Enough?"

VoteTrustUSA E-Voter Education Project, August 16, 2006. Published on the Internet at:

<http://www.votetrustusa.org/pdfs/VTTF/EVEPAuditing.pdf> .

Theisen, Ellen. 2005. Auditing Election Equipment --- The Real Scoop!, August 27, 2005, Published on

the Internet at: <http://www.votersunite.org/info/auditingissues.pdf>

Unknown authors. 2006. "The Titanium Standard for Election Verification & Security." October 1, 2006.

Published on the Internet at: <http://www.velvetrevolution.us/titanium.pdf>.

Wand, Jonathan. 2004. "Auditing an Election Using Sampling: The Impact of Bin Size on the Probability of Detecting Manipulation." Version of February 17, 2004. Published on the Internet at:

<http://wand.stanford.edu/elections/probability.pdf>.

Appendix A: Statistical Power and the Null Hypothesis

As explained in Appendix B, we can use the hypergeometric distribution, originally published by the French mathematician Siméon Denis Poisson [Poisson 1837], to compute the probability that a sample of n objects drawn without replacing any of them will contain at least B “bad” objects. In particular, if we sample n precincts from a total of N precincts containing B precincts with miscounts, the probability, P , that our sample of n precincts will contain at least one miscounted precinct is given by the formula:

$$P = 1 - \{ [(N - B)!(N - n)!] / [N! (N - B - n)!] \}$$

When P and N are known, n can be found by trial and error. In Appendix B we provide a more efficient method by Aslam, Popa and Rivest.

The SAFE audits described in this paper choose n so as to set this probability very high, in the case where B represents the *minimum* number of miscounted precincts sufficient to alter the outcome of the election, which we will call B_{min} . In this case, this probability is also the *statistical power* of a random sample of size n to reject the *null hypothesis* that the initial outcome (apparent winner) was correct.

In statistics, a null hypothesis is one that is assumed to be true by default, that is, in the absence of evidence in favor of an alternative hypothesis. Our analysis here posits that election audits are being used to verify the outcome of an original electronic vote count – that is, to confirm who won. The audit can thus begin with the assumption (null hypothesis) that the outcome is correct, as originally reported, and determine whether there is evidence to the contrary. The initially reported outcome could be *incorrect* if votes were miscounted in some number of precincts. The larger the originally reported margin, the more precincts would have to have been miscounted in order to reverse the outcome. Only if an audit of an adequate number of randomly sampled precincts reveals no substantively important miscounts can the audit offer strong evidence that decisive miscounts did not occur.

In general, statistical power is defined as the probability of (appropriately) rejecting a null hypothesis when it is false. This probability depends crucially on “how false” that null hypothesis is. For example, the power to reject the null hypothesis when there is only one miscounted precinct will generally be quite low, and it will increase as the number of miscounted precincts in the race increases. For our purposes, the statistical power of primary interest is the probability of finding at least one miscounted precinct in the audited sample *when* there are enough miscounted precincts in the race to change the result. If power is high, then failure to reject the null hypothesis is highly informative, since we have ensured that (when the preliminary outcome is wrong) the probability of failing to reject is low. A recount of a random sample of precincts that reveals no substantively important miscounts thus offers strong evidence that decisive miscounts did not occur.

Appendix B: How are Statistical Power and Sample Sizes Calculated?

The power of a procedure to detect a miscount can only be calculated in the context of specific assumptions about exactly how the problem is configured. In general, when computing the power of an audit to detect a miscount, we assume a highly simplified “worst plausible case” scenario. An audit procedure that has good power to detect a miscount in that setting has at least that much power to detect a more realistic and complex problem configuration. Note that if all precincts contained vote shifts in the same direction, an audit of any size ($n \geq 1$) would have 100% power to detect that a miscount existed. Clearly it is not safe to design the audit assuming that the pattern of miscounts would be that easy to detect.

At the other extreme, if an election-altering miscount occurs in a single precinct, the sample would have to include, say, 95% of the precincts to have 95% power to detect that. However, when there are enough irregularities to change the outcome, it is usually not plausible (and unless the race were extremely close, not even logically possible) for all the miscounts to be contained within a single precinct.

A simple model of how an election-altering miscount might occur is to assume:

- 1) every precinct with any miscounts involves a 20% shift of votes to the advantage of the winner over the runner-up, and
- 2) the total number of precincts with miscounts is the minimum number required to change the outcome, given that,
- 3) the average number of votes in precincts with miscounts is the same as the average number of votes in all precincts.

The early calculations in this paper (used in Figures 2, 3, 4 and 7) refer to the chance of detecting at least one miscounted precinct under these three (highly simplified) assumptions. The assumptions are reasonably conservative in the following sense: if audits are designed to have good power in this situation they will have even better power if smaller miscounts per precinct are spread across more precincts.

However, the third assumption (that miscounted precincts are of average size) could lead to an overstatement of the effectiveness of the audit. Anyone wanting to tamper with an election would realize that it is easier to avoid detection by shifting many votes in a few large precincts rather than a few votes in each of many precincts. Thus, it may be wise to design audits to have good power to detect miscounts concentrated in the largest districts. The worst-case alternative assumption would be that miscounts occur in as many of the largest precincts as necessary to change the outcome.

It is not possible to calculate power under the “largest-precinct assumption” in a perfectly general way, because the calculation requires knowing how many of the largest precincts are needed to hide enough miscounts to reverse an election outcome. Figures 5 and 6 display real data from a single House race (Ohio CD-5 in 2004) to show how variability in precinct size can allow miscounts to be hidden in fewer precincts. In Figures 7, 8 and 9 we estimate the power for detecting a miscount assuming that the miscounts are concentrated in the largest precincts – using the data from Ohio CD-5 to estimate power in other races as if they had the same variation in votes per precinct as occurred in Ohio CD-5 in 2004.

Power calculations assuming that errors reside in average-sized precincts

Power calculations used in Figures 2, 3, 4 and 7 assume 1), 2) and 3), above, and proceed as follows:

To estimate the probability that a random audit of size n will detect miscounting in at least one sampled precinct *if* miscounts in the election are sufficient to reverse the outcome, we first determine or estimate the number of precincts that would have to be miscounted in order to reverse the outcome (as explained below). Then we apply the hypergeometric distribution, which in effect answers questions of the form: if a set of N objects contains B “bad” objects, what is the probability that a sample of n objects drawn without replacing any of them will contain at least b “bad” objects? N is the total number of precincts; B is the number of precincts that would have to be miscounted in order to reverse the outcome; b equals 0 if a particular audit sample does *not* include a miscounted precinct, or otherwise the number of precincts it includes.

For instance, if there are 500 precincts of which 50 are miscounted, the probability that an audit of 20 precincts will *not* include a miscounted precinct equals the hypergeometric distribution where $N = 500$, $B = 50$, $n = 20$, and $b = 0$. This probability equals about 11.6%. Thus, the power of this audit sample to detect miscounts is about $1 - 11.6\% = 88.4\%$ or 0.884. In Microsoft Excel, this value can be entered as:

$$1 - \text{HYPGEOMDIST}(0, 20, 50, 500).$$

The number of miscounted precincts that could alter the outcome depends upon the possible extent of miscounts in each precinct. In this analysis, we assume that 20% of the total vote could be switched from one candidate to another in each precinct (a 40-point shift in the percentage margin within that precinct). We refer to this value as the Within-Precinct Miscount (WPM). Thus, if all precincts contained the same number of votes, a 10-percentage-point margin could be overcome by switching 20% of votes in each of 25% of the precincts. More generally, a 10-percentage-point margin can be overcome by switching 20% of the votes in precincts containing 25% of all votes (2.5 times the margin).

Power calculations assuming that errors reside in the largest precincts

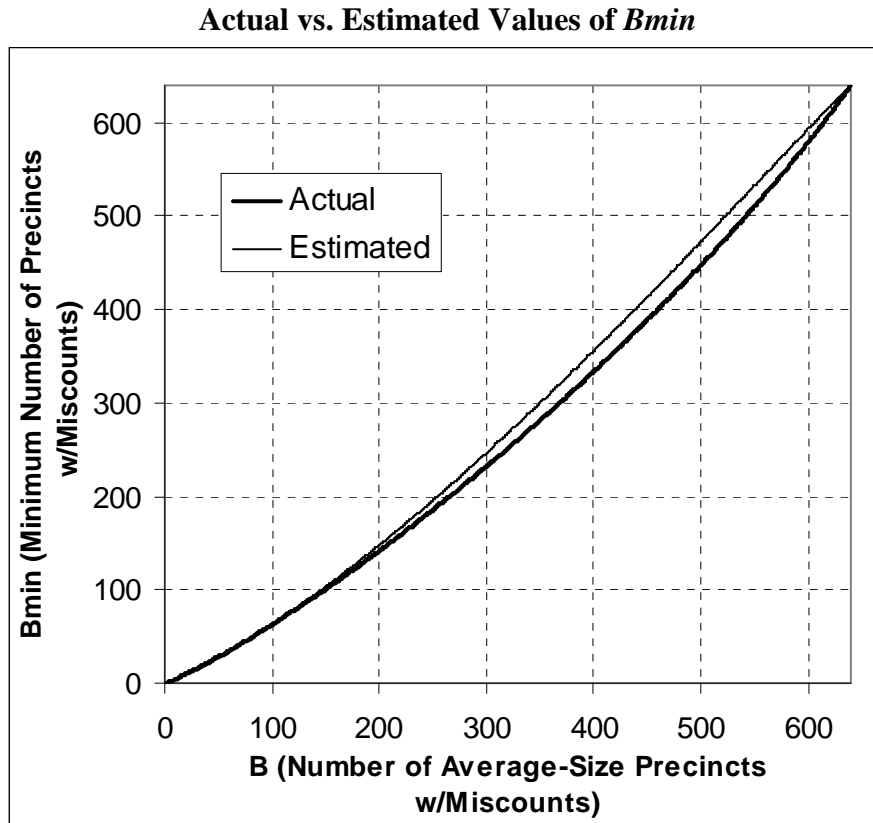
In Figure 7 we compare the power of various audit rules under the assumption that errors reside in average-sized precincts (as described just above), to the power of those audits assuming that the errors reside in the largest precincts. In Figures 8 and 9, the power calculations assume miscounts in the largest precincts. These calculations proceed as follows:

We first determine the *minimum* number of precincts containing a given percentage of all votes (in the case of the 20% WPM this would be 2.5 times the margin). It can be calculated directly if the distribution of precinct-level vote counts is available (from a preliminary report of precinct-level election returns), or it can be estimated based on the distribution of precinct-level votes in the previous election; a reference distribution (such as the Ohio CD-5 distribution shown in Figures 6 and 7); or using a heuristic approximation. Our calculations here use the formula:

$$B_{min} = \text{ceiling}(B / \log_{10}(1 / (B / N)) + 1)$$

to approximate the effect of the Ohio CD-5 distribution, where B_{min} is the *minimum* number of precincts that must be corrupted (assuming they are the *largest* ones), B and N are as above (thus, B / N would

approximately equal the proportion of votes in the miscounted precincts if they were average-sized), and in Microsoft Excel and several programming languages, ‘ceiling’ rounds the result up to the next whole number. The following graph shows values of B and B_{min} on the x-axis and y-axis respectively. True values of B_{min} obtained by using the actual precinct vote counts are shown by the thicker “Actual” line in the graph below, while the heuristic approximations using the above formula are shown by the narrower “Estimated” line.



Distributions that vary more or less in precinct size than Ohio CD-5 can be approximated by changing the base of the log in the above formula. A smaller base implies a greater concentration of votes in relatively few large precincts.

Sample sizes were calculated as follows:

Given B (or B_{min}), the minimum sample size n to achieve a given probability of sampling at least one miscounted precinct (i.e., a given power P) can be obtained experimentally by increasing n until the value of $1 - \text{HYPGEOMDIST}(0, n, B, N)$ is greater than or equal to P . However, there is an easier way that does not require software or a large lookup table.

Aslam, Popa and Rivest (2007) have derived a relatively simple computation. In our notation, the required sample size n equals approximately

$$\left\lceil \left(N - \frac{B-1}{2} \right) \cdot (1 - \exp(\ln(1-P)/B)) \right\rceil$$

where the outer half-brackets represent the ceiling (round-up) function. Aslam et al. report, “As a practical matter, this formula is essentially exact: we prove that it is never too small, and empirical testing for many representative values... never finds it more than one too large.”

Stanislevic has developed an open source JavaScript web page, using the above equation and the Ohio CD-5 heuristic approximation that can be used to calculate audit sample sizes from election results, assuming the above precinct size distribution: <http://mysite.verizon.net/evoter/AuditCalc.htm>

This equation can even be solved on a hand calculator, eliminating the need for software to calculate the sample size, once the value of *Bmin* has been determined from precinct-level vote counts, a reference distribution such as Ohio CD-5, or by using the above heuristic. A second equation that is identical can be used with a calculator without the *exp* or *log* functions (in either equation *Bmin* can be substituted for *B*):

$$\left\lceil \left(N - \frac{B-1}{2} \right) \cdot \left(1 - (1-P)^{1/B} \right) \right\rceil$$

Using the previous example and a hand calculator, if *N* = 500, *B* = 50 and we want a statistical power of 99% (.99), we could write and solve the above equation as:

$$\begin{aligned} n &= (500 - (50 - 1) / 2) \times (1 - (1 - .99)^{1/50}) \\ &= (500 - 24.5) \times (1 - (.01)^{.02}) \\ &= 475.5 \times .0878 \\ &= 42 \text{ precincts (when rounded up to next whole precinct)} \end{aligned}$$

Changing the statistical power *P* to 95% (.95) gives us:

$$\begin{aligned} n &= (500 - (50 - 1) / 2) \times (1 - (1 - .95)^{1/50}) \\ &= (500 - 24.5) \times (1 - (.05)^{.02}) \\ &= 475.5 \times .0582 \\ &= 28 \text{ precincts (when rounded up to next whole precinct)} \end{aligned}$$

Thus, once the correct value of *Bmin* is known, this method allows auditors, election officials, candidates and the public to ensure that the sample size of the audit is correct without depending on software.

Stanislevic has developed an Excel spreadsheet to automate this process using actual precinct-level vote count data for up to 5,000 precincts (expandable by Copy and Paste). The user must sort the precincts by vote count (Excel can be used to do this) and the worksheet generates the sample size and random precinct selections, including a dynamic chart. It can be downloaded here:

<https://vfv.jot.com/WikiHome/PublicDocuments/GraphicPaper/PrecinctDistUpTo5000.xls?cacheTime=1179818928947>

A web application by Dopp and Stenger that automatically sorts the precincts by vote count and calculates the sample size can be accessed online at <http://electionarchive.org/auditcalculator/eic.cgi>

Appendix C: A Brief Description of the SAFE Method for Auditing Vote Tabulations

Statutory guidance for audit teams*

Any procedure designed, adopted and implemented by the audit team shall be implemented to ensure, with at least 99% statistical power, that for each federal election held in the State, a 100% manual recount of the voter-verifiable paper records would not alter the electoral outcome reported by the audit. Such procedures designed, adopted and implemented by the audit team to achieve statistical power shall be based upon scientifically reasonable assumptions, with respect to each audited election, including but not limited to: the possibility that within any precinct up to twenty percent of the total votes cast may have been counted for a candidate other than the one intended by the voters; and that the number of votes cast in each precinct vary. Such procedures and assumptions shall be published prior to any given election, and the public shall have the opportunity to comment thereon.

* Adapted from New Jersey Bill No. S.507, as amended, June 8, 2007.

Summary implementation language

The following can be implemented by an audit team or included in regulations or legislation as long as the equation in section 2.2 of this summary can be included.

The following is a variable-percentage, fixed (99%) statistical-power-based audit applicable to all Federal [and possibly other] elections. If there are enough corrupt precincts, or, more generally, audit units (AUs), to alter the outcome of a contest, this protocol has 99% certainty of finding at least one. This shall be used in conjunction with trigger(s) for additional audits and/or full hand recounts:

2.1. The minimum number of AUs to change the outcome of any election as reported by the voting system shall be calculated as follows:

For each race in any federal election, the margin of victory between the two candidates receiving the largest number of votes as reported by the voting system shall be calculated. All AUs used in each race shall be sorted in descending order by the total number of votes counted in the race³⁴ in each AU. Beginning with the AU with the largest vote count, the minimum number of AUs containing at least one half the number of votes obtained by dividing the margin by the WPM parameter (i.e., 20% of the total vote count) – that is, containing at least 2.5 times the margin of votes – shall be determined. The resulting minimum number of corrupt AUs (Bmin) that could change the outcome of the race shall be used in the following equation, along with the total number of AUs used in each race (N), to determine the number of AUs to audit for each race.

2.2. The number of AUs to be audited in each race (n) shall be equal to:

$(N - (Bmin - 1) / 2) \times (1 - (1 - .99)^{1/Bmin})$ rounded up to the next greater integer,

where N is the total number of AUs and Bmin is the minimum number of corrupt AUs to change the outcome.

³⁴ It may be desirable to use the total number of *ballots cast* in each AU.

2.3. Each county or equivalent jurisdiction shall audit its pro rata share of the total number of AUs to be audited for each race (n) as determined by:

2.3.1. multiplying n by:

2.3.2 the quotient of: the number of AUs in the race which reside in each jurisdiction divided by the total number of AUs used in the race in total; and

2.3.3 rounding the product up to the next greater integer.