

Risk-limiting post-election audits: conservative P -values from common probability inequalities

Philip B. Stark

Abstract—Post-election audits of a random sample of batches of ballots against a trustworthy audit trail can limit the risk of certifying an incorrect electoral outcome to α , guaranteeing that—if the apparent outcome is wrong—the chance of a full hand count of the audit trail is at least $1 - \alpha$. Risk-limiting audits can be built as sequential tests that audit more batches until either (i) there is strong evidence that the outcome is correct, given the errors found, or (ii) there has been a complete hand count. The P -value of the hypothesis that the outcome is wrong is the largest chance, for all scenarios in which the outcome is wrong, that overstatements of the margins between winners and losers would be “no larger” than they were observed to be. Different definitions of “larger” give different P -values. A small P -value is strong evidence that the outcome is correct. This paper gives simple approaches to calculating a conservative P -value for several ways of summarizing overstatements and several ways of drawing the sample of batches to audit, emphasizing sampling with probability proportional to a bound u_p on the error in the p th audit batch (PPEB sampling). A P -value based on Markov’s inequality applied to a Martingale constructed from the data seems the most efficient among the methods discussed; there are plans to use it to audit contests in two California counties in November 2009.

Index Terms—Dvoretzky-Kiefer-Wolfowitz inequality, Hoeffding’s inequality, hypothesis test, Markov’s inequality, Martingale, monetary unit sampling (MUS), NEGEXP, PPEB, probability proportional to size, sequential test. EDICS: SEC-INTE, APP-CRIM, APP-INTE, APP-OTHE.

I. INTRODUCTION

THIS article concerns audits to assess whether the apparent winner of an election is the true winner. Statistical audits based on hand counts of an audit trail can assess, with pre-specified maximum risk, whether error caused the apparent outcome to differ from the outcome that a full hand count of the audit trail would show—regardless of the source of the error. Audits can serve other forensic purposes, including detecting anomalies characteristic of security breaches or equipment malfunctions.

The goal addressed here is to determine from a sample whether the apparent (semi-official) outcome agrees with the outcome that a full hand count of the audit trail would show. If it does not, there needs to be a full hand count to set

the record straight. Statistical evidence generally cannot be used to overturn an election outcome; if it could, sampling variability might result in disenfranchising the majority, which seems constitutionally intolerable. An audit can err. It can conclude that the apparent outcome is correct when a full hand count would show that the apparent outcome is wrong, or it can require a full hand count when that will merely confirm that the apparent outcome is right. The first error is more serious: The wrong candidate wins. The second error wastes resources. In the hypothesis-testing formulation of election auditing proposed by [1], the first error corresponds to a classical Type I error and the second to a Type II error.

An audit that has a known pre-specified minimum chance of requiring a full hand count whenever that would show a different outcome is called *risk-limiting*.¹ To the best of my knowledge, there have been only four risk-limiting election audits to date, all in California: an audit in Marin County in February 2008, and audits in Marin, Santa Cruz and Yolo counties, in November 2008 [2]. I designed and supervised all four. The first used a method introduced by [1], [3], based on stratified random sampling. The same general method was used in Yolo County in November 2008. The November audits in Marin and Santa Cruz counties used samples drawn with probability proportional to an error bound [4]; the auditing method is presented in [5]. See also section IV-C1 below.

Suppose that—given the reported election results, the sampling procedure, and the errors the audit found—we can say that either a full hand count would show the same outcome, or an event with probability no greater than P occurred. The number P is the P -value of the hypothesis that the outcome is wrong. Smaller P -values are stronger evidence that the outcome is right.

The P -value of the hypothesis that the outcome is wrong is a useful summary of any audit, and has a central role in risk-limiting audits. [1], [3], [6] give a general approach to constructing risk-limiting audits. The audit is conducted in stages, indexed by s . There is a sequence of thresholds $\{\alpha_s\}_{s=1}^S$. In each stage, additional batches of ballots are audited, and a P -value is computed.² The P -value is compared to α_s . If the P -value is less than α_s , the audit can stop, and the auditor can recommend that the outcome be certified. If not, more batches of ballots are audited. If, by stage $s = S$, the

Copyright (c) 2009 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org

Department of Statistics, University of California, Berkeley CA, 94720-3860, USA. e-mail: (see <http://statistics.berkeley.edu/users/stark>).

I am grateful to Steve Evans, Mike Higgins, Harold Kaplan, Mark Lindeman, John McCarthy, Luke Miratrix and Aviva Shimelman for conversations and comments on an earlier draft and to anonymous referees for helpful suggestions.

Manuscript received 19 February 2009; revised 8 September 2009.

¹See <http://www.electionaudits.org/bp-risklimiting> (last accessed 19 February 2009).

²In [1], the P -value is based on all the data collected. In [6], the P -value is based on the data collected in the current stage, conditional on the audit results at all prior stages, which can include audits of deliberately selected batches.

audit has not stopped, there is a complete hand count, and the result of that count is reported as the official outcome. [1], [3], [6] give details for simple and stratified random samples, but the general strategy works for any sampling plan and family of tests for which one can calculate a P -value. The present paper gives methods to calculate a conservative P -value for several sampling plans using tests based on the maximum relative overstatement of pairwise margins [3], focusing on sampling with probability proportional to an error bound (PPEB) [4]. “Conservative” means that if the outcome is wrong, the P -value might be bigger than the probability of observing no more error than was observed, but never smaller: The methods described here never overstate the strength of the evidence that the outcome is correct.

A. Connections to Financial auditing

Advantages of using statistical methods to draw financial audit samples have been known for more than fifty years (e.g., [7], [8], [9]). It has been known for almost as long that standard statistical techniques can be grossly inaccurate for analyzing accounting errors because accounting errors tend to be quite skewed—most errors are near zero, but some are very large. See, e.g., [10], [11].

Populations of vote-counting errors are similar to populations of accounting errors: Typically, the error in the count of a precinct-size batch of ballots is at most a few votes. However, fraud, bugs, and problems such as miscalibrated optical scanners, ballot definition errors, database or memory card “glitches,” lost memory cards, and lost (or “found”) boxes of ballots can produce discrepancies of thousands of votes.

In both financial and electoral auditing, the key question is not whether there is any error at all in the population audited—there usually is—but rather, whether the total error is *material*. Error in a financial report is typically considered material if a reasonable person relying on the report would have acted differently, but for the error. Materiality in electoral auditing is more straightforward: Error in reported election results is material if it makes the winning candidate(s) or position appear to lose.³

To test whether the total value of a financial account is materially overstated, it is common to sample items with probability proportional to their book value, rather than with equal probabilities. This is called *monetary unit sampling* (MUS) or *dollar unit sampling*. See, e.g., [12], [13], [14], [15], [10].⁴ [10] seems to be the first to report using MUS in financial auditing; see [11] for the history of MUS. In inferences about accounting overstatements, it is common to assume that the book value of an item is an upper bound on

³However, in determining whether a direct-recording electronic voting machine (DRE) is functioning correctly, *any* discrepancy between a complete, readable paper trail and the electronic record is material: Even a single error in DRE results indicates a bug or fraud. With voter-marked paper ballots, optical scanners will occasionally misread voter intent, if only because of variability in voters’ marks. If the margin is small enough, that can change the apparent outcome.

⁴Some methods rely on parametric approximations [16], [17], Bayesian prior distributions [18], [19], [20], [21], [22] or numerical simulations [23], [24], [25]; the reliability of such methods rests on assumptions that are largely untestable.

the amount by which the value of the item was overstated: At worst, the item is worthless. MUS is PPEB: Items that can have more error have proportionately larger probability of being drawn than items that can have less error.

B. Sampling schemes for election auditing

Existing legislation for post-election audits draws batches by simple random sampling (SRS) or stratified random sampling [2]. Sampling schemes that give more scrutiny to batches that can hide more error can be more efficient [4]. Suppose that the ballots are divided into N auditable batches and that we have an a priori upper bound u_p on the error e_p in batch p .⁵ Let $U \equiv \sum_{p=1}^N u_p$. [4] proposes two ways of drawing probability samples of batches for post election audits: NEGEXP, where the probability of auditing batch p is $1 - \exp(-\gamma u_p)$ and batches are selected independently; and PPEB, which makes n independent draws from the pool of batches, with probability u_p/U of drawing batch p each time.⁶ PPEB leads to test statistics that have attractive theoretical features; see section IV-C.

In November 2008, PPEB was field-tested in Marin and Santa Cruz counties, California, [5] and NEGEXP was field-tested in Boulder County, Colorado.⁷ If PPEB or NEGEXP sampling is sanctioned by audit laws, substantial gains in efficiency are possible. Nonetheless, there are arguments against methods other than simple or stratified random sampling for election audits. Such methods are less transparent to the public and jurisdictional users. Drawing a PPEB or NEGEXP sample requires first computing error bounds, which can require knowing the apparent outcome in every batch in the contest, across jurisdictions. Using dice or other simple physical sources of randomness to draw a NEGEXP or PPEB sample is more complicated than it is for simple or stratified random sampling; it typically requires the use of software or look-up tables. There are legal issues, too: If batches are sampled using simple random sampling or using stratified random sampling with equal sampling fractions in every stratum, every ballot has the same chance of being audited. That is not the case for NEGEXP and PPEB audits: Ballots cast in batches with larger error bounds are more likely to be audited. This could raise questions of equal protection or of differing chances of disenfranchisement in different batches.

⁵How to measure error and find an a priori bound on the error is addressed below. This paper uses the *maximum relative overstatement of pairwise margins* [3], the largest amount by which error inflated the margin of any apparent winner over any apparent loser, expressed as a fraction of the apparent margin between them.

⁶See also [26]. The sample size calculations in [4] for PPEB are a special case of formulae in [10].

⁷See <http://bcn.boulder.co.us/~neal/elections/boulder-audit-08-11/> (last accessed 19 February 2009). That site claims that the Boulder County audit was risk-limiting, but it does not meet the definition used here and in <http://www.electionaudits.org/bp-risklimiting> (last accessed 19 February 2009). It did not “have a large, pre-determined minimum chance of leading to a full recount whenever a full recount would show a different outcome,” for reasons including: (i) It did not have *any* rule for determining when to do a full hand count. (ii) It audited only the Boulder County portion of statewide contests. (iii) It used an ad hoc value for u_p . (iv) For local contests, the chance of finding any error at all on the assumption that the outcome was wrong varied by contest, as a result of how the auditing effort was allocated across contests. See [2] for more discussion of the Boulder County audit.

C. Previous work: detecting error versus limiting risk

Most research on election auditing has asked how large a sample is needed to have a high chance of detecting at least one error, on the assumption that the apparent outcome is wrong [27], [28], [29], [30], [4]. This is the “detection” formulation of auditing. In practice, auditing even a handful of precinct-size batches of voter-marked ballots often finds at least one discrepancy, so answering that question is not so helpful.

A more useful question to answer is, “Given the reported results of the contest, the design of the sample, and the discrepancies the audit found, how strong is the evidence that the apparent outcome is correct?” This is the “risk” formulation of auditing.

[1], [3], [6] answer the latter question, focusing on audits that select batches using simple or stratified random sampling. [1] argues that an audit procedure should always either recommend certifying the election outcome or require a full recount; otherwise, the procedure is incomplete. The procedure should have an error rate that can be quantified. For instance, it could guarantee that if the apparent outcome is wrong, the probability that the method will require a full hand count is at least $1 - \alpha$. The methods of [27], [28], [29], [30], [4] have this property only if a full hand count is conducted whenever the audit discovers any error at all.

D. Organization of the paper

This paper shows how to use common (and some less common) probability inequalities to find conservative P -values for the hypothesis that the outcome differs from the outcome a full hand count would show, given the discrepancies found in SRS, NEGEXP, or PPEB samples. As described in [1], [3], [6], a risk-limiting audit can be built by wrapping the P -value calculation in an iterative procedure—sample, assess evidence, stop or sample more.

Section II explains P -values. Section III sets out notation and gives a necessary condition for the election outcome to be incorrect. Section IV sketches three basic strategies for testing the hypothesis that the necessary condition holds; draws connections between SRS, NEGEXP and PPEB-based methods for detecting at least one error and how they relate to the method of [1], [3]; and gives a simple method for accommodating errors uncovered in SRS, NEGEXP or PPEB sampling. Section IV-C focuses on PPEB sampling and methods that use the observed distribution of errors, connecting election auditing to tests of hypotheses about the expected value of a nonnegative random variable from an iid sample. Section IV-C2 shows how to use Markov’s inequality to find a P -value for the hypothesis that the outcome is wrong, given the discrepancies uncovered in a PPEB sample. Section IV-C3 shows how to use the Kiefer-Dvoretzky-Wolfowitz inequality for that purpose and section IV-C4 shows how to use Hoeffding’s inequality.

Section IV-D finds a P -value using a method due to [31] for constructing lower confidence bounds for the mean of a nonnegative random variable. Section IV-E compares several

P -value calculations from PPEB samples in different hypothetical scenarios. The “Kaplan-Markov” P -value described in section IV-D is uniformly the best in the tests—it gives the smallest P -values. Section V summarizes the findings.

II. P -VALUES

This paper is about P -values, so a definition is in order. P -values are a way of summarizing the evidence against a hypothesis: The smaller the P -value, the less credible the hypothesis. To assign a numerical P -value requires quite a bit of structure. Here is one formal setup that allows P -values to be defined.

There is a hypothesis about the world. Datum Y will be collected; Y is a random variable that takes values in a measurable space \mathcal{Y} . The datum could be multivariate. If the hypothesis is true, the probability distribution G of Y is in some set \mathcal{G} of probability distributions. We identify the hypothesis with the assertion $G \in \mathcal{G}$; if the hypothesis is false, $G \notin \mathcal{G}$. For simplicity, we suppose that all the distributions in \mathcal{G} are defined on the same σ -algebra, denoted Σ . For every value $\alpha \in (0, 1)$, we have a set $R_\alpha \in \Sigma$ such that

$$\sup_{G \in \mathcal{G}} \mathbb{P}_G\{Y \in R_\alpha\} \leq \alpha. \quad (1)$$

That is, if the hypothesis $G \in \mathcal{G}$ is true, the chance that the datum Y will be in the set R_α is no greater than α . We call R_α the *rejection region* of a significance-level α test of the hypothesis. The hypothesis is rejected at significance level α if $Y \in R_\alpha$. If we conclude that the hypothesis is false whenever $Y \in R_\alpha$, the chance of erroneously concluding that the hypothesis is false if it is in fact true is at most α .

To define P -values, we shall also insist that the rejection regions nest: If $0 < \beta < \gamma < 1$, then

$$R_\beta \subset R_\gamma. \quad (2)$$

This ensures that if the family of tests rejects the hypothesis at significance level α , it also rejects the hypothesis at every significance level greater than α . We also define $R_1 \equiv \mathcal{Y}$.

Given this structure, the P -value of the hypothesis given the observation $Y = y$ is

$$P \equiv \inf\{\alpha \in (0, 1] : y \in R_\alpha\}. \quad (3)$$

That is, the P -value is the smallest significance level at which the family of tests rejects the hypothesis. Because of the nesting, the family rejects the hypothesis for all significance levels greater than the P -value.

The P -value depends not only on the hypothesis \mathcal{G} and the observation that $Y = y$, but also on the family of hypothesis tests (the sets R_α , $\alpha \in (0, 1]$). Different tests in general give different P -values for the same hypothesis given the same data. One test is better than another if (i) no matter what value y the datum Y takes, the first test assigns a P -value no larger than the second test does, and (ii) there is some $y \in \mathcal{Y}$ such that the first test assigns a smaller P -value than the second when $Y = y$.

In this paper, the hypothesis is that the reported election outcome is wrong. That hypothesis corresponds to a family \mathcal{G} of probability distributions for the datum Y , a vector of

errors observed in a random sample of reported election results normalized in a particular way—the maximum overstatement of pairwise errors [3], described below. Each component of Y represents an audited batch. A component is positive if error in the batch resulted in overstating the margin between an apparent winner and an apparent loser. The probability distribution G of Y depends on how the sample is drawn and on the true distribution of votes in every batch.

Many of the tests discussed below are defined by a test statistic—a single-number summary $\phi(Y)$ of the vector Y , where ϕ is a Σ -measurable function from \mathcal{Y} to \mathbb{R} . The value $\phi(Y)$ measures the overall “size” of the errors that the audit uncovers. For instance, $\phi(Y)$ might be the largest component of Y . The rejection region is then expressed in terms of $\phi(Y)$: The hypothesis is rejected if $\phi(Y) \leq f_\alpha$, where f_α is chosen to satisfy

$$\sup_{G \in \mathcal{G}} \mathbb{P}_G\{\phi(Y) \leq f_\alpha\} \leq \alpha. \quad (4)$$

The nesting condition requires that for $0 < \beta < \gamma < 1$,

$$\{y \in \mathcal{Y} : \phi(y) \leq f_\beta\} \subset \{y \in \mathcal{Y} : \phi(y) \leq f_\gamma\}; \quad (5)$$

i.e., f_α should increase monotonically with α . The P -value of the hypothesis for the observation $Y = y$ is then

$$P = \inf\{\alpha : \phi(y) \leq f_\alpha\}. \quad (6)$$

The following sections specialize this definition to election auditing and consider particular tests of the hypothesis that the apparent electoral outcome is wrong.

III. ASSUMPTIONS AND NOTATION

The audit trail is the legal record; a hand count of the audit trail determines the true outcome.⁸ We want to determine whether a full hand count of the audit trail would show a different result than the apparent result—the result that would become official but for the audit—and we want to know how reliable that determination is.

We consider one contest at a time, of the form “vote for up to f candidates.”⁹ There are K candidates in the contest. Election results are subtotaled separately in N auditable batches, $p = 1, \dots, N$.¹⁰ Let a_{kp} denote the number of votes for candidate k in batch p that an audit would show. The total number of votes a complete hand count would show for candidate k is $A_k \equiv \sum_{p=1}^N a_{kp}$.¹¹ Let v_{kp} be the reported number of votes for candidate k in batch p . The total number of votes reported for candidate k is $V_k \equiv \sum_{p=1}^N v_{kp}$. Let $V_{w\ell}$ be the apparent margin of candidate w over candidate ℓ :

$$V_{w\ell} \equiv V_w - V_\ell. \quad (7)$$

⁸If the audit trail is so incomplete or inaccurate that it does not reflect the true electoral outcome, then confirming that the apparent outcome is the outcome a full hand count would show is not helpful. Using voting systems that create a durable and accurate audit trail and maintaining proper chain of custody of that audit trail are obviously crucial.

⁹See [32] for a method to audit a collection of contests simultaneously.

¹⁰The symbol p is mnemonic for *precinct*, although batches need not correspond to precincts. A batch might consist of the ballots tabulated by a particular machine on election day, or a “deck” of ballots cast by mail that are run through optical scanners as a group. But subtotals must be reported for every batch, and it must be possible to isolate and hand count the audit record for any batch.

¹¹The symbols a and A stand for *audit* or *actual*.

Let \mathcal{W} be the set of indices of the f apparent winners and let \mathcal{L} be the set of indices of the $K - f$ apparent losers. If $w \in \mathcal{W}$ and $\ell \in \mathcal{L}$, then $V_{w\ell} > 0$. Candidate w really did beat candidate ℓ if $A_{w\ell} \equiv A_w - A_\ell > 0$.

We summarize error using the maximum relative overstatement of the margin between any apparent winner and any apparent loser [3], which we now define. Let $e_{w\ell p}$ be the relative overstatement of the margin between apparent winner $w \in \mathcal{W}$ and apparent loser $\ell \in \mathcal{L}$ in batch p :

$$e_{w\ell p} = \frac{v_{wp} - v_{\ell p} - (a_{wp} - a_{\ell p})}{V_{w\ell}}. \quad (8)$$

The outcome of the race is correct if for every apparent winner $w \in \mathcal{W}$ and apparent loser $\ell \in \mathcal{L}$,

$$\sum_{p=1}^N e_{w\ell p} < 1. \quad (9)$$

Let

$$e_p \equiv \max_{w \in \mathcal{W}, \ell \in \mathcal{L}} e_{w\ell p}. \quad (10)$$

This is the maximum relative overstatement of pairwise margins in batch p . The apparent outcome must agree with the outcome a full hand count would show if

$$E \equiv \sum_{p=1}^N e_p < 1. \quad (11)$$

The condition $E \geq 1$ is necessary but not sufficient for the apparent outcome to be wrong. Hence, testing the hypothesis $E \geq 1$ gives a conservative test of the hypothesis that the apparent outcome is wrong.

How strong is the evidence that $E < 1$, given the errors $\{e_p\}$ discovered by the audit? As noted in section II, the P -value of the hypothesis $E \geq 1$ depends not only on the data but also on how the sample is drawn and on the family of tests under consideration. Moreover, to have strong evidence that the outcome of the contest is correct without auditing the majority of batches requires an upper bound on the N values $\{e_p\}$.

We can derive an upper bound on e_p from independent information. Let b_p be a bound on the number of votes cast for any single candidate in batch p . For example, if p represents a precinct, b_p might be the number of voters registered in the precinct, the number of ballots sent to the precinct, or the number of pollbook signatures. The overstatement of the margin between winner w and loser ℓ is largest when all b_p votes were for ℓ , but were not reported that way:

$$\begin{aligned} e_{w\ell p} &\leq \frac{v_{wp} - v_{\ell p} - (0 - b_p)}{V_{w\ell}} \\ &= \frac{b_p + v_{wp} - v_{\ell p}}{V_{w\ell}}. \end{aligned} \quad (12)$$

Define

$$u_p \equiv \max_{w \in \mathcal{W}, \ell \in \mathcal{L}} \frac{b_p + v_{wp} - v_{\ell p}}{V_{w\ell}}. \quad (13)$$

Then

$$e_p \leq u_p. \quad (14)$$

Define

$$t_p \equiv e_p/u_p \leq 1. \quad (15)$$

We call t_p the *taint* in batch p : It is the ratio of the largest relative overstatement of any margin of an apparent winner over an apparent loser in batch p , divided by the maximum possible relative overstatement of the margin of an apparent winner over an apparent loser in batch p . The bounds $\{u_p\}_{p=1}^N$ can be calculated from $\{b_p\}$ and the reported election results, but the error e_p and the taint t_p are known only if batch p is audited. The fact that $t_p \leq 1$ will be exploited in section IV-C.

The total of the maximum possible relative overstatements is

$$U \equiv \sum_{p=1}^N u_p. \quad (16)$$

Auditing is superfluous if $U < 1$, because then there cannot be enough error to cause the wrong candidate to appear to win. Therefore, we assume that $U \geq 1$. The total attained error is

$$E \equiv \sum_{p=1}^N e_p \equiv \sum_{p=1}^N u_p t_p. \quad (17)$$

We now examine several nonparametric approaches to testing the hypothesis $E \geq 1$.

IV. AUDIT STRATEGIES

A risk-limiting audit can be performed by testing the hypothesis that $E \geq 1$. If the hypothesis is rejected, the audit stops without a full hand count. We would like a *conservative* test—one that does not understate the chance of rejecting the hypothesis when in fact $E \geq 1$. We shall discuss a variety sampling plans and test statistics.

Current auditing approaches use one of three basic strategies to test whether $E \geq 1$. The first strategy looks for *any* error that overstated a margin. It requires a full hand count if any observed $e_p > 0$. The second strategy is a refinement of the first: It looks for an error that overstates the margin by some threshold. That threshold can be larger than zero and can vary by batch. If an error exceeding the threshold is found, the method requires further auditing, but not necessarily a full hand count. These two strategies are based on *detection*: If no batch in an appropriately drawn sample has a (large) positive error and the sample is drawn appropriately, one can infer that there is no set of batches for which the total error is greater than or equal to one; hence, $E < 1$. The third strategy uses more details of the distribution of error in the sample. This can reduce the sample size needed to stop the audit and certify the outcome; in particular, errors that understated the margin can help. Examples are given in section IV-C.

Because audits typically find error, the first strategy is very likely to result in a full hand count, even when a full count agrees with the apparent outcome. The methods of [27], [29], [4] are in this category.¹² The second strategy permits some

¹² Those papers do not address risk-limiting audits. They calculate the probability of finding one or more errors when the outcome is wrong, rather than assess the evidence that $E < 1$ when the audit finds error, the typical case. To use those methods for risk-limiting audits, one must demand a full hand count if the audit finds any error at all.

errors provided there is strong evidence that the total error is not large enough to account for the entire apparent margin. Because it tolerates some error, it requires larger samples than the first strategy: It needs evidence that despite the error in the sample, the total error in all batches is still small. The methods of [1], [3], [6] are in this category. This paper extends the strategy to PPEB and NEGEXP sampling.

The third strategy is most closely related to approaches used in financial auditing [10], [15], [11]. It uses the observed distribution of errors—not merely the number above some threshold—to draw inferences about E . This paper gives several ways of using the distribution of error in PPEB samples and points to other approaches.

A. Strategy 1: Look for any error at all

If $E \geq 1$, there must be some set of “tainted” batches \mathcal{T} such that

$$\sum_{p \in \mathcal{T}} e_p \geq 1 \text{ and } e_p > 0, \forall p \in \mathcal{T}. \quad (18)$$

If \mathcal{T} exists, it must satisfy

$$\sum_{p \in \mathcal{T}} u_p \geq 1, \quad (19)$$

since $e_p \leq u_p$. If no such \mathcal{T} exists, the apparent outcome of the election is the outcome a full hand count would show.

This strategy tries to find strong statistical evidence that $E < 1$ by drawing a sample in such a way that if \mathcal{T} did exist, the sample would be very likely to contain at least one batch $p \in \mathcal{T}$. If the sample contains no batch with $e_p > 0$, that is evidence that $E < 1$. In this test, the datum Y is an n -vector of observed taints and the test statistic is

$$\phi(Y) \equiv \max_{j=1}^n Y_j. \quad (20)$$

The test rejects the hypothesis that the outcome is incorrect if $\phi(Y) \leq 0$. Details depend on how the sample is drawn.

1) *SRS sampling*: If the sample is a simple random sample of n of the N batches, the chance that the sample contains at least one batch with $e_p > 0$ is

$$1 - \frac{\binom{N - \#\{p: e_p > 0\}}{n}}{\binom{N}{n}}. \quad (21)$$

This increases monotonically as the number of batches p for which $e_p > 0$ increases; therefore, if \mathcal{T} exists, the chance of finding one or more batches with $e_p > 0$ is smallest if \mathcal{T} concentrates the errors in as few batches as possible. Let $d = d(u)$ be the smallest integer k for which the sum of the largest k elements of the multiset $\{u_p\}_{p=1}^N$ is greater than or equal to 1. If \mathcal{T} exists, necessarily $\#\mathcal{T} \geq d(u)$. Hence, if $E > 1$, then the chance that a simple random sample of n of the N batches contains no element of \mathcal{T} is at most

$$\frac{\binom{N - d(u)}{n}}{\binom{N}{n}}. \quad (22)$$

For this strategy, if $e_p \leq 0$ for every batch p in the SRS sample, (22) is the P -value of the hypothesis that $E \geq 1$.

2) *NEGEXP sampling*: Consider selecting batches for audit independently, with probability $1 - \exp(-\gamma u_p)$ of selecting batch p , where $\gamma > 0$ is a fixed constant. Suppose \mathcal{T} exists. What is the chance that the sample will contain one or more batches p with $e_p > 0$? The chance that no element of \mathcal{T} is in the sample is

$$\begin{aligned} \prod_{p \in \mathcal{T}} (1 - (1 - \exp(-\gamma u_p))) &= \prod_{p \in \mathcal{T}} \exp(-\gamma u_p) \\ &= \exp(-\gamma \sum_{p \in \mathcal{T}} u_p) \\ &\leq \exp(-\gamma), \end{aligned} \quad (23)$$

by (19). Hence, for this strategy, if $e_p \leq 0$ for every batch p in the sample, the P -value of the hypothesis that $E \geq 1$ is $\exp(-\gamma)$.

3) *PPEB sampling*: We draw n times at random, independently (with replacement), from the set of N batches, with probability u_p/U of selecting batch p in each draw. Suppose \mathcal{T} exists. Then the chance that the j th draw is an element of \mathcal{T} is

$$\sum_{p \in \mathcal{T}} \frac{u_p}{U} \geq 1/U. \quad (24)$$

Since the draws are independent, the chance that none of the n draws gives an element of \mathcal{T} is at most

$$(1 - 1/U)^n. \quad (25)$$

So, for this strategy, if $e_p \leq 0$ for every batch p in the sample, the P -value of the hypothesis $E \geq 1$ is $(1 - 1/U)^n$.

B. Strategy 2: Look for taint that exceeds a threshold

This section sketches the second method in its simplest form; in particular, it does not consider stratified samples or general weight functions for the error in each batch. For a more thorough treatment, see [1], [6].

Let $t \in [0, 1/U)$. Recall that $t_p \equiv e_p/u_p$. We will answer the question, ‘‘What is the chance that the audit would have found no batch p with t_p larger than t if $E \geq 1$, given the design and ‘size’¹³ of the sample?’’ This is the P -value of the hypothesis $E \geq 1$ for a test that rejects when $\phi(Y) \leq t$. A small P -value is strong evidence that $E < 1$ and hence that the apparent outcome of the race is correct.

The ability to calculate P -values also lets us answer the question, ‘‘If we would like to be able to certify the outcome of the race provided that the audit finds $t_p \leq t$ for every batch p in the sample, how big a sample do we need to guarantee that the chance of a full hand count is at least $1 - \alpha$ if the outcome is wrong?’’ By wrapping the P -value calculation in an iterative expansion of the sample as described above, we can construct risk-limiting audits. See [1], [3], [6].

There are many possibilities at this juncture. Here is a simple choice that allows a unified treatment of different sampling plans and reduces the problem to the problem solved

¹³ The measure of ‘‘size’’ depends on how the sample is drawn: For SRS, it is the number of batches. For NEGEXP, it is the scale factor γ . For PPEB, it is the number of draws, which tends to be larger than the number of batches drawn because duplicates are possible.

in section IV-A. The basic idea is to reduce the apparent margin to allow for a ‘‘background’’ taint of t in every batch, then to apply the method of the previous section to the reduced margin by looking for taints that exceed t .

For any set $\mathcal{T} \subset \{1, \dots, N\}$, $\sum_{p \in \mathcal{T}} u_p \leq U$. Hence, if $E \geq 1$, there must be some set of batches \mathcal{T}_t such that

$$\sum_{p \in \mathcal{T}_t} (e_p - t u_p) \geq 1 - tU; \quad (26)$$

i.e.,

$$\sum_{p \in \mathcal{T}_t} \frac{e_p - t u_p}{1 - tU} \geq 1. \quad (27)$$

Define

$$\tilde{e}_p \equiv \frac{e_p - t u_p}{1 - tU} \quad (28)$$

and

$$\tilde{u}_p \equiv \frac{u_p - t u_p}{1 - tU} = u_p \cdot \frac{1 - t}{1 - tU}. \quad (29)$$

Then $\tilde{e}_p \leq \tilde{u}_p$, $p = 1, \dots, N$. Define

$$\tilde{U} \equiv \sum_{p=1}^N \tilde{u}_p = U \frac{1 - t}{1 - tU}. \quad (30)$$

If the apparent outcome differs from the outcome a full hand count would show, there must exist some $\mathcal{T}_t \subset \{1, \dots, N\}$ such that

$$\sum_{p \in \mathcal{T}_t} \tilde{e}_p \geq 1 \text{ and } \tilde{e}_p > 0, \quad \forall p \in \mathcal{T}_t. \quad (31)$$

If \mathcal{T}_t exists,

$$\sum_{p \in \mathcal{T}_t} \tilde{u}_p \geq 1. \quad (32)$$

This is structurally the same condition tested in section IV-A. Moreover, because $\tilde{u}_p \propto u_p$, NEGEXP sampling using \tilde{u}_p is the same as NEGEXP sampling using u_p , but with a different value of γ . And PPEB sampling using error bounds $\{\tilde{u}_p\}_{p=1}^N$ is identical to PPEB sampling using error bounds $\{u_p\}_{p=1}^N$.

Hence, the calculations in section IV-A apply, provided we substitute \tilde{e}_p for e_p , \tilde{u}_p for u_p (thereby substituting \tilde{U} for U and $d(\tilde{u})$ for $d(u)$). If we fix t , those calculations tell us how large a sample to take to have chance at least $1 - \alpha$ of finding at least one batch p with $t_p \geq t$ if the outcome is wrong.

If instead we take t to be the maximum value of t_p in the sample, we can use the calculations in section IV-A to solve for α . That value of α is the maximal probability that no taint larger than t would be observed, if $E \geq 1$: It is a conservative P -value for the hypothesis that the outcome is wrong.

C. P -values using the distribution of observed error in PPEB sampling

We come to the third strategy, which uses more details of the observed distribution of taint. Part of the magic of PPEB sampling is that the taint is bounded above by 1 and that expected value of the taint in each draw is E/U , as we shall see presently. These facts are exploited by the Stringer bound and its sharpened version [10], [13], [14], by the multinomial bound [15], and by the trinomial bound [5].

They also make it possible to find P -values from PPEB samples using standard probability inequalities, as illustrated in the next three subsections.

The total error E can be written

$$E \equiv \sum_{p=1}^N e_p = \sum_{p=1}^N t_p u_p = U \sum_{p=1}^N t_p \frac{u_p}{U}. \quad (33)$$

Suppose we draw a batch at random, with chance u_p/U of drawing batch p : a PPEB sample of size 1. Let T be the taint t_p of the batch that is selected. Then

$$\mathbb{E}T \equiv \sum_{p=1}^N t_p \frac{u_p}{U} = \frac{E}{U}. \quad (34)$$

The outcome of the race must be correct if $\mathbb{E}T < 1/U$. The smallest significance level at which we can reject the hypothesis $\mathbb{E}T \geq 1/U$ is the P -value of the hypothesis $E \geq 1$.

Suppose we draw a PPEB sample of size n : We draw a batch n times at random, independently, with probability u_p/U of drawing batch p in each draw. Let $\{T_j\}_{j=1}^n$ be the taints of the n not necessarily distinct batches that are drawn. Then $\{T_j\}$ are independent and identically distributed with the same distribution as T (they are iid T). Moreover, since $t_p \leq 1$, $\mathbb{P}\{T_j \leq 1\} = 1$: The variables $\{T_j\}$ are all bounded above by 1. The next few subsections use these facts to derive nonparametric upper confidence bounds for $\mathbb{E}T$ from $\{T_j\}$.

If we define $X_j = 1 - T_j$ then $\{X_j\}$ are iid nonnegative random variables. We can reject the hypothesis $\mathbb{E}T \geq 1/U$ if we can reject the hypothesis $\mathbb{E}X_j \leq 1 - 1/U$. Thus, methods for testing hypotheses about the mean of a nonnegative random variable can be used to solve the auditing inference problem from a PPEB sample.

In what follows, we will use the definitions

$$\begin{aligned} T^- &\equiv \min_{j=1}^n T_j \\ T^+ &\equiv \max_{j=1}^n T_j \\ \bar{T} &\equiv \frac{1}{n} \sum_{j=1}^n T_j. \end{aligned}$$

1) *P-values from the Binomial*: Consider replacing T_j by a random variable that is stochastically larger:

$$\tilde{T}_j = \begin{cases} t, & T_j \leq t, \\ 1, & \text{otherwise.} \end{cases} \quad (35)$$

This variable is at least as large as T_j with probability 1. Hence, $\mathbb{E}\tilde{T}_j \geq \mathbb{E}T_j$. Now

$$\mathbb{E}\tilde{T}_j = t + (1-t)\mathbb{P}\{T > t\}, \text{ so} \quad (36)$$

$$\mathbb{E}\tilde{T}_j \geq 1/U \text{ iff } \mathbb{P}\{T > t\} \geq \frac{1/U - t}{1-t} \equiv \pi_U(t). \quad (37)$$

Define the test statistic

$$S = \phi(Y) \equiv \#\{j : T_j > t, j = 1, \dots, n\}; \quad (38)$$

i.e., S is the number of observed taints in the sample that exceed t . Whether each observation T_j is greater than t is a Bernoulli trial with probability $\mathbb{P}\{T > t\}$ of success; the n

trials are independent. Thus S has a binomial distribution with parameters n and $\mathbb{P}\{T > t\}$. We would reject the hypothesis $E \geq 1$ at significance level α if $S \leq s_\alpha$, where s_α is the largest integer for which, on the assumption that $S \sim \text{Bin}(n, \pi_U(t))$, $\mathbb{P}\{S \leq s_\alpha\} \leq \alpha$; i.e., s_α is the largest integer such that

$$\sum_{k=0}^{s_\alpha} \binom{n}{k} \pi_U^k(t) (1 - \pi_U(t))^{n-k} \leq \alpha. \quad (39)$$

Suppose that $S = s$ is observed. For this family of tests, the P -value of the hypothesis $E \geq 1$ is

$$P_{\text{Bin}} \equiv \sum_{k=0}^s \binom{n}{k} \pi_U^k(t) (1 - \pi_U(t))^{n-k}, \quad (40)$$

the largest probability that there would be so few large values of t_p in the sample if $E \geq 1$.

The binomial method constructs a random variable from T by discretizing the possible values of T into two categories—not greater than t and greater than t —and treating every value in a category as if it had the largest value possible in that category. It then finds a P -value for the hypothesis that the expected value of the new variable is bigger than $1/U$. This is a P -value for the hypothesis that the expected value of the original variable is bigger than $1/U$ because the new variable is, by construction, stochastically larger than T . Thus it is a P -value for the hypothesis $E \geq 1$.

The multinomial method of [15] and the related trinomial method of [5] are similar to the binomial method: They bound the mean of a random variable that is constructed to be stochastically larger than T by discretizing the values of T can take. But instead of dividing the values of T into two bins, they use three or more bins.

2) *P-values from Markov's Inequality*: Markov's inequality says that if $\mathbb{P}\{X \geq 0\} = 1$, then

$$\mathbb{P}\{X \geq \mathbb{E}X/\tau\} \leq \tau. \quad (41)$$

Suppose $\{X_j\}_{j=1}^n$ are iid with $\mathbb{P}\{X_1 \geq 0\} = 1$. Then the chance all n are greater than $\mathbb{E}X_1/\tau$ is

$$\mathbb{P}\{\cap_j \{X_j \geq \mathbb{E}X_1/\tau\}\} \leq \tau^n. \quad (42)$$

Let $\bar{X} = \frac{1}{n} \sum_{j=1}^n X_j$; then $\mathbb{E}\bar{X} = \mathbb{E}X_1$. Inequalities (41, 42) yield two tests of the hypothesis $\mathbb{E}X_1 \leq \mu$, the first based on \bar{X} and the second on $X^- \equiv \min_{j=1}^n X_j$: (i) Reject the hypothesis $\mathbb{E}X_1 \leq \mu$ at significance level α on observing $\bar{X} = x$ if

$$x \geq \mu/\alpha. \quad (43)$$

(ii) Reject the hypothesis $\mathbb{E}X_1 \leq \mu$ at significance level α on observing $X^- = x$ if

$$x \geq \mu/\alpha^{1/n}. \quad (44)$$

To connect these two tests to election auditing, define $X_j \equiv 1 - T_j$. Then $\{X_j\}_{j=1}^n$ are iid nonnegative random variables, and the outcome of the race must be correct if $\mathbb{E}X_1 > 1 - 1/U$. A test statistic for test (i), based on the sample mean, is $\phi(Y) = 1 - \bar{T}$. If we observe $\bar{T} = t$, the P -value of the hypothesis $E \geq 1$ is

$$P_{\text{Markov-mean}} = \frac{1 - 1/U}{1 - t}. \quad (45)$$

A test statistic for test (ii) is $\phi(Y) = \min_{j=1}^n (1 - T_j) = 1 - T^+$. If we observe $T^+ = t$, the P -value of the hypothesis $E \geq 1$ is

$$P_{\text{Markov-max}} \equiv \left(\frac{1 - 1/U}{1 - t} \right)^n. \quad (46)$$

The first test turns out to be too weak to be useful in practice, in part because generally $U \geq 2$ and the first test is insensitive to sample size: For $U = 2$, if all the observed taints are zero, $P_{\text{Markov-mean}} = 0.5$, no matter how large the sample, and the P -value grows with U and with the mean taint.

3) *P-values from the Massart-Dvoretzky-Kiefer-Wolfowitz Inequality*: Let $F(t)$ denote the cumulative distribution function of T :

$$F(t) = \mathbb{P}\{T \leq t\}, \quad t \in \mathbb{R}. \quad (47)$$

so that

$$\mathbb{E}T = \int_{-\infty}^{\infty} t dF(t). \quad (48)$$

Since $T_j \leq 1$, we know that $F(1) = 1$.

Let $\hat{F}_n(t)$ denote the empirical cumulative distribution function of T :

$$\hat{F}_n(t) \equiv \frac{1}{n} \sum_{j=1}^n 1_{t \geq T_j}. \quad (49)$$

Consider the one-sided Kolmogorov-Smirnov statistic

$$D_n^- = \sup_t (F(t) - \hat{F}_n(t)). \quad (50)$$

The distribution of this statistic does not depend on F if F is continuous. Dvoretzky, Kiefer and Wolfowitz [33] (DKW) showed that

$$\mathbb{P}\{D_n^- > \chi\} \leq C \exp(-2n\chi^2) \quad (51)$$

for some constant C . Massart [34] showed that $C = 1$ is sharp when $\exp(-2n\chi^2) \leq \frac{1}{2}$. I will call inequality (51) with $C = 1$ the MDKW inequality. It follows from the MDKW inequality that

$$\mathbb{P}\left\{D_n^- > \sqrt{-\frac{\ln \alpha}{2n}}\right\} \leq \alpha. \quad (52)$$

The distribution of D_n^- is stochastically larger when F is continuous than when F has jumps [34]; thus the MDKW inequality is conservative for iid sampling from finite populations.¹⁴

Inequality (52) can be used to test the hypothesis $E \geq 1$. Let $\mathcal{F}(\hat{F}_n; \beta)$ be the set of distribution functions G for which $G(1) = 1$ and

$$\sup_t (G(t) - \hat{F}_n(t)) \leq \beta. \quad (53)$$

If there is no $G \in \mathcal{F}(\hat{F}_n; \beta)$ for which $\int tdG(t) \geq 1/U$, we can reject the hypothesis $E \geq 1$ at significance level $\alpha = \exp(-2n\beta^2)$.

¹⁴Moreover, D_n^- is stochastically larger for sampling with replacement than for sampling without replacement, so the MDKW inequality is conservative for sampling from a finite population without replacement as well.

To use the MDKW inequality to find a P -value for the hypothesis $E \geq 1$, define

$$\mu^+(\beta) \equiv \sup_{G \in \mathcal{F}(\hat{F}_n; \beta)} \int tdG(t). \quad (54)$$

Then the P -value of the hypothesis $E \geq 1$ (taking \hat{F}_n as given) is

$$P_{\text{MDKW}} \equiv \sup_{\beta} \{\exp(-2n\beta^2) : \mu^+(\beta) \leq 1/U\}. \quad (55)$$

Solving the optimization problem (55) is straightforward: Move mass from the lowest observations to 1 until the mean of the resulting distribution is $1/U$. If the amount of mass moved is β , the P -value is $\exp(-2n\beta^2)$ —provided that is no greater than $1/2$. That restriction is not a problem, because the interesting case is when the P -value is small.

Two cases are particularly simple. Suppose we observe $(T_j = t_j)_{j=1}^n$, $\bar{T} = \bar{t}$, and $T^- = t^-$. (i) If $\bar{t} \geq 1/U$, the P -value is 1. (ii) Let

$$\epsilon \equiv \frac{1/U - \bar{t}}{1 - t^-}. \quad (56)$$

If $\bar{t} < 1/U$ and at least $[n\epsilon]$ of the values $\{t_j\}$ are equal to t^- , then the P -value is $\exp(-2n\epsilon^2)$, provided that is not greater than $1/2$.

4) *P-values from Hoeffding's inequality*: Hoeffding [35] establishes that if $\{X_j\}_{j=1}^n$ are iid random variables that take values in $[0, 1]$, then

$$\mathbb{P}\{\bar{X} - \mu \geq \chi\} \leq e^{-2n\chi^2}. \quad (57)$$

Let $x^+ = \max(x, 0)$. Then T_j^+ is stochastically larger than T_j and takes values in $[0, 1]$. Define $\tilde{T} \equiv \frac{1}{n} \sum_{j=1}^n T_j^+$. The Hoeffding Inequality P -value for the hypothesis $E \geq 1$ if we observe $\tilde{T} = t$ is

$$P_{\text{Hoeffding}} \equiv \begin{cases} 1, & t \geq 1/U \\ \exp(-2n(1/U - t)^2), & t < 1/U. \end{cases} \quad (58)$$

D. Kaplan-Markov P -value

As we have seen, methods for testing the hypothesis that the expected value of a nonnegative random variable is below a threshold using an iid sample can be adapted to test whether $E \geq 1$ from a PPEB sample. [31] gives two methods for finding a lower confidence bound for the expected value of a nonnegative random variable from an iid sample. Here is the simpler of the two, expressed as a P -value.

Suppose $\{X_j\}_{j=1}^n$ are iid with $\mathbb{P}\{X_j \geq 0\} = 1$. Form the nonnegative Martingale:

$$\left(X_1/\mathbb{E}X_1, (X_1/\mathbb{E}X_1) \cdot (X_2/\mathbb{E}X_1), \dots, \prod_{j=1}^n X_n/\mathbb{E}X_1 \right). \quad (59)$$

The expected value of each term is 1. [31] notes a result in [36, p88] that applies Markov's inequality to an optionally stopped nonnegative martingale Z_1, Z_2, \dots . For any $z > 0$, $\mathbb{P}\{\max_{j=1}^n Z_j > z\} \leq \mathbb{E}Z_n/z$. In the case at hand, that gives:

$$\mathbb{P}\left\{\max_{j=1}^n \prod_{i=1}^j X_j/\mathbb{E}X_1 > 1/\alpha\right\} \leq \alpha. \quad (60)$$

We can reject the hypothesis that $\mathbb{E}X_j \leq \mu$ at significance level α if

$$\max_{j=1}^n \prod_{i=1}^j X_i / \mu > 1/\alpha. \quad (61)$$

If we observe $(X_j = x_j)_{j=1}^n$, the P -value of the hypothesis that $\mathbb{E}X_1 \leq \mu$ is

$$\left(\max_{j=1}^n \prod_{i=1}^j x_j / \mu \right)^{-1}. \quad (62)$$

The transformation $X_j = 1 - T_j$ then gives a P -value for the hypothesis $E \geq 1$ if we observe $(T_j = t_j)_{j=1}^n$:

$$P_{\text{KM}} \equiv \min_{j=1}^n \prod_{i=1}^j \frac{1 - 1/U}{1 - t_i}. \quad (63)$$

This Kaplan-Markov P -value is equal to $P_{\text{Markov-max}}$ in section IV-C2 when all $\{T_j\}$ are equal and less than $1/U$; when they are not, P_{KM} can be much smaller. The P -value P_{KM} can depend on the order in which the data are collected, which is discomfiting unless the data are in fact collected sequentially. If every taint T_j is less than $1/U$ —typical in election auditing, unless there is a serious problem—the terms in the products are all less than 1. Then the minimum occurs for $j = n$ and every permutation of the data gives the same P -value.

E. Comparison

Table I compares P -values for PPEB samples for several of the methods in scenarios with sample sizes between $n = 10$ and $n = 40$ and total error bounds between $U = 2$ and $U = 15$. The binomial P -value is given for thresholds $t = 0.01$ and $t = 0.02$. For each combination of sample size and error bound, there are six hypothetical data sets with different taints. These hypotheticals are intended to mimic the taints one might find in batches of hand-marked optically scanned ballots when the outcome is correct: Most batches will show little or no error, but a few might show errors of a few votes out of a few hundred. In the table, rows where every entry is 0.5 or larger are suppressed.

The value of $P_{\text{Hoeffding}}$ is almost identical to (and just as bad as) that of P_{MDKW} . The value of P_{KM} is uniformly smallest, often by a quite a bit.

As discussed in [5], [2], in the November 2008 race for Measure B in Marin County, California, $U < 9.782$; 14 batches were selected by PPEB. All observed taints were zero. The Kaplan-Markov P -value for this race is 22.1%. The November 2008 Santa Cruz County race for Supervisor, District 1, had $U = 13.461$; 19 PPEB draws gave 16 distinct batches. The nonzero taints were 0.036, 0.007, -0.002 , -0.003 , -0.005 , -0.007 and -0.012 ; the other 12 taints were zero. The Kaplan-Markov P -value is 23.4%. The Marin and Santa Cruz audits were designed and analyzed using the trinomial method at risk limit 25%. Both were certified.

Method	Observed Taints					
	clean	0.01	0.01 0.01	0.02	0.01 0.03	-0.05×5 0.05×5
$n = 10, U = 2$						
Bin, $t = 0.01$	0.001	0.001	0.001	0.012	0.012	0.635
Bin, $t = 0.02$	0.001	0.001	0.001	0.001	0.013	0.648
Markov-max	0.001	0.001	0.001	0.001	0.001	0.002
MDKW	0.007	0.007	0.007	0.007	0.007	0.011
KM	0.001	0.001	0.001	0.001	0.001	0.001
$n = 10, U = 5$						
Bin, $t = 0.01$	0.119	0.119	0.119	0.401	0.401	0.995
Bin, $t = 0.02$	0.131	0.131	0.131	0.131	0.427	0.996
Markov-max	0.107	0.119	0.119	0.131	0.146	0.179
MDKW	0.449	0.453	0.457	0.457	0.464	0.484
KM	0.107	0.108	0.110	0.110	0.112	0.109
$n = 20, U = 5$						
Bin, $t = 0.01$	0.014	0.014	0.014	0.081	0.081	0.830
Bin, $t = 0.02$	0.017	0.017	0.017	0.017	0.095	0.854
Markov-max	0.012	0.014	0.014	0.017	0.021	0.032
MDKW	0.202	0.204	0.205	0.205	0.208	0.234
KM	0.012	0.012	0.012	0.012	0.012	0.012
$n = 20, U = 10$						
Bin, $t = 0.01$	0.149	0.149	0.149	0.446	0.446	0.993
Bin, $t = 0.02$	0.182	0.182	0.182	0.182	0.506	0.996
Markov-max	0.122	0.149	0.149	0.182	0.224	0.339
KM	0.122	0.123	0.124	0.124	0.127	0.123
$n = 30, U = 10$						
Bin, $t = 0.01$	0.057	0.057	0.057	0.229	0.229	0.950
Bin, $t = 0.02$	0.078	0.078	0.078	0.078	0.285	0.968
Markov-max	0.042	0.057	0.057	0.078	0.106	0.198
KM	0.042	0.043	0.043	0.043	0.044	0.043
$n = 40, U = 15$						
Bin, $t = 0.01$	0.095	0.095	0.095	0.324	0.324	0.975
Bin, $t = 0.02$	0.142	0.142	0.142	0.142	0.426	0.989
Markov-max	0.063	0.095	0.095	0.142	0.214	0.493
KM	0.063	0.064	0.065	0.065	0.066	0.064

TABLE I
 P -VALUES FOR PPEB SAMPLES USING SEVERAL TESTS

P -values for several tests of the hypothesis $E \geq 1$ based on taints of the maximum relative overstatement of margins in PPEB samples. The P -value is the maximum chance that the observed error would be no “larger” than it was, on the assumption that a full hand count would show a different outcome. Different tests use different definitions of “large.” Rows are suppressed if every entry is > 0.5 . The number of batches in the sample is n ; U is the a priori upper bound on the total overstatement error, as a multiple of the amount of error required to alter the apparent outcome. “Bin” is the binomial P -value of section IV-C1. The Markov-max P -value is in section IV-C2; MDKW is in section IV-C3. “KM” is the Kaplan-Markov P -value of section IV-D. Columns 2–7 list the taints in the sample: Column 2, “clean,” means no errors were found; column 3 is a single taint of 0.01; column 4 is two taints of 0.01; column 5 is a single taint of 0.02; column 6 is a one taint of 0.01 and one of 0.03; column 7 is five taints of -0.05 and five of 0.05.

V. CONCLUSIONS

The maximum probability that the audit would find “no more” error than it did find, if the outcome is wrong, is the P -value of the hypothesis that a full hand count would contradict the apparent outcome. (What “more” means depends on the test.) The smaller the P -value, the stronger the evidence that the apparent outcome is correct. Wrapping the calculation of a P -value in a cycle of expanding the audit sample and comparing the P -value to a sequence of thresholds makes it possible to construct risk-limiting post-election audits—audits with a known minimum chance of requiring a full hand count when the outcome of that count would show that the apparent outcome is wrong [6], [32].

To construct a risk-limiting audit, it helps to summarize the errors the audit discovers as the *maximum relative overstate-*

ment of pairwise margins [3], which expresses those errors as the fractions by which they inflated the margins of apparent winners over apparent losers. A necessary condition for the apparent outcome to be wrong is that E , the total of the maximum relative overstatements, is 1 or larger. *Taint* is the ratio of the maximum relative overstatement e_p in batch p , to the maximum possible relative overstatement u_p in batch p . Expressing overstatement as taint can simplify the analysis.

Sampling batches with probability proportional to the error bound u_p (PPEB) connects election auditing with financial auditing and with the problem of testing hypotheses about the expected value of a nonnegative random variable. If batches are drawn for audit independently, so that in each draw, batch p is selected with probability proportional to u_p , the observed taints are independent and identically distributed and their expected value is $E = \sum_p e_p$, the total overstatement error, divided by $U = \sum_p u_p$, an a priori bound on the total overstatement error.

For PPEB samples, in examples with observed taints like those one might expect to see for voter-marked optically scanned ballots, the Kaplan-Markov P -value is the smallest among those presented here and is very simple to compute. Audits using the Kaplan-Markov P -value and a new method for auditing several contests simultaneously are planned for November 2009 in Marin and Yolo counties, California [32].

REFERENCES

- [1] P. Stark, "Conservative statistical post-election audits," *Ann. Appl. Stat.*, vol. 2, pp. 550–581, 2008.
- [2] J. L. Hall, L. W. Miratrix, P. B. Stark, M. Briones, E. Ginnold, F. Oakley, M. Peadar, 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: USENIX, August 2009.
- [3] P. Stark, "A sharper discrepancy measure for post-election audits," *Ann. Appl. Stat.*, vol. 2, pp. 982–985, 2008. [Online]. Available: <http://arxiv.org/abs/0811.1697>
- [4] J. Aslam, R. Popa, and R. Rivest, "On auditing elections when precincts have different sizes," in *2008 USENIX/ACCURATE Electronic Voting Technology Workshop, San Jose, CA, 28–29 July, 2008*.
- [5] L. Miratrix and P. Stark, "The trinomial bound for post-election audits," *IEEE Transactions on Information Forensics and Security*, vol. accepted, p. tbd, 2009.
- [6] P. Stark, "CAST: Canvass audits by sampling and testing," *IEEE Transactions on Information Forensics and Security, Special Issue on Electronic Voting*, vol. accepted, p. tbd, 2009.
- [7] L. Vance, *Scientific Method for Auditing: Applications of Statistical Sampling Theory to Auditing Procedure*. Berkeley and Los Angeles: University of California Press, 1950.
- [8] R. Trueblood and W. Cooper, "Research and practice in statistical applications to accounting, auditing, and management control," *The Accounting Review*, vol. 30, pp. 221–229, 1955.
- [9] E. Gaynor, "Reliability of sampling plans in auditing," *The Accounting Review*, vol. 31, pp. 253–257, 1956.
- [10] K. Stringer, "Practical aspects of statistical sampling in auditing," in *Proceedings of the Business and Economic Statistics Section*. Washington, D.C.: American Statistical Association, 1963, pp. 405–411.
- [11] P. on Nonstandard Mixtures of Distributions, *Statistical models and analysis in auditing: A study of statistical models and methods for analyzing nonstandard mixtures of distributions in auditing*. Washington, D.C.: National Academy Press, 1988.
- [12] R. Anderson and A. Teitlebaum, "Dollar-unit sampling. A solution to the audit sampling dilemma," *Canadian Chartered Accountant*, pp. 30–39, 1973.
- [13] P. Bickel, "Inference and auditing: The Stringer bound," *Intl. Stat. Rev.*, vol. 60, pp. 197–209, 1992.
- [14] —, "Correction: Inference and auditing: The Stringer bound," *Intl. Stat. Rev.*, vol. 61, p. 487, 1993.
- [15] S. Fienberg, J. Neter, and R. Leitch, "Estimating total overstatement error in accounting populations," *J. Am. Stat. Assoc.*, vol. 72, pp. 295–302, 1977.
- [16] A. Kvanli, Y. Shen, and L. Deng, "Construction of confidence intervals for the mean of a population containing many zero values," *J. Bus. Econ. Stat.*, vol. 16, pp. 362–368, 1998.
- [17] H. Tamura and P. Frost, "Tightening CAV (DUS) bounds by using a parametric model," *J. Accounting Res.*, vol. 24, pp. 364–371, 1986.
- [18] D. Cox and E. Snell, "On sampling and the estimation of rare errors," *Biometrika*, vol. 66, pp. 125–132, 1979.
- [19] J. Godfrey and J. Neter, "Bayesian bounds for monetary unit sampling in accounting and auditing," *J. Accounting Res.*, vol. 22, pp. 497–525, 1984.
- [20] D. Laws and A. O'Hagan, "Bayesian inference for rare errors in populations with unequal unit sizes," *Appl. Stat.*, vol. 49, pp. 557–590, 2000.
- [21] W. Smieliauskas, "A note on a comparison of Bayesian with non-Bayesian dollar-unit sampling bounds for overstatement errors of accounting populations," *The Accounting Review*, vol. 61, pp. 118–128, 1986.
- [22] K.-W. Tsui, E. Matsumura, and K.-L. Tsui, "Multinomial-dirichlet bounds for dollar-unit sampling in auditing," *The Accounting Review*, vol. 60, pp. 76–96, 1985.
- [23] U. Menzefricke and W. Smieliauskas, "A simulation study of the performance of parametric dollar unit sampling statistical procedures," *J. Accounting Res.*, vol. 22, pp. 588–603, 1984.
- [24] H. Clayton, "A combined bound for errors in auditing based on Hoeffding's inequality and the bootstrap," *J. Bus. Econ. Stat.*, vol. 12, pp. 437–448, 1994.
- [25] R. Helmers, "Inference on rare errors using asymptotic expansions and bootstrap calibration," *Biometrika*, vol. 87, pp. 689–694, 2000.
- [26] D. Jefferson, K. Alexander, E. Ginnold, A. Lehmkuhl, K. Midstokke, and P. Stark, "Post election audit standards report—evaluation of audit sampling models and options for strengthening California's manual count," www.sos.ca.gov/elections/peas/final_peaswg_report.pdf, 2007.
- [27] R. Saltman, "Effective use of computing technology in vote-tallying," National Bureau of Standards, Washington, DC, Tech. Rep. NBSIR 75-687, 1975.
- [28] J. McCarthy, H. Stanislevic, M. Lindeman, A. Ash, V. Addona, and M. Batchner, "Percentage-based versus statistical-power-based vote tabulation audits," *The American Statistician*, vol. 62, pp. 11–16, 2008.
- [29] R. Rivest, "On estimating the size of a statistical audit," people.csail.mit.edu/rivest/Rivest-OnEstimatingTheSizeOfAStatisticalAudit.pdf, 2006.
- [30] —, "On auditing elections when precincts have different sizes," <http://people.csail.mit.edu/rivest/Rivest-OnAuditingElectionsWhenPrecinctsHaveDifferentSizes.pdf>, 2007.
- [31] H. Kaplan, "A method of one-sided nonparametric inference for the mean of a nonnegative population," *The American Statistician*, vol. 41, pp. 157–158, 1987.
- [32] P. Stark, "Efficient post-election audits of multiple contests: 2009 California tests," Social Science Research Network, Tech. Rep., 2009, submitted to the 2009 Conference on Empirical Legal Studies. [Online]. Available: <http://ssrn.com/abstract=1443314>
- [33] A. Dvoretzky, J. Kiefer, and J. Wolfowitz, "Asymptotic minimax character of the sample distribution function and classical multinomial estimator," *Ann. Math. Stat.*, vol. 27, pp. 642–669, 1956.
- [34] P. Massart, "The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality," *Ann. Probab.*, vol. 18, pp. 1269–1283, 1990.
- [35] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Am. Stat. Assoc.*, vol. 58, pp. 13–30, 1963.
- [36] L. Breiman, *Probability*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 1992.



PLACE
PHOTO
HERE

Philip B. Stark is Professor of Statistics, University of California, Berkeley. He served on the 2007 Post Election Audit Standards Working Group for California Secretary of State Debra Bowen, and designed and conducted the first four risk-limiting post election audits ever performed. For a more complete biography, see <http://statistics.berkeley.edu/~stark/bio.htm>.