Election Audits Using a Trinomial Bound

Luke W. Miratrix and Philip B. Stark

***Abstract*—In November 2008, we audited contests in Santa Cruz and Marin counties, California. The audits were risk-limiting: they had a prespecified minimum chance of requiring a full hand count if the outcomes were wrong. We developed a new technique for these audits, the trinomial bound. Batches of ballots are selected for audit using probabilities proportional to the amount of error each batch can conceal. Votes in the sample batches are counted by hand. Totals for each batch are compared to the semiofficial results. The "taint" in each sample batch is computed by dividing the largest relative overstatement of any margin by the largest possible relative overstatement of any margin. The observed taints are binned into three groups: less than or equal to zero, between zero and a threshold d, and larger than d. The number of batches in the three bins have a joint trinomial distribution. An upper confidence bound for the overstatement of the margin in the election as a whole is constructed by inverting tests for trinomial category probabilities and projecting the resulting set. If that confidence bound is sufficiently small, the hypothesis that the outcome is wrong is rejected, and the audit stops. If not, there is a full hand count. We conducted the audits with a risk limit of 25%, ensuring at least a 75% chance of a full manual count if the outcomes were wrong. The trinomial confidence bound confirmed the results without a full count, even though the Santa Cruz audit found some errors. The trinomial bound gave better results than the Stringer bound, which is commonly used to analyze financial audit samples drawn with probability proportional to error bounds.**

***Index Terms*—Dollar unit sampling, monetary unit sampling, probability proportional to error bound sampling (PPEB), risk-limiting audit, Stringer bound.**

I. Introduction

ELECTRONIC voting machines and vote tabulation software are complex and opaque, raising concerns about their reliability and vulnerability. Audits can provide a measure of "software independence," controlling the risk that errors—whatever their source—cause the apparent outcome to differ from the outcome a full hand count would show [1]–[4]. Several states have laws mandating election audits, and others are considering such laws [5].[1] It is crucial to ensure that the audit trail is accurate, durable, and complete from its creation through the audit. If there is no audit trail, there can be no audit. If there is an audit trail, but no audit, there is no assurance of accuracy. If there is an audit trail and an audit, but the audit trail does not reflect the electoral outcome, there is still no assurance.

Henceforth, we assume that the audit trail is complete and accurate. When we say "the apparent outcome is correct," we mean the apparent outcome is the same that a full hand count of the audit trail would show. "The apparent outcome is wrong" means a full hand count would show a different outcome.

An election outcome can be checked by hand counting the entire audit trail. This, however, is expensive and time-consuming, and unnecessary unless the outcome is wrong. The goal of a *statistical* audit, which compares a hand count of a random sample of batches of ballots to the audit trail for those batches, is to ensure that the outcome is correct without a full hand count—unless the outcome is wrong. If the outcome is wrong, a full hand count is needed to set the record straight. A *risk-limiting* audit has a minimum prespecified chance $1-\alpha$ of requiring a full hand count whenever the apparent outcome is wrong.[2] The *risk* α is the largest possible chance that there will not be a full hand count when the outcome is wrong, no matter what caused the discrepancies between the apparent outcome and the audit trail. (We assume that $\alpha < 1$; otherwise, an audit would be unnecessary.)

In statistical language, a risk-limiting audit is a significance-level α test of the null hypothesis "the outcome is wrong" against the alternative hypothesis "the outcome is right." Commonly, tests are formulated so that the null hypothesis that things are "good"; here, it is that things are "bad." The reason is that, in the Neyman–Pearson paradigm, the chance of incorrectly rejecting the null hypothesis is controlled to be at most α. We want to control the chance that an incorrect outcome will go undetected, i.e., the chance that there is not a full hand count when there should be.

Not rejecting the null hypothesis entails a full hand count. A good test simultaneously limits the chance of incorrectly rejecting the null hypothesis to at most α and has high power. That is, a good test has chance at least $1-\alpha$ of requiring a full hand count when the outcome is wrong and is very likely to conclude that the outcome is right, with a minimum of hand counting, when the outcome is indeed right.

The outcome can be right even when there are some errors, and audits of voter-marked paper ballots generally find errors at a rate of a few tenths of a percent.[3] For a test to have good

Manuscript received February 22, 2009; revised August 24, 2009. First published October 09, 2009; current version published November 18, 2009. The work of L. W. Miratrix was supported by the National Science Foundation under Graduate Research Fellowship 2007058607. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Poorvi L. Vora.

The authors are with the Department of Statistics, University of California, Berkeley, CA 94720-3860 USA.

Digital Object Identifier 10.1109/TIFS.2009.2034189

[1]See also http://www.verifiedvoting.org/article.php?id=5816 (last visited February 18, 2009).

[2]See http://www.electionaudits.org/bp-risklimiting (last visited February 19, 2009).

[3]We have seen much better accuracy than this, for instance, in the audit of the race in Marin county described here and in a November 2008 audit in Yolo County, CA, we participated in. If something goes wrong—a ballot definition error, miscalibrated scanner, bug, or fraud—errors can be much larger. Direct-recording electronic voting machines (DREs) should be perfectly accurate, and any errors in DRE results are cause for alarm and should be thoroughly investigated.

1556-6013/$26.00 © 2009 IEEE

Authorized licensed use limited to: Univ of Calif Berkeley. Downloaded on December 14, 2009 at 13:13 from IEEE Xplore. Restrictions apply.

TABLE I
SUMMARY OF THE TWO RACES AUDITED

County	Total Ballots	Winner	Loser	Margin	Precincts	Batches	Batches Audited	# Ballots Audited	% Ballots Audited
Santa Cruz	26,655	45%	37%	8%	76	105	16	7,105	27%
Marin	121,295	51%	35%	16%	189	544	14	3,347	3%

power, it needs to have a large probability of rejecting the null hypothesis even when some errors are observed, provided the outcome of the race is right. The issue is whether, in light of the errors found in the sample, there is still compelling statistical evidence that the outcome of the race is correct.

Audits compare hand counts of a random sample of batches to reported totals for those batches.[4] The sampling design used in this paper is sampling with probability proportional to an error bound (PPEB) [7], [4]. Suppose the error in batch p can be no larger than u_p. Let $U = \sum_p u_p$ be the total of all the error bounds. In PPEB, there are n independent draws from the set of N batches. In each draw, the chance of selecting batch p is u_p/U. This makes it more likely that batches that can conceal more error will be audited.

Sampling proportional to an error bound is common in financial auditing, where it is called *dollar unit sampling* or *monetary unit sampling* (MUS) [8]. A standard problem in financial auditing is to find an upper confidence bound for the total overstatement of a set of accounts. Each account has a "book value" in dollars; the real value—the value an audit would reveal—might be lower. The overstatement is the book value minus the real value. The overstatement can be no larger than the book value. Thus, book value is an error bound and MUS is PPEB.

Methods used to analyze MUS data generally convert the overstatement to *taint,* which is the overstatement divided by the book value. For instance, if an account with a book value of $\$1000$ has an audited value of $\$900$, the overstatement is $\$100$ and the taint is $\$100/\$1000 = 0.1$, i.e., ten cents per dollar.

Working with taint in PPEB samples has theoretical advantages; see [9]–[12] and [4]. The expected taint of each PPEB draw is the overall error in the population divided by the total of the error bounds for the population. Moreover, the observed taints are independent and identically distributed. Those features make it straightforward to use the taint in a PPEB sample to find an upper confidence bound on the total overstatement error.

There is an extensive literature on confidence bounds for overstatement from PPEB samples [8]. Apparently, [13] developed the first such confidence bound, based on nesting binomial confidence bounds. That bound turns out to be quite conservative in practice; the multinomial bound of [11] and [12] is sharper (see Section V). The multinomial bound bins the taint into pennies (zero cents per dollar, one cent per dollar, ..., 100 cents per dollar) and uses the multinomial distribution of the counts in each bin to make a confidence bound on the population taint by inverting hypothesis tests. References [9] and [10] develop a different improvement of the bound in [13], and [4] shows how some common probability inequalities can be used with the taint in a PPEB sample to test hypotheses about the overall error. Those tests can be converted into confidence bounds as well.

We present here a simplified variant of the multinomial bound, the trinomial bound. It divides the taint into three bins and constructs an upper confidence bound for the expected taint by inverting a set of hypothesis tests. The acceptance regions for the trinomial bound differ from those of the multinomial bound.[5] For the kind of data that typically arise in election audits, computing the trinomial bound is straightforward.[6] The trinomial confidence bound for the taint can be small even when some errors are observed. When that happens, the audit stops short of a full hand count and the risk is still limited to at most α.

We used the trinomial bound to audit two November 2008 races: one in Santa Cruz County and one in Marin County, California. Table I summarizes the election results. The Santa Cruz County contest was for County Supervisor in the 1st District. The competitive candidates were John Leopold and Betty Danner. According to the semiofficial results provided to us by the Santa Cruz County Clerk's office, Leopold won with votes on 45% of the 26 655 ballots. Danner received the votes on 37% of the ballots. The remaining ballots were undervoted, overvoted, or had votes for minor candidates.[7]

The Marin County race was for Measure B, a county-wide contest that required a simple majority. According to the semiofficial results, provided to us by the Marin County Registrar of Voters office, 121 295 ballots were cast in the race. Fifty-one percent of the ballots recorded "yes" votes; 35% said "no." The remaining 14% had undervotes or overvotes.

Both audits were designed to limit the risk to $\alpha = 0.25$. That is, the chance of a full hand count was at least 75% if the outcome was wrong. Both outcomes were confirmed without a full hand count.

This paper is organized as follows. Section II reviews notation and points to other work for details. Section III develops the trinomial confidence bound and a method for selecting the bins and the sample size. Section IV explains how the trinomial bound was used to audit contests in Marin and Santa Cruz counties and presents the audit results. Section V compares the

[4]The design of the sample matters for the probability calculations and for efficiency. Some methods, such as SAFE [6], use a simple random sample of batches. Others use stratified simple random samples [1]–[3]. States, including California and Minnesota, require drawing random samples stratified by county; batches are ballots for a single precinct. Stratifying on the method of voting—by mail, early, in-precinct, or provisional—can have logistical advantages.

[5]The multinomial bound bases the hypothesis tests on "step-down sets," which partially order the set of possible outcomes. We order outcomes by sample mean of the binned taints, which is more intuitive. Using the sample mean to order outcomes for the 101-bin multinomial would be combinatorially complex, but since the trinomial has only three bins, it turns out to be simple.

[6]The Kaplan–Markov bound [4] seems to be comparable but easier to compute; there has been no extensive comparison so far.

[7]In calculating the confidence bound on the error, the audit took every ballot into account, not just the ballots with votes.

Authorized licensed use limited to: Univ of Calif Berkeley. Downloaded on December 14, 2009 at 13:13 from IEEE Xplore. Restrictions apply.

trinomial bound to the Stringer bound. Section VI presents conclusions.

II. Notation and Assumptions

We generally follow the notation in [2]–[4]. There are K candidates; voters may vote for up to $f \geq 1$ of them (the contest has f winners). There are N batches of ballots, indexed by p. There are v_{kp} votes reported for candidate k in batch p. There are actually a_{kp} votes cast for candidate k in batch p. The total vote reported for candidate k is $V_k = \sum_p v_{kp}$, the sum of the votes reported for candidate k in the N batches. The total actual vote for candidate k is $A_k = \sum_p a_{kp}$. The set $\mathcal{W}$ comprises the indexes of the apparent winners, so $\#\mathcal{W} = f$. The set $\mathcal{L}$ comprises the indexes of the apparent losers, so $\#\mathcal{L} = K - f$.

If $w \in \mathcal{W}$ and $\ell \in \mathcal{L}$, then

$$V_{w\ell} \equiv V_w - V_\ell > 0. \tag{1}$$

The outcome of the election is right if, for every $w \in \mathcal{W}$ and $\ell \in \mathcal{L}$,

$$A_{w\ell} \equiv A_w - A_\ell > 0. \tag{2}$$

Define

$$e_{w\ell p} \equiv \frac{v_{wp} - v_{\ell p} - (a_{wp} - a_{\ell p})}{V_{w\ell}}. \tag{3}$$

That is the amount by which error in batch p overstated the margin between candidate w and candidate ℓ, expressed as a fraction of the reported margin between them.

If the outcome of the race is wrong, there is some pair $w \in \mathcal{W}, \ell \in \mathcal{L}$ for which

$$\sum_p e_{w\ell p} \geq 1. \tag{4}$$

Define

$$e_p \equiv \max_{w \in \mathcal{W}, \ell \in \mathcal{L}} e_{w\ell p}. \tag{5}$$

Reference [2] shows that a sufficient condition for the outcome to be correct is

$$E \equiv \sum_p e_p < 1. \tag{6}$$

This condition is sufficient but not necessary; tightening the condition could yield better tests.

We want to draw a statistical inference about E from a random sample of batches, making a bare minimum of assumptions about $\{e_p\}$. We do assume that we have a bound b_p on the total number of ballots in batch p. [2] shows that from such a bound, we can deduce that

$$e_p \leq u_p \equiv \max_{w \in \mathcal{W}, \ell \in \mathcal{L}} \frac{b_p + v_{wp} - v_{\ell p}}{V_{w\ell}}. \tag{7}$$

Let

$$U \equiv \sum_p u_p. \tag{8}$$

We call e_p the overstatement error in batch p, E the overstatement error, u_p the maximum overstatement error in batch p, and U the maximum overstatement error.

The sample is selected as follows. We draw n times independently (with replacement) from the set of N batches. In each draw, the probability of selecting batch p is u_p/U. This is called a PPEB sample [7]; it is equivalent to monetary unit sampling and dollar unit sampling in financial auditing [8].

This paper gives a method to compute an upper $1-\alpha$ confidence bound E_α^+ for E from a PPEB sample. One general strategy for risk-limiting audits, described in [1]–[3], is to test the hypothesis that the outcome is wrong sequentially. The auditor draws a sample, then assesses whether there is sufficiently strong evidence that the outcome is correct. If there is, the audit stops. If there is not, the audit sample is enlarged and the new evidence is assessed. Eventually, either there is strong evidence that the outcome is right, or there will have been a full hand count.

Stage s of a sequential audit can be viewed as a test at significance level α_s. In this paper, we focus on a single stage. The hypothesis that the outcome is wrong is rejected at significance level α_s if $E_{\alpha_s}^+ < 1$. That might be the only stage of an audit that takes a sample, then either stops or conducts a full hand count; or it might be one of the stages of a multistage audit that could expand the sample once or more before demanding a full hand count.

The two audits we conducted using the new method were single-stage audits. We drew an initial sample of n batches and calculated an upper 75% confidence bound for E from the errors the hand counts uncovered in those batches. If that upper confidence bound $E_{0.25}^+$ had been greater than one, the election officials would have conducted complete hand counts.

III. The Trinomial Confidence Bound

Our method for constructing a $1-\alpha$ upper confidence bound E_α^+ for E is similar to the multinomial bound with clustering [11], [12].

The *taint* t_p of batch p is the ratio of the actual overstatement in batch p to the maximum overstatement in batch p

$$t_p \equiv \frac{e_p}{u_p} \leq 1. \tag{9}$$

Now

$$E = \sum_p e_p = \sum_p \frac{e_p}{u_p} u_p = \sum_p t_p u_p. \tag{10}$$

Suppose we draw a PPEB sample of size n. Let T_j denote the taint of the jth draw. Then the expected value of T_j is

$$\mathbb{E}T_j = \sum_p t_p u_p / U = E/U. \tag{11}$$

Multiplication by U transforms an upper $1-\alpha$ confidence bound for $\mathbb{E}T_j$ into an upper $1-\alpha$ confidence bound for E. See also [4].

Authorized licensed use limited to: Univ of Calif Berkeley. Downloaded on December 14, 2009 at 13:13 from IEEE Xplore. Restrictions apply.

Let $d \in (0,1)$. Define

$$Y_j \equiv \begin{cases} 0, & T_j \le 0 \\ d, & 0 < T_j \le d \\ 1, & T_j > d \end{cases}. \quad (12)$$

For any $d \in (0,1)$,[8] Y_j is stochastically larger than T_j (i.e., $\mathbb{P}[Y_j \ge T_j] = 1$), so

$$\mathbb{E}T_j \le \mathbb{E}Y_j. \quad (13)$$

Let

$$\pi_0 \equiv \mathbb{P}[Y_j = 0]$$
$$\pi_d \equiv \mathbb{P}[Y_j = d]$$
$$\pi_1 \equiv \mathbb{P}[Y_j = 1]$$

and let $\pi \equiv (\pi_0, \pi_d, \pi_1)$. Define $\mu = \mu(d) \equiv (0, d, 1)$. Then

$$\mathbb{E}Y_j = 0\pi_0 + d\pi_d + 1\pi_1 = \mu \cdot \pi. \quad (14)$$

Define

$$Z \equiv (\#\{j : Y_j = 0\}, \#\{j : Y_j = d\}, \#\{j : Y_j = 1\}). \quad (15)$$

This is a random 3-vector. Its first component is the number of observed taints that are no bigger than zero; its second is the number of observed taints that are strictly positive but no bigger than d; and its third is the number of observed taints that exceed d. It has a trinomial distribution with category probabilities π.

We will use Z to find a set $S_\alpha(Z)$ such that

$$\mathbb{P}_\pi[S_\alpha(Z) \ni \pi] \ge 1 - \alpha. \quad (16)$$

That is, $S_\alpha(Z)$ is a $1-\alpha$ confidence set for π. Then

$$t_\alpha^+ \equiv \max_{\gamma \in S_\alpha(Z)} \mu \cdot \gamma \quad (17)$$

is the upper endpoint of a $1-\alpha$ upper confidence interval for $\mathbb{E}Y_j$ and hence for $\mathbb{E}T_j$. It follows that Ut_α^+ is the upper endpoint of a $1-\alpha$ upper confidence interval for E.

We construct $S_\alpha(Z)$ by inverting hypothesis tests about π. We are ultimately interested in inferring that $\mu \cdot \pi$ is not large, so it makes sense to reject the hypothesis $\pi = \gamma$ when

$$\mu \cdot Z \le z_\gamma \quad (18)$$

with

$$z_\gamma = z_\gamma(\alpha) \equiv \max\{z : \mathbb{P}_\gamma[\mu \cdot Z \le z] \le \alpha\} \quad (19)$$

so that the test has level α.

The test statistic $\mu \cdot Z$ orders the possible values of Z by the sample mean of the values of Y_j from which Z was constructed.[9] To find a confidence bound for $\mathbb{E}T_j$, we invert the hypothesis tests to find the confidence set $S_\alpha(z)$ of trinomial category probabilities γ for which the hypothesis $\pi = \gamma$ would not be rejected if we observed $Z = z$. That set is

$$\begin{aligned} S_\alpha(z) &\equiv \{\gamma = (\gamma_0, \gamma_d, \gamma_1) \in \Re^3 : \gamma \ge 0, \gamma_0 + \gamma_d + \gamma_1 = 1 \\ &\qquad \mu \cdot z > z_\gamma\} \\ &= \{\gamma = (\gamma_0, \gamma_d, \gamma_1) \in \Re^3 : \gamma \ge 0, \gamma_0 + \gamma_d + \gamma_1 = 1 \\ &\qquad \mathbb{P}_\gamma[\mu \cdot Z \le \mu \cdot z] > \alpha\}. \end{aligned} \quad (20)$$

The corresponding confidence bound for $\mathbb{E}T_j$ is the largest value of $\mu \cdot \gamma$ over $\gamma \in S_\alpha(z)$

$$t_\alpha^+ = t_\alpha^+(z) = \max_{\gamma : \mathbb{P}_\gamma[\mu \cdot Z \le \mu \cdot z] > \alpha} \mu \cdot \gamma. \quad (21)$$

We now characterize the solution to the optimization problem (21) in some useful cases. If $\mu \cdot Z/n \ge 1/U$, we certainly will not be able to conclude that $E < 1$. The question is how much smaller than $1/U$ the "sample mean" $\mu \cdot Z/n$ must be to provide strong evidence that $E < 1$. Because $\alpha < 1$ by assumption,

$$\mathbb{P}_{(0,0,1)}[\mu \cdot Z \le \mu \cdot z] < \alpha \quad (22)$$

unless $z = (0, 0, n)$. If $z = (0, 0, n)$, $t_\alpha^+ = 1$.

If $z \ne (0, 0, n)$, then the maximum in (21) is attained for some γ for which $\mathbb{P}_\gamma[\mu \cdot Z \le \mu \cdot z] = \alpha$.[10] Suppose no observed taints are greater than d and $k < 1/d$ taints are strictly positive. Then $z = (n-k, k, 0)$ and

$$\begin{aligned} \mathbb{P}_\gamma[\mu \cdot Z \le \mu \cdot z] &= \sum_{j=0}^{k} \mathbb{P}_\gamma[Z = (n-j, j, 0)] \\ &= \sum_{j=0}^{k} \binom{n}{j} \gamma_0^{n-j} \gamma_d^j. \end{aligned} \quad (23)$$

Hence

$$t_\alpha^+ = 1 + \max_{\gamma_0, \gamma_d} \left\{ (d-1)\gamma_d - \gamma_0 : \gamma_0, \gamma_d \ge 0, \quad \gamma_0 + \gamma_d \le 1, \right. \\ \left. \sum_{j=0}^{k} \binom{n}{j} \gamma_0^{n-j} \gamma_d^j = \alpha \right\}. \quad (24)$$

The two-dimensional optimization problem (24) can be solved using an ascent method or by searching. The R package "elec," available through CRAN, implements the computation.

A. Selecting n and d

No matter what values we select for n and d, the upper confidence bound for E will be conservative. However, if we choose n very small or d very large, the audit will not be able to provide

[8] Some papers on the multinomial bound in financial auditing suggest that d can be chosen after the data are collected. We have seen no proof that post hoc selection of d results in a valid confidence bound. We select d before the data are collected.

[9] This test statistic generally results in a different test from the "step-down set" acceptance region used by [11], [12].

[10] To see this, note that i) $\mu \cdot \gamma$ increases continuously and monotonically as mass is moved either from γ_0 to γ_1 or from γ_d to γ_1 and ii) $\mathbb{P}_\gamma[\mu \cdot Z \le \mu \cdot z]$ decreases monotonically and continuously as mass is moved either from γ_0 to γ_1 or from γ_d to γ_1.

Suppose the maximum in (21) were attained for some $\delta = (\delta_0, \delta_d, \delta_1)$, with $\mathbb{P}_\delta[\mu \cdot Z \le \mu \cdot z] > \alpha$. By assumption, $\delta \ne (0, 0, 1)$. Hence, either $\delta_0 > 0$ or $\delta_d > 0$. Moving an infinitesimal amount mass from either of those components to δ_1 increases $\mu \cdot \delta$ and decreases $\mathbb{P}_\delta[\mu \cdot Z \le \mu \cdot z]$. Hence, δ cannot be optimal.

[11] http://cran.r-project.org.

Authorized licensed use limited to: Univ of Calif Berkeley. Downloaded on December 14, 2009 at 13:13 from IEEE Xplore. Restrictions apply.

TABLE II
SANTA CRUZ AUDIT DATA

			Leopold		Danner				
Batch ID	b_p	u_p	Reported	Actual	Reported	Actual	MOV	t_p	Times
1002 VBM	573	0.28	251	252	227	227	-1	-0.002	1
1005 PCT	556	0.32	292	304	166	170	-8	-0.012	1
1005 VBM	436	0.23	208	208	150	150	0	0	1
1007 PCT	692	0.40	367	382	205	216	-4	-0.005	1
1007 VBM	630	0.33	311	311	240	240	0	0	1
1013 VBM	557	0.28	261	261	216	216	0	0	2
1017 VBM	399	0.21	191	191	139	139	0	0	1
1019 PCT	448	0.25	218	223	128	137	4	0.007	1
1019 VBM	378	0.20	186	186	128	128	0	0	1
1027 VBM	232	0.11	107	107	98	98	0	0	1
1028 VBM	365	0.15	136	137	174	174	-1	-0.003	1
1037 VBM	758	0.33	261	261	309	309	0	0	2
1053 VBM	18	0.01	10	10	4	4	0	0	1
1060 PCT	322	0.17	142	145	105	108	0	0	2
1073 VBM	20	0.01	11	11	3	4	1	0.036	1
1101 PCT	721	0.35	312	321	275	279	-5	-0.007	1

strong evidence that $E < 1$, even when the outcome of the election is correct. The confidence bound E_α^+ will be greater than one, and the audit will progress—either to the next stage or to a full hand count. On the other hand, setting n large entails a lot of auditing in the first stage, perhaps more than necessary to confirm the outcome when the outcome is in fact correct.

We select d and n iteratively, using simulation to estimate the power of the test against a "realistic" alternative hypothesis under which there is error, but not enough error to alter the outcome of the contest. In the alternative, the error is randomly distributed. Batches are tainted with probability τ, independently. If batch p is tainted, it has an overstatement of (up to) η votes, and the error is $\min\{\eta/V_{w\ell}, u_p\}$. The amount of taint that the η votes represent thus depends on the batch. For batches with small u_p, an overstatement of η votes is a large taint, while for batches with large u_p, it is a small taint. Because the chance of drawing batch p is smaller for batches with small u_p, it is less likely that the sample will include the larger taints.

We adjust d and n iteratively until the chance is approximately $1-\beta$ that the $1-\alpha$ trinomial confidence bound for E is less than one. The chance is estimated by simulation. The confidence level is always at least $1-\alpha$. Adjusting n and d only affects the power.

In the simulations to select d and n for the Marin and Santa Cruz County audits, which were conducted at level $\alpha = 0.25$, we used $\tau = 0.05,\ \eta = 10$ votes, and $1 - \beta = 0.9$. These choices resulted in using $d = 0.047,\ n = 19$ for Santa Cruz and $d = 0.038,\ n = 14$ for Marin.

IV. NOVEMBER 2008 AUDITS IN MARIN AND SANTA CRUZ COUNTIES

In November 2008, we audited races in Marin and Santa Cruz counties, using the trinomial bound,[12] as follows. The elections officials provided us the semiofficial results $\{v_{kp}\}$ and the number of ballots cast in each batch, which we took as $\{b_p\}$. From $\{v_{kp}\}$ and $\{b_p\}$, we calculated $\{u_p\}$ and U. We selected the number of draws n as described in Section III-A.

[12]We audited a race in Yolo County, CA, using a different method.

The elections officials rolled dice to generate six-digit seeds, which they sent to us.[13] We used the seeds in the R implementation of the Mersenne Twister algorithm to make n PPEB draws to select batches for audit. The batches selected were counted by hand by members of the staffs of the Santa Cruz County Clerk's office and the Marin County Registrar of Voters office. They reported the hand-count results to us. We calculated confidence bounds for E from the observed discrepancies and U using the trinomial bound. In both cases, the 75% upper confidence bounds were less than one, so no further counting was required.

Section IV-A describes the Santa Cruz County audit in some detail. Section IV-B summarizes the Marin County audit.

A. Santa Cruz County Supervisor, 1st District

There were 152 batches containing 0 to 855 ballots (median 66). The maximum potential error per batch ranged from $u_p = 0\%$ to 49% of the margin. Some individual batches could hide enough error to account for nearly half the margin. The distribution of the $\{u_p\}$ was heavily skewed to the right. The total possible margin overstatement across all batches was $U = 13.46$.

As described in Section III-A, we used $d = 0.047$ and $n = 19$ in this audit. Since the draws are independent, they need not yield distinct batches. The expected number of distinct batches in 19 PPEB draws is

$$\sum_p \left(1 - \left(1 - \frac{u_p}{U}\right)^n\right) = 16.3 \tag{25}$$

and the expected number of ballots in the sample is

$$\sum_p b_p \left(1 - \left(1 - \frac{u_p}{U}\right)^n\right) = 7214. \tag{26}$$

A simple random sample would have required a much larger audit to control the risk to the same level.[14] The 19 draws produced 16 distinct batches containing 7105 ballots in all. Even with PPEB, a high proportion of ballots needed to be audited,

[13]The Santa Cruz seed was 541 227; the Marin seed was 568 964.

[14]For example, the method in [1], [3] would have required a simple random sample of $n = 38$ batches, with the expectation of counting 13 017 ballots, on the order of twice the effort required by the trinomial bound with PPEB sampling.

Authorized licensed use limited to: Univ of Calif Berkeley. Downloaded on December 14, 2009 at 13:13 from IEEE Xplore. Restrictions apply.

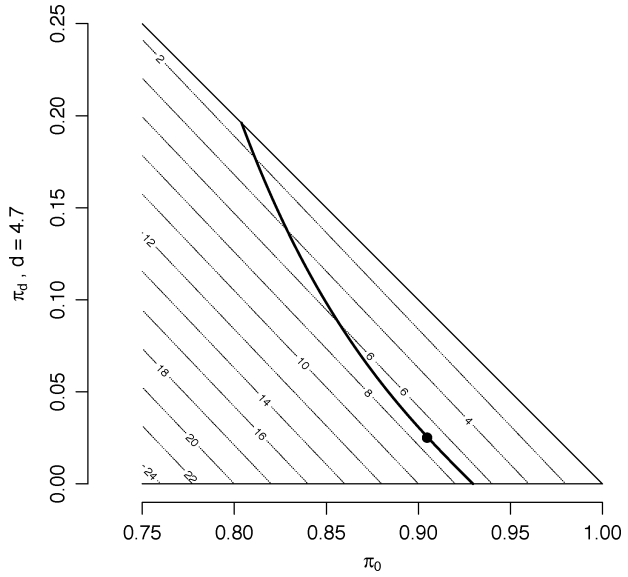

Fig. 1. The optimization problem over trinomial category probabilities for the Santa Cruz audit. The heavy line is the set $\{\gamma : \mathbb{P}_\gamma[\mu \cdot Z \leq \mu \cdot z] = \alpha = 0.25\}$. The parallel lines are the contours of $100 \times \mu \cdot \gamma$. The points to the right of the heavy line compose the confidence set. The heavy dot is the category probability vector with the largest value of $\mu \cdot \gamma$ among parameters in the confidence set. For this contest, $U = 13.46$, so the audit can stop if the confidence set excludes $1/13.46 \approx 0.074$, corresponding to a contour line at 7.4 in the units of this figure.

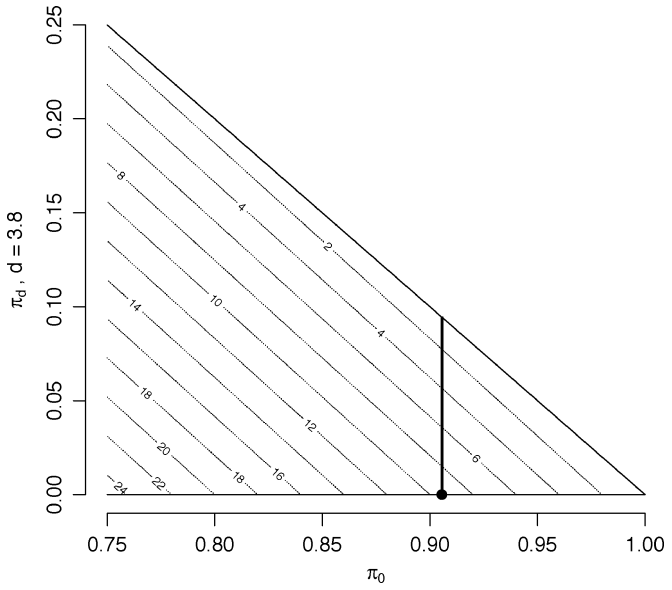

Fig. 2. The optimization problem over trinomial category probabilities for the Marin audit. The heavy line is the set $\{\gamma : \mathbb{P}_\gamma[\mu \cdot Z \leq \mu \cdot z] = \alpha = 0.25\}$. The parallel lines are contours of $100 \times \mu \cdot \gamma$. The confidence set consists of the points to the right of the heavy line. The point is the category probability vector in the confidence set with the largest value of $\mu \cdot \gamma$. Because no errors were found, the maximum lies on the boundary. For this contest, $U = 9.78$, so the audit can stop if the confidence set excludes $1/9.78 \approx 0.102$, corresponding to a contour line at 10.2 in the units of this figure.

which is typical for small races. The sample size needed to control the risk does not depend directly on the size of the race. Wide variations in the error bounds $\{u_p\}$ also contribute to the need for a larger sample. Table II gives the audit results.

While analyzing the data, we learned that, although the audit data included provisional ballots, the original totals on which we had based the audit did not.[15] This increased the number of ballots in several audited batches and changed the margins in some of them. The audit also showed a difference of one in the number of ballots in some vote by mail (VBM) batches. We attribute that difference to ballots that needed special treatment. To ensure that the audit remained statistically conservative, we treated every change to the reported margins—including changes produced by provisional ballots—as error in the reported counts, i.e., as error uncovered by the audit.[16] The change in b_p, the number of ballots in a batch, affects u_p. If u_p is still an upper bound on e_p, the audit remains valid. Since the bound u_p is extremely conservative (calculated by assuming that *all* the votes in batch p are actually for the loser) and there are so few provisional ballots in all, it is implausible that $e_p > b_p$ in any batch.

The largest observed taint, 0.036, was a one-vote overstatement in a tiny precinct. The largest absolute overstatement, four votes, was in a much larger precinct; that taint was only 0.007. "Error" was as large as eight votes in some batches, an atypically high rate for voter-marked optically scanned ballots. As far as we can tell, this discrepancy was due to miscommunication, not an error in the counts per se. This experience underscores the importance of clear communication among the auditors and election officials and their staff.

Apparently, the majority of the provisional ballots in the sample were for the winner, so including them among the ballots in the audited batches only strengthened the evidence that the outcome was right. Despite treating changes caused by including provisional ballots as errors, only two batches had margin overstatements, both less than $d = 0.047$. (If any of the three batches that were drawn twice had positive taint, the taint of that batch would count twice.)

The trinomial observation was thus $z = (17, 2, 0)$. The calculation of the trinomial confidence bound is illustrated in Fig. 1. The upper confidence bound for $\mathbb{E}T_j$ is $t^+_{0.25} = 0.072$, which yields the upper confidence bound

$$E^+_{0.25} = U t^+_{0.25} = 13.46 \times 0.072 = 0.97 < 1. \qquad (27)$$

This allowed us to reject the hypothesis that the outcome was wrong and stop the audit without a full manual count.[17]

B. Marin County Measure B

Table I summarizes the results of the race. In Marin, "decks" of VBM ballots are run through the scanner as a group. Decks usually contain about 250 ballots, sometimes from several precincts. To collect all the ballots for a single precinct could require sorting through several decks of ballots. This is laborious and prone to error; for a race as large as Measure B, the effort is prohibitive. For this reason, we used the decks as batches.

[15]Apparently 806 provisional ballots had been cast in the race in all. Among the audited batches, precinct 1005 had 37; 1007 had 30; 1019 had 32; 1060 had 11; and 1101 had 39.

[16]It would also have been conservative to treat all the provisional ballots as error, but we had no way to separate the votes for the provisional and original ballots, so it was impossible to isolate the error in the original counts.

[17]On the basis of the trinomial bound, the P-value of the hypothesis that the outcome is wrong is 0.24.

Authorized licensed use limited to: Univ of Calif Berkeley. Downloaded on December 14, 2009 at 13:13 from IEEE Xplore. Restrictions apply.

There was a complication. While the total number of ballots b_p in each deck is known, the number of votes for each candidate or position is not. (The vote tabulation software would not generate such subtotals without extensive hand editing.) To calculate a rigorous upper bound u_p for decks, we made extremely conservative assumptions: $v_{wp} = b_p$ but $a_{\ell p} = b_p$. That is, to find an upper bound on the margin overstatement in batch p, we assumed that every ballot was reported as cast for the apparent winner but that in reality every ballot was cast for the reported loser. That leads to the bound $u_p = 2b_p$. While this is extremely conservative, the resulting sample size was still manageable. The sample size n was larger than it would have been had we known v_{wp} and $v_{\ell p}$, but that was balanced by the labor saved in not having to generate vote totals for the decks manually.[18] The bound would have been effectively much more conservative if only a subset of the ballots in a deck included Measure B, but Measure B was county-wide.

There were 544 batches in all—189 batches of ballots cast in precinct and 355 decks. Using (small) decks as batches reduces the expected workload because the more batches there are, the smaller the size of each. The number of draws required does not depend directly on the number of batches in the population, so dividing the ballots into many small batches usually leads to less counting than dividing the population into fewer large batches.

The total error bound was $U = 9.78$. The distribution of error bounds was roughly bell-shaped, with a spike at 0.025 because many decks were about the same size (roughly 250 ballots each). In this election, no batch could hold error of more than 3% of the margin. In contrast, in the Santa Cruz race, some batches could hold errors of up to 49% of the margin.

As described in Section III-A, we chose $d = 0.038$ and $n = 14$ draws, which were expected to yield 13.8 distinct batches and 3424 ballots. The expected number of batches is close to the number of draws because the error bounds u_p are reasonably uniform and no u_p is very large, in contrast to the bounds in Santa Cruz. With simple random sampling, the audit would have required roughly 22 batches to control the risk to the same level ($\alpha = 0.25$). The expected number of ballots to audit would have been about 4900, 44% more than with PPEB and the trinomial bound.

Once the decks to audit were selected, subtotals for those decks were produced in order to have semiofficial figures to audit. This involved replicating the database and generating a special report for each audited precinct by manually deleting every batch but one and generating a report for the remaining batch, an arduous and error-prone procedure. Those subtotals were then audited by hand-counting paper ballots. Table III lists the reported votes in the 14 batches in the sample, which included 3347 ballots. Remarkably, the audit found no errors. The vector of trinomial counts was thus $z = (14, 0, 0)$. The 75% confidence bound for taint was $t^+_{0.25} = 0.094$, and the 75% confidence bound for E was

$$E^+_{0.25} = 0.0943 \times 9.78 = 0.922 < 1 \tag{28}$$

[18]If the vote tabulation software had been able to report v_{wp} and $v_{\ell p}$ for each deck, we would not have had to use such a conservative bound. Data export from vote-tabulation systems is a serious bottleneck for election auditing.

TABLE III
MARIN AUDIT RESULTS

Batch ID	b_p	Yes	No	u_p
D-31	91	50	33	0.009
D-43	108	59	40	0.011
D-104	40	16	16	0.004
D-191	217	137	57	0.022
D-255	246	156	67	0.025
D-286	258	144	88	0.026
D-301	245	129	88	0.025
D-339	248	134	80	0.025
IP-1002	316	151	110	0.018
IP-1017	362	186	133	0.021
IP-3013	277	125	102	0.015
IP-3014	498	256	152	0.030
IP-3017	318	154	111	0.018
IP-3020	123	64	39	0.007

so the audit stopped without a full hand count. The corresponding P-value was about 0.22.

C. Late Problems in Marin County

We discovered in late July 2009, after this paper was accepted and long after the end of the canvass period, that while Marin County had not found any discrepancies in any audited batches, the totals they audited were not identical to the totals on which we had based the audit calculations. In Marin County, voters in precincts with fewer than 250 registered voters are required to vote by mail, and VBM ballots are reported as if they were in-person (IP) ballots. For larger precincts, the IP results were final by November 7, but for precincts with fewer than 250 registered voters, the "nominal" IP results were not final until November 14: it takes longer for the VBM ballots to be sorted and tallied.[19] We based our audit calculations on the IP results in the November 7 statement of vote, understanding—incorrectly—that those were final. They were final for larger precincts but not for VBM-only precincts. Marin County audited the November 14 statement of vote. Again, this emphasizes the importance of clear communication between auditors and elections officials and shows the value of pilot studies.

V. COMPARISON WITH THE STRINGER BOUND

The Stringer bound [13] has long been used in financial auditing to find an upper confidence bound on the overstatement of a group of accounts using a PPEB sample. It is generally—though not always—quite conservative, more so than the multinomial bound [14]. If there are M nonzero taints, $t_1 > \cdots > t_M$, the Stringer bound is

$$t^+_{S,\alpha} \equiv \pi^+_\alpha(0) + \sum_{j=1}^{M} \left[\pi^+_\alpha(j) - \pi^+_\alpha(j-1)\right] t_j \tag{29}$$

where $\pi^+_\alpha(k)$ is the exact $1-\alpha$ upper confidence bound for π from datum $X \sim \mathrm{Bin}(n, \pi)$ when the observed value of X is k.

Table IV compares the 75% upper confidence bound for E based on the Stringer bound and the trinomial bound for the Santa Cruz and Marin audit data. For the Santa Cruz data, the

[19]The VBM ballots for VBM-only precincts get special treatment: They are segregated from the other VBM ballots and sorted by precinct.

Authorized licensed use limited to: Univ of Calif Berkeley. Downloaded on December 14, 2009 at 13:13 from IEEE Xplore. Restrictions apply.

TABLE IV
75% UPPER CONFIDENCE BOUNDS FOR E

County	n	positive taints	Stringer	Trinomial
Santa Cruz	19	0.036, 0.007	0.984	0.956
Marin	14	none	0.922	0.922

Stringer bound is larger but still below one, so it would have permitted the audit to stop. When all the taints are nonpositive, as they are for the Marin data, the Stringer bound equals the trinomial bound. The Kaplan–Markov bound [4] can be sharper, especially if there are negative taints.

VI. CONCLUSION

We used a novel method to audit two November 2008 contests in California: one in Santa Cruz County and one in Marin County. The audits were conducted in a way that guaranteed at least a 75% chance of a full hand count if the outcome of the contest were wrong. Neither audit resulted in a full hand count.

The method we used, the trinomial bound, constructs an upper confidence bound for the total overstatement error E in the race. For the apparent outcome of the race to be wrong, it is necessary that $E \geq 1$. Hence, if the confidence bound for E is less than one, the audit can stop. If the confidence bound is one or greater, there is a full manual count. This results in a risk-limiting audit, i.e., an audit with a guaranteed minimum chance of a full manual count whenever the apparent outcome is wrong.

The trinomial bound relies on a sample drawn with probability proportional to a bound on the overstatement error in each batch of ballots (PPEB sampling), a technique long used in financial auditing but new to election auditing [7]. There are other ways of using PPEB samples to draw inferences about E [4], [15]. The trinomial bound constructs a confidence set for the category probabilities for a trinomial variable from the *taints* observed in the PPEB sample, then projects and scales that confidence set to find a confidence bound for E.

The audit in Marin county posed unusual logistic challenges because ballots were not sorted by precinct. We used batches defined by "decks" of ballots that were fed through scanners as a group. The inability of the vote tabulation software to produce batch subtotals made it necessary then to use extremely conservative bounds on the possible error in each batch: twice the number of ballots.

Election audits face considerable logistic challenges. The time and effort of counting votes by hand is one. The lack of good "data plumbing" is another. Current vote tabulation systems do not seem to export data in formats that are convenient for audits, necessitating hours of error-prone hand editing. Elections officials and legislators interested in promoting post-election audits could help by demanding this functionality. Embracing standard data formats would also help considerably.

ACKNOWLEDGMENT

The authors are extremely grateful to Marin County Registrar of Voters E. Ginnold, Santa Cruz County Clerk G. Pellerin, and their staffs for their generous cooperation and the considerable time and effort they spent counting ballots by hand. They thank A. Shimelman for comments.

REFERENCES

[1] P. Stark, "Conservative statistical post-election audits," *Ann. Appl. Stat.* vol. 2, pp. 550–581, 2008 [Online]. Available: http://arxivorg/abs/0807.4005

[2] 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

[3] P. B. Stark, "CAST: Canvass audits by sampling and testing," *Special Issue on Electronic Voting, IEEE Trans. Inf. Forensics Security*, vol. 4, no. 4, pp. 708–717, Dec. 2009.

[4] P. B. Stark, "Risk-limiting postelection audits: Conservative P-values from common probability inequalities," *IEEE Trans. Inf. Forensics Security*, vol. 4, no. 4, pp. 1005–1014, Dec. 2009.

[5] J. L. Hall, L. W. Miratrix, P. B. Stark, M. Briones, E. Ginnold, F. Oakley, M. Peaden, G. Pellerin, T. Stanionis, and T. Webber, "Implementing risk-limiting post-election audits in California," in *Proc. 2009 Electron. Voting Technol. Workshop/Workshop Trustworthy Elections (EVT/WOTE'09)*, Montreal, PQ, Canada, Aug. 2009.

[6] J. McCarthy, H. Stanislevic, M. Lindeman, A. Ash, V. Addona, and M. Batcher, "Percentage-based versus statistical-power-based vote tabulation audits," *Amer. Stat.*, vol. 62, pp. 11–16, 2008.

[7] J. Aslam, R. Popa, and R. Rivest, "On auditing elections when precincts have different sizes," in *Proc. 2008 USENIX/ACCURATE Electron. Voting Technol. Workshop*, San Jose, CA, Jul. 28–29, 2008.

[8] *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, Panel on Nonstandard Mixtures of Distributions.

[9] P. Bickel, "Inference and auditing: The Stringer bound," *Intl. Stat. Rev.*, vol. 60, pp. 197–209, 1992.

[10] P. Bickel, "Correction: Inference and auditing: The stringer bound," *Intl. Stat. Rev.*, vol. 61, pp. 487–, 1993.

[11] S. Fienberg, J. Neter, and R. Leitch, "Estimating total overstatement error in accounting populations," *J. Amer. Stat. Assoc.*, vol. 72, pp. 295–302, 1977.

[12] J. Neter, R. Leitch, and S. Fienberg, "Dollar unit sampling: Multinomial bounds for total overstatement and understatement errors," *Account. Rev.*, vol. 53, pp. 77–93, 1978.

[13] 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.

[14] R. Plante, J. Neter, and R. Leitch, "Comparative performance of multinomial, cell and stringer bounds," *Auditing*, vol. 5, pp. 40–56, 1985.

[15] P. Stark, Efficient post-election audits of multiple contests: 2009 California tests Social Science Research Network, Tech. Rep., 2009 [Online]. Available: http://ssrn.com/abstract=1443314

Luke W. Miratrix studied mathematics and computer science at Reed College, Portland, OR; California Institute of Technology, Pasadena; and Massachusetts Institute of Technology, Cambridge. He currently is pursuing the Ph.D. degree in the Department of Statistics, University of California, Berkeley.

After teaching for several years, he returned to school to study the interplay of numerical information and human beings. He is primarily interested in creating tools for investigating issues in the social sciences in a robust and defensible manner.

Philip B. Stark received the A.B. degree in philosophy from Princeton University and the Ph.D. degree in earth science from the University of California, San Diego. He was a postdoctoral fellow at the Scripps Institution of Oceanography, University of California, San Diego, and in the Department of Statistics, University of California, Berkeley.

He 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.

Authorized licensed use limited to: Univ of Calif Berkeley. Downloaded on December 14, 2009 at 13:13 from IEEE Xplore. Restrictions apply.