A Bayesian Method for Auditing Elections

Ronald L. Rivest
Computer Science and Artificial Intelligence Lab,
MIT, Cambridge, MA 02139
rivest@mit.edu

Emily Shen
Computer Science and Artificial Intelligence Lab,
MIT, Cambridge, MA 02139
eshen@csail.mit.edu

Abstract

We propose an approach to post-election auditing based on Bayesian principles, and give experimental evidence for its efficiency and effectiveness. We call such an audit a "Bayes audit". It aims to control the probability of miscertification (certifying a wrong election outcome). The miscertification probability is computed using a Bayesian model based on information gathered by the audit so far.

A Bayes audit is a single-ballot audit method applicable to any voting system (e.g. plurality, approval, IRV, Borda, Schulze, etc.) as long as the number of ballot types is not too large. The method requires only the ability to randomly sample single ballots and the ability to compute the election outcome for a profile of ballots. A Bayes audit does not require the computation of a "margin of victory" in order to get started.

Bayes audits are applicable both to ballot-polling audits, which work just from the paper ballots, and to comparison audits, which work by comparing the paper ballots to their electronic representations. The procedure is quite simple and can be described on a single page.

The Bayes audit uses an efficient method (which may be based on the use of gamma variates or on Pólya's Urn) for simulating a Bayesian posterior distribution on the tally of a profile of ballots.

A Bayes audit is very similar to single-ballot risk-limiting audits. However, since Bayes audits are based on different principles, the precise relationship between risk-limiting audits and Bayes audits remains open.

We provide some initial experimental results indicating that Bayes audits are quite efficient, requiring few ballots to be examined, and that the miscertification rate is indeed kept small, even for very close elections.

1 Introduction

This section provides a quick introduction to post-election audits and our notation. Section 2 then presents our proposed Bayes audit procedure. Section 3 gives the results of our initial experiments using this method on simulated and real election data. Section 4 considers some extensions and variations of the basic method, and Sections 5 and 6 discuss and summarize what we have learned about the Bayes audit. Appendix A provides some additional technical details on efficient implementation methods.

1.1 Post-election audits

Informally, the purpose of a post-election audit is to check that the reported election outcome is correct, by auditing enough randomly chosen ballots.

Absolute certainty isn't required of an audit (the only way to achieve absolute certainty is to audit by hand all, or nearly all, of the ballots), but a good audit should have a high probability of exposing (and correcting) an incorrect reported outcome.

The number of ballots audited is typically variable, depending on factors such as the margin of victory (close elections require more work), the random sampling process, whether the audit is a ballot-polling audit or a comparison audit, and (if a comparison audit) the number and nature of errors found. The audit may proceed in stages, auditing more and more ballots until an audit result can be announced with sufficient statistical confidence.

Norden et al. [15] give an excellent introduction to post-election audits. VerifiedVoting.org [24] summarizes U.S. legislative requirements for post-election audits. Philip Stark has pioneered many of the most recent and powerful post-election audit methods; he provides a web page [19] listing key papers and talks. Checkoway et al. [2] give an interesting information-theoretic single-ballot audit method.

Since election equipment and procedures are complex, and since the initially reported election outcome may be incorrect due to various errors or even fraud, post-election audits are strongly recommended as a means of providing a high degree of confidence—to the voters, to the election officials, and to the candidates—that the election outcome is indeed correct.

This paper provides election officials with a new post-election auditing tool, one that is efficient and effective.

1.2 Framework and notation

Voters. We consider a single-contest election with n voters, each of whom casts a single paper ballot.

Ballot types. We assume that each ballot has one of t "types" (or "signatures"). In a plurality election, t may equal the number m of candidates, and each ballot indicates the voter's preferred choice. If "overvote" and/or "undervote" are allowed ballot types, t may be larger than m by one or two. With ranked-choice voting, a ballot lists candidates in order of preference, so $t = m!$ (or even larger, if overvotes and undervotes are allowed, or if ballots may be partial).

Let $[a..b]$ denote $\{a, a+1, a+2, \ldots, b\}$. Each ballot is treated as an element of $[1..t]$, that is, as an integer in the range 1 to t, inclusive.

Reported ballot types and profile. Assume that the n paper ballots are scanned, producing a sequence of n *reported ballot types*, each in $[1..t]$. We let r_i denote the reported type of the ith ballot, for $i = 1, 2, \ldots, n$. The *reported election profile* is

$$\boldsymbol{r} = (r_1, r_2, \ldots, r_n) \ .$$

Some authors used the adjective "apparent" rather than "reported". Reported results are initial, preliminary, and unofficial.

Denote the set of *possible profiles* for an election with n voters and t ballot types as:

$$P_{t,n} = \{(x_1, x_2, \ldots, x_n) \mid (\forall i)(x_i \in [1..t])\} \ ;$$

These are sequences of length n, of which each element is a ballot type (an integer in $[1..t]$).

Tallies. Let R_j denote the number of ballots of reported type j, for $j = 1, 2, \ldots, t$. The R_j's are nonnegative integers whose sum is n. Let

$$\boldsymbol{R} = (R_1, R_2, \ldots, R_t)$$

be the *reported tally* vector.

For any list $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ with positive integer elements, let tally($\boldsymbol{x}$) denote the *tally* of $\boldsymbol{x}$: the list $(X_1, X_2, \ldots)$ whose kth component X_k is the number of elements of $\boldsymbol{x}$ that are equal to k. Thus

$$\boldsymbol{R} = \text{tally}(\boldsymbol{r}) \ .$$

Let

$$T_{t,n} = \{(X_1, \ldots, X_t) \mid (\forall i) X_i \geq 0 \text{ and } \sum_{i=1}^{t} X_i = n\}$$

denote the set of *possible tally vectors* for an election with n voters and t ballot types. Each such tally vector has length t; the ith element X_i denotes the number of ballots of ballot type i. The sum of the tally elements is n.

It is convenient for us in this definition to admit arbitrary nonnegative real numbers as tally components, rather than just arbitrary nonnegative integers, even though in reality there will never be "fractional votes" cast. This detail is not consequential here. Note that social choice functions are almost always based on comparisons between tallies, or sums of tallies, so they work just fine on an input "tally vector" consisting of arbitrary real numbers.

Voting system election outcome. We assume that the reported election outcome is determined by applying the appropriate social choice function f to the reported tally: $f(\boldsymbol{R})$ is the *reported election outcome*. The function f depends on whether the voting system is plurality, approval, IRV, etc. We assume that f is deterministic (that is, not randomized); f always reports the same outcome for the same given input tally. (In practice, this just means that ties are broken in some pre-determined way.)

We let M denote the number of possible election outcomes, and represent the outcome as an integer in $[1..M]$. In a plurality election, M equals m, the number of candidates. In an election with multiple winners, M equals the number of possible winning combinations. For example, suppose there are four candidates vying for two open city council seats; in this case $m = 4$ and $M = 6$.

Audit types. We are concerned about miscertification—where election officials certify (accept) an incorrect election outcome. The risk of such miscertification can be dramatically reduced by conducting an effective *post-election audit* that involves examining, by hand, some or all of the paper ballots.

The audit may be of one of two types: a *ballot-polling audit* or a *comparison audit* [9]. In both cases, a hand examination of the audited ballots determines their actual ballot types. The comparison audit also then compares

these actual types, one by one, with their corresponding reported types. This requires voting systems that can track this correspondence. A comparison audit is usually significantly more efficient, requiring the auditing of fewer ballots, but in either case an efficient audit for a typical election may need to examine only a tiny fraction of the paper ballots. The Bayes method works for both audit types. Lindeman et al. [10] argue for the utility of ballot-polling audits in today's environment.

Errors and actual ballot types. Some reported ballot types may be erroneous, due to scanning errors, scanner limitations at interpreting voter intent, misprogramming of scanners, recording errors, or even some form of fraud. Such errors are usually inconsequential, but they may in some cases cause election officials to accept (certify) the wrong election outcome.

We assume that a hand examination (audit) of the ith ballot yields its *actual type* (or true type) a_i. Let

$$\boldsymbol{a} = (a_1, a_2, \ldots, a_n)$$

denote the *actual election profile*.

Let A_j denote the number of ballots of actual type j, and let

$$\boldsymbol{A} = (A_1, A_2, \ldots, A_t) = \text{tally}(\boldsymbol{a})$$

be the *actual tally*. Thus, $f(\boldsymbol{A})$ is the *actual election outcome*, which is by definition *correct*.

When a ballot of reported type j has actual type k, we say that ballot i has "error type" (j,k). Of course, there was an error only if $j \neq k$, but it is convenient to use the term "error type" even in this case when there is no actual error.

Correct election outcomes and upsets. If there are no errors, then $r_i = a_i$ for all i, and the reported tally $\boldsymbol{R}$ equals the actual tally $\boldsymbol{A}$. Errors may cause an *incorrect* reported election outcome:

$$f(\boldsymbol{R}) \neq f(\boldsymbol{A}) \ .$$

We call such an event an "outcome error" or, for brevity, an "upset." That is, an "upset" happens when the reported election outcome was incorrect. An audit should not only detect upsets with high probability, but also provide the correct election outcome in such cases. A Bayes audit will stop examining ballots when the probability of an upset is estimated to be sufficiently small.

2 Bayes Audit Method

This section presents details of the "Bayes audit." We begin with an intuitive overview of the approach, and then dive into the technical details.

Intuitive overview. A Bayes audit examines randomly selected paper ballots one at a time. When a selected paper ballot is audited, its actual type is determined by hand examination. The actual type of an audited ballot may or may not be compared to the reported type for that ballot—this is the difference between a comparison audit and a ballot-polling audit.

The Bayes approach can be viewed as providing an "oracle" or "magic box" that can answer the question:

> What is the probability that each candidate would be determined to be the actual winner, if the auditing process continued to examine all the ballots, and the remaining ballots audited showed results similar to what we've seen in the ballots audited so far?

This is an informally stated question; the precise interpretation is given later in this section.

Thus, as more and more ballots are examined, the "winning probability" for each candidate, as computed by the oracle, will vary. Figure 2 of Section 3.1 illustrates how the estimated winning probabilities might vary with the number of ballots audited, for an example election.

As the number of audited ballots increases, however, one election outcome—the correct election outcome—will see its winning probability increase towards 1. When its winning probability exceeds a given threshold, such as 0.95, then the Bayes audit may stop with high confidence that the correct election outcome has been determined. Otherwise, the audit continues.

As a matter of policy, however, it seems politic to only stop the election early, before all of the ballots have been audited, when it is the *reported* winner whose winning probability exceeds the given threshold. If some other election outcome has an estimated winning probability in excess of the threshold, then the audit process has detected an *upset*, and the audit may continue, more for political reasons than for statistical reasons. An election official may reasonably require an audit of all of the ballots in order to overturn the initially reported winner and announce that another outcome is indeed the correct outcome. The Bayes audit described here takes this approach: the audit stops early only when the winning probability of the *reported winner* exceeds the given threshold. Equivalently, the Bayes audit stops when the probability of an upset drops below a given upset probability threshold.

The ability of the Bayes approach to provide an answer to the question, "What is the chance that candidate X will be the winner if we proceed to a full audit?" distinguishes the Bayes approach from other post-election auditing procedures, such as risk-limiting audits. Such information can be helpful in estimating how close the audit is to stopping, or to understanding what the initial

auditing results mean if the audit has to be stopped partway through for budgetary or scheduling reasons.

The Bayes audit also provides a framework for *several* auditors to participate, each of whom has a different approach to answering the key "winning probability" question. For example, each candidate might approach this question with slightly different biases, suspicions, or prejudices. This variability can be reflected by having each such auditor use a different Bayesian prior. The audit stops when *all* such auditors agree to stop. We include discussion of such multi-auditor methods below, but recommend only using a single "nonpartisan" auditor.

Relation to prior work on Bayesian audits. Our basic approach is only slightly new—it follows established paradigms in Bayesian analysis—for example, see Meeden [13], Ghosh and Meeden [4], and Meeden and Sargent [14] for very similar approaches to the auditing of financial books and other applications; also see Lazar et al. [8], Ghosh et al. [3].

What is perhaps new here is mostly just the application of these methods to post-election audits, and their experimental evaluation. However, since Bayesian methods appear to be very effective (see Section 3), and since Bayesian methods have not (to our knowledge) been proposed before for election audits, it seems worthwhile to explore the details, pros, and cons of this approach.

Auditing process. A Bayes audit works in a sequential decision-making manner; Stark [22] gives for a framework for performing post-election audits in this manner. However, because our approach is Bayesian, our details differ from his.

Figure 1 presents the Bayes (comparison) audit; the following section gives a detailed explanation.

Single-ballot auditing. We consider only audits that examine single ballots one at a time in a randomly chosen order—we call these "single-ballot audits." Such an audit operates in *stages*: the sth stage examines ballot s.

Auditing randomly chosen single ballots may be much more efficient than auditing randomly selected batches of ballots (e.g. by precinct, see Stark [18]) but requires (at least for comparison audits) individual marking of the paper ballots so that they may be associated with their scanned representations, an operation that many existing scanners do not support. Stark [23] gives a framework for single-ballot audits.

We assume, without loss of generality, that the ballots are well-shuffled to begin with (equivalently, the ballots are numbered $1,2,\ldots,n$ in a random order), so we may audit ballot 1 first, then ballot 2, up to ballot n if necessary. Of course, a paper ballot and its electronic representations are reordered or randomly numbered in the same way. (See Stark's site[1] for an approach to determining how to access the ballots in a random order, based on the use of the cryptographic hash function SHA-256.)

For a comparison audit, after ballot s is audited (i.e., after a_s is determined by a hand examination), the state of knowledge of the auditor can be represented as follows:

i	1	2	$\cdots$	s	$s+1$	$\cdots$	n
r_i	r_1	r_2	$\cdots$	r_s	r_{s+1}	$\cdots$	r_n
a_i	a_1	a_2	$\cdots$	a_s	**?**	**?**	**?**

(1)

The auditor knows the reported ballot types $r[1..n]$ and also the sampled actual ballot types $a[1..s]$; the values $a[s+1..n]$ are as yet unknown to the auditor. Let D_s denote the data available to the auditor after s ballots have been audited:

$$D_s = ((r_1, r_2, \ldots, r_n), (a_1, a_2, \ldots, a_s)) \; . \tag{2}$$

For a ballot-polling audit, we have

$$D_s = ((a_1, a_2, \ldots, a_s)) \; , \tag{3}$$

since the reported types corresponding to the actual types of the audited ballots are not available. (Equivalently, one could implement a ballot-polling audit within a comparison-audit framework where one pretends that the reported ballot types are all equal to some fixed value.)

Note that the auditor knows in either case the reported tally $\boldsymbol{R}$, and so he knows the reported election outcome $f(\boldsymbol{R})$.

The auditor does not know the actual tally $\boldsymbol{A}$, and so he does not know the actual election outcome $f(\boldsymbol{A})$ (typically, although as s nears n the actual outcome may be no longer depend on the actual ballot types of the unaudited ballots).

Stopping. At the end of any stage the auditor may do one of three things:

- **OK:** Stop and declare that the reported election outcome is correct.
- **NOT OK:** Stop and declare that the reported election outcome is incorrect.
- **Don't know yet:** Announce that the auditing so far is inconclusive, and more auditing should be done.

An audit may make errors of two types:

- *(False OK)* Declaring that the reported outcome is correct when it is not. We also call this error a *miscertification*.

- *(False NOT OK)* Declaring that the reported outcome is incorrect when it is correct.

We assume an audit may not make "False NOT OK" errors—the audit must examine *all* of the ballots before an output of "NOT OK" is produced.

One may view an audit as testing the hypothesis

$$H = \text{"The reported election outcome is incorrect."}$$

(see Stark [22]). Hence a "False OK" error is "False negative" error, while a "False NOT OK" error is a "False positive" error.

Risk-limiting audits. An audit is *risk-limiting* if the probability (over the random ordering of the ballots and the random choices of the audit) of a "False OK" error is at most a pre-specified value α (the "risk limit"); here α might be 0.01, 0.05, or 0.10. Lindeman and Stark [9] give an introduction to risk-limiting audits. Bayes audits are not known to be risk-limiting, although experimental results make this plausible.

Multiple auditors "The purpose of an election is not to name the winner, it is to convince the losers that they lost." [2]

"Indeed, it is often said that the main purpose of election fairness is to convince the loser that he or she lost the election fair and square – winners rarely complain about the fairness of an election. Perhaps more important, these comments apply even more strongly to the electorate supporting the losing candidate." [3]

In support of this principle, we suggest that post-election audits might be run in the following manner. There is at least one *nonpartisan* auditor $\mathscr{N}$ (representing, say, the election officials) and perhaps several *partisan* "auditors" $\mathscr{P}_k$ (with one partisan auditor representing each candidate, say). The audit proceeds to examine ballots one by one until *all* of the auditors are convinced that the reported outcome is correct (or that their preferred candidate has lost), or until all of the ballots have been examined. We call such a Bayes audit by one nonpartisan auditor and one partisan auditor for each outcome as an $\mathscr{NP}$ audit. While our recommend audit is an $\mathscr{N}$ audit (with a single nonpartisan auditor), the $\mathscr{NP}$ audits are nonetheless well-motivated and intriguing; we study their performance as well.

A post-election audit is an ideal situation for running multiple Bayesian audits in parallel, since the interests and prior probabilities of various candidates are likely not well aligned. The data provided by the sequential hand auditing of the ballots can be used by all such auditors to determine whether they are "ready to throw in the towel." When all do, the audit stops.

Modeling uncertainty; upset probability. The true values of the actual profile $\boldsymbol{a}$ and associated actual tally $\boldsymbol{A}$ are unknown to the auditor. It is reasonable for the auditor to represent this uncertainty regarding $\boldsymbol{a}$ after s ballots have been audited using a probability distribution π_s on $P_{t,n}$.

In a Bayesian framework π_s might be called a "subjective probability distribution," even though "subjective" seems inappropriate for an entirely mechanical auditing procedure.

Intuitively, $\pi_s(\boldsymbol{x})$ is the (subjective) probability that the true profile $\boldsymbol{a}$ equals $\boldsymbol{x}$, where $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, and where the probability is taken over the various ways the initial known segment $(a_1, a_2, \ldots, a_s)$ may be completed to yield the actual profile $a = (a_1, a_2, \ldots, a_n)$.

Let $\boldsymbol{b}_s$ be a $P_{t,n}$-valued random variable distributed according to π_s. Likely values of $\boldsymbol{b}_s$ represent ways of completing $a[1..s]$ to a profile $a[1..n]$ that seem likely to the auditor, given what is known so far.

Let $\boldsymbol{B}_s$ be the corresponding random variable taking values in the set $T_{t,n}$ of possible tallies, so that

$$\boldsymbol{B}_s = \text{tally}(\boldsymbol{b}_s) \ .$$

Likely values of $\boldsymbol{B}_s$ should represent actual election tallies that seem likely, given what is known so far. We may say for convenience that both the profile $\boldsymbol{b}_s$ and its associated tally $\boldsymbol{B}_s$ "have distribution π_s"; we hope this causes no confusion.

Finally, likely values of $f(\boldsymbol{B}_s)$ should represent actual election outcomes that seem likely, given what is known so far.

At the end of the sth stage the auditor can compute from π_s the *winning probability* $w_{s,\ell}$ for each possible election outcome $\ell \in [1..M]$:

$$w_{s,\ell} = \Pr[f(\boldsymbol{B}_s) = \ell] \ .$$

Let $\boldsymbol{w}_s$ denote the vector of winning probabilities:

$$\boldsymbol{w}_s = (w_{s,1}, w_{s,2}, \ldots, w_{s,M}) \ .$$

Simulation of posterior; estimating winning probabilities. It is important that the auditor be able to compute $\boldsymbol{w}_s$ (at least approximately), so that the auditor can determine whether to stop auditing. To do so, it suffices for the auditor to be able to *sample* the random variable $\boldsymbol{B}_s$ (or equivalently, be able to sample $\boldsymbol{b}_s$); the method we propose makes use of this approach.

One very nice property of this approach is that *it does not restrict the social choice function f in any way*. (This is reminiscent of the way in which bootstrap or other resampling methods may be used to provide estimates of the sampling distribution of complex statistics.)

Stopping rule. An auditor might wish to stop auditing when (1) the estimated probability of an upset becomes very low, or (2) the estimated probability of his favored candidate winning becomes very low.

Here "very low" means "at most ε," where ε is a value selected by the auditor. Values such as $\varepsilon = 0.01$, $\varepsilon = 0.05$, or $\varepsilon = 0.10$ might be appropriate for a typical audit. Here ε corresponds to α in a risk-limiting audit.

Given $\boldsymbol{w}_s$, the auditor can easily estimate the probability $u_{s,\ell}$ of an "*upset by* ℓ" (where $\ell \neq f(\boldsymbol{R})$) as the probability of ℓ winning:

$$u_{s,\ell} = w_{s,\ell} \; ; \tag{4}$$

and can estimate the overall *upset probability*

$$\begin{aligned} u_s &= \Pr[f(\boldsymbol{B}_s) \neq f(\boldsymbol{R})] \\ &= 1 - w_{s,f(R)} \\ &= \sum_{\ell \neq f(R)} w_{s,\ell} \; . \end{aligned}$$

When $s = n$, the upset probability u_n is either 0 or 1, as no uncertainty remains about the unaudited actual ballot types—there are no unaudited ballots.

We assume that the auditor really cares about either the overall upset probability u_s (for a nonpartisan auditor) or an individual upset probability $u_{s,\ell}$ (for a partisan auditor). Let v_s denote the value the auditor cares about; we may call this a generalized upset probability. (In general, $v_s = v(\boldsymbol{w}_s)$ for a suitable function v.)

The auditor will wish to stop auditing at the end of stage s if

$$v_s \leq \varepsilon \; .$$

That is, the auditor will be willing to stop and accept the reported election outcome (i.e., say "OK") when he estimates that the probability that auditing all of the remaining ballots will yield an upset (of the sort cared about) with probability no more than ε.

For those familiar with auditing methods based on p-values (e.g., Stark [22]), we have, as Meeden [13] notes, a situation where "it might be somewhat of a surprise that the posterior probability studied here behaves approximately like a p-value."

Selecting π_s. How should the auditor determine the probability distribution π_s on $\boldsymbol{b}_s$ (equivalently, on $\boldsymbol{B}_s$)? How to determine the probability of an actual profile (or its tally), given the ballots audited so far (and also the reported profile, for a comparison audit)?

We recommend a Bayesian approach. That is, π_s has the form of a posterior distribution, given the data D_s available and a suitable prior π_0; the distributions π_s follow Bayes' Rule:

$$\begin{aligned} \pi_s(\boldsymbol{x}) &= \Pr[B_s = \boldsymbol{x} \mid D_s] && (5) \\ &= \text{const} \cdot \Pr[D_s \mid \boldsymbol{B}_s = \boldsymbol{x}] \cdot \pi_0(\boldsymbol{x}) && (6) \end{aligned}$$

Of course, the Bayesian auditor needs to start with a suitable prior distribution on profiles (or tallies).

Our recommendation is for a nonpartisan auditor to start with a prior distribution that is noninformative—it gives equal prior probabilities to all possible tallies—while our recommendation for a partisan auditor is to assume that all unseen actual ballot types favor the auditor's favored outcome.

Our general approach is then that of Bayesian Sequential Analysis (see Berger [1, Ch. 7]).

We thus may call our auditors "skeptical Bayesians" that choose a prior for the Bayesian approach not to be realistic, but to be rather pessimistic (about confirming the reported outcome). While he is using a Bayesian framework, the choice by $\mathscr{N}$ of a noninformative prior helps to ensure that an incorrect reported election outcome will not be easily accepted as correct. In this manner we attempt to avoid many of the usual concerns about using a Bayesian approach arising from the choice of prior. A partisan auditor $\mathscr{P}_k$ doesn't start with a balanced noninformative prior, but rather a highly unbalanced one that says, "I suspect my candidate was cheated out of this election; show me evidence to the contrary!"

Ballot-polling versus comparison audits. The sample set of s audited ballots are used to form a Bayesian model that provides a probability distribution π_s on the set of possibilities for the "completed" actual profile—that is, what the profile would look like if all of the ballots were audited.

For a ballot-polling audit, the probability distribution π_s is based only on the distribution of actual ballot types in the sample of audited ballots. A ballot-polling audit is simpler to describe, so we shall do that first.

For a comparison audit, the probability distribution π_s also depends on the reported ballot types $\boldsymbol{r}$. Comparison audits are usually much more efficient than ballot-polling audits, since the actual ballot types are very strongly correlated with the reported ballot types. A comparison audit is roughly comparable to t ballot-polling audits, one for each reported ballot type (although only a single decision is made, not a decision for each reported ballot type).

As noted, we shall focus first on ballot-polling audits as they are simpler to describe.

Dirichlet distributions. We choose to use Dirichlet distributions for the prior probability distribution π_0 and the distributions π_s, for four reasons:

- The Dirichlet distributions form a rich class that are well-matched to our application.
- It is possible to choose a noninformative prior distribution π_0 from within this class, that gives all possi-

ble tallies equal likelihood (for the nonpartisan auditor).

- The Dirichlet distributions are conjugate to the multinomial distribution, so that the posterior distribution at each step is Dirichlet, given that the prior is Dirichlet and the data is multinomial. This is exactly our situation (see equation (6)).

- It is possible to efficiently sample from a Dirichlet distribution (see Appendix A).

A (scaled) Dirichlet probability distribution $\mathrm{Dir}(\boldsymbol{\alpha}, n)$ on t distinct possibilities (i.e., of order t) is defined by a vector of t nonnegative "*hyperparameters*"

$$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_t)$$

and a scale factor n. It is defined on the $(t-1)$-dimensional simplex $T_{t,n}$ with density function

$$p_{\mathrm{Dir}}(\boldsymbol{x}) \propto \prod_{\substack{i=1 \\ x_i \neq 0}}^{t} x_i^{\alpha_i - 1}$$

(where $x_i = 0$ if $\alpha_i = 0$; this is an "improper prior").

Of special interest is the case $\boldsymbol{\alpha} = \mathbf{1}^t = (1, 1, \ldots, 1)$, when the Dirichlet distribution is *uniform*; we suggest it as a prior for the nonpartisan auditor $\mathscr{N}$:

$$\pi_0 = \mathrm{Dir}(\mathbf{1}^t, n) \; ;$$

this is a common choice when using the Dirichlet distribution in Bayesian analysis. For example, in an election of $n = 600$ votes and $m = 3$ candidates, a tally of $(200, 200, 200)$ is as likely with π_0 as one of $(150, 392, 58)$, which is as likely as one of $(0, 0, 600)$, which is as likely as any other tally.

Of course, choosing a prior is at the heart of a Bayesian approach, and the use of a prior distribution makes a Bayes audit not obviously "risk-limiting" in the usual sense. However, a Bayes audit has a very similar goal: making the subjective (i.e. posterior) probability of making a "False OK" error (an "upset") less than a given upset probability limit ε.

For a partisan auditor $\mathscr{P}_k$, the prior only envisions the possibility of ballots of a particular actual type k, that is, the prior has hyperparameters $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_t)$ where $\alpha_k = 1$ and $\alpha_j = 0$ if $j \neq k$. (This is an "improper prior" because of the zeros, but no technical difficulties arise.)

In any case, we feel it is interesting and perhaps useful to follow this line of thought to the end, to see what possible benefits such "Bayes audits" may have, compared to non-Bayesian risk-limiting audits.

Applying Bayes' Rule. Because the Dirichlet distribution is conjugate to the multinomial distribution, applying Bayes' Rule is exceptionally simple: if your prior distribution on the actual tally for a set of n ballots is $\mathrm{Dir}(\boldsymbol{\alpha}, n)$, and you draw (without replacement) for a ballot-polling audit a sample $C = (c_1, c_2, \ldots, c_s)$ of size s with tally $\boldsymbol{C} = (C_1, C_2, \ldots, C_t)$, then your posterior distribution on the actual tally for the entire set of n ballots will be [11]:

$$\pi_s = \boldsymbol{C} + \mathrm{Dir}(\boldsymbol{\alpha} + \boldsymbol{C}, n - s) \; .$$

The first term $\boldsymbol{C}$ represents the actual tally for the sample; the second term represents the Bayesian posterior distribution for the tally of the as-yet-unaudited $n - s$ ballots, where the tally $\boldsymbol{C}$ has been simply added to the hyperparameters for the Dirichlet distribution.

The operation of sampling from a Dirichlet distribution is standard and quite easy; details of drawing a sample are given in Appendix A. So the operation of drawing a simulated tally $\boldsymbol{B}_s$

$$\boldsymbol{B}_s \leftarrow \boldsymbol{C} + \mathrm{Dir}(\boldsymbol{\alpha} + \boldsymbol{C}, n - s)$$

can be simulated on the computer numerous times to develop an accurate estimate of the upset probability v_s—the probability that $f(\boldsymbol{B}_s) \neq f(\boldsymbol{R})$ (for a nonpartisan auditor) or that $f(\boldsymbol{B}_s) \neq \ell$ (for a partisan auditor favoring outcome $\ell \neq f(R)$). Note that the value of C here is fixed—it is the sample taken—so the repetitions are merely computational, involving repeated sampling from $\mathrm{Dir}(\boldsymbol{\alpha} + \boldsymbol{C}, n - s)$ to obtain a simulated tally vector for the remaining $n - s$ unaudited ballots, which is added to the vector $\boldsymbol{C}$ to obtain a simulated tally $\boldsymbol{B}_s$ for the entire n-ballot profile. In particular, the computation of v_s for a given s does not require the examination of additional ballots by hand.

Comparison audits. For a comparison audit, the reported ballot types help form an "error model" for the scanning process, as a simple extension of the ballot-polling procedure described above.

Recall that the "error type" of a ballot with reported type j and actual type k is the pair (j, k).

Auditing the first s ballots of the profile determines the count $C_{jk}^{(s)}$ of ballots of each error type among these first s ballots. These counts, together with the prior hyperparameters, determine the error model.

This error model will then be used to estimate the chance that the $n - s$ unaudited ballots, if subject to the same sorts of errors as seen so far, together with the s audited ballots (as corrected) will yield a NOT OK judgment (an upset).

The auditor's prior probability model is fully characterized by t^2 positive integer values $\alpha_{j,k}$, one such parameter for each error type (j, k). For the nonpartisan

Procedure for a Bayes comparison audit

1. **[Input reported ballot types]** Number the ballots $1, 2, \ldots, n$ in a random order. Let $\boldsymbol{r} = (r_1, r_2, \ldots, r_n)$ be the profile of n reported ballot types (each given as an integer in $[1..t]$). Let $\boldsymbol{R} = (R_1, \ldots, R_t)$ denote the reported tally per ballot type. Let f denote the desired social choice function, and $f(\boldsymbol{R})$ the reported outcome.

2. **[Outer audit loop]** Let ε denote the desired upset probability limit.
For each s between 1 and n, inclusive:

   - Determine actual type a_s of ballot s by hand examination.
   - If $s < n$ and you wish to consider stopping early, or if $s = n$:
     - Compute the "upset probability" v_s from $\boldsymbol{r}$ and $(a_1, \ldots, a_s)$.
     - If $v_s \leq \varepsilon$: Stop and report "OK".

3. **[Upset]** Stop and report "NOT OK".

Procedure for computing upset probability

1. **[Compute many simulated outcomes]** Compute outcomes for many (e.g. 5000) simulated profiles $\boldsymbol{b}_s$ generated using Pólya's Urn, or (more efficiently) for many simulated tally vectors $\boldsymbol{B}_s$ generated using Dirichlet.

2. **[Return upset fraction]** Return the fraction of simulated outcomes that are "upsets".

Procedure for generating a simulated outcome using Pólya's Urn

1. **[Initialize urns]** Create an urn for each reported type j, and add $\alpha_{j,k}$ balls to it of each type k, $1 \leq k \leq t$.

2. **[Condition by audited sample of size s]** For each i, $1 \leq i \leq s$:

   - Add a ball of type a_i to urn r_i.
   - Let $b_i = a_i$.

3. **[Draw from urns $(n-s)$ times]** For each i, $s < i \leq n$:

   - Remove a randomly selected ball from urn r_i and let x denote its type.
   - Let $b_i = x$.
   - Return *two* balls of type x to urn r_i.

4. **[Return outcome]** Return $f(\boldsymbol{B}_s)$ where $\boldsymbol{B}_s = \text{tally}(\boldsymbol{b}_s))$ and $\boldsymbol{b}_s = (b_1, b_2, \ldots, b_n)$.

Procedure for generating a simulated outcome using Dirichlet

1. **[Generate simulated tally]** For each reported type j, compute a simulated tally $\boldsymbol{B}_{s,j}$ by sampling from

$$\boldsymbol{C} + \text{Dir}(\boldsymbol{\alpha}_j + \boldsymbol{C}, R_j - C_j)$$

   where $\boldsymbol{C} = (C_1, \ldots, C_t)$ is the distribution of actual ballot types among the ballots audited so far having reported type j and $\boldsymbol{\alpha}_j = (\alpha_{j,1}, \alpha_{j,2}, \ldots, \alpha_{j,t})$. See Section A.1.

2. **[Return outcome]** Return $f(\boldsymbol{B}_s)$ where $\boldsymbol{B}_s = \boldsymbol{B}_{s,1} + \boldsymbol{B}_{s,2} + \cdots + \boldsymbol{B}_{s,t}$ is the overall simulated tally.

Figure 1: A Bayesian comparison audit method for a single auditor with prior hyperparameters $\alpha_{j,k}$.

auditor $\mathscr{N}$, we propose setting each such hyperparameter to 1:

$$\alpha_{j,k} = 1 \ .$$

We also consider a partisan auditor $\mathscr{P}_k$ for each actual ballot type k, who has $\alpha_{j,k} = 1$ and $\alpha_{j,k'} = 0$ for $k' \neq k$. We let $\mathscr{P}$ denote the auditor consisting of all t such partisan auditors, and $\mathscr{NP}$ denote the auditor consisting of the nonpartisan auditor $\mathscr{N}$ and all t auditors in $\mathscr{P}$. All see Section 4.

For each reported type j, then, we have a posterior probability distribution $\pi_{s,j}$ on the possible actual tallies for the set of all ballots of reported type j. For each such reported type, the posterior distributions are updated just as for a ballot-polling audit. More precisely, the posterior for the actual tallies of the ballots of reported type j has the form

$$\pi_{s,j} = \boldsymbol{C}_j + \mathrm{Dir}(\boldsymbol{\alpha}_j + \boldsymbol{C}_j, R_j - C_j) \ .$$

where $\boldsymbol{C}_j = (C_{j,1}, C_{j,2}, \ldots, C_{j,t})$ is the actual tally vector for the ballots in the sample of reported type j (which sums to C_j), where R_j is the total number of ballots of reported type j in the entire profile of size n, and C_j is the number of ballots of reported type j within the sample, and where

$$\boldsymbol{\alpha}_j = (\alpha_{j,1}, \alpha_{j,2}, \ldots, \alpha_{j,t})$$

are the hyperparameters for the Dirichlet prior distribution on the actual tallies for the ballots of reported type j.

Since we can sample from these posterior distributions $\pi_{s,j}$ for all j, by adding the results together we can easily obtain a sample from the posterior distribution π_s for the actual tally vector for the entire profile.

Given this ability to sample from π_s we can as before use repeated sampling to compute (accurately estimate) the upset probability v_s.

Determining upset probability. The Bayes audit has the structure of two nested loops: an outer, "real-world" loop that is auditing the ballots one by one, and an inner "simulation" loop that is using Monte Carlo simulation to estimate the probability that the Bayesian model would generate an election upset (compared to the originally reported election outcome).

Since computers are fast, and human time is precious, this seems like a reasonable structure for an audit.

Although the computers are doing Monte Carlo simulation, their use of randomness can be made reproducible by using deterministic pseudo-random generators based on a seed generated at the time of the audit that is unpredictable to an adversary (see [9]). Therefore, all the computations in a Bayes audit are checkable by others.

The simulations may involve a fair amount of computation, but not an undue amount for today's computers; less than a second of computation is required to estimate each upset probability. However, this computation is not something that can be performed reasonably by hand.

Our hope is that the proposed structure and approach would yield audits that are noticeably more efficient in terms of the number of ballots that need to be audited when the reported election outcome is in fact correct. The next section provides some experimental support for this goal.

3 Experimental results

This section gives experimental evidence regarding the effectiveness and efficiency of a Bayes post-election audit.

We start in Section 3.1 with a simple example election, just to get a feeling as to how the winning probabilities might evolve in a Bayes audit of a three-candidate election with 10,000 voters.

Section 3.2 gives experimental results studying the miscertification rate of the Bayes audit for a very close election (10,000 votes with a margin of 2 votes) with an incorrect reported election outcome. The results show that the Bayes audit can be effective at controlling the miscertification rate to a low level.

Section 3.3 compares the efficiency (number of ballots audited) of the Bayes audit to the single-ballot comparison audit method of Checkoway et al. [2]. The results show that the Bayes audit is an order of magnitude more efficient. (Caveat: the Bayes audit and the Checkoway method are both post-election audit methods, but are solving somewhat different problems.)

Section 3.4 compares the efficiency of the Bayes audit to the single-ballot comparison auditing method of Stark [20] for real election data from Stanislaus Oakdale. The Bayes audit appears to be somewhat more efficient (but again, these audits are not solving exactly the same problem, so it is a bit of apples versus oranges).

Finally, Section 3.5 compares the efficiency of the Bayes audit to the single-ballot ballot-polling audit of Stark [21] on real election data from Monterey. The Bayes audit was several times more efficient (although this is still apples versus oranges, and may suffer from small-sample effects).

We find these initial experimental results tremendously encouraging. Further experiements are needed to assess how well the Bayes audit would do on a larger variety of election data.

Audit types used in experiments. We consider in our experiments three types of Bayes audits:

- The nonpartisan auditor $\mathscr{N}$,

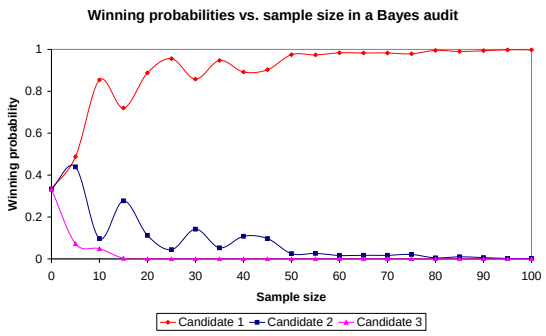

Figure 2: Winning probabilities vs. sample size, for a three-candidate example election of size $n = 10,000$, where candidate 1 received 5000 votes, candidate 2 received 4000 votes, and candidate 3 received 1000 votes, with the $\mathscr{N}$ auditor. Candidate 3's winning probability drops to zero after 15 ballots have been audited, and candidate 2's winning probability drops below 5% after 55 ballots have been audited, and below 1% once 80 ballots are audited. It is assumed for this experiment that the actual ballot types are all equal to the corresponding reported ballot types (i.e., there were no errors in the scanning).

- The collection $\mathscr{P}$ of partisan auditors, consisting of one partisan auditor $\mathscr{P}_k$ for each actual type k. The auditor $\mathscr{P}$ stops only when all component auditors agree to stop. The $\mathscr{P}_k$ auditors agree to stop when they agree that the probability of an upset is at most ε.

- The auditor $\mathscr{N}\mathscr{P}$, consisting of one nonpartisan auditor $\mathscr{N}$ and also one partisan auditor $\mathscr{P}_k$ for each actual type k. As with $\mathscr{P}$, the auditor $\mathscr{N}\mathscr{P}$ stops only when all component auditors agree to stop. The auditor $\mathscr{N}\mathscr{P}$ is just the union of the auditor $\mathscr{N}$ and the auditors in $\mathscr{P}$.

The Python software we used for these experiments, or for running additional experiments, is available from the authors.

3.1 How subjective winning probabilities change with sample size

This section illustrates how the subjective winning probabilities evolve as the sample size grows, using as an example a three-candidate plurality election with no errors, with candidate 1 receiving 5000 votes, candidate 2 receiving 4000 votes, and candidate 3 receiving 1000 votes (a total of $n = 10,000$ ballots). The $\mathscr{N}$ prior was used.

Figure 2 illustrates how the winning probabilities evolve as the sample size grows. We observe that these probabilities have large fluctuations for very small sample sizes but converge nicely as the sample size grows: to 1 for candidate 1, and to 0 for candidates 2 and 3.

	Actual type	
Reported type	1	2
1	4999	2
2	0	4999

Table 1: Example two-candidate election used to evaluate miscertification rates.

ε	$\mathscr{N}$	$\mathscr{P}$	$\mathscr{N}\mathscr{P}$
0.00	0.008	0.002	0.003
0.01	0.037	0.015	0.015
0.02	0.042	0.037	0.035
0.05	0.075	0.042	0.042
0.07	0.099	0.042	0.042
0.10	0.103	0.078	0.085

Table 2: Miscertification rates for various ε and Bayes audit types $\mathscr{N}$, $\mathscr{P}$, $\mathscr{N}\mathscr{P}$, on the election of Table 1.

3.2 Miscertification Rate

We experimentally evaluated how the miscertification rate of a Bayes comparison audit compares with the upset probability limit ε. We used the example election shown in Table 1.

The reported totals are 5001 votes for candidate 1 and 4999 for candidate 2, so candidate 1 is the reported winner. There were 2 votes for candidate 2 that were misreported as votes for candidate 1, so candidate 2 is the actual winner. (Of course, for an election with such a small reported margin of victory, a full hand count would in any case be a good idea!)

For each of five different values of upset probability limit ε, and for each auditing type $\mathscr{N}$, $\mathscr{P}$, $\mathscr{N}\mathscr{P}$, we counted the number of miscertifications (i.e., the number of times the audit declared the election outcome "OK") out of 1000 trials. See Table 2.

Thus, we observe that the empirical miscertification rates are roughly equal to the upset probability limit ε. We ran similar experiments on several other elections, including those used to test miscertification rates in Checkoway et al. [2], and obtained similar results: the empirical miscertification rate was usually less than or close to the upset probability limit ε.

We leave proving or disproving the following conjecture as a very interesting open problem.

Conjecture 1 *There is a small positive constant c such that for any election and for any $\varepsilon > 0$, a Bayes audit with upset probability limit ε is a risk-limiting audit for*

Reported type	Actual type 1	2	3
1	$0.1n$	0	0
2	0	$(0.45-m/2)n$	0
3	0	0	$(0.45+m/2)n$

Table 3: Checkoway example, with $n = 10,000$ ballots and margin m varying between 0.5% and 5%.

Margin m	Number of ballots examined: Checkoway	$\mathscr{N}$	$\mathscr{P}$	$\mathscr{NP}$
0.005	22,080	2000	2850	2940
0.010	10,949	921	1950	1900
0.015	7099	627	745	745
0.020	5022	486	590	581
0.025	3839	398	473	475
0.030	3206	325	400	400
0.035	2740	297	343	345
0.040	2391	273	300	300
0.045	2110	218	298	290
0.050	1889	203	274	258

Table 4: Average number of ballots audited for election given by Table 3, compared to Checkoway et al. [2].

risk limit $\alpha = c \cdot \varepsilon$. That is, for any election where the reported outcome is incorrect, the probability that a Bayes audit will miscertify the election is at most $c \cdot \varepsilon$.

3.3 Comparisons with Checkoway et al. results

This section compares the efficiency of a Bayes audit with that of the single-ballot audit method of Checkoway et al. [2]. We use the same elections as [2], shown in Table 3. The elections have two candidates (2 and 3), and undervotes (candidate 1). Each election had $n = 10,000$ ballots, while the margin m varied between 0.5% and 5%. We measured the efficiency of a Bayes audit, in terms of the average number of ballots examined in order to complete an audit for an upset probability limit $\varepsilon = 0.01$, averaged over 100 simulated audits. We compared these numbers to the average number of ballots examined in an audit by Checkoway et al. for a risk limit $\alpha = 0.01$, averaged over 1000 simulated audits.

See Table 4. (Note: the number of ballots for Checkoway et al. for $m = 0.015$ was estimated by linearly interpolating between the numbers for $m = 0.01$ and $m = 0.02$.) We see that Bayes auditing is significantly more efficient than the method of Checkoway et al., requiring almost 10 times fewer ballot to be examined, for the election given in Table 3.

We also ran similar experiments comparing the efficiency of Bayes auditing to Checkoway et al., for the three other error models in [2]. The results are very similar to those presented above, and thus are omitted for brevity.

3.4 Comparison with Stark's single-ballot audit

We compare the efficiency of the Bayes single-ballot comparison audit $\mathscr{N}$ to a single-ballot method of Stark for the 2011 Stanislaus Oakdale Measure O Election, which Stark audited as part of the AB 2023 pilot [20]. There were 1728 reported "Yes" votes, 1392 "No" votes, and 32 undervotes. For our purposes, we ignore undervotes. Stark's audit at $\alpha = 0.10$ examined 49 ballots [20]. The Bayes audit method $\mathscr{N}$, at an upset probability limit of $\varepsilon = 0.10$, over 100 simulated audits, examined an average of 92 ballots, and a median of 39 ballots. Thus, Bayes auditing appears approximately as efficient than Stark's audit. (These results are difficult to compare, since our results are for 100 trials, while Stark's actual result was only for the single actual case; the variance of his method, over different possible sampling results, may be high, as was ours. Moreover, it isn't clear what it means to compare an ε of 0.10 in our case with an α of 0.10 in his.)

Further study is needed to more fully understand the relative efficiency and effectiveness of Bayes audits and existing risk-limiting audits.

3.5 Comparison with Stark's ballot-polling audit

We compare the efficiency of the ballot-polling variant of the Bayes ballot-polling audit method (with audit type $\mathscr{N}$) to that of Stark's ballot-polling audit on the 2011 election for Monterey Peninsula Water Management District Director, Division 1 [21]. There were two candidates, Lewis and Mancini, and write-ins. The reported totals were 1353 votes for Lewis, 742 votes for Mancini, 13 write-ins, and 3 undervotes or overvotes. Stark's audit ignored undervotes and overvotes, and treated Mancini and write-ins as one candidate with a reported total of 755. Stark's audit looked at 89 ballots to complete an audit with a risk limit of 10%. The expected number of ballots for his method to examine, if there were no errors in the reported data, is 58, according to Stark [21].

We ran the Bayes ballot-polling audit method on the above data, ignoring under- and over-votes, treating Mancini and write-ins as one candidate, and assuming the reported data to be error-free (a reasonable assumption, given that Stark's single-ballot audits have never found any discrepancies).

For an upset probability limit ε of 10%, over 100 trials, the Bayes audit examined an average of 23 ballots, and

a median number of 11 ballots. Thus, with similar parameters the Bayes audit method examined fewer ballots than both the actual and the expected number of ballots examined using Stark's method.

4 Extensions and Variations

Multiple priors; other priors. One could use other priors different than, or in addition to, the priors $\mathscr{N}$ and $\mathscr{P}_k$ described above.

For example, one could set $\alpha_{j,k} = 0.1$ for all j,k; this would decrease the significance of the prior relative to the audit data, while still asserting that all error types are possible.

One could also consider a prior where correct machine operation is believed to be more likely than incorrect operation, so that $\alpha_{j,k} = 1$ if $j = k$ but $\alpha_{j,k} = 0.01$ if $j \neq k$.

However, we fear that a prior that places too much belief in the correct operation of the machine may give misleading results exactly when the machine is misbehaving.

So, for now we prefer the prior $\mathscr{N}$ with $\alpha_{j,k} = 1$ for all j,k. We believe that this prior will tend to give reasonably conservative results, possibly requiring more ballots to be audited than a more "realistic" prior. This conservatism is appropriate for an adversarial situation, such as our situation. The auditor method $\mathscr{N}\mathscr{P}$ is by design even more conservative. One could even envisage priors $\mathscr{N}(c)$ or $\mathscr{N}\mathscr{P}(c)$ which are just like $\mathscr{N}$ and $\mathscr{P}_k$ except all hyperparameters are multiplied by a constant $c > 1$.

As noted, for simplicity we suggest using the Bayes audit $\mathscr{N}$ for now, but further research may suggest the advisability of using other or additional priors.

Voting systems other than plurality. The Bayes audit method applies as well to other voting methods, as long as t is not too large.

We have tested Bayes audits on the following voting systems, in addition to plurality: IRV, Borda, Schulze, minimax. These are ranked-choice voting systems, so the test data was for $m = 3$ candidates, giving a total of $t = 3! = 6$ ballot types. No problems were noted, and the audit method seemed to work well (details omitted here).

If t becomes very large for a voting system, then modifications to the Bayes approach suggested here might be workable. But this is future research, and we do not explore such possibilities further here.

Multiple contests. Bayes audits can be applied in a straightforward manner to auditing multiple simultaneous contests. Ballots can be selected in a random order and examined by hand if and only if the audit for any of its contests is still in progress. The auditing for any individual contest can cease once the audit develops sufficient confidence in the reported election outcome for that particular contest, independent of what happens in the other contests.

Separating by reported ballot type. A Bayes audit can also be performed by separating the ballots into piles according to reported ballot type, and having auditing for each pile proceeding at its own rate, since the error model evolves separately for each reported ballot type. (Details omitted.)

Parallel with another method. Note that the Bayes audit computations can be done in parallel with any other single-ballot audit, say as a pilot. The data being input to both methods would be the same; the interesting question would be which method recommends stopping first.

5 Discussion

A Bayes comparison audit does not require the computation of a "margin of victory" in order to get started. This is particularly helpful for voting systems other than plurality, where the computation of a margin of victory may be difficult [12, 25].

We note that a Bayes audit *may* also be "risk-limiting" in the standard sense [9]; we leave this as a fascinating open question (Conjecture 1); perhaps the most interesting open question in this paper.

A Bayes audit is very nicely sensitive to any regularities in the errors seen. For example, if votes for Alice and for Bob are always switched, and votes for Charlie are frequently scanned as votes for David, but not the reverse, then our error model will take such regularities into account. This is not easy to do with many other auditing methods.

A nice feature of a Bayes audit is that you always have a (Bayesian) estimate of the likelihood that there will be an upset. Thus, you can decide whether to switch over to a more efficient method for counting *all* of the ballots, rather than the continuing with the single-ballot audit, if the upset probability is large. You also have an estimate as to how things look, if you have to stop auditing earlier than one might like (e.g. for budgetary reasons).

You also have at each stage a ("subjective") estimate of the probability that each candidate would win if the auditing "continues in more the less the way it has been going so far." This might be useful in situations where the auditing budget is limited, but candidates could pay out of pocket for additional auditing.

Can this work be extended from single-ballot auditing to auditing by precincts? The handling of such "stratification" is an interesting open problem.

It would also be of interest to extend the Bayesian approach to elections with many more ballot types (e.g. ranked-choice ballots on 20 candidates).

It also seems possible (although computationally a bit expensive) to extend this approach to be able to estimate the amount of additional auditing to be expected (details omitted here).

6 Conclusions

We have presented Bayesian methods for conducting single-ballot post-election audits. These methods are attractive for the following reasons:

- A Bayes audit appears to have very well-controlled and small miscertification rates.
- The efficiency of Bayes audits appears to be extremely good (in the sense that very few ballots need to be examined by hand during a Bayes audit).
- A Bayes audit is very simple in structure, and is easily implemented.
- A Bayes audit is applicable to both comparison audits and to ballot-polling audits.
- A Bayes audit is based on well-understood technology (Bayesian statistics, Dirichlet distributions, etc.).
- A Bayes audit does not require the computation of a margin of victory in order to get started.
- A Bayes audit is very flexible, and can be applied to any voting system, as long as the number of ballot types is not too large.
- A Bayes audit returns meaningful results—the estimated probabilities that each possible outcome is in fact the correct outcome—even when the audit is stopped early.
- A Bayes audit can accommodate multiple auditors, each with his own prior.

On the other hand, the Bayes audit method has the following possible shortcomings:

- It is not (yet) known how to apply Bayesian techniques to auditing "by precincts" (non-single-ballot auditing); single-ballot auditing is required.
- The relationship of Bayes audits to risk-limiting audits is unclear. While we conjecture that Bayes audits are in fact risk-limiting, this is only a conjecture at this point, although our experimental evidence supports this conjecture.
- As with all Bayesian methods, the results depend to some extent on the choice of the prior.
- The computations required during the audit are sufficiently extensive that they can not reasonably be done by hand; a computer program is required during the audit to compute the "winning probabilities."
- While the average or median number of ballots examined may be quite low, the variance may be fairly high.
- While Bayesian methods are frequently used and well understood in other fields, their use in post-election audits is still nascent, and better-tested methods (such as risk-limiting post-election audits) are available.
- A number of states are beginning to draft legislation that specifies certain forms of post-election audits (such as risk-limiting audits); a Bayes audit may not qualify.

Further study is needed to clarify the comparative advantages and disadvantages of Bayes audits relative to other methods, such as risk-limiting audits, but we suspect that for practical purposes the differences between Bayes audits and risk-limiting audits may not be large; perhaps Bayes audits will turn out to be slightly more efficient.

For election officials trying to figure out which auditing method to use, the most important question is almost certainly "Are you auditing your elections?" and not "Which auditing method are you using?"

7 Acknowledgments

This work partially supported by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370. Additional support provided by Andrew and Erna Viterbi.

Thanks also to Lirong Xia, Maya R. Gupta, Glen Meeden, and Philip Stark for helpful suggestions.

References

[1] James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, 2010.

[2] S. Checkoway, A. Sarwate, and H. Shacham. Single-ballot risk-limiting audits using convex optimization. In D. Jones, J.-J. Quisquater, and E. Rescorla, editors, *Proceedings 2010 EVT/WOTE Conference*. USENIX/ACCURATE/IAVoSS, August 2010.

[3] Jayanta K. Ghosh, Mohan Delampady, and Tapas Samanta. *An Introduction to Bayesian Analysis: Theory and Methods*. Springer, 2010.

[4] M. Ghosh and G. Meeden. *Bayesian Methods for Finite Population Sampling*. Chapman, 1997.

[5] Marian Grendar and Robert K. Niven. The Pólya Urn: Limit theorems, Pólya divergence, maximum entropy and maximum probability. arXiv:cond-mat/0612697v1, December 29, 2006.

[6] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Discrete Multivariate Distributions*. Wiley, 1997.

[7] Norman L. Johnson and Samuel Kotz. *Urn Models and Their Application*. Wiley, 1977.

[8] Radu Lazar, Glen Meeden, and David Nelson. A noninformative Bayesian approach to finite population sampling using auxiliary variables. *Survey Methodology*, 34(1):51–64, June 2008.

[9] Mark Lindeman and Philip B. Stark. A gentle introduction to risk-limiting audits, 2012. To appear in *IEEE Computing Now*. Preprint: `http://statistics.berkeley.edu/~stark/Preprints/gentle12.pdf`.

[10] Mark Lindeman, Philip B. Stark, and Vincent S. Yates. BRAVO: Ballot-polling risk-limiting audits to verify outcomes. In Alex Halderman and Olivier Pereira, editors, *Proceedings 2012 EVT/WOTE Conference*, 2012.

[11] Albert Y. Lo. Bayesian statistical inference for sampling a finite population. *Annals of Statistics*, 14(3):1226–1233, 1986.

[12] Thomas R. Magrino, Ronald L. Rivest, Emily Shen, and David Wagner. Computing the margin of victory in IRV elections. In Hovav Shacham and Vanessa Teague, editors, *Proceedings 2011 EVT/WOTE Conference*, 2011.

[13] Glen Meeden. A Bayesian solution for a statistical auditing problem. *J. Amer. Statist. Assn.*, 98:735–740, July 2003.

[14] Glen Meeden and Dawn Sargeant. Some Bayesian methods for two auditing problems. *Communications in Statistics: Theory and Methods*, 36(15):2741–2760, 2007.

[15] Lawrence Norden, Aaron Burstein, Joseph Lorenzo Hall, and Margaret Chen. Post-election audits: Restoring trust in elections. Technical report, Brennan Center for Justice and Samuelson Law, Technology & Public Policy Clinic, 2007.

[16] Carl-Erik Särndal. A unified derivation of some nonparametric distributions. *J. Amer. Statist. Assn.*, 59(308):1042–1053, Dec. 1964.

[17] Carl-Erik Särndal. Derivation of a class of frequency distributions via Bayes's theorem. *J. Royal Statist. Soc., Ser. B*, 27(2):290–300, 1965.

[18] P. B. Stark. Risk-limiting vote-tabulation audits: The importance of cluster size. *Chance*, 23(3):9–12, 2010.

[19] Philip Stark. Papers, talks, video, legislation, software, and other documents on voting and election auditing. `http://statistics.berkeley.edu/~stark/Vote/index.htm`.

[20] Philip B. Stark. Personal communication, May 9, 2012.

[21] Philip B. Stark. Philip stark: Report on second risk-limiting audit under ab 2023 in monterey county california. `http://blog.verifiedvoting.org/2011/05/07/1370`.

[22] Philip B. Stark. CAST: Canvass audits by sampling and testing. *IEEE Trans. Inform. Forensics and Security*, 4(4):708–717, Dec. 2009.

[23] Philip B. Stark. Super-simple simultaneous single-ballot risk-limiting audits. In *Proc. 2010 EVT/WOTE Workshop*, 2010. `http://www.usenix.org/events/evtwote10/tech/full_papers/Stark.pdf`.

[24] Verified Voting. `http://www.verifiedvoting.org/article.php?id=5816`.

[25] Lirong Xia. Computing the margin of victory for various voting rules. In *Proceedings of the 13th ACM Conference on Electronic Commerce (EC-12)*, 2012.

Notes

[1] `http://statistics.berkeley.edu/~stark/Vote/auditTools.htm`

[2] Dan Wallach, in "Some Texas counties are clinging to the chad," Dallas Morning News, March 8, 2004

[3] Asking the Right Questions about Electronic Voting (CSTB, National Academies Press, 2005), Section 2.2 Legitimacy in a Democracy.

A Appendix. Sampling the Dirichlet and Pólya Distributions

We give two methods for sampling from the posterior distribution; the first samples from the (scaled and shifted) Dirichlet distribution:

$$\pi_s = \boldsymbol{C} + \mathrm{Dir}(\boldsymbol{\alpha} + \boldsymbol{C}, n-s) \tag{7}$$

given a tally $\boldsymbol{C} = (C_1, C_2, \ldots, C_t)$ for a sample C of size s from a population of size n, where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_t)$ is the hyperparameter vector for the prior distribution π_0; the second method samples from the (equivalent for our purposes) Pólya-Eggenberger distribution.

A.1 Sampling with Gamma variates.

This is the most efficient procedure, and is recommended. It is slightly less exact than the Pólya's Urn method, as it yields simulated tallies that sum to n but which need not be integers. We feel this difference is inconsequential.

Let Y_k have distribution $\mathrm{Gamma}(\alpha_t, 1)$ for $k = 1, 2, \ldots, t$, and let $Z_k = Y_k/Y$ for $k = 1, 2, \ldots, t$, where $Y = \sum_{k=1}^{t} Y_k$. Let

$$\begin{aligned} B_k &= C_k + (n-s)Z_k \\ \boldsymbol{B}_s &= (B_1, B_2, \ldots, B_t)\ . \end{aligned}$$

Then $\boldsymbol{B}_s$ has the desired distribution (7). Wikipedia[4] gives Python code for generating a Dirichlet distribution this way in time $O(t)$.

A.2 Simulation with Pólya's Urn.

Urn models are well studied; see for example Johnson et al. [7], Grendar et al. [5], or Johnson et al. [6].

A Bayes audit may use the lovely simulation procedure provided by sampling from Pólya's Urn for exact posterior distribution sampling. The simulated tally counts will be integers and the distribution is exactly the desired posterior. The method is efficient, but not as efficient as the Gamma method—the running time is $O(n)$ rather than $O(t)$. The distribution this procedure produces, aside from having support on sequences of integers rather than sequences of reals, is essentially indistinguishable from the corresponding Dirichlet distribution with the same hyperparameters (see references from [7, p. 196]).

A Pólya-Eggenberger distribution $PE(\boldsymbol{\alpha}, n)$ is a distribution on $[0..n]^t$ determined by the following procedure. (This is the most basic PE model.) Here $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_t)$ is a sequence of t positive integers whose sum is α.

- An urn is initialized to contain α_k balls of color k, for $k = 1, 2, \ldots, t$.
- For n times in a row, a ball is drawn at random from the urn, then returned to the urn, along with a second ball of the same color.
- Output $\boldsymbol{X} = (X_1, X_2, \ldots, X_t)$, where X_k is the number of balls drawn of color k; the elements of $\boldsymbol{X}$ sum to n.

This randomized procedure defines a probability distribution on $[0..n]^t$; so $PE(\boldsymbol{\alpha}, n)(\boldsymbol{X})$ denotes the probability that the above PE procedure yields $\boldsymbol{X}$.

Here is an equivalent description. Place in a row an initialization sequence of α balls, with α_k balls of each color k. Next, for n times, place a new ball at the right end of the row, where the new ball has the same color as a randomly chosen ball to its left.

For example, if we have $t = 3$ colors A, B, C , with $\boldsymbol{\alpha} = \mathbf{1}_3 = (1, 1, 1)$, we might obtain for $n = 6$

$$A\ B\ C \mid B\ A\ A\ C\ B\ A \tag{8}$$

where "|" separates the initialization sequence of length $a = 3$ from the $n = 6$ subsequent draws. We obtain color counts $\boldsymbol{X} = (3, 2, 1)$.

This procedure has been thoroughly analyzed, and the PE distributions are well understood [7, Ch. 4].

The most important (and surprising!) property of the PE urn model is *exchangeability*: the probability of drawing a given sequence of colors depends only on the *number* drawn of each color, and not on the *order* in which they are drawn—all permutations of a given sequence are equally likely [7, p. 97]. This is true even though the draws are not independent.

The probability of a given distribution $\boldsymbol{X}$ of colors drawn is:

$$PE(\boldsymbol{\alpha}, n)(\boldsymbol{X}) = \frac{n!}{\prod_{k=1}^{t} X_i!} \cdot \frac{\prod_{k=1}^{t} \alpha_k^{[X_k]}}{\alpha^{[n]}} \tag{9}$$

where

$$x^{[y]} = x \cdot (x+1) \cdots (x+y-1)$$

is the "rising factorial" notation [7, p. 195].

An important special case is when $\boldsymbol{\alpha} = \mathbf{1}_t = (1, 1, \ldots, 1)$; then we obtain a uniform distribution over all possible sequences $\boldsymbol{X}$ of t nonnegative integers summing to n.

The expected number $E(X_k)$ of balls drawn of color k is $(\alpha_k/\alpha) \cdot n$. As the α_k's increase, the X_k concentrate around their expected values.

Looking now to our inference problem, we see that there is a duality between Pólya's urn procedure, which can grow an urn of initial size s to one of size n, and a sampling procedure that can produce a sample of size s from a population of size n.

Suppose from an urn of n balls with color counts $\boldsymbol{n} = (n_1, n_2, \ldots, n_t)$ we draw without replacement a sample $\boldsymbol{S}$ of size $s < n$: $\boldsymbol{S} = (x_1, x_2, \ldots, x_s)$. Let $\boldsymbol{s} = \text{tally}(S) = (s_1, s_2, \ldots, s_t)$.

We are now interested in the "inverse of sampling"—inferring what we can about $\boldsymbol{n}$ from $\boldsymbol{s}$. Of course, there is uncertainty, so it is natural (for a Bayesian!) to represent this uncertainty as a (subjective) probability distribution for $\boldsymbol{n}$. Särndal [16, 17] studies this inference question, but without reference to the Pólya-Eggenberger distributions.

Let $PE(\boldsymbol{\alpha}, n|\boldsymbol{s})$ denote the probability distribution $PE(\boldsymbol{\alpha}, n)$ conditioned on having the first s outputs have tally $\boldsymbol{s}$.

Theorem 1 *Given the above sampling process resulting in a sample with tally $\boldsymbol{s}$, and a prior distribution $PE(\boldsymbol{\alpha}, n)$, the posterior distribution on the original tally $\boldsymbol{n}$ is $PE(\boldsymbol{\alpha}, n|\boldsymbol{s})$.*

That is: drawing a sample from the original urn (discarding $n - s$ randomly drawn balls one by one until only s are left) is nicely "inverted" (in a Bayesian sense) by the following Pólya urn process (adding balls one by one until you have added $n - s$ balls):

1. start with α balls, with α_k of each color k,

$AB \mid BB \mid AA$	$(2A, 2B)$	$(1/4)\cdot(2/5) = 1/10$
$AB \mid BB \mid AB$	$(1A, 3B)$	$(1/4)\cdot(3/5) = 3/20$
$AB \mid BB \mid BA$	$(1A, 3B)$	$(3/4)\cdot(1/5) = 3/20$
$\underbrace{AB}_{t} \mid \underbrace{\underbrace{BB}_{s} \mid \underbrace{BB}_{n-s}}_{n}$	$(0A, 4B)$	$(3/4)\cdot(4/5) = 6/10$

Figure 3: $s = 2$ balls are drawn without replacement from an urn of $n = 4$ balls of $t = 2$ possible colors (A,B). Both balls drawn are color B. The PE process gives probability $1/10$ to the original tally (2A,2B), $3/10$ to tally $(1A, 3B)$ (arising in two ways), and $6/10$ to $(0A, 4B)$. The table shows the initial sequence A B, the sample B B, the remaining two balls, and the probability that the PE process selects the two remaining balls to be as shown.

2. add the s balls $(x_1, x_2, \ldots, x_s)$ from the sample,
3. add $n-s$ more balls $x_{s+1}, x_{s+2}, \ldots, x_n$, each of which has the color of a randomly selected ball already in the urn at that time.

The output is $\boldsymbol{X} = \text{tally}((x_1, x_2, \ldots, x_n))$, which has distribution $PE(\boldsymbol{\alpha}, n|\boldsymbol{s})$. This procedure is efficient, taking time $O(n)$.

Graphically, following the style of the example (8), and assuming $\boldsymbol{\alpha} = \mathbf{1}_t$:

$$\underbrace{\underbrace{1\ 2\ \cdots\ t}_{t} \mid \underbrace{x_1\ x_2\ \cdots\ x_s}_{s} \mid \underbrace{x_{s+1}\ x_{s+2} \cdots\ x_n}_{n-s}}_{n} \ ; \tag{10}$$

Proof: The exchangeability property yields the proof: since any permutation of a sequence $(x_1, x_2, \ldots, x_n)$ is equally likely according to prior $PE(\boldsymbol{\alpha}, n)$, considering sequences starting with the desired sample yields the desired conditioning. ■

See Figure 3 for an example.

The key point is that it is easy to sample from the distribution $PE(\boldsymbol{\alpha}, n|\boldsymbol{s})$ by sampling the expression

$$\boldsymbol{s} + PE(\boldsymbol{\alpha} + \boldsymbol{s}, n - s)\ .$$

The sample can be viewed as "part of the prior" for the remaining $n - s$ elements, but the sample tally is also added in as part of the final tally, since we want a distribution on length-n sequence tallies and not on length-$(n-s)$ sequence tallies.