Election audits by sampling with probability proportional to an error bound: dealing with discrepancies

Philip B. Stark
Department of Statistics
University of California
Berkeley, CA 94720-3860

DRAFT 20 February 2008, 1:10pm

Abstract

Sampling items using probability proportional to a bound on the possible error in the item (PPEB) has a long history in financial auditing, but has only recently been suggested for auditing elections. How large a PPEB sample should be drawn to have confidence $1 - \alpha$ that the election outcome is correct if the sample includes only "small" errors? What is the confidence that the outcome of the election is correct, given the discrepancies a PPEB audit uncovers, whatever their size? If one wants to end up with confidence level $1 - \alpha$, how should one increase the PPEB sample size if discrepancies are found, and when can one stop auditing? This paper develops a simple way to answer these questions, and shows how techniques from financial auditing can also be adapted to answer these questions. It develops a conservative sequential test of the hypothesis that the outcome of the election is incorrect based on the vote-counting errors found in a hand tally of PPEB sample of precincts. It also shows how to use PPEB with stratification.

Keywords: hypothesis test, sequential test, auditing, elections, probability proportional to size, monetary unit sampling.

Acknowledgments. I am grateful to Mark Lindeman and Luke Miratrix for helpful comments.

1 Introduction

Advantages of drawing financial audit samples using statistical methods have been known for more than fifty years (e.g., [31, 29, 9]). It has been known for almost as long that standard statistical techniques for analyzing accounting errors can be grossly inaccurate, because populations of accounting errors tend to be mostly zero, with rare large values. See, e.g., Stringer [26] and references in [19].

Populations of vote-counting errors are similar: typically, precinct-level discrepancies are at most a few votes, but fraud, bugs and other gross errors such as miscalibrated optical scanners or misplaced boxes of ballots can produce discrepancies of thousands of votes. This is essentially the “black swan” problem (Taleb [27]): rare large values are unlikely to turn up in samples of moderate size, and the normal approximation can grossly underestimate the incidence of extreme values in the population—and hence their effect on the population total.

In both financial and electoral audits, a key question is whether the total error is *material*. The notion of materiality in financial auditing is slippery. Error in a financial report is material if a reasonable person relying on the report would have acted differently but for the error. In electoral auditing, materiality is more straightforward: error is material if it changed the apparent outcome of the contest by making the true winning candidate or position appear to lose.

Several classes of methods have been developed to deal with statistical audit data. The most basic uses *attributes*: does an item in the sample have error or not? Election auditing methods such as those of Dopp and Stenger, McCarthy et al., Rivest and Saltman [7, 16, 21, 22] are based on attributes. These methods do not make use of the magnitude of any error that is found. In essence, they ask, "If the total error is large enough to affect the outcome, what is the chance that a sample of a given size will find any error at all?" These methods use a simple random sample (SRS) of precincts: every set of n of the N precincts in the race is equally likely to be selected for audit. As a consequence, every precinct has the same chance of being audited.

For determining whether the value of a financial account is materially overstated, it is common to sample items with probability proportional to their book value, rather than with equal probability. This is called *monetary* *unit sampling* (MUS) or *dollar unit sampling*. See, e.g., [1, 3, 4, 8, 26].[1] MUS is an example of sampling with probability proportional to size (PPS) Its use in accounting appears to have originated with Stringer [26].[2] In inferences about the total overstatement, it is common to assume that the overstatement of each item is between zero (i.e., the book value of the item is correct) and the book value of the item (i.e., the true value of the item is zero). The book value is then an upper bound on the overstatement error. MUS thus samples each item with probability proportional to an error bound for the item (PPEB).

Sampling precincts with probability proportional to a bound on the error in the precinct vote count (such as the number of ballots cast) has been proposed for election audits [12, 2, 24], but to my knowledge no jurisdiction currently allows PPEB sampling for post-election audits. Election integrity activists are drafting sample audit bills for several states, and sanctioning PPEB is appealing because, for a given level of confidence, it could lead to lower workloads than simple or stratified random sampling [2]. If PPEB audits are permitted, the research on MUS in financial auditing becomes relevant for election auditing—which contains methods that can be re-purposed to calculate the confidence that a full recount would not find a different win-

[1] Some methods rely on parametric approximations [14, 28], Bayesian prior distributions [6, 10, 15, 23, 30] or numerical simulations [17, 5, 11]; the reliability of such methods rests on assumptions that are largely untestable.

[2] The Stringer Bound is an example of a method based on *combined attributes and values*. It and other combined attributes and values methods are typically applied to systematic random samples with probability proportional to size, but they are analyzed as if the samples were random samples with replacement.

ner, given the discrepancies an audit finds.

Despite the efficiency of PPEB, there are reasonable arguments against using it for election audits. For example, PPEB is more complex and more difficult for the public and jurisdictional users to understand. Drawing a PPEB sample requires a first step of computing error bounds. Using dice rolls or other physical sources of randomness to draw a PPEB sample is more complicated than it is for SRS. These factors could decrease the transparency of election audits, decreasing their public value. There are legal issues as well. If precincts are sampled using simple random sampling or stratified random sampling with equal sampling fractions in all strata, every ballot is equally likely to be audited. That is not the case for PPEB audits: ballots cast in precincts with large error bounds are more likely to be audited. This could raise questions of equal protection and differing chances of disenfranchisement in different precincts.

Aslam et al. [2] propose drawing samples of precincts for post election audits using probability proportional to a bound on the error in each precinct, with replacement (PPEBWR). They calculate the minimum sample size required for such a sample to have chance at least $1 - \alpha$ of finding one or more precincts with a discrepancy if a full manual count would show a different winner than the preliminary count did. Their calculations are implicit in the work of Stringer [26] for financial auditing; see also Kaplan [13] and section 8 below for a discussion of the connections. Aslam et al. [2] show that PPEBWR sampling can reduce the workload compared with simple random sampling when precincts have varying error bounds. They determine how little PPEBWR sampling is too little. For PPEBWR samples smaller than the minimum they prescribe, the confidence in the outcome will be low even if the audit finds no discrepancies. In practice, auditing a reasonable number of precincts will find one or more precincts with discrepancies. The question elections officials need to answer is: "how much sampling is enough to give high confidence that the apparent outcome is correct, given the discrepancies the audit uncovers?" That will be more sampling than the minimum [2] prescribe, unless the audit finds no discrepancies.

Stark [25] shows how to deal with discrepancies that an audit uncovers, to end up with any desired confidence that a full recount would not show a different winner from the preliminary count, if precincts are selected using simple or stratified random sampling. Stark argues that to be complete, an audit procedure should always either (i) certify the election outcome, or (ii) demand a full recount. The procedure should have an error rate that can be quantified, for example, a guarantee that if the outcome is certified, either the outcome was right, or an event with probability no greater than α occurred. PPEBWR as proposed by Aslam et al. [2] has these properties only if a full manual count is conducted whenever the audit turns up any discrepancy whatsoever. This paper makes PPEBWR more useful by allowing one to certify even if some discrepancies are found, provided the sample size is large enough and the discrepancies that could have made the margin appear artificially large are small enough. It answers the question "what size sample can you stop with?" rather than "what size sample should you start with?" It develops a simple test using PPEB samples, and shows how tools from financial auditing give more powerful—though more complex—tests .

This paper also extends PPEB to allow stratification. The ability to use stratified samples is important for several reasons: In many states, counties independently draw precincts to audit, resulting in a stratified sample of precincts for races that cross county lines. Then the strata are counties. Even for races entirely contained in one county, it could be useful to stratify on a variety of variables, such as the type of technology used in the preliminary tally. Moreover, to implement audits efficiently, it can be useful to draw a sample of ballots cast in precincts on election day, then to draw an independent sample of provisional ballots and ballots cast by mail (VBM ballots). That can allow an audit to start while VBM and provisional ballots are still being counted. This was the approach used in the first election audit performed to attain a target level of confidence, the audit of Measure A in Marin County, California, after the 5 February 2008 election (Stark [24]).[3]

Section 2 sets out the notation. Section 3 explains how Aslam et al. [2] apply PPEBWR to the problem of detecting discrepancy, if the aggregate discrepancy is enough to eliminate the apparent margin. It draws connections between PPEBWR and SRS-based methods for detection, and how they relate to the method of Stark [25]. Section 4 explains how PPEBWR can be modified to provide a conservative test when there is a nonzero background rate of discrepancy. Section 5 explains how to use the maximum potential margin overstatement observed in a PPEBWR sample to find a P-value

[3] Measure A was audited to attain 75% confidence that the outcome was correct. The audit was performed in two stages. First, a simple random sample of 6 precincts was drawn; the votes cast in those precincts on election day were tallied by hand. Once the VBM and provisional ballots were counted, an independent simple random sample of 6 precincts was drawn; the (valid) provisional and VBM ballots from those precincts were tallied by hand.

for the hypothesis that the aggregate overstatement could have produced the apparent margin. Section 6 wraps the P-value calculation in a sequential test, following Stark [25]. Section 7 extends PPEB sampling to allow stratification. Section 8 makes connections to MUS audit sampling in finance and explains how several tools from financial auditing can be used to test whether the apparent outcome of an election is the same that a full manual recount would show—if PPEB sampling is sanctioned. Sections 9 and 10 present discussion and conclusions.

2 Assumptions and notation

We consider one contest at a time. There are N precincts in the contest. Each voter can vote for up to f of K candidates; there can be f winners. (Under some circumstances, it can help to "pool" the vote counts for some of the losing candidates, as discussed in Stark [25]; K is the number of candidates after any pooling.) The reported vote for candidate k in precinct p is v_{kp}. The vote that an audit would show for candidate k in precinct p is a_{kp}. The total reported vote for candidate k is $V_k = \sum_{p=1}^N v_{kp}$. The total true vote for candidate k is $A_k = \sum_{p=1}^N a_{kp}$. The indices of the set of apparent winners is $\mathcal{K}_w \subset \mathcal{K}$, where $\#\mathcal{K}_w = f$. The indices of the set of apparent losers is $\mathcal{K}_\ell \subset \mathcal{K}$. The apparent margin M is the total number of votes reported for the apparent winner with the fewest votes, minus the total number of votes reported for the apparent loser with the most votes:

$$M \equiv \wedge_{k \in \mathcal{K}_w} V_k - \vee_{k \in \mathcal{K}_\ell} V_k. \qquad (1)$$

The potential margin overstatement discrepancy in the vote in precinct p is

$$e_p \equiv \sum_{k \in \mathcal{K}_w} (v_{kp} - a_{kp})_+ + \sum_{k \in \mathcal{K}_\ell} (a_{kp} - v_{kp})_+ \tag{2}$$

Let $e \equiv (e_p)_{p=1}^N$. For $\mathcal{I} \subset \mathcal{N} \equiv \{1, 2, \ldots, N\}$ and $x \in \mathrm{I\!R}^N$, define

$$\sum_{\mathcal{I}} x \equiv \sum_{p \in \mathcal{I}} x_p, \tag{3}$$

$$\vee_{\mathcal{I}} x \equiv \max_{p \in \mathcal{I}} x_p, \tag{4}$$

and

$$\wedge_{\mathcal{I}} x \equiv \min_{p \in \mathcal{I}} x_p. \tag{5}$$

The total discrepancy is $E = \sum_{\mathcal{N}} e$.

3 The logic behind current election audit sampling procedures

The current crop of election auditing methods are based on the same underlying reduction: A necessary condition for the preliminary outcome to differ from the outcome a full manual count would show is that $E = \sum_{\mathcal{N}} e \geq M$; see Stark [25].

Since precincts for which $e_p = 0$ contribute nothing to $\sum_{\mathcal{N}} e$, if $\sum_{\mathcal{N}} e \geq M$, there must be some set $\mathcal{T}$ of precincts for which

$$\sum_{\mathcal{T}} e \geq M \text{ and } e_p > 0, \forall p \in \mathcal{T}. \tag{6}$$

If no such $\mathcal{T}$ exists, the preliminary outcome of the election must be the same that a full hand count would show. Audits try to find strong statistical evidence that no such $\mathcal{T}$ exists, which gives high confidence that the preliminary outcome is correct.

To do so, the auditing methods require knowing a vector u such that $e \le u$: u is a set of upper bounds on the discrepancy, precinct by precinct. Stark [25] derives u_p from an upper bound r_p on the number of valid votes in precinct p and $\{v_{kp}\}_{k\in\mathcal{K}}$, the reported votes for each candidate in precinct p:

$$u_p = r_p + \sum_{k\in\mathcal{K}_w} v_{kp} - \wedge_{k\in\mathcal{K}_\ell} v_{kp}. \tag{7}$$

A value for r_p might in turn come from voter registrations, pollbooks, or an accounting of paper ballots ("ballot reconciliation"). Aslam et al. [2] propose a bound that assumes there are only two candidates and that $\sum_{k=1}^K a_{kp} = \sum_{k=1}^K v_{kp}$, which need not be the case.[4] Some studies advocate taking $u = 0.4b$ [22, 21, 16, 7], which is, at best, ad hoc.

For any $\mathcal{I} \subset \mathcal{N}$, $\sum_{\mathcal{I}} e \le \sum_{\mathcal{I}} u$. If $M > \sum_{\mathcal{N}} u$, the apparent outcome of the election must agree with what a full manual tally would show—an audit is not required to confirm the outcome. Henceforth, we assume that $M \le \sum_{\mathcal{N}} u$. It is impossible that $\sum_{\mathcal{T}} e \ge M$ unless $\sum_{\mathcal{T}} u \ge M$. This observation is key to current statistical election auditing methods.

PPEBWR and methods based on simple random samples part ways here. PPEBWR uses the fact that $\sum_{\mathcal{T}} u \ge M$ directly: Let $\lambda_{p;u} \equiv u_p / \sum_{\mathcal{N}} u$, $p \in \mathcal{N}$. The vector $\lambda = (\lambda_{p;u})_{p=1}^N$ is a probability vector: $\wedge_{\mathcal{N}} \lambda \ge 0$ and $\sum_{\mathcal{N}} \lambda = 1$. If we draw precinct p with probability $\lambda_{p;u}$ then the probability of getting an element of $\mathcal{T}$ is

$$\begin{aligned}\mathbb{P}\{\text{ draw an element of } \mathcal{T}\ \} &= \sum_{p\in\mathcal{T}} \mathbb{P}\{\text{ draw precinct } p\ \} \\ &= \sum_{p\in\mathcal{T}} \lambda_{p;u}\end{aligned}$$

[4] For example, in the 2006 U.S. Senate race in Minnesota, that assumption was violated.

$$\begin{aligned} &= \sum_{\mathcal{T}} \lambda \\ &= \sum_{\mathcal{T}} u / \sum_{\mathcal{N}} u \\ &\geq M / \sum_{\mathcal{N}} u. \end{aligned} \tag{8}$$

The chance that in n independent draws, each using probability vector λ, no element of $\mathcal{T}$ is drawn is

$$(1 - \sum_{\mathcal{T}} u / \sum_{\mathcal{N}} u)^n \leq (1 - M / \sum_{\mathcal{N}} u)^n. \tag{9}$$

So the chance that n independent draws using λ yield at least one element of $\mathcal{T}$ is

$$1 - (1 - \sum_{\mathcal{T}} u / \sum_{\mathcal{N}} u)^n \geq 1 - (1 - M / \sum_{\mathcal{N}} u)^n. \tag{10}$$

Hence if we set

$$n_{PPS}(\alpha; u, M) = \left\lceil \frac{\ln(\alpha)}{\ln(1 - M / \sum_{\mathcal{N}} u)} \right\rceil, \tag{11}$$

the chance that $n_{PPS}(\alpha; u, M)$ independent draws from $\mathcal{N}$ using probabilities λ yield at least one element of $\mathcal{T}$ is at least $1 - \alpha$—if $\mathcal{T}$ exists.

Let $\mathcal{J}_n^\lambda$ denote the set of values that result from n independent draws from $\mathcal{N}$ according to the probability vector $\lambda = (u_p / \sum_{\mathcal{N}} u)_{p=1}^N$. The set $\mathcal{J}_n^\lambda$ is a subset of $\mathcal{N}$ with at most n elements. We have just shown that

$$\mathbb{P}\{\mathcal{J}_{n_{PPS}(\alpha;u,M)}^\lambda \cap \mathcal{T} \neq \emptyset\} \leq \alpha \tag{12}$$

if $\sum_{\mathcal{T}}(e) \geq M$.

When we look at the precincts in the sample, we can tell whether p *could be* an element of $\mathcal{T}$: if we see that $e_p = 0$, then $p \notin \mathcal{T}$, since every $p \in \mathcal{T}$ has $e_p > 0$. Hence if $e_p = 0$ for every p in the sample, no element of the sample could be a member of $\mathcal{T}$, and we have statistical evidence that $\mathcal{T}$ does not exist. The PPEBWR method rides on the fact that we have a lower bound on the probability $\sum_{\mathcal{T}} \lambda$ that each draw gives an element of $\mathcal{T}$: for any $\mathcal{T}$ for which $\sum_{\mathcal{T}} e$ could equal or exceed M, $\sum_{\mathcal{T}} \lambda \geq M / \sum_{\mathcal{N}} u$.[5]

Note that $n_{PPS}(\alpha)$ is the number of draws with replacement, not the attained sample size, because the same precinct p could be drawn more than once. Let I_p be the indicator of the event that precinct p is selected:

$$I_p \equiv \begin{cases} 1, & \mathcal{J}^{\lambda}_{n_{PPS}(\alpha;u,M)} \ni p \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

The attained sample size is $\sum_{p \in \mathcal{N}} I_p$. The expected sample size is

$$\begin{aligned} \mathbb{E} \sum_{p \in \mathcal{N}} I_p &= \sum_{p \in \mathcal{N}} \mathbb{P}\{I_p = 1\} \\ &= \sum_{p \in \mathcal{N}} (1 - (1 - u_p / \sum_{\mathcal{N}} u)^{n_{PPS}}) \\ &= N - \sum_{p \in \mathcal{N}} (1 - u_p / \sum_{\mathcal{N}} u)^{n_{PPS}}. \end{aligned} \tag{14}$$

The expected sample size is maximal when the elements of u are equal; the more the elements of u differ, the smaller the attained sample size is expected to be.

Methods based on simple random samples use the fact that $\sum_{\mathcal{T}} u \geq M$ indirectly: they use u to find a lower bound on $\#\mathcal{T}$, the number of elements of $\mathcal{T}$. By starting with the largest element of u and working down, one can find the smallest number of elements of u required for the sum of those elements to wipe out the margin. Suppose that the minimum number is $k(M, u)$. We

[5] Other sampling probability vectors could also yield useful bounds; independent sampling using λ is particularly simple to analyze.

know that, if $\mathcal{T}$ exists, $\#\mathcal{T} \ge k(M,u)$. Hence, if $\mathcal{T}$ exists, the chance that a simple random sample of n elements of $\mathcal{N}$ will contain at least one element of $\mathcal{T}$ is

$$1 - \frac{\binom{N-\#\mathcal{T}}{n}}{\binom{N}{n}} \ge 1 - \frac{\binom{N-k(M,u)}{n}}{\binom{N}{n}}. \tag{15}$$

If we pick $n_{SRS}(\alpha; u, M)$ so that

$$\frac{\binom{N-k(M,u)}{n_{SRS}(\alpha;u,M)}}{\binom{N}{n_{SRS}(\alpha;u,M)}} \le \alpha \tag{16}$$

then chance that a simple random sample of size $n_{SRS}(\alpha; u, M)$ from $\mathcal{N}$ contains at least one element of $\mathcal{T}$ is at least $1-\alpha$—if $\mathcal{T}$ exists. If p is selected, the audit can tell whether p *could be* an element of $\mathcal{T}$: if we see that $e_p = 0$, then $p \notin \mathcal{T}$, since every $p \in \mathcal{T}$ has $e_p > 0$. So if $e_p = 0$ for every p in the sample, we have statistical evidence that $\mathcal{T}$ does not exist.

The approach of Stark [25] is related through its indirect use of the fact that $\sum_{\mathcal{T}} u \ge M$. Let $\{w_p\}_{p\in\mathcal{N}}$ be a set of N monotonic functions. Let $w_p^{-1}(t) = \sup\{q \in \mathbb{R} : w_p(q) \le t\}$ and define $w^{-1}(t) = (w_p^{-1}(t))_{p=1}^N$. Define

$$\mathcal{T}_{w,t} \equiv \{p \in \mathcal{N} : w_p(e_p) > t\}. \tag{17}$$

(If $e_p > 0$ implies that $w_p(e_p) > t$, $p \in \mathcal{N}$, then $\mathcal{T}_{w,t} = \mathcal{T}$.) A necessary condition for $\sum_{\mathcal{T}} u \ge M$ is that $\sum_{\mathcal{T}_{w,t}} u \ge M - \sum_{\mathcal{N}} w^{-1}(t)$.

Instead of using u to find a lower bound on $\#\mathcal{T}$, Stark [25] uses u to find a lower bound on $\#\mathcal{T}_{w,t}$. Suppose that lower bound is $k(M,u,w,t)$. Then if $\mathcal{T}$ exists, the chance that a simple random sample of n elements of $\mathcal{N}$ will contain at least one element of $\mathcal{T}_{w,t}$ is

$$1 - \frac{\binom{N-\#\mathcal{T}_{w,t}}{n}}{\binom{N}{n}} \ge 1 - \frac{\binom{N-k(M,u,w,t)}{n}}{\binom{N}{n}}. \tag{18}$$

If we pick $n_{SRS}(\alpha; u, M, w, t)$ so that

$$\frac{\binom{N-k(M,u,w,t)}{n_{SRS}(\alpha;u,M,w,t)}}{\binom{N}{n_{SRS}(\alpha;u,M,w,t)}} \le \alpha \tag{19}$$

then the chance that a simple random sample of size $n_{SRS}(\alpha; u, M, w, t)$ from $\mathcal{N}$ includes at least one element of $\mathcal{T}_{w,t}$ is at least $1-\alpha$—if $\mathcal{T}$ exists. As before, if p is selected, the audit can tell whether p *could be* an element of $\mathcal{T}_{w,t}$: if we see that $w_p(e_p) \le t$, then $p \notin \mathcal{T}_{w,t}$. So if $w_p(e_p) \le t$ for every p in the sample, we have statistical evidence that $\mathcal{T}$ does not exist.

For all these methods, the requirement that $\sum_{\mathcal{T}} u \ge M$ is used to find a lower bound on the probability that the sample drawn in a particular way will turn up an element p with a recognizable property, if $\mathcal{T}$ exists. For most of the methods, the property is that $e_p > 0$, a property shared by all elements of $\mathcal{T}$. For the method of Stark [25], the property is that $w_p(e_p) > t$. The methods conclude that $\mathcal{T}$ does not exist—that $\sum_{\mathcal{N}} e < M$—if the sample contains no precinct with the property in question.

4 Extending PPEBWR to account for discrepancies

If $\mathcal{T}$ exists, $\sum_{\mathcal{T}_{w,t}} u - w^{-1}(t) \ge M - \sum_{\mathcal{N}} w^{-1}(t)$. If we draw precinct p from $\mathcal{N}$ with probability $\lambda_{p;u,w,t} = (u_p - w_p^{-1}(t))/\sum_{\mathcal{N}} u - w^{-1}(t)$, then

$$\begin{aligned} \mathbb{P}\{\text{ draw an element of } \mathcal{T}_{w,t} \} &= \sum_{p \in \mathcal{T}_{w,t}} \mathbb{P}\{\text{ draw precinct } p \} \\ &= \sum_{p \in \mathcal{T}_{w,t}} \frac{u_p - w_p^{-1}(t)}{\sum_{\mathcal{N}} u - w^{-1}(t)} \end{aligned}$$

$$
\begin{aligned}
&= \frac{\sum_{\mathcal{T}_{w,t}} u - w^{-1}(t)}{\sum_{\mathcal{N}} u - w^{-1}(t)} \\
&\geq \frac{M - \sum_{\mathcal{N}} w^{-1}(t)}{\sum_{\mathcal{N}} [u - w^{-1}(t)]}. \qquad (20)
\end{aligned}
$$

The chance we select at least one element of $\mathcal{T}_{w,t}$ in n independent draws, each using the same sampling probability vector $(\lambda_{p;u,w,t})_{p=1}^N$, is

$$
\begin{aligned}
& 1 - \left(1 - \textstyle\sum_{\mathcal{T}_{w,t}} [u - w^{-1}(t)] / \sum_{\mathcal{N}} [u - w^{-1}(t)]\right)^n \\
& \geq 1 - (1 - (M - \textstyle\sum_{\mathcal{N}} w^{-1}(t)) / \sum_{\mathcal{N}} [u - w^{-1}(t)])^n \quad . \qquad (21)
\end{aligned}
$$

Hence if we set

$$
n_{PPS}(\alpha; u, M, w, t) = \left\lceil \frac{\ln(\alpha)}{\ln(1 - (M - \sum_{\mathcal{N}} w^{-1}(t)) / \sum_{\mathcal{N}} [u - w^{-1}(t)])} \right\rceil, \qquad (22)
$$

the chance that $n_{PPS}(\alpha; u, M, w, t)$ independent draws from $\mathcal{N}$ using probabilities $(\lambda_{p;u,w,t})_{p=1}^N$ contains at least one element of $\mathcal{T}_{w,t}$ is at least $1 - \alpha$—if $\mathcal{T}$ exists. When we look at the precincts in the sample, we can tell whether p *could be* an element of $\mathcal{T}_{w,t}$: if we see that $w_p(e_p) < t$, then $p \notin \mathcal{T}_{w,t}$. So if $w_p(e_p) < t$ for every p in the sample, no element of the sample could be a member of $\mathcal{T}_{w,t}$, which is statistical evidence that $\mathcal{T}$ does not exist. Note that this approach essentially assumes that for every $p \in \mathcal{N}$, the discrepancy is at least $w_p^{-1}(t)$, even if the data show that some precincts have less error than that. The methods developed in financial auditing use more information about the observed distribution of $w_p(e_p)$ and can lead to less auditing, although the computations are more complex; see section 8.

As with PPEBWR, $n_{PPS}(\alpha; u, M, w, t)$ is the number of draws, not the attained sample size. The expected sample size is maximal when the elements of $u - w^{-1}(t)$ are equal; the more variable the elements of $u - w^{-1}(t)$ are, the smaller the attained sample size is expected to be.

5 P-values

The previous derivation shows how many times to draw to have a big chance ($> 1 - \alpha$) of finding at least one precinct p with $w_p(e_p) > t$ if the total discrepancy $E \equiv \sum_{\mathcal{N}} e \geq M$. In this section, we turn things around to ask what the chance is that $w_p(e_p) \leq t$ for every p in a PPS sample of n draws if $\sum_{\mathcal{N}} e \geq M$. The wrinkle is that we need to be able to choose t after the sample is drawn, so we need a sampling probability vector λ and a vector of functions w that let us bound the chance of finding no p with $w_p(e_p) \leq t$ for any choice of t, assuming $\mathcal{T}$ exists. The easiest way to accomplish this is to choose w so that

$$\lambda_{p;u,w,t} = \frac{u_p - w_p^{-1}(t)}{\sum_{\mathcal{N}}[u - w^{-1}(t)]} \tag{23}$$

does not depend on t. One simple choice of w that works is $w_p^*(q) = q/u_p$. That is, $w_p^*(e_p)$ scales the discrepancy e_p in precinct p to be a fraction of the error bound u_p for precinct p; thus $0 \leq t \leq 1$. This is related to the notion of *taint* in financial audits, discussed in section 8.

With this choice, $u_p - w_p^{*-1}(t) = (1 - t)u_p$, so

$$\begin{aligned} \lambda_{p;u,w^*,t} &= \frac{u_p - w_p^{*-1}(t)}{\sum_{\mathcal{N}}[u - w^{*-1}(t)]} \\ &= \frac{(1-t)u_p}{\sum_{\mathcal{N}}(1-t)u} \\ &= u_p / \sum_{\mathcal{N}} u \\ &= \lambda_{p;u}, \end{aligned} \tag{24}$$

the same sampling probability vector used in PPEBWR. If $\mathcal{T}$ exists, the chance a single PPEBWR draw will yield an element of $\mathcal{T}_{w^*,t}$ is

$$\begin{aligned}\sum_{p \in \mathcal{T}_{w,t}} \lambda_{p;u,w^*,t} &= \sum_{p \in \mathcal{T}_{w,t}} u_p / \sum_{\mathcal{N}} u \\ &\geq \frac{M - \sum_{\mathcal{N}} w^{*-1}(t)}{\sum_{\mathcal{N}} u} \\ &= \frac{M}{\sum_{\mathcal{N}} u} - t. \qquad (25)\end{aligned}$$

The chance n independent PPEB draws include no member of $\mathcal{T}_{w^*,t}$ is no greater than

$$P(t; u, M, w^*, n) = (1 - M/\sum_{\mathcal{N}} u + t)^n \qquad (26)$$

if $\mathcal{T}$ exists, for $t \in [0, 1]$. I.e., if $\sum_{\mathcal{T}} e \geq M$,

$$\mathbb{P}\{\mathcal{J}_n^\lambda \cap \mathcal{T}_{w^*,t} = \emptyset\} \leq (1 - M/\sum_{\mathcal{N}} u + t)^n. \qquad (27)$$

Hence $P(t; u, M, w^*, n)$ is a conservative P-value for the hypothesis that the apparent election outcome differs from the outcome a full manual count would show, if the largest observed value of $w_p^*(e_p) \equiv e_p/u_p$ in n PPEBWR draws is t.

This calculation uses only the maximum observed value of e_p/u_p. By using information about all of the nonzero observed values, one can get a more powerful test. See section 8.

In PPEB draws with replacement, the conditional chance of drawing an element of $\mathcal{T}$ is the same for every draw, while in PPEB draws without replacement, the conditional probability of drawing an element of $\mathcal{T}$ given that none has yet been drawn increases monotonically with each unsuccessful draw. Hence, if precincts are drawn by PPEB without replacement, updating λ before each draw to reflect only those elements of $\mathcal{N}$ not yet in the sample, the chance of drawing at least one element of $\mathcal{T}$ in n draws is increased.

6 A sequential test that the preliminary outcome is wrong

This P-value can be used in the sequential testing approach proposed by Stark [25] to determine whether, given the discrepancies observed in a PPEBWR sample, to certify the election outcome or to expand the sample. Here is the full procedure. Recall that $w_p^*(q) \equiv q/u_p$. Let $\mathcal{J}_n^\lambda$ denote the indices that result from n independent draws from $\mathcal{N}$ according to the probability vector $\lambda = (u_p / \sum_{\mathcal{N}} u)_{p=1}^N$. Recall that for any set $\mathcal{I} \subset \mathcal{N}$ and $x \in \mathbb{R}^N$, $\vee_{\mathcal{I}} x \equiv \max_{p \in \mathcal{I}} x_p$. Define $w^*(e) \equiv (w_p^*(e_p))_{p=1}^N$.

1. Select an overall significance level α and a sequence (α_s) so that sequential tests at significance levels $\alpha_1, \alpha_2, \ldots$, give an overall significance level no larger than α. For example, set $\alpha_s \equiv \alpha/2^s$, $s = 1, 2, \ldots$.

2. Group apparent losing candidates using the pooling rule given by Stark [25].

3. Compute error bounds u and the apparent margin M.

4. Select an initial sample size n_1 and a rule for selecting n_s when the hypothesis $E \ge M$ is not rejected at stage $s-1$. One can take $n_1 = n_{PPS}(\alpha_1; u, M)$ The only requirement is that $n_1 \ge 0$ and $n_s - n_{s-1} \ge 1$.

5. Set $s = 1$, $n_0 = 0$ and $\mathcal{J}_0 = \emptyset$.

6. Draw $n_s - n_{s-1}$ times independently from $\mathcal{N}$ using probability vector λ to form $\mathcal{J}_{n_s - n_{s-1}}^\lambda$. Set $\mathcal{J}_s = \mathcal{J}_{s-1} \cup \mathcal{J}_{n_s - n_{s-1}}^\lambda$.

7. Tally the votes in any precincts in $\mathcal{J}_s$ that were not in $\mathcal{J}_{s-1}$.

8. If $\mathcal{J}_s = \mathcal{N}$, the correct outcome is known: certify the election if the outcome was correct, and stop.

9. If there are still precincts not in the sample, calculate $t_s = \vee_{\mathcal{J}_s} w^*(e)$.

10. If $P(t_s; u, M, w^*, n_s) \le \alpha_s$, certify the election and stop. Otherwise, increment s and return to step 6.

7 Stratified PPEB sampling

Suppose the population of N precincts is divided into C strata, with N_c precincts in stratum c, $c \in \mathcal{C} = \{1, \ldots, C\}$. As mentioned in the introduction, it can be desirable to sample independently from each stratum.

More to come ...

8 Methods from financial auditing

This section shows how PPEB-based methods (MUS methods) for estimating financial overstatement error could be applied to election auditing, and could lead to sharper inferences at the cost of more complex and less transparent calculations.

The total overstatement of a ledger is analogous to the total vote discrepancy in an election. If a $1-\alpha$ upper confidence bound for the total vote discrepancy is less than M, there is $1-\alpha$ confidence that a full recount would show the same result as the preliminary count.

The first step in applying MUS methods to elections is to think of PPEB as sampling units of possible error, rather than sampling precincts or votes.

The discrepancy in precinct p is between 0 and u_p. We think of each of the u_p possible units of discrepancy in precinct p as having *taint* $t_p \equiv e_p/u_p \in [0, 1]$. The total discrepancy is

$$E = \sum_{\mathcal{N}} e = \sum_{p \in \mathcal{N}} t_p u_p. \tag{28}$$

Thus a necessary condition for the apparent outcome to differ from the outcome a full hand count would show is $\sum_{p\in\mathcal{N}} t_p u_p \geq M$. The *(weighted) mean taint* τ is

$$\tau = \frac{\sum_{\mathcal{N}} t_p u_p}{\sum_{\mathcal{N}} u} = \frac{E}{\sum_{\mathcal{N}} u}. \tag{29}$$

So the apparent outcome must be the same that a full hand count would show if $\tau < M/\sum_{\mathcal{N}} u$.

In MUS, we draw a random sample from $\{t_p\}_{p\in\mathcal{N}}$ such that the probability of selecting t_p is $u_p/\sum_{\mathcal{N}} u$. In practice, the draws are made without replacement, but MUS methods are typically analyzed as if the draws are with replacement; i.e., as a PPEBWR sample.

More to come . . .

8.1 The Stringer Bound and the Bickel Bound

Stringer [26] proposed an upper confidence bound for the weighted mean taint τ based on the number of nonzero observed taints and their values. Let T_j be the taint observed on the jth draw, $j = 1, \ldots, n$. Suppose that P of $\{T_j\}_{j=1}^n$ are strictly positive. Let (z_j) be the P nonzero observed taints in decreasing order, so that $1 \geq z_1 \geq \ldots \geq z_P > 0$, and define $z_0 \equiv 1$. Suppose $X \sim \mathit{Binomial}(n, \pi)$ Let $\pi_{\alpha,n}(j)$ denote the upper $1 - \alpha$ confidence bound for π when $X = j$; i.e., $\pi_{\alpha,n}(j)$ satisfies

$$\left(\begin{array}{c} \sum_{\ell=0}^{j} n \\ \ell \pi_{\alpha,n}(j)^{\ell}(1-\pi_{\alpha,n}(j))^{n-\ell} = \alpha. \end{array}\right) \tag{30}$$

Let $\pi_{\alpha,n}(-1) \equiv 0$. The Stringer Bound is

$$\tau_\alpha(z) \equiv \sum_{j=0}^{P} [\pi_{\alpha,n}(j) - \pi_{\alpha,n}(j-1)] z_j. \tag{31}$$

There is considerable evidence that the Stringer Bound is conservative; Bickel [3] proves that it is essentially always conservative for finite samples, and asymptotically quite conservative.

The special case that all the observed taints are zero gives

$$\tau_\alpha = \pi_{\alpha,n}(0). \tag{32}$$

Note that $\pi_{\alpha,n}(0)$ solves

$$(1 - \pi_{\alpha,n})^n = \alpha, \tag{33}$$

i.e., $n = \ln(\alpha)/\ln(1-\pi_{\alpha,n})$. Recall that $E = \tau \sum_{\mathcal{N}} u$. To have $1-\alpha$ confidence that $E < M$ when all observed taints are zero—so that the outcome of the election is not in doubt—we need $1-\alpha$ confidence that $\tau < M/\sum_{\mathcal{N}} u$, i.e., $\tau_\alpha = \pi_{\alpha,n} < M/\sum_{\mathcal{N}} u$. The smallest n that suffices is

$$n_\alpha = \left\lceil \frac{\ln \alpha}{\ln(1 - M/\sum_{\mathcal{N}} u)} \right\rceil = n_{PPS}(\alpha; u, M). \tag{34}$$

That is, the result of Aaslam et al. [2] is a special case of the Stringer Bound when all observed taints are zero. See also Kaplan [13].

The Stringer Bound uses all the observed taints, not just the maximum observed taint. As a result, it generally leads to smaller P-values than the statistic proposed in section 5 and to a more powerful test than that proposed in section 6.

[19, 3, 4]

8.2 Multinomial Bounds

Suppose we divide the interval of possible taints, $[0, 1]$, into bins. For example, we might use the 100 bins $[(j-1)/100, j/100)$ for $j = 1, \ldots, 99$, and $[0.99, 1]$. For with PPEB sampling, the joint distribution of the number of observed taints that fall in each bin is multinomial. Fienberg et al. [8] exploit this to construct an upper confidence bound for τ.

...

Like the Stringer Bound, the multinomial bounds developed by Fienberg et al. [8] use the values and frequencies of all the observed taints. It can also lead to smaller P-values and a more powerful test, but it is quite complex to compute when the data fall into more than a few bins.

[8, 18, 20]

More to come ....

9 Discussion

9.1 Room for improvement

The condition $E \geq M$ is necessary but not sufficient...

Still to come ...

10 Conclusions

Still to come ...

References

[1] R.G. Anderson and A.D. Teitlebaum, *Dollar-unit sampling. a solution to the audit sampling dilemma*, Canadian Chartered Accountant (1973), 30–39.

[2] J.A. Aslam, R.A. Popa, and R.L. Rivest, *On auditing elections when precincts have different sizes*, people.csail.mit.edu/rivest/AslamPopaRivest-OnAuditingElectionsWhenPrecinctsHaveDifferentSizes.pdf, 2007.

[3] P.J. Bickel, *Inference and auditing: the Stringer bound*, Intl. Stat. Rev. **60** (1992), 197–209.

[4] ______, *Correction: Inference and auditing: the Stringer bound*, Intl. Stat. Rev. **61** (1993), 487.

[5] H.R. Clayton, *A combined bound for errors in auditing based on Hoeffding's inequality and the bootstrap*, J. Bus. Econ. Stat. **12** (1994), 437–448.

[6] D. Cox and E.J. Snell, *On sampling and the estimation of rare errors*, Biometrika **66** (1979), 125–132.

[7] K. Dopp and F. Stenger, *The election integrity audit*, uscountvotes.org/ucvInfo/release/ElectionIntegrityAudit-release.pdf, 2006.

[8] S.E. Fienberg, J. Neter, and R.A. Leitch, *Estimating total overstatement error in accounting populations*, J. Am. Stat. Assoc. **72** (1977), 295–302.

[9] E.W. Gaynor, *Reliability of sampling plans in auditing*, The Accounting Review **31** (1956), 253–257.

[10] J. Godfrey and J. Neter, *Bayesian bounds for monetary unit sampling in accounting and auditing*, J. Accounting Res. **22** (1984), 497–525.

[11] R. Helmers, *Inference on rare errors using asymptotic expansions and bootstrap calibration*, Biometrika **87** (2000), 689–694.

[12] D. Jefferson, K.Alexander, E. Ginnold, A. Lehmkuhl, K. Midstokke, and P.B. Stark, *Post election audit standards report–evaluation of audit sampling models and options for strengthening californias manual count*, www.sos.ca.gov/elections/peas/final_peaswg_report.pdf, 2007.

[13] R.S. Kaplan, *Sample size computations for dollar-unit sampling*, J. Accounting Res. **13** (1975), 126–133.

[14] A.H. Kvanli, Y.K. Shen, and L.Y. Deng, *Construction of confidence intervals for the mean of a population containing many zero values*, J. Bus. Econ. Stat. **16** (1998), 362–368.

[15] D.J. Laws and A. O'Hagan, *Bayesian inference for rare errors in populations with unequal unit sizes*, Appl. Stat. **49** (2000), 557–590.

[16] J. McCarthy, H.I. Stanislevic, M. Lindeman, A. Ash, V. Addona, and M. Batcher, *Percentage-based versus statistical-power-based vote tabulation audits*, The American Statistician **62** (2008), 11–16.

[17] U. Menzefricke and W. Smieliauskas, *A simulation study of the performance of parametric dollar unit sampling statistical procedures*, J. Accounting Res. **22** (1984), 588–603.

[18] J. Neter, R.A. Leitch, and S.E. Fienberg, *Dollar unit sampling: multinomial bounds for total overstatement and understatement errors*, The Accounting Review **53** (1978), 77–93.

[19] Panel on Nonstandard Mixtures of Distributions, *Statistical models and analysis in auditing: A stody of statistical models and methods for analyzing nonstandard mixtures of distributions in auditing*, National Academy Press, Washington, D.C., 1988.

[20] R. Plante, J. Neter, and R.A. Leitch, *Comparative performance of multinomial, cell and Stringer bounds*, Auditing: A Journal of Practice & Theory **5** (1985), 40–56.

[21] R.L. Rivest, *On estimating the size of a statistical audit*, people.csail.mit.edu/rivest/Rivest-OnEstimatingTheSizeOfAStatisticalAudit.pdf, 2006.

[22] R.G. Saltman, *Effective use of computing technology in vote-tallying*, Tech. Report NBSIR 75-687, National Bureau of Standards, Washington, DC, 1975.

[23] W. Smieliauskas, *A note on a comparison of Bayesian with non-Bayesian dollar-unit sampling bounds for overstatement errors of accounting populations*, The Accounting Review **61** (1986), 118–128.

[24] P.B. Stark, *A confidence-driven audit of Measure A in the February 2008 Marin County election*, 2008.

[25] ———, *Conservative statistical post-election audits*, Ann. Appl. Stat. **in press** (2008).

[26] K.W. Stringer, *Practical aspects of statistical sampling in auditing*, Proceedings of the Business and Economic Statistics Section (Washington, D.C.), American Statistical Association, 1963, pp. 405–411.

[27] N.N. Taleb, *The Black Swan: The Impact of the Highly Improbable*, Random House, New York, 2007.

[28] H. Tamura and P.A. Frost, *Tightening CAV (DUS) bounds by using a parametric model*, J. Accounting Res. **24** (1986), 364–371.

[29] R.M. Trueblood and W.W. Cooper, *Research and practice in statistical applications to accounting, auditing, and management control*, The Accounting Review **30** (1955), 221–229.

[30] K.-W. Tsui, E.M. Matsumura, and K.-L. Tsui, *Multinomial-dirichlet bounds for dollar-unit sampling in auditing*, The Accounting Review **60** (1985), 76–96.

[31] L.L. Vance, *Scientific method for auditing: Applications of statistical sampling theory to auditing procedure*, University of California Press, Berkeley and Los Angeles, 1950.