On Estimating the Size and Confidence of a Statistical Audit

Javed A. Aslam
College of Computer and Information Science
Northeastern University
Boston, MA 02115
jaa@ccs.neu.edu

Raluca A. Popa and Ronald L. Rivest
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139
{ralucap, rivest}@mit.edu

Abstract

We consider the problem of statistical sampling for auditing elections, and we develop a remarkably simple and easily-calculated upper bound for the sample size necessary for determining with probability at least c if a given set of n objects contains fewer than b "bad" objects. While the size of the optimal sample drawn without replacement can be determined with a computer program, our goal is to derive a highly accurate and simple formula that can be used by election officials equipped with only a hand-held calculator. We actually develop several formulae, but the one we recommend for use in practice is:

$$\begin{aligned} & U_3(n, b, c) \\ = & \left\lceil \left(n - \frac{(b-1)}{2}\right) \cdot \left(1 - (1-c)^{1/b}\right) \right\rceil \\ = & \left\lceil \left(n - \frac{(b-1)}{2}\right) \cdot \left(1 - \exp(\ln(1-c)/b)\right) \right\rceil \end{aligned}$$

As a practical matter, this formula is essentially exact: we prove that it is never too small, and empirical testing for many representative values of $n \leq 10,000$, and $b \leq n/2$, and $c \leq 0.99$ never finds it more than one too large. Theoretically, we show that for all n and b this formula never exceeds the optimal sample size by more than 3 for $c \leq 0.9975$, and by more than $(-\ln(1-c))/2$ for general c.

1 Introduction

Given the increased popularity of voting systems with voter-verified paper ballots, there is increased need for effective audits to confirm that those paper ballots agree with their electronic counterparts (which might be the result of scanning those ballots). Since auditing is expensive (it is typically done by hand), it is important to minimize the expense by choosing a sample size for the audit that is as small as possible, while guaranteeing a desired level of statistical confidence. This paper addresses the question of determining the appropriate sample size and develops nearly exact approximations that can be evaluated easily on a hand-held calculator. We believe that these formulae will turn out to be useful in practice.

Given a universe of n objects, how large a sample should be tested to determine with high confidence if fewer than a given number b of them are bad? (In the voting context, these objects are typically voting precincts.)

As noted, our goal is to develop approximations that are both accurate and simple enough to be usable, if not by hand, then at least with the use of only a calculator, with no computer needed. (Your calculator must be a "scientific" one, though, so that you can compute arbitrary powers.[1])

We first present a simple approximate "rule of thumb" (the "Rule of Three") for estimating how big such a statistical sample should be, when using sampling *with replacement*.

This "Rule of Three" is simple and known, although perhaps not particularly well-known. Jovanovic and Levy [15] discuss the Rule of Three, its derivation, and its application to clinical studies. See also van Belle [24].

We then address the question of sampling *without replacement*, which is the desired procedure for an election audit, of course, and provide improved formulae for sample size when sampling without replacement.

This paper justifies and improves approximations originally developed by Rivest [20], who attempted to correct for the bias in the Rule of Three due to sampling with replacement instead of sampling without replacement, by only sampling (now without replacement) the expected number of *distinct* elements that the Rule of Three sample (with replacement) would have contained. While that may be an interesting approach, the current paper derives its approximation formulae more directly, and provides rigorous upper and lower bounds on their approximation error.

Finally, in Section 5, we address three related questions: (1) determining the confidence level for a given audit size and level of fraud one wishes to detect, (2) determining the minimum amount of fraud one can detect for a given audit size with a given confidence level, and (3) auditing with constraints, such as the requirement that at least one precinct in each county be audited.

1.1 Related Work

Saltman [22, Appendix B] was the first to study sample size (for sampling without replacement) in the context of voting; the basic formulae he develops for the optimal sample size are the ones we are trying to approximate here.

(There is much earlier relevant work on sampling theory, particularly the notion of "lot acceptance sampling" in statistical quality control. For example, the Dodge-Romig Sampling Inspection Tables [6], developed in the 1930's and first published in 1940, provide generalizations of the simple sampling methods used here.)

Previous work by Neff [17] is noteworthy, particularly with regard to the economies resulting from having a larger universe of many smaller, easily-testable, objects. Brennan Center report [3, Appendix J] gives some simple estimation formula, based on sampling with replacement. An excellent report [8] on choosing appropriate audit sizes by Dopp and Stenger from the National Election Data Archive Project is now also available; there is also a nice associated audit size calculation utility on a web site [16]. Stanislevic [23] also examines the issue of choosing a sufficient audit size; he gives a particularly nice treatment of handling varying precinct sizes.

Some states, such as California, mandate a certain level (e.g. 1%) of auditing [19]. As we shall see, using a fixed level of auditing is not a well justified approach; sometimes one may need more auditing, and sometimes less, to obtain a given level of confidence that no fraud has occurred.

2 Auditing Model

Suppose we have n "objects." In a voting context, such an "object" might typically correspond to a precinct; it could also correspond to a particular voting machine or even an individual ballot, depending on the situation; the math is the same.

In this paper, we assume an adversarial situation, where an adversary may have corrupted some of the objects. For example, the adversary might have tampered with the results of some precincts in a state.

Thus, after the adversary has acted, each object is either "good" (that is, clean, untampered with, uncorrupted), or "bad" (that is, tampered with, corrupted).

We now wish to test a sample of the objects to determine with high confidence that the adversary has not committed a "large" amount of fraud.

(As we shall see, this is related to the following standard combinatorial problem: We have an urn containing n balls, b of which are black and $n - b$ of which are white; we wish to sample enough balls to have a sufficiently high probability of sampling at least one black ball.)

We assume that each object is independently auditable. That is, there is a test or audit procedure that can determine whether a given object is good or bad. We assume this procedure is always correct.

For example, testing the results in a precinct may involve comparing the electronic results from the precinct with a hand count of the corresponding voter-verified paper ballots. The precinct may be judged to be good if the results are equal (or perhaps if they are "sufficiently close").

Of course, there may easily be explanations for a discrepancy other than malicious behavior on the part of some "adversary." Indeed, one of the normal goals of such a post-election audit is to assess if there were systematic errors in the results due to undetected equipment or procedural problems, such a misprogramming ballot style or other configuration data, in addition to assessing if fraud was present. In general, electronic officials will try to determine the cause of any discrepancies found, whether it be due to malicious causes or not, and take appropriate corrective or remedial action as necessary (which might involve further auditing and investigations).

While discrepancies found are typically not due to adversarial behavior (fraud), in this paper we nonetheless focus on the problem of detecting fraud in election results through the use of appropriate post-election auditing. This is in part because this is the most difficult case; we feel that systematic problems not due to malicious behavior are likely to be uncovered pretty well anyway during an audit designed to detect fraud. (Although we do favor requirements such as ensuring that at least one precinct in each county participate in a post-election audit.)

To continue with our modeling: we'll now simply assume that each object tested is found to be "good" or "bad." In the voting scenario, we assume each precinct whose results have been in any way manipulated by an adversary will test as "bad," and each precinct which the adversary has not touched will test as "good."

This is of course a bit of a simplification, since an adversary may try to influence an election by making very small modifications to the results a large number of precincts, hoping that each such modification will be judged as too small to cause the precinct to be flagged as "bad" during the audit.

We note in this context that Ansolabehere and Reeves [2] determined from historical recount data in New Hampshire that there is typically a difference in range of 0.5%–1.0% between an initial machine count of a set of opscan ballots and a hand recount of those ballots; adversarial manipulation in a similar range might pass as non-anomalous. Similar results were found by Alvarez, Katz, and Hill [1] in their study of the recounting of punch-card ballots in California. (Another interesting study of recounts in New Hampshire by Herron and Wand [12] examines the question as to whether the choice of different voting technologies by different precincts was a source of partisan bias in the 2004 election results.)

Nonetheless, in spite of the clear existence of a small level of such "measurement noise," we'll continue to make the simplifying assumption for our purposes in this paper that each precinct measures cleanly as "good" or "bad."

For more general discussions of fraud in elections, see the references collected by Hill [13], the report by the EAC [4], and the report by the Brennan Center [3]. For more general discussions of auditing election results, see the case study by `electionline.org` [9], and the Brennan Center web site [10] (including the House testimony by Norden [18]).

To determine whether *any* fraud at all occurred, we would need to test *all* objects. Here we give up the ability to detect *any* fraud, and test only a sample of the objects in order to determine, with high confidence, that a *large* amount of fraud has not occurred. We lose a bit of confidence in return for a large increase in efficiency, as is usually the case for a statistical test.

Let b denote the number of "bad" objects we wish to detect, where b is a given constant, $1 \leq b \leq n$. That is, we wish to determine, with high confidence, that the number of corrupted objects is not b or greater.

Since the adversary wishes to escape detection, he will corrupt as few objects as possible, consistent with achieving his evil goals. We assume that corrupting b objects suffices, and so the adversary corrupts exactly b objects. (For voting, this implies that all precincts are assumed to have roughly the same size; see Section 2.1.)

We let c denote our desired "confidence level"—that is, we want the probability of detecting corruption of b or more objects to be at least c, where c is a given parameter, $0 \leq c \leq 1$ (e.g. $c = 0.95$).

We let

$$f = b/n \tag{1}$$

denote the fraction of bad objects we wish to detect; we call f the "fraud rate." Given one of b or f, the other is determined, via equation (1).

We will be considering samples drawn both with replacement and without replacement. For mnemonic convenience, we use t to denote sample sizes when the sample is drawn with replacement, and u to denote sample sizes when the sample is drawn without replacement. (Think of "u" for "unique" or "distinct".)

The auditing process can be cast in the terminology of a conventional "hypothesis-testing" framework. We set the *null hypothesis* to be the hypothesis we wish to refute,

$$H_0 = \text{ the reported election outcome is incorrect}$$

(i.e., there was fraud or other error in the electronic totals sufficient to change the election outcome), while the alternative hypothesis is its complement,

$$H_1 = \text{ the reported election outcome is correct}$$

(i.e., the electronic totals give the correct result).

We then randomly select a sample of precincts to audit and hand-count the corresponding paper ballots. If the hand-counts all match the electronic totals in those precincts, we have event:

$$D_0 = \text{ no sampled precincts were ``bad''}$$

(where "bad" means "showed evidence of possible fraud"); otherwise, we have event:

$$D_1 = \text{ at least one sampled precinct was ``bad'' .}$$

Our statistical test is designed in such a way that if the null hypothesis were true (sufficient fraud to change the election outcome exists), then it is very unlikely (probability at most $1 - c$) that no "bad" objects would be detected (event D_0) due to the random nature of sampling. Thus, the absence of any "bad" objects in our sample permits us to reject the null hypothesis with high confidence, and in this case, we would typically declare the winner to be the winner as shown by the electronic totals.

However, if we do detect "bad" objects (event D_1), we cannot reject the null hypothesis. This does not necessarily imply that the null hypothesis is true; rather, we simply do not have sufficient evidence to reject the null hypothesis and declare a winner. (Indeed, if $b-1$ objects were corrupted, a quantity posited to be insufficient to alter the election outcome, then it is quite likely nonetheless that a "bad" object would be detected.) In the case where we have evidence that some fraud may have occurred, one would typically proceed with a wider investigation of the election. Thus, our statistical test is inherently one-sided: the absence of any "bad" objects in our sample allows us to reject the null hypothesis with high confidence and typically declare a winner, while the presence of any "bad" objects would typically trigger a wider investigation.

In hypothesis testing terms, one considers two types of errors. A Type I error occurs when the null hypothesis is rejected incorrectly. For a given statistical test, the probability that a Type I error occurs is denoted by α; in our case, this type of error occurs with probability

$$\alpha = \mathbf{Pr}[D_0 \mid H_0] \ .$$

Here we have failed to detect fraud (or other significant error) sufficient to change the outcome of the election.

Similarly, a Type II error occurs when the null hypothesis is accepted incorrectly. For a given statistical test, the probability that a Type II error occurs is denoted by β; in our case, this type of error occurs with probability

$$\beta = \mathbf{Pr}[D_1 \mid H_1] \ .$$

Here we have detected errors or fraud in some precincts even though the election outcome is correctly determined by the electronic totals.[2]

The quantity α is the statistical *significance level* of the test, and the quantity $1 - \beta$ is the statistical *power* of the test. We are primarily concerned with the significance level of our test, i.e, the probability that we fail to detect an incorrect election outcome due to the nature of random sampling, a Type I error that occurs with probability α. In this paper, for historical reasons, we also refer to the *confidence level* c of our test, where $c = 1 - \alpha$. (In the final version of this paper we may stick with the more typical usage.)

If we choose b appropriately (as, for example, suggested in the next section), then we have that

$$\alpha = \mathbf{Pr}[D_0 \mid H_0] \leq \mathbf{Pr}[D_0 \mid H_0']$$

where

$$H_0' = \ b \text{ precincts are ``bad''} \ .$$

By determining an appropriate value of b, and then choosing the appropriate sample size to make $\mathbf{Pr}[D_0 \mid H_0']$ sufficiently small, we will make the probability of a Type I error at most $\alpha = 1 - c$. Thus, the probability of reporting an incorrect election outcome will be at most $1 - c$.

2.1 Deriving b from the margin of victory

We now explain how a suitable value for b might be determined for an election audit from the apparent margin of victory, using a method suggested by Dopp and Stenger [8]. Here, b is the number of precincts that an adversary would have had to corrupt to swing the election. If we assume (as is suggested by Dopp and Stenger to be reasonable) that the adversary wouldn't dare to change more than a fraction $s = 0.20$ (i.e. 20%) of the votes in a precinct, and that the "winner" won by a margin of m of the votes (where $0 \leq m \leq 1$), then the adversary would have had to have corrupted a fraction

$$f = m/(2s) = 2.5m \qquad (2)$$

of the precincts—or, equivalently,

$$b = mn/(2s) = 2.5mn \qquad (3)$$

precincts.

(We assume all precincts have the same size. If all of the votes changed had been moved from the actual winner to the alleged winner, then a margin of victory of a fraction m of the votes cast for the alleged winner must have involved at least a fraction $m/(2 * 0.20) = 2.5m$ of the precincts, since each precinct corrupted changed the difference in vote count between the top two candidates by $2s = 40\%$ of the vote count of that precinct.) If the apparent winner has won by $m = 1\%$ in a county with 400 precincts, you would want to test for $b = 2.5mn = 10$ or more bad precinct counts.

This approach can be modified in various ways (e.g. by adjusting s) to suit particular situations; in any case a value of b is determined that represents the minimum number of precincts that an adversary "must" corrupt in order to have changed the election outcome.

See Saltman [22], Stanislevic [23], or Dopp et al. [8] for further examples and excellent treatment of the issue of computing appropriate target values b (or f) given a set of election results and possibly varying precinct sizes. Rivest [21] also treats the case of varying precinct sizes.

3 Sampling with replacement and the Rule of Three

We begin by examining sampling *with replacement* (where the sample may contain an element more than once). Although this wouldn't be used in practice for auditing an election, it is a useful starting point for our analyses, and provides some reasonably accurate estimation formulae that can be easily computed in one's head.

For sampling *with* replacement, we use t to denote the sample size, and $t_*(n, b, c)$ to denote the optimal sample size (when sampling a set of size n with replacement, in order to find at least one bad element, with probability at least c, when b bad elements are present). We'll later use the analogous notation $u_*(n, b, c)$ for the optimal sample size for sampling *without replacement*.

Here now is a simple "rule of thumb" for sampling with replacement.

Rule of Three:
Test, using sampling with replacement, enough objects so that you expect to see *at least three* corrupted objects. That is, ensure that:

$$ft = \frac{bt}{n} \geq 3. \tag{4}$$

or equivalently:

$$t \geq 3n/b \; . \tag{5}$$

(Where t is the number of objects to be tested, b is the number of bad objects one wishes to detect, and $f = b/n$, at a 95% confidence level.)

As a simple example: to detect a 1% fraud rate ($f = 0.01$) (with 95% confidence), you then need to test $t = 300$ objects.

Note that for a given fraud rate f, the rule's sample size is *independent* of the universe size n. This may seem counter-intuitive, but is to be expected. If you have some well-mixed sand where most sand grains are white, but a fraction f are black, you need only sample a handful to be confident of obtaining a black grain, no matter whether the amount of sand to be examined is a cupful, a bucketfull, or a beach.

The sample size t may even be greater than n (if $b < 3$); this is OK since we are sampling with replacement, and it may take more than n samples (when sampling with replacement) to get adequate coverage when b is so small.

3.1 A Sampling with Replacement Bound

We now justify the Rule of Three, and then generalize it to handle an arbitrary confidence level (not just $c = 0.95$). Let $f = b/n$ be the *fraud rate*, and let t be the sample size.

We first justify the Rule of Three for a confidence level of 95%; this analysis follows that given by Jovanovic and Levy [15].

The probability that a fraud rate of f or greater goes *undetected* (when drawing a sample of size t with replacement) is:

$$(1 - b/n)^t = (1 - f)^t \; . \tag{6}$$

so t must be large enough so that

$$(1 - f)^t \leq 0.05$$

or equivalently:

$$t \geq \frac{\ln(0.05)}{\ln(1 - f)} \tag{7}$$

Since

$$\ln(0.05) = -\ln(20) = -2.9957 \approx -3$$

—isn't it so very nice that $\ln(20)$ is almost exactly 3?—equation (7) is implied by

$$t \geq \frac{-3}{\ln(1 - f)} \; . \tag{8}$$

Using the well-known approximation

$$\ln(1 - f) \approx -f \; , \tag{9}$$

which is quite accurate for small values of f (and $-f$ is an lower bound on $\ln(1 - f)$), equation (8) is implied by:

$$t \geq \frac{3}{f}$$

which can be rewritten as

$$t \geq \frac{3n}{b} \tag{10}$$

or equivalently as

$$ft \geq 3 \; . \tag{11}$$

Equation (11) has a very nice and intuitive interpretation. Since t is the number of objects tested, and f is the fraud rate, then ft *is the number of objects among the test objects that we would expect to find corrupted.*

The sample should be large enough so that you expect it to contain at least three corrupted objects. If you sample enough so that you expect to see at least three corrupted objects on the average, then you'll see at least one corrupted object almost always (i.e., at least 95% of the time).

(Similarly, a random variable X distributed according to the Poisson distribution with mean $\lambda > 3$ satisfies $\mathbf{Pr}[X = 0] = e^{-\lambda} < e^{-3} = 0.04978\ldots$)

As a running example, suppose that $n = 400$, $b = 10$, and $f = b/n = 0.025$; the Rule of Three says to pick a sample of size $3n/b = 3 * 400/10 = 120$.

(We shall see that the optimal sample size for sampling *without replacement* for these parameters is a little smaller—103—, so considering sample size with replacement may be a good first-cut approximation to the sample size needed for sampling without replacement.This "Rule of Three" ($t \geq 3n/b$) is simple enough for some practical guidance.)

The Rule of Three is also easily generalized to handle other confidence levels. For a general confidence level c, $0 \leq c \leq 1$, we need that

$$(1 - f)^t \leq (1 - c) \tag{12}$$

so we obtain the following formulae for the optimal sample size $t_*(n, b, c)$, when sampling with replacement:

$$\begin{aligned} t_*(n, b, c) &= \frac{\ln(1 - c)}{\ln(1 - f)} &(13)\\ &= \frac{\ln(1 - c)}{\ln(1 - b/n)} \; . &(14) \end{aligned}$$

We note that equation (14) may give "optimal" values for t_* that are non-integral, while in practice the sample size must be an integer. Of course, the optimal integral sample size is then just t_* rounded up to the next integer, yielding T_*:

$$T_*(n,b,c) = \lceil t_*(n,b,c) \rceil \ .$$

Using equation (9), we obtain the generalized form of the Rule of Three as an approximation:

$$t_1(n,b,c) = \frac{-n \ln(1-c)}{b} \ . \qquad (15)$$

This completes our discussion of sample sizes for sampling with replacement.

4 Sampling without replacement

Suppose we pick u objects to test, where $0 < u \leq n$. These u objects are chosen independently at random, without replacement—the objects are distinct.[3]

In an election, if *any* of the u tested objects (e.g. precincts or voting machines) turns out to be "bad," then we may declare that "evidence of possible fraud is detected" (i.e., at least one bad object was discovered). Otherwise, we report that "no evidence of fraud was detected." When a bad object is detected, additional investigation and further testing may be required to determine the actual cause of the problem.

We wish it to be the case that if a large amount of fraud has occurred (i.e., if the number of corrupted objects is b or greater), then we have a high chance of detecting at least one bad object.[4]

Given that we are drawing, without replacement, a sample of size u from a universe of size n containing b bad objects, the chance that no bad objects are detected (i.e. all bad objects escape detection) is:

$$\begin{aligned} e(n,b,u) &= \binom{n-b}{u} \Big/ \binom{n}{u} && (16) \\ &= \frac{(n-b)!}{(n-b-u)!} \cdot \frac{(n-u)!}{n!} && (17) \\ &= \prod_{k=0}^{u-1} \frac{n-b-k}{n-k} \ ; && (18) \end{aligned}$$

the chance that at least one bad object is detected is:

$$\begin{aligned} d(n,b,u) &= 1 \ - \ e(n,b,u) && (19) \\ &= 1 \ - \ \prod_{k=0}^{u-1} \frac{n-b-k}{n-k} \ . && (20) \end{aligned}$$

We note here the convenient duality between b and u, which we shall use later:

$$\begin{aligned} e(n,b,u) &= \frac{(n-b)!}{(n-b-u)!} \cdot \frac{(n-u)!}{n!} && (21) \\ &= \frac{(n-u)!}{(n-u-b)!} \cdot \frac{(n-b)!}{n!} && (22) \\ &= e(n,u,b) \ . && (23) \end{aligned}$$

(If we think of the b bad objects as the "sample" and the u audit objects as the targets to be detected, then we are just switching the role of the bad objects and the audited objects.) This duality gives us another expression for $e(n,b,u)$, dual to equation (18):

$$e(n,b,u) = \prod_{k=0}^{b-1} \frac{n-u-k}{n-k} \ . \qquad (24)$$

For a given confidence level c (e.g. $c = 0.95$), the optimal sample size $u_* = u_*(n,b,c)$ is the least value of u making $d(n,b,u)$ at least c:

$$\begin{aligned} u_*(n,b,c) &= \min\{u \mid d(n,b,u) \geq c\} && (25) \\ &= \min\{u \mid e(n,b,u) \leq 1-c\} \ . && (26) \end{aligned}$$

We now address again the issue of non-integral sample sizes. Although of course sample sizes are integral in practice, our formulae work perfectly well for non-integral sample sizes, and it is convenient for us to work with them: note that $e(n,b,u)$ equation (24) is well defined when u is any real number, and so $d(n,b,u) = 1 - e(n,b,u)$ is also well defined when u is any real number. In practice, a non-integral optimal sample size $u_*(n,b,c)$ would be rounded up to the next integer $\lceil u_*(n,b,c) \rceil$, which we denote as $U_*(n,b,c)$.

Equations (16)–(20) and (25)–(26) are not new here; they have been given and studied by others (e.g. [22, 17, 8]).

In our running example, we have $n = 400$ and $b = 10$; we wish to determine if a set of 400 objects contains 10 or more bad ones. Using a computer program to try successive values of u yields the result:

$$U_*(400, 10, 0.95) = 103 \ ; \qquad (27)$$

we need to test a sample (drawn without replacement) of size at least 103 in order to determine if our set of 400 objects contains 10 or more bad objects, with probability at least 95%.

In some sense, this completes the analysis of the problem; it is easy for a computer program to determine the optimal sample size $U_*(n,b,c)$, given n, b, and c. (See `http://uscountvotes.org` where such a program may be posted.)

However, it is useful to find simple but accurate approximations for this optimal value $U_*(n,b,c)$ of u, that can be easily calculated without the use of a computer. That is the main purpose of this paper—to derive accurate and rigorously justified approximations for U_* that can be evaluated by election officials using only a pocket calculator.

The formulae of the previous section for T_* (for sampling with replacement) are of course crude estimates for U_* (sampling without replacement); they are overestimates.

To see this, note that equation (18) implies that

$$e(n,b,u) \leq \left(1 - \frac{b}{n}\right)^u \tag{28}$$

Now $(1 - b/n)^u$ is the probability of drawing a multiset of size u *with replacement* having no bad objects. Thus, for a fixed sample size, the probability of failure when drawing samples without replacement is, as one would expect, upper bounded by the probability of failure when drawing samples with replacement. The quality of this upper bound is a function of the difference between the right-hand sides of equation (18) and inequality (28). Note that this difference grows as u increases, and for high probability results with large n and small b, u can be quite large. (Indeed, when $b = 1$ and c very large, $t_*(n, b, c)$ is approximately $n \ln(n)$ — this is the "coupon collector's problem" — while $u_*(n, b, c)$ is clearly no larger than n.)

Thus, we can in fact use the Rule of Three or other formulae from the preceding section to get an upper bound on the sample size needed for sampling without replacement; in many cases this may give a satisfactory first-cut answer. But we can do better, as the next section shows.

4.1 Upper Bounds on Optimal Sample Size for Sampling without Replacement

We now develop an upper bound on the optimal sample size when sampling without replacement to detect at least one of b bad objects in a universe of size n with probability at least c.

From equation (24), one can derive (analogous to the derivation of equation (28) from equation (18)), the following bound:

$$e(n,b,u) \leq \left(1 - \frac{u}{n}\right)^b \tag{29}$$

Our goal is to determine a value u is sufficiently large to guarantee that $e(n, b, u)$ is at most $1 - c$; from the bound (29) we can obtain such a sufficiently large u:

$$\begin{array}{rrcl} & \left(1 - \frac{u}{n}\right)^b & \leq & 1 - c \\ \Leftrightarrow & 1 - u/n & \leq & (1-c)^{1/b} \\ \Leftrightarrow & u/n & \geq & 1 - (1-c)^{1/b} \\ \Leftrightarrow & u & \geq & n(1 - (1-c)^{1/b}) \qquad (30) \end{array}$$

Since (29) holds for any u satisfying (30), $u_*(n, b, c)$ is no larger than the right hand side of (30). This upper bound on $u_*(n, b, c)$ is our first major result for sampling without replacement; it is a formula that is easy to calculate, yet remarkably accurate.

We designate this bound as u_1:

First Upper Bound on $u_*(n, b, c)$:

$$u_*(n,b,c) \leq u_1(n,b,c) \tag{31}$$

where

$$\begin{array}{rcl} u_1(n,b,c) & = & n(1-(1-c)^{1/b}) \qquad (32) \\ & = & n(1 - \exp(\ln(1-c)/b)) \end{array}$$

The formula for $u_1(n, b, c)$ is the same as the that proposed by Rivest [20] as an approximation for $u_*(n, b, c)$; however, that paper only justified u_1 as an approximation heuristically and empirically; here we have shown that it is a firm upper bound for $u_*(n, b, c)$.

Of course, if we round up $u_1(n, b, c)$ to obtain $U_1(n, b, c)$, we obtain an integer upper bound on the optimal integral sample size:

$$\begin{array}{rcl} U_1(n,b,c) & = & \lceil u_1(n,b,c) \rceil \\ & \geq & \lceil u_*(n,b,c) \rceil = U_*(n,b,c) \ . \end{array}$$

A Tighter Upper Bound: We can obtain a tighter upper bound by analyzing the product in equation (24) directly. Using the following well-known inequalities relating the harmonic, geometric, and arithmetic means for non-negative values x_i [11]

$$\frac{k}{\sum_{i=1}^k 1/x_i} \leq \sqrt[k]{\prod_{i=1}^k x_i} \leq \frac{\sum_{i=1}^k x_i}{k} \tag{33}$$

we proceed as follows, where H_k is the k-th harmonic number, i.e., $H_k = 1 + 1/2 + \cdots + 1/k$.

$$\begin{array}{rcl} e(n,b,u) & = & \prod_{k=0}^{b-1} \left(1 - \frac{u}{n-k}\right) \\ & = & \left(\sqrt[b]{\prod_{k=0}^{b-1} \left(1 - \frac{u}{n-k}\right)}\right)^b \\ & \leq & \left(\frac{1}{b}\sum_{k=0}^{b-1} \left(1 - \frac{u}{n-k}\right)\right)^b \qquad (34) \\ & = & \left(1 - \frac{u}{b} \cdot \sum_{k=0}^{b-1} 1/(n-k)\right)^b \\ & = & \left(1 - u \cdot \frac{H_n - H_{n-b}}{b}\right)^b \end{array}$$

As before, our goal is to determine a u sufficient to guarantee that the above quantity is at most $1-c$. Solving the inequality

$$\left(1 - u \cdot \frac{H_n - H_{n-b}}{b}\right)^b \leq 1 - c$$

in much the same manner as the derivation of inequality (30), we obtain

$$u \geq \frac{b}{H_n - H_{n-b}} \cdot (1 - (1-c)^{1/b}) \qquad (35)$$

Note that the bound obtained in inequality (35) was derived using only one approximation, inequality (34) above. The right-hand side of inequality (35) is our second upper bound on the optimal sample size required for sampling without replacement. We call this upper bound $u_2(n, b, c)$; we also let $U_2(n, b, c) = \lceil u_2(n, b, c) \rceil$; this is of course an upper bound on $U_*(n, b, c)$.

Second Upper Bound on u_*

$$u_*(n, b, c) \leq u_2(n, b, c) \qquad (36)$$

where

$$\begin{aligned} & u_2(n, b, c) \\ = \; & \frac{b}{H_n - H_{n-b}} \cdot (1 - (1-c)^{1/b}) \qquad (37) \\ = \; & \frac{b}{H_n - H_{n-b}} \cdot (1 - \exp(\ln(1-c)/b)) \end{aligned}$$

Unfortunately, most calculators don't have a "harmonic number" button, so inequality (35) isn't so useful in practice!

To fix this situation, without weakening our bound too much, we note that

$$\frac{b}{H_n - H_{n-b}} = \frac{b}{\sum_{k=0}^{b-1} \frac{1}{n-k}}$$

is the harmonic mean of the set of values $\{n, \ldots, n - b + 1\}$; thus, we can obtain a simpler though slightly weaker bound by employing inequality (33) and replacing this harmonic mean by the corresponding (and somewhat larger) arithmetic mean $(n - \frac{(b-1)}{2})$, which yields

$$u \geq \left(n - \frac{(b-1)}{2}\right) \cdot \left(1 - (1-c)^{1/b}\right) \qquad (38)$$

This gives our third and final upper bound:

Third Upper Bound on u_*

$$u_*(n, b, c) \leq u_3(n, b, c) \qquad (39)$$

where

$$\begin{aligned} & u_3(n, b, c) \\ = \; & \left(n - \frac{(b-1)}{2}\right) \cdot \left(1 - (1-c)^{1/b}\right) \qquad (40) \\ = \; & \left(n - \frac{(b-1)}{2}\right) \cdot \left(1 - \exp(\ln(1-c)/b)\right) \end{aligned}$$

Note the similarity of inequalities (30) and (38): the factor n has been replaced with $(n - \frac{(b-1)}{2})$. Thus, the new inequality (38) (and inequality (35) which precedes it) is a strict improvement over inequality (30) for all $b > 1$ (and the same for $b = 1$).

We let $U_3(n, b, c) = \lceil u_3(n, b, c) \rceil$; this is of course also an upper bound on $U_*(n, b, c)$.

Inequality (38) is our third (and final) upper bound on the optimal sample size required for sampling without replacement; it is the inequality that we recommend for actual use in practice.[5] As we see in the next section, it should never give a sample size that is more than 3 too large, assuming that $c \leq 0.9975$.

4.2 Lower Bounds on Optimal Sample Size for Sampling without Replacement

Here is a simple proof that our bound (38) does not exceed $u_*(n, b, c)$ by too much. Interestingly, the amount that it exceeds $u_*(n, b, c)$ is largely independent of both n and b.

We now give a lower bound on our probability of failure, derived from equation (24), complementary to our previous upper bound (29):

$$\begin{aligned} e(n, b, u) &= \prod_{k=0}^{b-1} \frac{n-u-k}{n-k} \\ &= \prod_{k=0}^{b-1} \left(1 - \frac{u}{n-k}\right) \\ &\geq \left(1 - \frac{u}{n-b+1}\right)^b . \end{aligned}$$

Thus, our probability of failure is at least $1-c$ if

$$\left(1 - \frac{u}{n-b+1}\right)^b \geq 1 - c \, .$$

Solving for u, this is equivalent to

$$u \leq (n - (b-1)) \cdot (1 - (1-c)^{1/b}) \, .$$

Thus,

$$u_*(n,b,c) \geq (n-(b-1)) \cdot (1-(1-c)^{1/b}) \qquad (41)$$

Note the resemblance of this lower bound on u_* to the upper bound of inequality (38):

$$u_*(n,b,c) \leq (n-(b-1)/2) \cdot (1-(1-c)^{1/b}).$$

Now we can show that the bound (38) does not exceed $u_*(n,b,c)$ by much; the difference is at most

$$\frac{(b-1)}{2} \cdot (1-(1-c)^{1/b}). \qquad (42)$$

Note that this is independent of n. It is also effectively independent of b: Using elementary calculus, one can show that the difference (42) above is monotonically increasing in b and that

$$\lim_{b\to\infty} \left[\tfrac{b-1}{2} \cdot (1-(1-c)^{1/b})\right] = \frac{-\ln(1-c)}{2}$$

Thus, our bound $u_3(n,b,c)$ never exceeds $u_*(n,b,c)$ by more than $(-\ln(1-c))/2$, independent of n and b, and this quantity is less than 3 for all $c \leq 0.9975$. (It follows that $U_3(n,b,c) - U_*(n,b,c)$ is at most 3.)

Similar reasoning shows that our bound $u_1(n,b,c)$ never exceeds $u_*(n,b,c)$ by more than twice as much as $u_3(n,b,c)$ does: it is off by no more than $(-\ln(1-c))$, independent of n and b, and this quantity is less than 6 for all $c \leq 0.9975$.

In conclusion, we have a sample size

$$u_3(n,b,c) = \left(n - \frac{(b-1)}{2}\right) \cdot \left(1-(1-c)^{1/b}\right)$$

that is

- simple,
- provably "conservative" (an upper bound on $u_*(n,b,c)$),
- provably accurate (exceeding u_* by no more than $(-\ln(1-c))/2$ for all n, b, c), and
- empirically very accurate (see next table).

The following table demonstrates the accuracy of our formula for $n = 500$ (slightly larger than the number of precincts in a typical U.S. Congressional district), for $c = 0.95$ and $c = 0.99$, and various values of b. The "low" column gives the lower bound of equation (41), the "opt" column gives the optimal sample size $U_*(n,b,c)$, and the "up" column gives our upper bound $u_3(n,b,c)$.

		$c = 0.95$			$c = 0.99$		
n	b	low	opt	up	low	opt	up
500	1	475	475	475	495	495	495
500	2	388	388	388	450	450	450
500	5	224	225	225	299	300	300
500	10	128	129	129	182	183	183
500	20	67	69	69	99	101	101
500	50	27	28	28	40	42	42
500	100	12	14	14	19	21	21
500	200	5	6	6	7	9	10

5 Related Questions

This paper has largely been concerned with determining the size of a statistical audit u for a given universe of size n, desired fraud detectability level b, and desired confidence c. However, there are related "inverse" questions which are frequently asked that our bounds and techniques can usefully address.

For example, the size u of a statistical audit may be mandated by law (e.g., $u = 0.02n$ for 2% audit), and one may wish to know for this u and a given b what confidence level c one has in detecting corruption of b (or more) objects. This is the "confidence level" question.

Or, one may wish to know for this u and a given c the smallest number b of corrupted objects one can detect with confidence at least c. This is the "level of fraud detectability" question.

These two questions can be effectively answered using the bounds or techniques developed above. Essentially, the four variables n, u, b, and c are related by the equation

$$\binom{n-b}{u} \Big/ \binom{n}{u} = \binom{n-u}{b} \Big/ \binom{n}{b} = 1-c$$

and fixing any three of these variables, one can approximate the fourth.

Besides the "inverse" questions, our techniques can be adapted to handle various restrictions imposed on the auditing process. The proposed Holt bill [14], for instance, specifies that at least one precinct from each county needs to be audited.

We show how to answer the "inverse" as well as the restriction questions above using our bounds and techniques.

5.1 Estimating Confidence Levels

Given a universe of size n and a given audit size u, what confidence can one have in being able to detect one (or more) of b "bad" objects?

This confidence is given exactly by

$$c = d(n,b,u) = 1 - e(n,b,u). \qquad (43)$$

Much of Section 4 was effectively devoted to proving the following bounds on $e(n, b, u)$:

$$\begin{aligned}\left(1-\frac{u}{n-(b-1)}\right)^b &\leq e(n,b,u)\\ &\leq \left(1-\frac{u}{n-(b-1)/2}\right)^b .\end{aligned}$$

Applying these inequalities to equation (43), we obtain:

Upper and Lower Bounds on c

$$\begin{aligned} c &\geq 1-\left(1-\frac{u}{n-(b-1)/2}\right)^b\\ c &\leq 1-\left(1-\frac{u}{n-(b-1)}\right)^b\end{aligned}$$

The above inequalities may be useful, say, when considering legislation that mandates some fixed level u of auditing (see [7] as one example of this sort of consideration). This scenario is also useful for a "tiered" auditing approach [18]. In the tiered approach, thresholds for the margin of victory correspond to particular auditing percentages. For example, for a margin of victory of 1.75%, one should audit 5% of the precincts because it offers a confidence of 61% compared to 43% as given by a 3% auditing strategy [18]. In this case, our techniques are useful to compute the confidence level achieved when auditing a certain percent of precincts for a given margin of victory.

5.2 Estimating Level of Detectable Fraud

Given a universe of size n, a fixed audit size u, and a confidence level c, what is the smallest b for which can one detect one (or more) of b "bad" objects with confidence at least c?

While our original problem was solved by approximating the quantity

$$e(n,b,u) = \binom{n-u}{b} \Big/ \binom{n}{b},$$

this dual problem is best solved by approximating the equivalent quantity

$$e(n,b,u) = \binom{n-b}{u} \Big/ \binom{n}{u}.$$

Using the techniques developed in Section 4, one can derive the following analogous bounds on $e(n, b, u)$:

$$\begin{aligned}\left(1-\frac{b}{n-(u-1)}\right)^u &\leq e(n,b,u)\\ &\leq \left(1-\frac{b}{n-(u-1)/2}\right)^u .\end{aligned}$$

Setting $e(n, b, u) = 1 - c$ and solving for b, we obtain:

Upper and Lower Bounds on b

$$\begin{aligned} b &\geq (n-(u-1))\cdot(1-(1-c)^{1/u})\\ b &\leq (n-(u-1)/2)\cdot(1-(1-c)^{1/u})\end{aligned}$$

As before, one can show that these bounds are never different by more than $(-\ln(1 - c))/2$, which is less than 3 for all $c \leq 0.9975$.

One could then apply these results using relationship (3) to estimate what is the corresponding smallest margin of victory that one could confirm with an audit of the given size u, to the given confidence level c, in a straightforward manner.

5.3 Auditing with constraints

Some election systems might place constraints on the auditing process. In this section, we illustrate how one can employ our techniques in the case of the Holt bill when at least one precinct from each county needs to be audited (as specified in [14]).

Let z denote the number of counties and a_i the number of precincts in county i. We are still aiming for a total confidence of c, when sampling n precincts out of which b are corrupted. There are at least two ways to audit the precincts:

1. Select one precinct from each county uniformly at random and then employ our techniques on the rest of the precincts.

2. Employ our technique first, and then audit one precinct from each county that did not have any precincts audited.

In this case, we prefer the first approach because the latter can audit as many as $u(n, b, c) + z - 1$ if all the precincts audited with our technique belong to the same county.

Specifically, we propose the following procedure:

1. Audit at random one precinct from each county. If any corrupted vote is detected or the confidence achieved at this step (and computed below) is larger or equal to the overall confidence needed, we stop auditing.

2. Otherwise, we audit $u(n^*, b^*, c^*)$ from the remaining precincts using our formula (for instance, equation (32)), where n^*, b^*, and c^* are computed below.

At step 2, $b^* = b$ because no fraud was detected so far and $n^* = n - z$ because z precincts were already audited.

We show that $c^* = 1 - \frac{1-c}{(1-1/a_{max})^b}$ guarantees a global confidence level of at least c, where a_{max} is the maximum number of precincts in a county. For brevity, let us define the following events: A = "fraud goes undetected after conducting the procedure above," A_1 = "fraud is undetected at step 1," and A_2 = "fraud is undetected at step 2." We now have:

$$\mathbf{Pr}[A] = \mathbf{Pr}[A_1 A_2] = \mathbf{Pr}[A_1]\mathbf{Pr}[A_2 \mid A_1] \tag{44}$$

$$\begin{aligned} \mathbf{Pr}[A_1] &= \prod_{i=1}^{z}(1 - b_i/a_i) \\ &\leq \prod_{i=1}^{z}(1 - 1/a_i)^{b_i} \leq (1 - 1/a_{max})^b, \end{aligned}$$

where b_i is the number of corrupted precincts from the i-th county. We also used the fact that $(1 - b_i/a_i) \leq (1 - 1/a_i)^{b_i}$.

According to our formula in Section 4,

$$\mathbf{Pr}[A_2 \mid A_1] \leq 1 - c^* \leq \frac{1-c}{(1-1/a_{max})^b},$$

if we audit $u(n^*, b^*, c^*)$.

Therefore, using equation (44), the probability that corruption goes undetected is less than $1 - c$.

Note that the formulae for c^* makes sense because because when $1 - c > (1 - 1/a_{max})^b$ step 1 suffices to guarantee the desired confidence. Also, since $n^* < n$ and $c^* < c$, $u(n^*, b^*, c^*) + z < u(n, b, c) + z$; hence, this method is more efficient method than the first.

Furthermore, the number of precincts audited beyond the $u(n, b, c)$ that we would audit when no constraints exist is small as follows. We make use of the formulae in equation (32) for simplicity.

$$\begin{aligned} & z + u(n^*, b^*, c^*) - u(n, b, c) \\ &\quad = z + n - z - (n - z) \cdot \frac{(1-c)^{1/b}}{1 - 1/a_{max}} \\ &\qquad - n + n(1-c)^{1/b} \\ &\quad = \frac{(1-c)^{1/b}}{1 - 1/a_{max}} \cdot (n - n/a_{max} - n + z) \\ &\quad \leq z - n/a_{max}, \end{aligned} \tag{45}$$

where the inequality comes from the fact that $1-c \leq 1-(1 - 1/a_{max})^b$ since otherwise one would stop auditing at step 1. For a balanced distribution of the precincts in counties, $z - n/a_{max}$ is close to 0 which means that the additional number of precincts we audit with the new constraint is small.

A related question that may be encountered in elections is when the candidate that lost the election is allowed to audit z precincts of his/her choice. In this scenario, we have a similar choice: to use our techniques before or after the loser's auditing. If the loser audits first, we will next audit an additional number of $u(n - z, b, c)$ precincts. The sample space is reduced by z while the confidence required, c, stays the same because we do not know the loser's probability of detecting fraud (and thus assume it to be 0). On the other hand, if the loser picks after we employ our auditing techniques, we will have to audit $u(n, b, c)$. The sample space remains the same because the loser's strategy may be so bad that he/she selects only uncorrupted precincts. The first procedure turns out to be better once again.

6 Discussion

We note (as other authors have as well) that overly simple rules, such as "sample at a 1% rate", are not statistically justified in general. Using the Rule of Three, we see that a 1% sample rate is appropriate only when

$$t \leq 0.01n$$

or

$$3n/b \leq 0.01n$$

or

$$b \geq 300\ .$$

Since b is the total number of corrupted objects, we see that a 1% sampling rate may be inadequate when n is small, or the fraud rate is small... (Of course, the Rule of Three is only for sampling with replacement, but the intuition it gives carries over to the case of sampling without replacement.)

We hope that the rules presented here will provide useful guidance for those designing sampling procedures for audits.

Indeed, since the formula

$$U_3(n, b, c) = \lceil (n - (b-1)/2)(1 - (1-c)^{1/b}) \rceil \tag{46}$$

is so simple, so accurate, and always conservative, one could imagine just always using this sample size (instead of the optimal value), or writing this formula into election law legislation mandating audit sample sizes. Along with this formula, one could perhaps mandate use of equation (3) deriving the number b of bad objects to test from the apparent margin of victory m of the winner. (But it would probably be best to merely mandate a sample size sufficient to detect, with a specified level of confidence, any election fraud sufficient to have changed the outcome. In addition, one may wish to ensure that objects (e.g. precincts) with surprising or suspicious results also get examined.)

7 Acknowledgments

We thank Mike Alvarez and the anonymous reviewers for their helpful comments. We also gratefully acknowledge the support provided by the CalTech/MIT Voting Technology Project and NSF grants CCF-0418390 and IIS-0534482.

References

[1] Alvarez, R. M., Katz, J. N., and Hill, S. A. Machines versus humans: The counting and recounting of prescored punchcard ballots. Tech. Rep. VTP Working Paper #32, CalTech/MIT Voting Technology Project, September 2005. Available at: `http://www.vote.caltech.edu/media/documents/wps/vtp_wp32.pdf`.

[2] Ansolabehere, S., and Reeves, A. Using recounts to measure the accuracy of vote tabulations: Evidence from new hampshire elections 1946–2002. Tech. Rep. VTP Working Paper #11, CalTech/MIT Voting Technology Project, January 2004. Available at: `http://www.vote.caltech.edu/media/documents/wps/vtp_wp11.pdf`.

[3] Brennan Center Task Force on Voting System Security (Lawrence Norden, Chair). The machinery of democracy: Protecting elections in an electronic world, 2006. Available at: `http://www.brennancenter.org/programs/downloads/Full%20Report.pdf`.

[4] Commission, U. E. A. Election crimes: An initial review and recommendations for future study, December 2006. Available at: `http://www.eac.gov/docs/Voter%20Fraud%20&%20Intimidation%20Report%20-POSTED.pdf`.

[5] Cordero, A., Wagner, D., and Dill, D. The role of dice in election audits — extended abstract, June 16 2006. To appear at IAVoSS Workshop on Trustworthy Elections (WOTE 2006). Preliminary version available at: `http://www.cs.berkeley.edu/~daw/papers/dice-wote06.pdf`.

[6] Dodge, H. F., and Romig, H. G. *Sampling Inspection Tables: Single and Double Sampling (2nd ed).* Wiley, 1944.

[7] Dopp, K. Federal election audit costs based on 2002 and 2004 u.s. house and senate races, 2007. Available at: `http://electionarchive.org/ucvAnalysis/US/paper-audits/FederalAuditCosts.pdf`.

[8] Dopp, K., and Stenger, F. The election integrity audit, 2006. Available at: `http://electionarchive.org/ucvAnalysis/US/paper-audits/ElectionIntegrityAudit.pdf`.

[9] electionline.org. Case study: Auditing the vote, March 2007. Available at: `http://www.electionline.org/Portals/1/Publications/EB17.pdf`.

[10] for Justice, B. C. Policy brief on electronic voting systems, 2007. Available at: `http://www.brennancenter.org/stack_detail.asp?key=97&subkey=38339`.

[11] Hardy, G. H., Littlewood, J. E., and Pólya, G. *Inequalities*, second ed. Cambridge University Press, 1952.

[12] Herron, M. C., and Wand, J. Assessing partisan bias in voting technology: The case of the 2004 new hampshire recount, January 16 2006. Forthcoming in *Electoral Studies*, Available at: `http://www.dartmouth.edu/~herron/nh.pdf`.

[13] Hill, S. Election fraud references. Tech. rep., CalTech/MIT Voting Technology Project, July 2006. Available at: `http://vote.caltech.edu/media/documents/wps/vtp_wp50.pdf`.

[14] Holt, R. H.r. 811 – voter confidence and increased accessibility act of 2007, May 16 2007.

[15] Jovanovic, B. D., and Levy, P. S. A look at the rule of three. *American Statistician 51*, 2 (1997), 137–139.

[16] NEDA. Election integrity audit calculator. Available at: `http://electionarchive.org/auditcalculator/eic.cgi`.

[17] Neff, C. A. Election confidence—a comparison of methodologies and their relative effectiveness at achieving it (revision 6), December 17 2003. Available at: `http://www.votehere.net/papers/ElectionConfidence.pdf`.

[18] Norden, L. D. Testimony before the committee on house administration, subcommittee on elections, u.s. house of representatives, March 2007. Available at: `http://www.brennancenter.org/dynamic/subpages/download_file_48698.pdf`.

[19] (press release), C. V. F. Governor signs landmark bill to require public audits of software vote counts, October 11 2005. Available at: `http://www.calvoter.org/news/releases/101105release.html`.

[20] Rivest, R. L. On estimating the size of a statistical audit, November 14, 2006. Available at: `http://theory.csail.mit.edu/~rivest/Rivest-OnEstimatingTheSizeOfAStatisticalAudit.pdf`.

[21] Rivest, R. L. On auditing elections when precincts have different sizes, 2007. Available at: `http://theory.csail.mit.edu/~rivest/Rivest-OnAuditingElectionsWhenPrecinctsHaveDifferentSizes.pdf`.

[22] Saltman, R. G. Effective use of computing technology in vote-tallying. Tech. Rep. NBSIR 75–687, National Bureau of Standards (Information Technology Division), March 1975. Available at: `http://csrc.nist.gov/publications/nistpubs/NBS_SP_500-30.pdf`.

[23] Stanislevic, H. Random auditing of e-voting systems: How much is enough?, revision August 16, 2006. Available at: `http://www.votetrustusa.org/pdfs/VTTF/EVEPAuditing.pdf`.

[24] van Belle, G. *Statistical Rules of Thumb.* Wiley, 2002.

Notes

[1] For example, the calculator must be able to compute x^y given real numbers x and y or equivalently be able to do so with the logarithm and exponential functions via $x^y = \exp(\ln(x) \cdot y)$.

[2] While our test does not explicitly accept the null hypothesis, i.e., that fraud sufficient to change the outcome of the election has occurred, one could view the unnecessary triggering of a wider investigation when a "bad" object is sampled as a kind of Type II error.

[3] The question of how to pick objects "randomly" in a publicly verifiable and trustworthy manner is itself a very interesting one; see Cordero et al. [5] for an excellent discussion of this problem.

[4] This is in keeping with the hypothesis testing framework described in Section 2. One may also ask the question of *how many* "bad" objects are likely to be seen in such a sample. The number of sampled "bad" objects follows the well-known hypergeometric distribution.

[5] We also developed other formulae – such as

$$n \cdot (1 - (1 - c)^{-1/(n \ln(1 - \frac{b}{n}))}) + 1$$

which we could prove to be an upper bound on optimal sample size; the current paper only reports on the most useful such bounds.