CALTECH / MIT
VOTING TECHNOLOGY PROJECT
A multi-disciplinary, collaborative project of
the California Institute of Technology – Pasadena, California 91125 and
the Massachusetts Institute of Technology – Cambridge, Massachusetts 02139

ON ESTIMATING THE SIZE AND CONFIDENCE OF A STATISTICAL AUDIT

Ronald L. Rivest, Raluca A. Popa
MIT
Javed A. Aslam
Northeastern University

Key words: ***statistical audit, statistical sampling, auditing elections***

VTP WORKING PAPER #54
April 2007

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

April 22, 2007

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 whether a given set of n objects contains b or more "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 simple 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 whether a given number b of them (or more) 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, al-

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

though perhaps not particularly well-known. Jovanovic and Levy [7] discuss the Rule of Three, its derivation, and its application to clinical studies. See also van Belle [15].

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 [11], who attempted to correct for the bias in the Rule of Three due to sampling with replacement instead 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 two related "inverse" questions: determining the confidence level for a given audit size and level of fraud one wishes to detect, and determining the minimize amount of fraud one can detect for a given audit size with a given confidence level.

1.1 Related Work

Saltman [13, 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 [3], developed in the 1930's and first published in 1940, provide generalizations of the simple sampling methods used here.)

Previous work by Neff [9] is noteworthy, particularly with regard to the economies resulting from having a larger universe of many smaller, easily-testable, objects. Brennan Center report [1, Appendix J] gives some simple estimation formula, based on sampling with replacement. An excellent report [5] 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 [8]. Stanislevic [14] 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 [10]. 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 be a precinct; it could also be a voting machine or an individual ballot, depending on the situation; the math is the same.

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 whether the adversary has committed a "large" amount of fraud.

(With another standard formulation, 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 recount of the corresponding voter-verified paper ballots. The precinct is judged to be good if the results are equal. Of course, there may easily be explanations for a discrepancy other than malicious behavior; such explanations might be determined with further investigation. Nonetheless, for our purposes, we'll simply assume that each object tested is found to be "good" or "bad."

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, whether a *large* amount of fraud has 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, if the number of corrupted objects is 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".)

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 [5]. Here, b is the number of precincts that an adversary would have had to corrupt to swing the election. If we assume (as is 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

$$f = m/(2s) = 2.5m \tag{2}$$

of the precincts—or, equivalently,

$$b = mn/(2s) = 2.5mn \tag{3}$$

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.

See Saltman [13], Stanislevic [14], or Dopp et al. [5] 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 [12] 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 [7].

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:

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

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.[2]

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.

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:

$$d(n,b,u) \;=\; 1 \;-\; e(n,b,u) \tag{19}$$

[2] 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. [2] for an excellent discussion of this problem.

$$= 1 - \prod_{k=0}^{u-1} \frac{n-b-k}{n-k} \ . \qquad (20)$$

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. [13, 9, 5]).

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 an overestimate.

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

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

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:

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

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 which is 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

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

The formula for $u_1(n,b,c)$ is the same as the that proposed by Rivest [11] 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:

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

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 [6]

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

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

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

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

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

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_*

$$u_*(n,b,c) \leq u_2(n,b,c) \tag{36}$$

where

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

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

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

This gives our third and final upper bound:

Third Upper Bound on u_*

$$u_*(n,b,c) \leq u_3(n,b,c) \tag{39}$$

where

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

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.[3] 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

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

[3] 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.

Thus,

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

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

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

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

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

that is

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

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 b 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. We show how to answer the two 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

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

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 [4] as one example of this sort of consideration).

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.

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

$$U_3(n,b,c) = \lceil (n-(b-1)/2)(1-(1-c)^{1/b}) \rceil \quad (44)$$

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 objects 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.)

References

[1] 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`.

[2] Arel Cordero, David Wagner, and David Dill. 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`.

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

[4] Kathy Dopp. 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`.

[5] Kathy Dopp and Frank Stenger. The election integrity audit, 2006. `http://electionarchive.org/ucvAnalysis/US/paper-audits/ElectionIntegrityAudit.pdf`.

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

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

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

[9] C. Andrew Neff. 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`.

[10] California Voter Foundation (press release). 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`.

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

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

[13] Roy G. Saltman. Effective use of computing technology in vote-tallying. Technical Report 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`.

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

[15] Gerald van Belle. *Statistical Rules of Thumb*. Wiley, 2002.

A Appendix

A.1 Description of empirical results

In this appendix, we illustrate the use of our final upper bound on the optimal sample size, and we compare it to the optimal number of objects in a sample as well as to the lower bound we derived in Section 4. All numbers in this appendix refer to sampling without replacement.

Each table contains values for a different number n of objects, where n takes representative values from 2 to $10,000$. Within a table, each row considers a different value of b, the number of "bad" objects. In each table there are two sections, one for a confidence level of 0.95 and the other for a confidence level of $c = 0.99$. Within each section, there are three columns containing the lower bound on the optimal number of objects in a sample

$$\lceil (n - (b-1)) \cdot (1 - (1-c)^{1/b}) \rceil,$$

the optimal number $U_*(n, b, c)$ of elements in a sample, and our final upper bound

$$\lceil (n - (b-1)/2) \cdot (1 - (1-c)^{1/b}) \rceil,$$

respectively.

Note the accuracy of our final upper bound, the formula we suggest be used in practice. Over the entire range of n, b, and c values shown, this upper bound exceeds the optimal value in only four out of 156 cases, and in each of those four cases, it exceeds the optimal value by only 1.

A.2 Charts of optimal and estimated sample sizes

		$c = 0.95$			$c = 0.99$		
n	b	low	opt	up	low	opt	up
2	1	2	2	2	2	2	2

		$c = 0.95$			$c = 0.99$		
n	b	low	opt	up	low	opt	up
5	1	5	5	5	5	5	5
5	2	4	4	4	4	4	5

		$c = 0.95$			$c = 0.99$		
n	b	low	opt	up	low	opt	up
10	1	10	10	10	10	10	10
10	2	7	8	8	9	9	9
10	5	3	4	4	4	5	5

		$c = 0.95$			$c = 0.99$		
n	b	low	opt	up	low	opt	up
20	1	19	19	19	20	20	20
20	2	15	16	16	18	18	18
20	5	8	9	9	10	11	11
20	10	3	4	5	5	6	6

		$c = 0.95$			$c = 0.99$		
n	b	low	opt	up	low	opt	up
50	1	48	48	48	50	50	50
50	2	39	39	39	45	45	45
50	5	21	22	22	28	29	29
50	10	11	12	12	16	17	17
50	20	5	6	6	7	9	9

		$c = 0.95$			$c = 0.99$		
n	b	low	opt	up	low	opt	up
100	1	95	95	95	99	99	99
100	2	77	78	78	90	90	90
100	5	44	45	45	58	59	59
100	10	24	25	25	34	36	36
100	20	12	13	13	17	19	19
100	50	3	5	5	5	7	7

		$c = 0.95$			$c = 0.99$		
n	b	low	opt	up	low	opt	up
200	1	190	190	190	198	198	198
200	2	155	155	155	180	180	180
200	5	89	90	90	118	120	120
200	10	50	51	51	71	73	73
200	20	26	27	27	38	40	40
200	50	9	11	11	14	16	16
200	100	3	5	5	5	7	7

		$c = 0.95$			$c = 0.99$		
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

		$c = 0.95$			$c = 0.99$		
n	b	low	opt	up	low	opt	up
1000	1	950	950	950	990	990	990
1000	2	776	777	777	900	900	900
1000	5	449	450	450	600	601	601
1000	10	257	258	258	366	368	368
1000	20	137	138	138	202	204	204
1000	50	56	57	57	84	86	86
1000	100	27	29	29	41	43	43
1000	200	12	14	14	19	21	21
1000	500	3	5	5	5	7	7

		$c = 0.95$			$c = 0.99$		
n	b	low	opt	up	low	opt	up
2000	1	1900	1900	1900	1980	1980	1980
2000	2	1553	1553	1553	1800	1800	1800
2000	5	900	901	901	1202	1203	1203
2000	10	516	517	517	735	737	737
2000	20	276	277	277	408	410	410
2000	50	114	115	115	172	174	174
2000	100	57	58	58	86	88	88
2000	200	27	29	29	41	44	44
2000	500	9	11	11	14	16	17
2000	1000	3	5	5	5	7	7

		$c = 0.95$			$c = 0.99$		
n	b	low	opt	up	low	opt	up
5000	1	4750	4750	4750	4950	4950	4950
5000	2	3882	3882	3882	4500	4500	4500
5000	5	2252	2253	2253	3008	3009	3009
5000	10	1292	1294	1294	1842	1844	1844
5000	20	693	695	695	1025	1027	1027
5000	50	288	290	290	436	438	438
5000	100	145	147	147	221	223	223
5000	200	72	73	73	110	112	112
5000	500	27	29	29	42	44	44
5000	1000	12	14	14	19	21	21
5000	2000	5	6	6	7	10	10

		$c = 0.95$			$c = 0.99$		
n	b	low	opt	up	low	opt	up
10000	1	9500	9500	9500	9900	9900	9900
10000	2	7764	7764	7764	9000	9000	9000
10000	5	4506	4507	4507	6017	6018	6018
10000	10	2587	2588	2588	3688	3689	3689
10000	20	1389	1390	1390	2053	2055	2055
10000	50	579	581	581	876	878	878
10000	100	293	294	294	446	448	448
10000	200	146	148	148	224	226	226
10000	500	57	59	59	88	90	90
10000	1000	27	29	29	42	44	44
10000	2000	12	14	14	19	21	21
10000	5000	3	5	5	5	7	7