On Estimating the Size of a Statistical Audit

Ronald L. Rivest
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139
rivest@mit.edu

November 14, 2006*

Abstract

We develop a remarkably simple and easily-calculated estimate for the sample size necessary for determining whether a given set of n objects contains b or more "bad" objects:

$$n(1 - \exp(-3/b)) \tag{1}$$

(This is for sampling without replacement, and a confidence level of 95%). The basis for this estimate is the following procedure: (a) estimate the sample size t needed if sampling were to be done with replacement, (b) estimate the expected number u of distinct elements seen in such a sample, and finally (c) draw a sample of size u without replacement. This formula is also remarkably accurate: experiments show that for $n < 5000$, this formula gives results that are never too small (with some exceptions when $b = 1$), but are never too large by more than 4 (additively).

*The latest version of this paper can always be found at http://theory.csail.mit.edu/~rivest/Rivest-OnEstimatingTheSizeOfAStatisticalAudit.pdf

1 Introduction

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?

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 [5] discuss the Rule of Three, its derivation, and its application to clinical studies. See also van Belle [12].)

We then consider the question of how many *distinct* elements such a sample really contains.

We finally provide an Improved Rule of Three for use with sampling without replacement; it corrects for the bias in the Rule of Three due to sampling with replacement rather than 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.

Saltman [10, 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 [8] 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 [4] 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 [7]. Stanislevic [11] 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 [9].

This note is written for a fairly general audience.

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 that we are in 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, we assume the availability of a test or audit procedure that can determine whether a given object is good or bad. We assume this test procedure is always correct. For example, testing the results of a given voting machine may involve comparing the electronic results from the voting machine with a hand recount of voter-verified paper ballots. If the comparison turns out to be equal, then the machine is judged to be good; otherwise it is judged to be bad. Of course, there may easily be explanations for the discrepancy other than malicious behavior; such explanations might be determined with further investigation. Nonetheless, for our purposes here, we'll keep it simple and assume that each object tested is found to be "good" or "bad."

If we are trying to determine whether *any* fraud occurred, then we would clearly need to test *all* objects in the worst case. Here we are willing to sacrifice the ability to detect *any* fraud, and only detect with high confidence whether or not a *large* amount of fraud has occurred, in return for having to examine only a statistical sample of the objects; this is the usual notion of 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 is trying to avoid detection, he will corrupt as few object as possible, consistent with achieving his evil goals. We assume that corrupting a number b of objects is sufficient to achieve his goals, and so we may assume for our analysis that the adversary doesn't try to corrupt more than b objects.

We let

$$f = b/n \tag{2}$$

denote the fraction of bad objects we wish to detect. In this note we call f the "fraud rate." Given one of b or f, the other is determined, via equation (2).

In a voting context, the value of b might be the requisite number of precincts that the adversary would have to corrupt to swing the election. If, for example, you assume (as is reasonable) that the adversary wouldn't dare to change more than 20% of the votes in any one precinct, and that the "winner" won by a margin of v of the votes (where $0 \leq v \leq 1$), then the adversary would need to change a fraction

$$f = 2.5v \tag{3}$$

of the precincts—or, equivalently,

$$b = 2.5vn \tag{4}$$

precincts. (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 v of the votes cast by the alleged winner must have involved at least a fraction $v/(2 * 0.20) = 2.5v$ of the precincts, since each precinct corrupted changes the difference in vote count between the top two candidates by 40% of the vote count of that precinct.) If the apparent winner has won by $v = 1\%$ in a county with 400 precincts, you would want to test for $b = 2.5vn = 10$ or more bad precinct counts. See Saltman [10], Stanislevic [11], or Dopp et al. [4] 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.

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".)

3 Sampling without replacement

We begin by reviewing the (well-known) math for determining the proper sample size, when sampling without replacement. Although this math is not complicated, the results seem to require the use of a computer program to determine the optimal sample size.

The rest of this paper is then devoted to determined ways of accurately estimating this optimal size, simple enough to be usable, if not by hand, at least with the use of only a calculator, with no computer needed. (Your calculator must be a "scientific" one, though, so you can compute the exponential function $\exp(x) = e^x$.)

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.

(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.)

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 at least one bad object is detected is:

$$\begin{aligned} d(n,b,u) &= 1 - \binom{n-b}{u} \Big/ \binom{n}{u} && (5) \\ &= 1 - \prod_{k=0}^{u-1} \frac{n-b-k}{n-k} \ . && (6) \end{aligned}$$

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 $d(n,b,u)$ greater than c:

$$u_*(n,b,c) = \min\{u \mid d(n,b,u) > c\} \ . \qquad (7)$$

Equations (5)–(7) are not new here; they have been given and studied by others (e.g. [10, 8, 4]).

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

$$u_*(400, 10, 0.95) = 103 \; ; \tag{8}$$

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 may nonetheless be valuable 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.

The rest of this note is devoted to this purpose.

We do so by first considering the same problem, but when sampling *with* replacement. We then "correct" for the changes made by this assumption.

4 Sampling with replacement and the Rule of Three

Here now is a simple "rule of thumb" that is easily remembered; it applies when we are sampling with replacement. Since we are now 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—this is analogous to the optimal sample size $u_*(n, b, c)$ for sampling without replacement).

Rule of Three:
Test enough objects so that, for the fraud level you are trying to detect, you expect to see *at least three* corrupted objects among those examined (via sampling with replacement). That is, ensure that:

$$ft \geq 3. \tag{9}$$

or equivalently, ensure that:

$$t \geq 3n/b\ . \tag{10}$$

(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 at first, but is really to be expected. If you have available some well-mixed sand where most of the sand grains are white, but a fraction f of the grains are black, you may only need to sample a handful of the sand 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.

We now justify the Rule of Three (for a confidence level of 95% that a fraud rate of f or greater will be detected). (This analysis follows that given by Jovanovic and Levy [5].)

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

$$(1-f)^t\ . \tag{11}$$

If we want the chance that significant fraud goes undetected to be 5% or less, then we want

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

or equivalently:

$$t \geq \frac{\ln(0.05)}{\ln(1-f)} \tag{12}$$

Since

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

—isn't it so very nice that ln(20) is almost exactly 3?—equation (12) becomes

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

Using the well-known approximation

$$\ln(1-f) \approx -f \ , \tag{13}$$

which is quite accurate for small values of f, we can rewrite the bound on t from equation (12) as:

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

which can be rewritten as

$$t \geq \frac{3n}{b} \tag{14}$$

or equivalently as

$$ft \geq 3 \ . \tag{15}$$

Equation (15) 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.*

You want the test set to be big enough that you expect to see that at least three corrupted test 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 $\mathbf{Pr}[X = 0] = e^{-\lambda} < e^{-3} = 0.04978\ldots$.)

With our running example, we have $n = 400$, $b = 10$, and thus $f = b/n = 0.025$; the Rule of Three says to pick a sample of size $3n/b = 3{*}400/10 = 120$. While this estimate is about 17% larger than the optimal value of 103 that we computed earlier for sampling without replacement, it is nonetheless not too bad for an estimate you can compute in your head.

This "Rule of Three" ($t > 3n/b$) is thus simple enough for some practical guidance.

It is also easily adjusted. For example, for a 99% chance of detecting fraud, we can similarly use the "**Rule of Five**":

$$\begin{aligned} ft &\geq 5 \\ &\geq -\ln(0.01) \approx 4.6 \ . \end{aligned}$$

We could call it the "Rule of 4.6", but the name "Rule of Five" is easier to remember.... For a confidence level of 99%, we should thus test enough objects so that, for the fraud level we are trying to detect, we expect to see *at least five* corrupted objects among those examined.

However, in practice one samples without replacement, instead of sampling with replacement, so a sample size derived by assuming sampling with replacement is going to be an overestimate—in some cases a serious overestimate. Nonetheless the Rule of Three, which can be applied in one's head, provides an easy "first rough guess" of the sample size that might be needed in practice.

To summarize this section, for general c, we have the following formula for the optimal sample size $t_*(n, b, c)$, when sampling with replacement:

$$\begin{aligned} t_*(n, b, c) &= \frac{\ln(1-c)}{\ln(1-f)} && (16) \\ &= \frac{\ln(1-c)}{\ln(1-b/n)} && (17) \end{aligned}$$

or, using equation (13), we get the generalized form of the Rule of Three as an approximation:

$$t_1(n, b, c) = \frac{-n\ln(1-c)}{b} \ . \qquad (18)$$

(Here we ignore the fact that a sample size must be integral; in practice one can just round the values up to the next integer if necessary.)

5 Adjusting for Sampling without Replacement

In this section we propose a means for "correcting" the estimate given by the Rule of Three, to account for the fact that in practice one samples without replacement, instead of with replacement. The “correction” replaces the estimate from the Rule of Three with an estimate of the number of distinct objects seen when drawing (with replacement) a sample of the size suggested by the Rule of Three. We call this modification the Improved Rule of Three. There is no rigorous justification for the accuracy of this heuristic, but it seems intuitively well-motivated, and experiments show it to be in fact very accurate.

Suppose that we draw with replacement a sample of size t from a universe of size n; how many *distinct* elements do we expect to see? Let $s(n, t)$ denote this number. (As a mnemonic, $s(n, t)$ is the size of the $\underline{s}$et that is the support for the multiset drawn of size t.)

The function $s(n, t)$ is well studied. (The usual way of formulating the question in the literature is: suppose one throws t balls into n bins uniformly at random; how many bins remain empty? The expected value of this quantity is $n - s(n, t)$ in our notation.) Kolchin et al. [6, page 5,Theorem 1] give the results (where $r = t/n$):

$$s(n,t) \geq n(1 - \exp(-r)) \text{ , and} \tag{19}$$

$$s(n,t) = n(1 - \exp(-r)) + \frac{r}{2}\exp(-r) - O(\frac{r(1+r)\exp(-r)}{n}) \ . \tag{20}$$

We will ignore the last two terms of this equation, as they are small, and let

$$\hat{s}(n,t) = n(1 - \exp(-t/n)) \ . \tag{21}$$

Then $\hat{s}(n, t)$ is our estimate of the expected number $s(n, t)$ of distinct elements in a sample of size t drawn (with replacement) from a set of size n.

The approximation (21) is very accurate, but is always a slight underestimate. The largest term in the power series for the error of this approximation is $t/2n$. The estimation error is never more than $1/e = 0.367\ldots$; this occurs when $n = t = 1$. For a given n, the maximum error occurs when $t = n$; as n gets large, this maximum error converges to $1/2e = 0.1839\ldots$. Thus, the approximate formula could be improved slightly by adding $0.1839\ldots$, while still remaining an underestimate, or instead by adding $0.3678\ldots$, to yield an upper bound:

$$\hat{s}(n,t) + 0.1839\ldots \leq s(n,t) \leq \hat{s}(n,t) + 0.3678\ldots \ . \tag{22}$$

For example, when drawing (with replacement) a sample of size $t = 120$ from a universe of size $n = 400$ (following our running example), we don't expect to see 120 distinct objects; there will be some repetitions. The number of distinct objects we expect to see is approximately

$$\begin{aligned} \hat{s}(400, 120) &= 400(1 - \exp(-120/400)) \\ &= 103.67 \; ; \end{aligned}$$

approximately $120 - 104 = 16$ of the 120 objects will have been "repeats."

The qualitative behavior of $s(n, t)$ for fixed t as n increases is quite simple. When n is much smaller than t, $s(n, t)$ is approximately n—we expect to see almost all of the elements of the universe in the sample. When $n = t$, we are at the knee of the curve, and $s(n, t) \approx 0.632n$. As n increases beyond t, $s(n, t)$ approaches t asymptotically:

$$\lim_{n \to \infty} s(n, t) = t \; . \tag{23}$$

Indeed, once n is larger than t^2 or so, we don't expect to see repeats in our sample, and so $s(n, t)$ is then very nearly t.

Given then a universe of size n, and a target number b of bad objects we wish to detect, instead of using the Rule of Three sample size (for a confidence level of 95%):

$$t = 3n/b$$

we suggest "correcting" this by using instead the number of *distinct* elements we expect to see in such a sample of size t, viz.:

$$\begin{aligned} u_1(n, b, 0.95) &= \hat{s}(n, t_1(n, b, 0.95)) \\ &= n(1 - \exp(-t_1(n, b, 0.95)/n)) \\ &= n(1 - \exp(-3/b)) \; . \end{aligned}$$

For a general confidence level c, we have

$$\begin{aligned} u_1(n, b, c) &= \hat{s}(n, t_1(n, b, c)) \\ &= n(1 - \exp(\ln(1 - c)/b)) \; . \end{aligned}$$

This then gives us our Improved Rule of Three:

Improved Rule of Three:
Test enough objects so that:

$$u \geq n(1 - \exp(-3/b)) \ , \tag{24}$$

where n is the size of the universe, u is the number of objects to be tested, and b is the number of bad objects one wishes to detect, at a 95% confidence level, using sampling without replacement.

Although perhaps less "intuitive" than the original Rule of Three, the Improved Rule of Three gives very accurate guidance, and it provides significant savings when t/n gets large (say, more than 10%). There is no rigorous justification for the Improved Rule of Three, although we shall provide some empirical justification for its accuracy.

When $b < 3\sqrt{n}$, the original Rule of Three in equation (14) gives that $t_1(n, b, 0.95) > \sqrt{n}$, so that repetitions may be common, and Improved Rule of Three (u_1) will indeed be an improvement over the original Rule of Three. However, when b exceeds $3\sqrt{n}$, then the Improved Rule of Three and the original Rule of Three will give nearly identical results.

For our running example ($n = 400$,$b = 10$), equation (24) says to use

$$\begin{aligned} u_1(400, 10, 0.95) &= 400(1 - \exp(-3/10)) \\ &= 103.67 \ . \end{aligned}$$

This is within 0.67 of the correct answer (equation (8))!

For a confidence level of 99%, use a sample of size:

$$u_1(n, b, 0.99) = n(1 - \exp(-4.6/b)) \ . \tag{25}$$

Empirically, for $1 \leq n \leq 5000$ and all b, $1 \leq b \leq n$, we always have

$$u_*(n, b, 0.95) - 1 \leq u_1(n, b, 0.95) \leq u_*(n, b, 0.95) + 4 \tag{26}$$

so that the Improved Rule of Three is seen to be exceptionally accurate (always within 4 of the correct value for this range of n values). Moreover, it appears almost never to *under*estimate the required sample size; there are a few exceptional cases when $b = 1$ when the estimate u_1 is one less than u_*. (These exceptional cases go away if we add 0.3678...to the estimate on the right hand side of inequality (24) in line with the bound in (22).)

For $c = 0.95$ and $n \leq 5000$, this approximation is one too small about 0.0007% of the time, correct about 0.09% of the time, one too large about 29.96% of the time, two too large about 65.14% of the time, and three too large about 4.79% of the time. The approximation is only occasionally too small, and then only when $b = 1$; by increasing the approximation by adding 0.3678, in line with bound (22) the approximation never seems to be too small, empirically.

A closely related estimate

$$\begin{aligned} u_2(n, b, 0.95) &= s(t_*(n, b, 0.95)) \\ &\approx \min(n, \lceil n * (1 - \exp(-3/\ln(1 - b/n))) + 1 \rceil) \end{aligned}$$

yields very similar experimental results—the estimate always appears to be conservative, but may be a little too large. (This estimate is based on plugging equation (12) instead of equation (14) into equation (21), after adding 1 to ensure that the estimate is conservative, and taking the minimum with n to ensure that the result is not too large.)

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.)

The analysis of this paper doesn't take into account the possibility that different objects have different size or weight. For example, different voting precincts may have different numbers of voters. This complicates matters considerably. Stanislevic [11] has a good approach to handling this situation.

The empirical bounds given in equation (26) are not proven to hold in general; it would be interesting to provide proofs (if indeed the bounds hold in general).

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

Indeed, since the formula

$$n(1 - \exp(-3/b)) \tag{27}$$

is so simple, so accurate, and yet (almost) always appears to be 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. (To make it conservative, it appears to suffice to add 0.3678 to this formula.) Along with this formula, one could perhaps mandate use of equation (4) deriving the number of bad objects to test for from the apparent margin of victory; the result says to sample

$$n(1 - \exp(-1.2/vn)) \tag{28}$$

objects, where v is the apparent margin of victory 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.)

Acknowledgments

I'd like to thank Kathy Dopp, David Karger, Avi Rubin, Doug Jones, Harry Khamis, and Howard Stanislevic for helpful comments and pointers.

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

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

[6] Valentine F. Kolchin, Boris A. Sevast'yanov, and Vladimir P. Chistyakov. *Random Allocations*. V. H. Winston & Sons (Washington, D. C.), 1978. (translated from Russian) Distributed by Halsted Press, a Division of John Wiley & Sons, Inc.

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

[8] 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`.

[9] 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`.

[10] 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`.

[11] 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`.

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

Appendix A.

In this Appendix, we illustrate the use of our proposed estimate, and compare it to the optimal sample size.

Each table is for a different number n of objects, from $n = 2$ to $n = 10,000$.

Within a table, each row considers a different value of b, the number of bad objects we wish to detect.

There are in each table two sections of two columns each, one for a confidence level of $c = 0.95$ and one for a confidence level of $c = 0.99$. Within each column we give the optimal number $u_*(n, b, c)$ of elements in a sample (drawn without replacement), and also our estimate

$$u_1(n, b, c) = n(1 - \exp(\ln(1 - c)/b))$$

of the number of elements in a sample (again, drawn without replacement). Values are shown rounded up to the next integer, as necessary.

Note the accuracy of the proposed estimate, over the entire range of values n, b, and c. Note also that our estimate u_1 is almost always conservative (it is almost never less than the optimal value u_*); in the charts it is only too small for $b = 1$ and $n = 5000$, $n = 10000$. These exceptions disappear if we add 0.3678 to the estimate.

		$c = 0.95$		$c = 0.99$	
n	b	opt	est	opt	est
2	1	2	2	2	2

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

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

		$c = 0.95$		$c = 0.99$	
n	b	opt	est	opt	est
20	1	19	19	20	20
20	2	16	16	18	18
20	5	9	10	11	13
20	10	4	6	6	8

		$c = 0.95$		$c = 0.99$	
n	b	opt	est	opt	est
50	1	48	48	50	50
50	2	39	39	45	45
50	5	22	23	29	31
50	10	12	13	17	19
50	20	6	7	9	11

		$c = 0.95$		$c = 0.99$	
n	b	opt	est	opt	est
100	1	95	95	99	99
100	2	78	78	90	90
100	5	45	46	59	61
100	10	25	26	36	37
100	20	13	14	19	21
100	50	5	6	7	9

		$c = 0.95$		$c = 0.99$	
n	b	opt	est	opt	est
200	1	190	190	198	198
200	2	155	156	180	180
200	5	90	91	120	121
200	10	51	52	73	74
200	20	27	28	40	42
200	50	11	12	16	18
200	100	5	6	7	10

		$c = 0.95$		$c = 0.99$	
n	b	opt	est	opt	est
500	1	475	475	495	495
500	2	388	389	450	450
500	5	225	226	300	301
500	10	129	130	183	185
500	20	69	70	101	103
500	50	28	30	42	44
500	100	14	15	21	23
500	200	6	8	9	12

		$c = 0.95$		$c = 0.99$	
n	b	opt	est	opt	est
1000	1	950	950	990	990
1000	2	777	777	900	900
1000	5	450	451	601	602
1000	10	258	259	368	370
1000	20	138	140	204	206
1000	50	57	59	86	88
1000	100	29	30	43	46
1000	200	14	15	21	23
1000	500	5	6	7	10

		$c = 0.95$		$c = 0.99$	
n	b	opt	est	opt	est
2000	1	1900	1900	1980	1980
2000	2	1553	1553	1800	1800
2000	5	901	902	1203	1204
2000	10	517	518	737	739
2000	20	277	279	410	412
2000	50	115	117	174	176
2000	100	58	60	88	91
2000	200	29	30	44	46
2000	500	11	12	16	19
2000	1000	5	6	7	10

		$c = 0.95$		$c = 0.99$	
n	b	opt	est	opt	est
5000	1	4751	4750	4951	4950
5000	2	3882	3882	4500	4500
5000	5	2253	2254	3009	3010
5000	10	1294	1295	1844	1846
5000	20	695	696	1027	1029
5000	50	290	291	438	440
5000	100	147	148	223	226
5000	200	73	75	112	114
5000	500	29	30	44	46
5000	1000	14	15	21	23
5000	2000	6	8	10	12

		$c = 0.95$		$c = 0.99$	
n	b	opt	est	opt	est
10000	1	9501	9500	9901	9900
10000	2	7764	7764	9000	9000
10000	5	4507	4508	6018	6019
10000	10	2588	2589	3689	3691
10000	20	1390	1392	2055	2057
10000	50	581	582	878	880
10000	100	294	296	448	451
10000	200	148	149	226	228
10000	500	59	60	90	92
10000	1000	29	30	44	46
10000	2000	14	15	21	23
10000	5000	5	6	7	10