Sequential Sampling Inspection Could Save Money: A Case in Connecticut in Point

Nitis Mukhopadhyay

Department of Statistics
University of Connecticut
CLAS Building, , U-4120
215 Glenbrook Road
Storrs, CT 06269-4120, U.S.A.
(e-mail: mukhop@uconnvm.uconn.edu)

Abstract. The State of Connecticut bought 10,000 computers/servers from a contracted supplier. These were supposed to include some special internal hardware. The technology department inspected 4,000 pieces from the delivered batch and found only 58 "good" ones! It is shown that the inspection protocol that allowed checking 4,000 computers was at best outrageously wasteful. An appropriately designed strategy with fewer than 10% inspections could conclude with near certainty that the batch was far below expectation.
Keywords: Inspection protocol, Inspection sampling, Percentage saving, Sampling strategy.

1 Introduction

On Tuesday, June 8, 2004, the Hartford Courant's Connecticut section's headline read "Woman Accused of Bilking State" which drew my immediate attention. I became intrigued as I read the story, "... In March 2001, the Computer Plus Center won a $17.2 million state contract, making it the exclusive vendor of Dell computers and servers for all state agencies, the arrest affidavit states. In January 2003, the state Department of Information Technology filed a complaint about the company, ... announcing the arrest. The servers would not work, according to the affidavit." The article quoted Chief State's Attorney Christopher Morano saying, "The servers that were delivered did not have the amount of memory, or the quality memory, in them, that was required." The article then went on to report, "... The state's technology department took apart 4,000 of the 10,000 (computers/servers) delivered by the said company. Of those, Morano said, only 58 contained the required network interface cards, ... "

This note is not about allegations or legal posturing. I was struck by the fact that Connecticut's technology department took apart 4,000 from 10,000 computers/servers delivered by the said company whereas only 58 good items were found! Given this batch, a moot question is this: Was it possible to come to the conclusion of alleged fraud by inspecting a small fraction of $10,000$ items? The answer is, 'of course, yes'. I will substantiate this with appropriately designed random sampling strategies to gather just the right amount of information much sooner.

Now, suppose that the State's technology department could come to the same conclusion of alleged fraud by inspecting n items (computers) where n was decisively "small" compared with $4,000$. Also, suppose that the inspection per computer takes x minutes and a skilled technologist charges \$$a$ per x minutes of inspection. When an item is checked out, it is out of commission so that the State loses \$$b$ per piece per inspection. A technologist will probably be paid \$$c$ for the mileage and per diem on an average per inspection assuming that these 10,000 computers are scattered in different locations. Then, we have:

$$\begin{aligned} \text{Savings: } SV &= \$(a+b+c)(4000-n), \\ \text{Percentage Savings: } PSV &= \left(1-\tfrac{n}{4000}\right)\times 100. \end{aligned} \quad \%tag1 \qquad (1)$$

Let us throw in some realistic numbers. For example, suppose that $a = 150, b = 50, c = 10$ and the savings would amount to $\$735,000$ or $\$420,000$ if one could arrive at the conclusion of alleged fraud by inspecting only 500 or $2,000$ computers respectively. These savings could be in cash or kind, for example, in the form of savings from cost-share or overtime payments.

There are other expenses too when a computer is inspected. For example, there is cost for electricity and for storage of non-functioning computers. Also, the supplier was already paid and the State "lost" interest income from that fund! Then, waiting for a year or more to bring lawsuits against supplier(s) drains the State's resources even further. The term SV in (1) may not take into account all kinds of costs borne by the State. Yet, one cannot deny that the term PSV from (1) portrays a realistic quantification of percentage savings regardless of the magnitudes of a, b, c and other costs involved.

2 A statistical formulation

We face a large population of $10,000(= R)$ items where each item is either 'good' or 'bad'. When an item is randomly selected, suppose that the probability that it is good (or bad) is p (or $q = 1-p$), $0 < p < 1$. The percentage of good items ($= 100p\%$) is assumed unknown.

Clearly, I can set the following lower and upper bounds for p:

$$0.0058 \approx \tfrac{58}{10000} \le p \le \tfrac{6058}{10000} \approx 0.6058 \qquad (2)$$

The lower (upper) bound for p in (2) is obtained by assuming that there were no (all) good items among $6,000$ remaining uninspected items. On the other hand, it appears that p should be closer to $\frac{58}{4000} \approx 0.0145$ rather than the most pessimistic value 0.0058 or the overly optimistic value 0.6058.

To estimate p, one would inspect n items selected randomly from the batch to check how many items ($= X$) out of n items are indeed good. Having a large population on hand, I treat X as an approximately binomial random variable with n and p. An estimator of p is

$$\widehat{p}_n = \frac{\#\text{ good items in the random sample}}{n}. \tag{3}$$

This $\widehat{p}_n$ has the following variance and estimated variance:

$$Var\left(\widehat{p}_n\right) = p(1-p)/n,\ \widehat{Var\left(\widehat{p}_n\right)} = \widehat{p}_n\left(1-\widehat{p}_n\right)/n, \tag{4}$$

by disregarding the finite population correction factor $1 - nR^{-1}$. See Sukhatme *et al.* (1984, p. 43).

Now, how many items (that is, n) should one inspect so that the standard confidence interval $\widehat{p}_n \pm E$ for p would have $100(1-\alpha)\%$ confidence? By appealing to the central limit theorem, the required sample size n is then approximately given by

$$n \equiv n(p) = \left(z_{\alpha/2}/E\right)^2 p(1-p). \tag{5}$$

Since p is unknown, one may opt for the maximum possible value of $n(p)$ that would work for all possible values of $p, 0 < p < 1$. This maximum occurs when $p = \frac{1}{2}$ which motivates the following expression for n:

$$n \equiv n_{\max} = \tfrac{1}{4}\left(z_{\alpha/2}/E\right)^2. \tag{6}$$

α / $z_{\alpha/2}$	E						
	0.10	0.08	0.05	0.02	0.016	0.012	0.01
	$\mathbf{n}_{\max}$ **values:**						
0.10 / 1.645	68	108	271	1692	3007	4699	6766
	$\mathbf{n(p)}$ **values:**						
$\mathbf{p = 0.2}$	43.3	67.7	173.2	1082.4	1691.3	3006.7	4329.6
$\mathbf{p = 0.1}$	24.4	38.1	97.4	608.9	951.3	1691.3	2435.4
$\mathbf{p = 0.0145}$	3. 87	6.04	15.5	96.7	151.1	268.5	386.7
	$\mathbf{n}_{\max}$ **values:**						
0.05 / 1.96	97	151	385	2401	4269	6670	9604
	$\mathbf{n(p)}$ **values:**						
$\mathbf{p = 0.2}$	61.5	96.0	245.9	1536.6	24010	4268.4	6146.6
$\mathbf{p = 0.1}$	34.6	54.0	138.3	864.4	1350.6	2401.0	3457.4
$\mathbf{p = 0.0145}$	5.49	8.58	22.0	137.2	214.4	381.2	549.0

Table 1. Sample size $n(p)$ from (5) and $n_{\max}$ from (6).

The recommended expression $n_{\max}$ is used in many practical problems. For example, one may refer to Chase and Bown (2000, p. 330). But, $n_{\max}$ is rather conservative because it works for all p values across the board. In fact, $n_{\max}$ may be viewed as a minimax choice for the required sample size.

In table 1, the first (second) block corresponds to 90% (95%) confidence intervals with a particular E value. First, $n_{\max}$ values are provided and then for each fixed E, I provide $n(p)$ values for $p = 0.2, 0.1, 0.0145$. Note that $p = 0.0145$ corresponds to $\frac{58}{4000}$. From this table, I immediately summarize some features, namely:

(1) $n(p)$ goes up for fixed p if $E \downarrow$;

(2) $n(p)$ goes down for fixed E if $p \downarrow$;

(3) $n(p)$ goes down significantly compared with $n_{\max}$ for fixed E if $p \downarrow$.

3 Two-stage sampling to determine sample size

Recall that $\widehat{p}_n \pm E$ would have approximately $100(1-\alpha)\%$ confidence when the required sample size $n \equiv n(p)$ is approximated by the expression from (5). But, this expression involves unknown p to begin with! Hence, one must inspect items at least in two steps. This is called *two-stage* or *double sampling* strategy. See Stein (1945,1949), Ghosh *et al.* (1997, Chapter 6), and Mukhopadhyay and Solanky (1994, Chapter 2). Robbins and Siegmund (1974) and Mukhopadhyay and Cicconetti (2004) respectively proposed purely sequential and two-stage estimation strategies for p under various kinds of loss functions. One may also take a look at Corneliussen and Ladd (1970), Ghosh (1970), Wald (1947), and other sources.

I propose to inspect $m(\geq 2)$ pilot items and obtain only a preliminary estimate $\widehat{p}_m$ in the first stage. Here, $< u >$ stands for the largest integer $< u$. Now, I let

$$N = \max\left\{m, \left\{\left(\frac{t_{m-1,\alpha/2}}{E}\right)^2 \left(\widehat{p}_m + m^{-1}\right)\left(1 - \widehat{p}_m - m^{-1}\right)\right\} + 1\right\} \qquad (7)$$

where $t_{m-1,\alpha/2}$ is the upper $50\alpha\%$ point of the Student's t distribution with $m-1$ degrees of freedom. If one believes that p is rather small, then $\widehat{p}_m$ may be zero and hence $\widehat{p}_m$ is replaced by $\widehat{p}_m + m^{-1}$ in (7).

If $N = m$, there will be no need for more inspections beyond the pilot stage. But, if $N > m$, then I propose to inspect $N - m$ additional items and record the number of good items in the second stage. The final confidence interval estimator for p is going to be $\widehat{p}_N \pm E$ where

$$\widehat{p}_N = \frac{\#\text{ good items in the combined random sample from both stages}}{N}. \qquad (8)$$

p $\frac{58r}{4000}$	Observed sample sizes in ten replications	% Savings PSV $100\left(1-\frac{\text{Ave }N}{4000}\right)\%$
$r=1$	$378, 620, 620, 254, 378,$ $128, 500, 254, 254, 128$ $\overline{N}=351.4$, $\mathrm{S_N}=181.4$, $\widetilde{N}=316$	91.22%
$r=2$	$620, 966, 853, 620, 254,$ $620, 378, 378, 254, 620$ $\overline{N}=556.3$, $\mathrm{S_N}=240.0$, $\widetilde{N}=620$	86.1%
$r=3$	$737, 966, 500, 500, 128,$ $853, 378, 853, 853, 620$ $\overline{N}=638.8$, $\mathrm{S_N}=262.9$, $\widetilde{N}=678.5$	84.0%
$r=10$	$2165, 1986, 2251, 2251, 1893,$ $1893, 2498, 2251, 2417, 2165$ $\overline{N}=2177.0$, $\mathrm{S_N}=204.2$, $\widetilde{N}=2208.0$	45.6%
$r=41.779$	$3896, 3896, 3850, 3974, 3541,$ $3663, 3824, 3796, 3734, 3963$ $\overline{N}=3813.7$, $\mathrm{S_N}=136.1$, $\widetilde{N}=3837$	4.7%

Table 2. Values of N using (7) from ten replications with 2% over-sampling on an average compared with n(p) from (5) and $\alpha=0.05$, $m=124$.

By the way, Mukhopadhyay (2004) gave a practical way to determine the pilot sample size m as follows:

$$m = \text{smallest positive integer such that } t^2_{m-1,\alpha/2}/z^2_{\alpha/2} \leq 1+\varepsilon, \tag{9}$$

assuming that one can comfortably entertain $100\varepsilon\%$ over-sampling on an average compared with $n(p)$ from (5). Then, one arrives at the following choice for the pilot sample size m depending upon ε:

$$m = \left\langle \frac{1}{2\varepsilon}\left(\frac{1}{2}(z^2_{\alpha/2}+1)+\left\{2\left[\frac{1}{3}z^4_{\alpha/2}+\frac{23}{12}z^2_{\alpha/2}+\frac{5}{4}\right]\varepsilon\right\}^{1/2}\right)\right\rangle + 1. \tag{10}$$

Now, in order to have a feel for what one may face in practice, I decided to generate a Bernoulli population where $p=\frac{58r}{4000}$ with $r=1,2,3,10$, and 41.779. The case "$r=1$" simulates the situation on hand where we are told that 58 good items have been observed among 4000 inspected items and no more. The cases "$r=2,3,10$" respectively simulate situations where we may expect to see good items at the rate (p) of two, three or ten times the rate of what we have been told to have happened. The case "$r=41.779$" simulated a situation where one may expect to see good items at the most optimistic rate given what has happened in the situation on hand, that is with $p=0.6058(\equiv\frac{6058}{10000})$. We fixed $\alpha=0.05$, $\varepsilon=0.02$ and hence (10) suggested a pilot sample size $m = 124$. I determined N ten separate times from independent replications in each situation. Table 2 provides all ten observed N values along with their average $\overline{N}$, standard deviation S_N, and the median $\widetilde{N}$, in each case. The last column provides the estimated average percentage savings compared with $n = 4000$. One sees unbelievable percentage savings in sample size on an average when $r = 1, 2, 3$ and 10. Even in the most optimistic situation ($r = 41.779$) described in the last block of table 2, we note 4.7% average savings in sample size compared with $n = 4000$. This saving may appear insignificant, but then one should consider this: After observing only 58 good items among 4000 inspected items, what is the likelihood that all remaining 6000 uninspected items would be judged good if inspected? The point is that even under such rarest of rare occurrence, the present sampling strategy could have saved us by inspecting nearly 3800 items on an average instead of 4000 items!

4 Sequential testing to determine a sample size

I continue with random sampling from a Bernoulli(p) population where p is the fraction of good items in a very large population having $R(= 10,000)$ items. The inspection team must have certain high value $p_0, 0 < p_0 < 1$, in mind that it expects the vendor to comply with in good faith. The State may hope that $p_0 \approx 1.0$. Obviously, a small percentage of items may turn out bad, but those bad items would be expected to be properly 'corrected' by the supplier. So, the inspection team could set up a sampling strategy for the following testing problem:

$$H_0\colon\ p \geq p_0 \text{ versus } H_1\colon\ p < p_0 \tag{11}$$

Suppose that one fixes $p_0 = 0.95$ or 0.99 and it means that the State considers 9500 or 9900 good items found among 10,000 items is within reason. But, if $p < p_0$ where p_0 is a set number, then the inspection team will 'raise a flag' in favor of possible suspicion of receiving lesser than expected quality. I would like to clear one important point. The number p_0 ought to be specified by the State. Such specification may take into consideration the inspection team's mindset that is consistent with the State's budgetary constraints plus other protocols as required.

One may feel tempted to use customary normal approximation to a binomial distribution and hence having $n(\geq 30)$ observations, one would reject the null hypothesis H_0 if and only if

$$\widehat{p}_n < p_0 - z_\alpha \sqrt{\tfrac{p_0(1-p_0)}{n}}, \tag{12}$$

with the level of significance, $\alpha = P\{\text{ Rejecting} H_0 \mid H_0 \text{ is true}\}$.

But, what should be the appropriate sample size, n? Now, suppose that one asks that the power of the test (12) when $p = p_1(< p_0)$ be at least $1 - \beta$

where $0 < \beta < 1$ is a small and fixed number. In many investigations, one fixes power 80% (that is, $\beta = 0.20$ to detect a certain "effect size" (that is, $p_1 - p_0$). Let me denote $\sigma_i = \sqrt{p_i(1-p_i)}, i = 0, 1$. For large n, the power of the test (12) when $p = p_1 (< p_0)$ can be expressed as

$$P\{\text{Rejecting } H_0 \mid p = p_1\} \approx P\left\{Z < \tfrac{\sqrt{n}(p_0 - p_1)}{\sigma_1} - z_\alpha \tfrac{\sigma_0}{\sigma_1}\right\}. \qquad (13)$$

Now, it ought to be clear that the power given in (13) would be at least $1 - \beta$ provided that the sample size n is chosen as follows:

$$n \geq \frac{(z_\alpha \sigma_0 + z_\beta \sigma_1)^2}{(p_1 - p_0)^2} = n^*(p_0, p_1).$$

I define

$$n \equiv n(p_0, p_1) = \max\{30, \langle n^*(p_0, p_1)\rangle + 1\}, \qquad (14)$$

so that customary normal approximation to a binomial distribution will be expected to work (since $n \geq 30$).

In table 3, I provide values of $\langle n^*(p_0, p_1)\rangle + 1$ for $p_0 = 0.95, 0.99$ with $\alpha = 0.05$ and $\beta = 0.05, 0.10, 0.20$. When the null hypothesis tests a large p_0 value, naturally the required sample size becomes rather too small in order to detect p_1 far away from p_0 whether the power is set at 80%, 90% or 95%. It is clear that one needs to inspect nearly 30 or so items while testing a null hypothesis with large $p_0 (= 0.95, 0.99)$ when the true fraction of good items is indeed only $0.0145 (\approx \frac{58}{4000})$ or close to the most optimistic value $0.6058 (\approx \frac{6058}{10000})$.

p_0	β	p_1					
		0.90	0.80	0.60	0.50	0.25	0.0145
0.99	0.20	22	7	3	2	1	1
	0.10	38	13	5	3	1	1
	0.05	54	19	7	5	2	1
		p_1					
		0.90	0.80	0.60	0.50	0.25	0.0145
0.95	0.20	150	22	5	3	2	1
	0.10	221	34	8	5	2	1
	0.05	291	46	12	7	3	1

Table 3. Values of $\langle n^*(p_0, p_1)\rangle + 1$ from (14) with 5% level and power $1 - \beta$ with $\beta = 0.05, 0.10$, and 0.20. Sample size n is max{table entry,30}.

It is obvious that the best fixed-sample-size test (12) could arrive at a decision with very few inspections if p was indeed as small as it was in the supplied batch. It is also well known, however, that the test (12) is not really 'optimal' in a larger class of sequential tests. Wald's (1947) *sequential* *probability ratio test* (SPRT) which is optimal [Wald, 1947; Wald and Wolfowitz, 1948] would have required the least number of inspections N on an average with comparable error rates α and β for testing H_0: $p = p_0$ versus H_1: $p = p_1 (< p_0)$.

In table 4, I summarize some findings obtained from 1000 independently run simulations in each case. I generated Bernoulli populations with $p = 0.1045$ and 0.6058, the most pessimistic and optimistic values of p respectively, that are possible for the supplied batch of computers. In no situation, I ended up accepting H_0 which postulated a higher p than that under H_1 as indicated by the entry '$\#H_0 = 0$'. The entries $\overline{N}, S_{\mathrm{N}}, N_{\min}, N_{\max}$ respectively stand for the average, standard deviation, the minimum and the maximum obtained from 1000 iterations. Even $N_{\max}$ ranged from merely 9 to 385! The rest of the numbers speak for themselves.

	$p_0 = 0.90, p_1 = 0.80$	$p_0 = 0.75, p_1 = 0.70$
$\mathbf{p = 0.0145}$	$\#H_0 = 0, \overline{N} = 7.11, S_{\mathrm{N}} = 0.34$ $N_{\min} = 7, N_{\max} = 9$	$\#H_0 = 0, \overline{N} = 26.37, S_{\mathrm{N}} = 0.64$ $N_{\min} = 26, N_{\max} = 30$
$\mathbf{p = 0.6058}$	$\#H_0 = 0, \overline{N} = 23.39, S_{\mathrm{N}} = 9.49$ $N_{\min} = 7, N_{\max} = 66$	$\#H_0 = 0, \overline{N} = 152.95, S_{\mathrm{N}} = 49.09$ $N_{\min} = 57, N_{\max} = 385$

Table 4. Summary from 1000 simulations in each case for Wald's SPRT with $\alpha = \beta = 0.01$.

5 Concluding thoughts

It is clear that the protocol that allowed inspecting $4,000$ computers to detect *only* 58 good ones was at best outrageously wasteful. This is a stunning example of the fleecing of taxpayer's money! An appropriately designed sampling strategy could conclude with near certainty (that is, $\alpha = \beta = 0.01$) that the supplied batch was far below any expected standard with fewer than 10% inspections. Hiring a qualified statistical consultant at the right time would have saved the State of Connecticut much wasted resources amounting to hundreds of thousands of dollars in this one project alone.

References

[Chase and Bown, 2000]W. Chase and F. Bown. *General Statistics*, 4^{th} edition. Wiley, New York, 2000.

[Corneliussen and Ladd, 1970]A.H. Corneliussen and D.W. Ladd. On sequential tests of the binomial distribution. *Technometrics*, 12, pages 635-646, 1970.

[Ghosh, 1970]B.K. Ghosh. *Sequential Tests of Statistical Hypotheses*. Addison-Wesley, Reading, 1970.

[Ghosh *et al.*, 1997]M. Ghosh, N. Mukhopadhyay, and P.K. Sen. *Sequential Estimation*. Wiley, New York, 1997.

[Hartford Courant, 2004]Hartford Courant. The lead article in "Connecticut" section B of the newspaper. Copyright 2004, The Hartford Courant Co. A Tribune Publishing Company, Tuesday, June 8, 2004.

[Mukhopadhyay, 2004]N. Mukhopadhyay. A new approach to determine the pilot sample size in two-stage sampling. *Communication in Statistics-Theory & Methods* (A special issue dedicated to Milton Sobel's memory), P. Chen, L. Hsu, and S. Panchapakesan, editors, 33, 2004, in press.

[Mukhopadhyay and Cicconetti, 2004]N. Mukhopadhyay and G. Cicconetti. How many simulations to run? In N. Mukhopadhyay, S. Datta, and S. Chattopadhyay, editors, *Applied Sequential Methodologies*, pages 261-292; Marcel Dekker: New York, 2004.

[Mukhopadhyay and Solanky, 1994]N. Mukhopadhyay and T.K.S. Solanky. *Multistage Selection and Ranking Procedures*. Marcel Dekker: New York, 1994.

[Robbins and Siegmund, 1974]H. Robbins and D. Siegmund. Sequential estimation of p in Bernoulli trials. In E.J. Williams, editor, *Studies in Probability and Statistics*, E.J.G. Pitman Volume, pages 103-107; Jerusalem Academic Press: Jerusalem, 1974.

[Stein, 1945]C. Stein. A two sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics*, 16, pages 243-258, 1945.

[Stein, 1949]C. Stein. Some problems in sequential estimation (abstract). *Econometrica*, 17, pages 77-78, 1949.

[Sukhatme *et al.*, 1984]P.V. Sukhatme, B.V. Sukhatme, S. Sukhatme, and C. Asok. *Sampling Theory of Surveys Applications*, 3^{rd} edition. Iowa State University Press, Ames, 1984.

[Wald, 1947]A. Wald. *Sequential Analysis*. Wiley, New York, 1947.

[Wald and Wolfowitz, 1948]A. Wald and J. Wolfowitz. Optimum character of the sequential probability ratio test. *Annals of Mathematical Statistics*, 19, pages 326-339, 1948.