USE OF LOSS FUNCTIONS TO DETERMINE SAMPLE SIZE IN THE SOCIAL SECURITY ADMINISTRATION

Thomas B. Jabine and Rudolph E. Schwartz, Social Security Administration

1. INTRODUCTION

In this paper, we illustrate the determination of optimum sample size by minimization of an appropriate loss function. The theory is straightforward and well known [3, 8]; however, opportunities to apply it do not appear frequently and examples in textbooks tend to be contrived. Surprisingly, the authors have recently found in the Social Security Administration (SSA) several applications of sampling which lend themselves to this technique. After a brief description of the basic idea and the general conditions required for its application, we describe in detail two applications in the newly enacted Supplemental Security Income (SSI) program. The paper concludes with a general discussion of some possible extensions of the technique, as well as some of the problems and limitations associated with it.

The traditional textbook approach to the determination of sample size starts with the specification of the desired variance of a sample estimate of some population parameter, expresses this variance as a function of the sample size and of other population parameters assumed to be known, i.e.

$$\sigma^2_{\hat{x}} = f(n, \theta_1, \ldots \theta_r) \quad (1)$$

and solves this equation to determine the required sample size.

Since there are normally a large number of alternate sampling and estimation procedures available, we may establish relationships like (1) for several possible sample designs, solve each for n, and choose the design which gives the smallest value of n. Often, the variable cost attached to each sample unit differs for different designs. When this is the case, we will choose the design with the lowest variable cost.

For some applications, the total budget is fixed. In that case, assuming that fixed costs and variable unit costs are known for each design, we will calculate the size of the sample that we can afford, calculate the variance from (1), and choose the design that minimizes the variance.

This procedure leaves unanswered, however, the fundamental question of how to establish the appropriate variance requirements or budget for a particular survey or other investigation carried out by sampling. Commonly, the sampling technician for a project proposes certain reliability requirements, based on his experience with similar applications, and if the corresponding costs look reasonable to the project manager, he accepts these specifications. Alternatively, the project manager may specify the budget and the sampling technician will try to maximize reliability within that cost.

People trained in decision theory, cost-benefit analysis and other tools of management find the whole process to be rather arbitrary; and, as a consequence, administrators of statistical programs often find themselves under pressure to develop more objective bases for allocating their resources among various data collection programs. Unfortunately for this end, no one has found, nor are they likely to find, any satisfactory way of quantifying the benefits of general-purpose statistical programs. We have no solutions to this dilemma. However, there are situations where a more objective approach can be used. These situations arise when

1. The purpose of the sample is to obtain estimates of one or more population parameters which will be used, according to some specified rules, to determine an amount of money to be disbursed or collected.

2. The cost of collecting and processing the necessary data for the sample units is known or can be estimated.

3. The loss resulting from estimates which differ from the population parameters being estimated can be defined in such a way that its expected value is not zero.

2. THE BASIC METHOD

Let $C = f(n)$ be the cost of collecting and processing the sample data. We will call this the cost function.

n = number of units in the sample.

Let $\hat{A} = g(\hat{Y}_1\ \hat{Y}_2, \ldots, \hat{Y}_s)$ be the amount to be disbursed ($\hat{A} > 0$) or collected ($\hat{A} < 0$) by an entity, with $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_s$ representing sample estimates of population parameters which enter into the determination of this amount, and

Let $L = h(\hat{A}-A)$ be a function representing the loss (or gain) to the entity when $\hat{A} \neq A$, where $A=g(Y_1, Y_2, \ldots, Y_s)$. We will call this the payment error function.

Then our loss function is

$$\theta = E(L) + C$$

If there were no constraints on the sample size n, we could determine its optimum i.e., the value that minimizes the loss function, by differentiating the loss function with respect to n and solving for n in

$$\frac{\partial\theta}{\partial n} = 0$$

However, in this instance, we will introduce the restriction that n be an integer in the interval

$$1 \leq n \leq N$$

where N = number of units in the population from which the sample is to be selected

so that in some instances the value of n which minimizes the loss function will have to be determined by other methods.

One obvious form of the payment error function would be

$$L_1 = \hat{A} - A$$

However, $E(L_1) = 0$ if $\hat{A}$ is an unbiased estimate of A, and this would lead to n = 1, which is not very helpful.

There are at least two ways to resolve this difficulty. One would be to define a function L_2 which is always positive when $\hat{A} \neq A$, for example, a function involving $|A-A|$ or $(\hat{A}-A)^2$. This would reflect a philosophy which says there is always some economic loss when we err in estimating the amount to be disbursed or collected, regardless of the direction of the error. This seems like a reasonable position, but it is very difficult to quantify. For example, suppose the issue to be resolved is how much of the cost of a particular program is to be borne by the Federal and State governments, respectively. If the Federal government pays more than its share, and the State governments pay less, or vice versa, how do we evaluate the overall loss to the economy? This question is almost as difficult to answer as one requiring the quantification of the consequences of errors in a general purpose statistical survey.

A second approach is to look at the situation from the point of view of one of the two parties involved, and to take the position that the principal concern is with losses resulting from errors of estimation which have unfavorable consequences for that party. For example, using the illustration from the previous paragraph, suppose we represent the Federal government, and we want to minimize losses from estimates which result in overpayments to the States. We may then define a payment error function

$$\begin{aligned} L_3 &= \hat{A}-A \text{ when } \hat{A} \geq A \\ &= 0 \quad \text{ when } \hat{A} < A \end{aligned}$$

This function has a positive expected value, and in the important case where we can assume $\hat{A}$ to be normally distributed, we have

$$E(L_3) = \sqrt{\frac{1}{2\pi}}\ \sigma_{\hat{A}}$$

where $\sigma^2_{\hat{A}}$ = population variance of the estimate $\hat{A}$

This is the kind of payment error function we have used to determine optimum sample size in the illustrations which follow.

3. ILLUSTRATIONS

We now present descriptions of two applications of this technique in the Supplemental Security Income (SSI) program [2]. In the first one, which we call estimation of the adjusted payment level (APL), the cost per unit of obtaining and processing the data was relatively low, and the amounts of potential overpayments were substantial. Use of the loss function technique led to the conclusion that there should be no sampling, i.e., that the calculation of adjusted payment level should be based on all eligible cases.

In the second application, which we call estimation of Federal fiscal liability (FFL), the cost per unit of obtaining data was quite large. Here, application of the method led to recommendations for sample sizes considerably smaller than had been proposed on other grounds.

Application No.1 - Estimation of Adjusted Payment Level for the Supplemental Security Income Program

The Supplemental Security Income Program (SSI), in effect since January 1, 1974, provides assistance to people with low incomes who are aged (65 and over), disabled or blind. Eligible individuals and couples receive basic Federal payments, the amounts depending on their living arrangements and on how much income, if any, they receive from other sources. The current basic payment is $146 per month for an individual living alone with no income and $219 for a couple.

SSI also provides for supplementary payments by the States. In some cases, these payments are mandatory, in order to assure that persons who had been receiving benefits from the prior Federal-State assistance programs will continue to receive benefits at essentially the same level. In other cases, the supplementary payments by the States are optional. In either case, the State may elect to have its supplementary payments administered by the Federal government.

A provision of the SSI legislation known as "hold harmless" assures that no State electing Federal administration of its supplemental payments will have to spend more on this program than its share of the total expenditures for public assistance to recipients in these categories in calendar 1972. All costs in excess of this amount will be borne by the Federal government.

However, in order to limit Federal liability under the hold harmless provision, it was further specified that State supplementary payments would be protected only to the extent that these payments did not exceed, on the average, an amount called the adjusted payment level (APL). The APL is defined as the average money payment by the State, in January 1972, to individuals who had no income and were living alone.

In general, the States did not have available tabulations or tape files from which they could readily calculate the APL, so it was necessary to obtain the data for eligible individuals (those living alone, with no income, in January 1972) from case folders. For States with small numbers of recipients this was not difficult, but for States with large workloads it appeared that locating the case folders for cases qualifying for inclusion in the APL calculation, transcribing the data and making the calculation would be a substantial undertaking.

This situation led to consideration of the possible use of sampling. Initially several States asked for and received permission to estimate APL from a sample of cases, with the requirement that the estimate be made with a standard deviation of not more than $2.50 or a coefficient of variation no greater than 1.5 percent, if the latter condition permitted a larger standard deviation.

Subsequently, however, concern developed about the possible effects of sampling error of the estimated APL on the size of the Federal liability under the hold harmless provision. It was at this point that the loss function approach was applied for the first time.

Without going into the details of the hold harmless calculation, we may assert that an appropriate payment error function of the type L_3 for this situation was given by:

$$L_3 = 12\, W\, (\hat{Y}-Y) \text{ for } \hat{Y} > Y$$
$$= 0 \text{ otherwise}$$

where Y = APL based on calculation including all eligible cases

$\hat{Y}$ = estimate of Y from a simple random sample without replacement

W = the program workload, i.e., the total number of persons currently receiving payments under the program.

The factor of 12 was used to convert the loss to an annual basis, since the APL is an average monthly payment. We arbitrarily chose one year to represent the loss, even though it can in theory go on for an indefinite period once the APL is established. As will be seen, extending this period would not have changed our conclusion.

Assuming the estimates $\hat{Y}$ from repeated samples to follow a normal distribution, we have

$$E(L_3) = 12\, W\sqrt{\frac{1}{2\pi}}\, \sigma_{\hat{Y}}$$
$$= 12\, W\sqrt{\frac{1}{2\pi}}\left(\frac{N-n}{Nn}\right)^{1/2} \sigma_Y$$

where N = number of persons eligible for inclusion in the APL calculation

n = number of eligible persons in the sample

σ_Y = population standard deviation of the January 1972 payment level for eligible persons.

In general, $N << W$, since W represents the entire current program workload, whereas N is the number of individuals living alone, with no income, in January 1972.

Then we have for our loss function

$$\Theta = K\left(\frac{N-n}{Nn}\right)^{1/2} + cn$$

where c = unit cost per eligible case of locating the case folder and transcribing the data

and $K = 12\, W\sqrt{\frac{1}{2\pi}}\, \sigma_{\hat{Y}}$ is independent of n

Differentiating with respect to n, we have

$$\frac{\partial\Theta}{\partial n} = g\,(n) = -\frac{K}{2n^2}\left(\frac{N-n}{Nn}\right)^{-1/2} + c$$

Keeping in mind the restriction that n be an integer in the interval $1 \leq n \leq N$, analysis of $g(n)$ shows

1.· For $N < \frac{4}{3}\left(\frac{K}{c}\right)^{2/3}$

   $g(n)$ is negative and Θ uniformly decreasing in the interval $1 \leq n \leq N$. Therefore, Θ is minimized by taking $n = N$, i.e., include all eligible cases in the sample.

2. For $N = \frac{4}{3}\left(\frac{K}{c}\right)^{2/3}$

   $g(n) = 0$ at $n = \frac{3}{4}N$, and is negative in the remainder of the interval, so $n = \frac{3}{4}N$ is an inflection point and the constrained minimum for Θ is at $n = N$.

3. For $N > \frac{4}{3}\left(\frac{K}{c}\right)^{2/3}$

   $g(n)$ has its maximum at $n = \frac{3}{4}N$. The equation $g(n) = 0$ has 2 real roots in the interval $0 \leq n \leq N$, one to the left of $n = \frac{3}{4}N$ and one to the right. The root to the left is a local minimum and the one to the right is a local maximum. In order to determine the constrained value of n which minimizes Θ we must calculate Θ for each of the 2 integer values of n surrounding the local minimum and for $n = N$ and select from this triad the one which minimizes Θ.

In practice the approximate optimum value of n was easily determined by computing Θ for all values of n spaced at some reasonable sized interval, say 500, between 0 and N.

Chart 1 illustrates the behavior of the loss function Θ for changing values of K. In this illustration, we have used fixed values

$$N = 36{,}000 \qquad c = \$10$$

and allowed K to vary in the interval $10 million to $80 million. For these values of N and c, we have

$$N = \frac{4}{3}\left(\frac{K}{c}\right)^{2/3}$$

at $K \doteq \$44.4$ million.

For each of 6 values of K, values of the loss function Θ were calculated for varying sample sizes, n, starting with $n = 1500$ and continuing at intervals of 750 to $n = N = 36{,}000$.

For the first 3 values of K, all in excess of $44.4 million, we observe that the loss function decreases monotonically, and the constrained minimum occurs at $n = N$. For $K = \$35$ million, there is a local minimum in the neighborhood of $n = 18{,}750$, but the constrained overall minimum continues to be at $n = N$.

For K = \$20 million and \$10 million, however, the overall constrained minima are the local minima in the neighborhoods of n = 11,250 and n = 6,750, respectively.

It is interesting to note that at $K \doteq \$34.15$ million there occurs a threshold value of K for which it is indifferent whether we choose n = N or $n \doteq 18{,}000$. For all smaller values of K, n_{opt} will be the local minimum, and this will decrease continuously with decreasing K.

It is also of interest to observe how n_{opt} behaves if we keep K and c constant and vary N. Up to the threshold point, we have $n_{opt} = N$, so that n_{opt} increases with N. Beyond this point, however, n_{opt} will be the local minimum to the left of $n = \frac{3}{4}N$, and we find that this decreases as N increases, with

$$\lim_{N \to \infty} n_{opt} = (\frac{K}{2c})^{2/3}$$

Coming back to the APL application, there were 5 of the larger States which had initially estimated APL from samples. For these States, we had reasonably good estimates of W, the program workload and, from the earlier sample calculations, N, the number of persons eligible for inclusion in the APL calculation and σ^2_y, the variance of their January 1972 payments. We had only a rough idea of c, the unit cost, so values of θ, the loss function, were calculated for values of c in the range \$1 to \$10.

These calculations showed that $n_{opt} = N$ for all States, categories (APL was estimated separately for aged and for disabled, including blind) and unit costs in the range considered. In other words, sampling should not be used to determine the APL. As a result, the APL's for these States estimated from samples were accepted only on a provisional basis, and arrangements were made for new calculations based on all eligible cases.

Application No.2 - Determination of Federal Fiscal Liability to States for Errors in the Administration of Supplementary Payments

Under the SSI program, 33 States and the District of Columbia have selected Federal administration of supplementary benefit payments. The Federal Government (SSA) determines eligibility and benefit amounts for both basic and supplementary payments, makes the payments and is reimbursed by the States for the amount of the supplementary payments not covered by the "hold harmless" provision. All administrative costs are borne by SSA.

Federal fiscal liability (FFL) is the liability which the Federal government, represented by SSA, has agreed to assume for excessive errors in making supplementary payments on behalf of the States. These errors are of two kinds

- Eligibility errors, which occur when an individual is incorrectly certified as eligible for supplementary payments.

  Overpayment errors, which occur when an individual has been correctly certified as eligible, but the amount of the supplementary payment has been set too high.

Errors may occur either because incorrect information was obtained about factors used to determine eligibility and benefit amounts, or because correct information was processed incorrectly.

Errors affecting supplementary payments are to be identified as part of a broad quality assurance program, in which a sample of cases will be reviewed for both substantive and procedural errors relating to the basic and/or supplementary payments. The reviews are extensive, comprising an examination of materials in the SSI files, interviews with recipients, and contacts with collateral sources of information, such as banks and insurance companies.

The unit cost of these reviews is high and the use of sampling is clearly indicated. For the 31 States with Federal administration of supplementary payments, where a primary objective of the sample reviews will be to determine the FFL, the cost-benefit approach has provided useful guidance in setting reasonable sample sizes.

The liability determination period for FFL has been set at 6 months. Total liability to a State during a 6-month period is estimated, using the results of the sample reviews, by

$$A = \frac{Y}{\bar{y}} [\bar{y}_e \; (p_e - t_e) \; a_e + \bar{z}_o \; (p_o - t_o) \; a_o]$$

where the terms are defined as follows:

Y = total State supplementary payments

$\bar{y}$ = mean supplementary payment in the sample

$\bar{y}_e$= mean supplementary payment to ineligible cases in the sample

p_e= proportion of ineligible cases in the sample

t_e= tolerance limit for eligibility errors

a_e= 1 if $p_e > t_e$
= 0 otherwise

$\bar{z}_o$= mean overpayment of the supplementary payment to cases with overpayments in the sample

p_o= proportion of cases with overpayments in the sample

t_o= tolerance limit for overpayments

a_o= 1 if $p_o > t_o$
= 0 otherwise

In this case our payment error function was simply

$$L_3 = \hat{A} - A \text{ for } \hat{A} > A$$
$$= 0 \text{ otherwise}$$

where A is analogous to $\hat{A}$, with population values substituted for sample estimates.

To simplify matters, we restricted our analysis to situations where the eligibility and overpayment error rates were substantially in excess of their respective tolerances ($t_e = .03$, $t_o = .05$), so that we could let $a_e = a_o = 1$. This was a conservative approach, since the potential loss from overestimating A is clearly greatest in this range.

With this restriction, we could assume estimates of A from repeated samples to be normally distributed so that

$$E(L_3) = \sqrt{\frac{1}{2\pi}}\, \sigma_{\hat{A}}$$

By the method of propagation of variances [4, p.585], we obtained:

$$\sigma^2_{\hat{A}} = Y^2 \left[\sigma^2_{\bar{y}} - \left(\frac{b^2}{\bar{Y}}\right) + \sigma^2_{y_e}\left(\frac{p_e - t_e}{\bar{Y}}\right)^2 + \sigma^2_{p_e}\left(\frac{\bar{Y}_e}{\bar{Y}}\right)^2 + \sigma^2_{\bar{z}}\left(\frac{P_o - t_o}{\bar{Y}}\right)^2 + \sigma^2_{P_o}\left(\frac{\bar{Z}_o}{\bar{Y}}\right)^2 + \text{covariance terms}\right]$$

where $b = \bar{Y}_e\,(P_e - t_e) + \bar{Z}_o\,(P_o - t_o)$

It seemed reasonable to neglect the covariance terms, since the variables involved are largely independent of each other. From our knowledge of the relevant population parameters, it appeared that $\sigma^2_{\hat{A}}$ would be dominated by the terms involving $\sigma^2_{p_e}$ and $\sigma^2_{P_o}$. Assuming the use of a simple random sample, with replacement, of n cases, and letting $\bar{Y}_{e/\bar{Y}} = .8$ and $\bar{Z}_{o/\bar{Y}} = .4$, we arrived at the approximation

$$\sigma_{\hat{A}} \doteq \frac{Y}{\sqrt{n}} \left[.64\, P_e(1-P_e) + .16\, P_o(1-P_o)\right]^{1/2}$$

The overall loss function was defined as

$$\theta = E(L_3) + C$$
$$= Y \left[\frac{.64\, P_e[(1-P_e) + .16\, P_o(1-P_o)}{2\pi n}\right]^{1/2} + nc$$

where c = unit cost of a case review.

Differentiating θ with respect to n, setting the result equal to 0 and solving for n, we found

$$n_{opt} = \left(\frac{Y}{2c}\right)^{2/3} \left[\frac{.64\, P_e(1-P_e) + .16\, P_o(1-P_o)}{2\pi}\right]^{1/3}$$

Thus, at a given level of error, the optimum sample size is directly proportional to the total amount of supplementary payments, raised to the two-thirds power, and inversely proportional to the cost per sample case, raised to the two-thirds power.

Table 1 shows, for a unit cost c = $200, the optimum sample sizes for various combinations of Y, P_e and P_o. Since the average supplemental payments per recipient over a six-month period are on the order of $400, it can be seen that the optimum sample size is a small fraction, generally less than 1 percent, of the population of recipients. This is in strong contrast to the APL application where, with a much smaller unit cost, the indicated solution was to include all eligible cases in the calculation.

The above findings were applied directly to determine sample sizes for 18 States with mandatory supplementation only under Federal administration. A minimum sample size of 100 cases per State was established, because we felt it would be difficult to persuade all parties involved that it made sense to estimate FFL from a sample any smaller than that.

For the 13 States with mandatory and optional supplementation under Federal administration, somewhat different criteria were adopted, in order to be consistent with procedures currently in use to control the level of error by States in making payments, partly financed by Federal funds, under the program of aid to families with dependent children (AFDC). Nevertheless, the assigned sample sizes, with one or two exceptions, did not differ greatly from those indicated by use of the approach just described.

4. DISCUSSION

One can imagine many other applications and extensions of the basic method presented here. With respect to areas of application, any situation where data are needed to determine amounts of money to be transferred from one entity to another lends itself to this approach. In SSA, for example, when there are indications that a provider of services to Medicare enrollees may have overcharged for these services, there are provisions for reviewing a sample of cases to estimate the total extent of overcharging and hence the amount that the provider must return to SSA. Present procedures for determining sample sizes for these reviews are being reexamined using the technique described in this paper.

There would appear to be no special difficulty in using the method for sample designs other than simple random sampling. For example, in a situation calling for stratified sampling to estimate a single variable, we could proceed as follows:

(1) Determine optimum allocation of the sample for fixed n in the form

$$n_h = w_h\, n, \quad \sum_h w_h = 1$$

where w_h is a function of various population parameters and costs for the hth stratum and is independent of n[1]

(2) Express the variance of $\hat{A}$ as a function of n and the stratum weights. Suppose, for example, that

$$\hat{A} = K\,\bar{x}$$

so that $\sigma_{\hat{A}} = K\,\sigma_{\bar{x}}$

We have, for stratified sampling without replacement

$$\sigma^2_{\bar{x}} = \sum_n \left(\frac{N_h}{N}\right)^2 \frac{N_h - n_h}{N_h} \frac{S^2_h}{n_h}$$

which becomes, when we substitute $w_h n$ for n_h

$$\sigma^2_{\bar{x}} = \frac{1}{n} \sum_h \left(\frac{N_h}{N}\right)^2 \frac{S^2_h}{w_h} - \sum_h \left(\frac{N_h}{N}\right)^2 \frac{S^2_h}{N_h} .$$

hence $\sigma_{\hat{A}} = K \left(\frac{a}{n} - b\right)^{1/2}$

(3) Solve for the optimum n in the usual way. In the illustration, we would have

$$\theta = K' \left(\frac{a}{n} - b\right)^{1/2} + n\,c$$

where $K' = K \left(\frac{1}{2\pi}\right)^{1/2}$

and $c = \sum_h w_h c_h$

and when we differentiate and set equal to 0, we find that n is the solution to a 4th degree equation.

(4) Allocate the sample of size n to the strata using the weights w_h. Since these weights were determined independently of n (within a specified range), we will have a solution which gives both the optimum sample size, n, for a stratified design and optimum allocation of the sample among the strata.

Similar solutions should be possible for designs involving cluster and multistage sampling.

Should we try to take into account the effects of nonsampling errors on the estimates of population parameters for use in payment or reimbursement formulas? This should offer no particular theoretical difficulty; however, in practical terms it would be much more difficult to develop a formal, predictable relationship between expected losses due to sampling and nonsampling error, and the resources expended to get the data. Nevertheless, for some applications nonsampling error may dominate total error, and every possible effort should be made to consider its effects.

What are the limitations of the cost-benefit technique for determining optimum sample size? Mostly they revolve around the difficulty of defining an explicit payment error function, i.e., the term that reflects the losses resulting from errors in estimating the relevant population parameters. The solution we have adopted is admittedly somewhat artificial, in that expected losses are based on errors in one direction only. It can be rationalized on the grounds that what we are trying to do is to protect the agency we represent against incurring losses which would expose it to the charge of failing to act in a prudent and responsible manner. Errors in other directions are considered as producing windfall "profits", and these should not enter into the determination of the optimum design. On the other hand, if we consider the interests of society as a whole, then clearly losses can result from errors in either direction. However, the losses to society will in most cases not be as large (assuming a value can be placed on them at all!) as the actual amounts of underpayment or overpayment.

It can be argued that the agency, even when given a fairly precise determination of the optimum sample size, may not wish to invest its resources in exactly that way. The administrator cannot consider this particular problem in isolation from all others confronting him. His problem is more likely to be one of considering relative opportunity costs of alternative uses of more or less fixed resources. Thus, it is possible that some of the resources that would be needed to take care of the optimum sample could be applied in another area where the payoff was greater. This kind of consideration might be built into the model explicitly by introducing utility functions, i.e.,

$$\theta = u_1\,[E(L)] + u_2[C]$$

This would allow us to give different weights to real dollars which would have to be spent to collect and process the sample data, and expected dollars lost from errors in estimating the payment amount.

Another question to be considered is the extent to which the concepts involved in determining the optimum sample size by this method can be readily explained to those responsible for making decisions. In our experience, the idea of the expected loss, especially for a non-symmetric function like L_3, has been somewhat difficult to explain. The use of tables showing values of the loss function and its components for various sample sizes and values of the relevant population parameters has been very helpful in providing a clearer view of the whole problem and the relationships of the variables. The general idea of applying a cost-benefit approach to determining sample size has been received enthusiastically, and we have been asked on various occasions why we can't apply the technique to general-purpose statistical surveys!

Despite the limitations we have described, we have found this method to be considerably more satisfactory than approaches used previously for similar types of problems. Even in cases where the solution is not precise because some of the parameters needed could not be estimated accurately, the results still provide useful guidelines to a general course of action.

FOOTNOTE

[1] If sampling without replacement, independence will hold only in the range of n for which $n_h < N_h$ for all strata.

REFERENCES

1. Blythe, R.H., "The Economics of Sample Size Applied to the Scaling of Sawlogs." Biometrics Bulletin, 1, (1945), 67-70.

2. Callison, James C., "Early Experience Under the Supplemental Security Income Program." Social Security Bulletin, 37 (June 1974), 3-11, 30.

3. Cochran, William G., Sampling Techniques. New York: John Wiley & Sons, Inc., 1953.

4. Kish, Leslie, Survey Sampling. New York: John Wiley & Sons, Inc., 1965.

5. Nordin, J.A., "Determining Sample Size." JASA, 39, (1944), 497-506

6. Sadowski, Wieslaw, The Theory of Decision Making, London: Pergamon Press, 1965.

7. Willis, Raymond E., "Confidence Procedures and the Cost of Sampling", The American Statistician, 27 (December 1973), 219-221.

8. Yates, Frank, Sampling Methods for Censuses and Surveys, London: Charles Griffin & Co., Ltd., 1949.

CHART 1 - Loss Function (Θ) for selected Values of K, with c = $10, N = 36,000

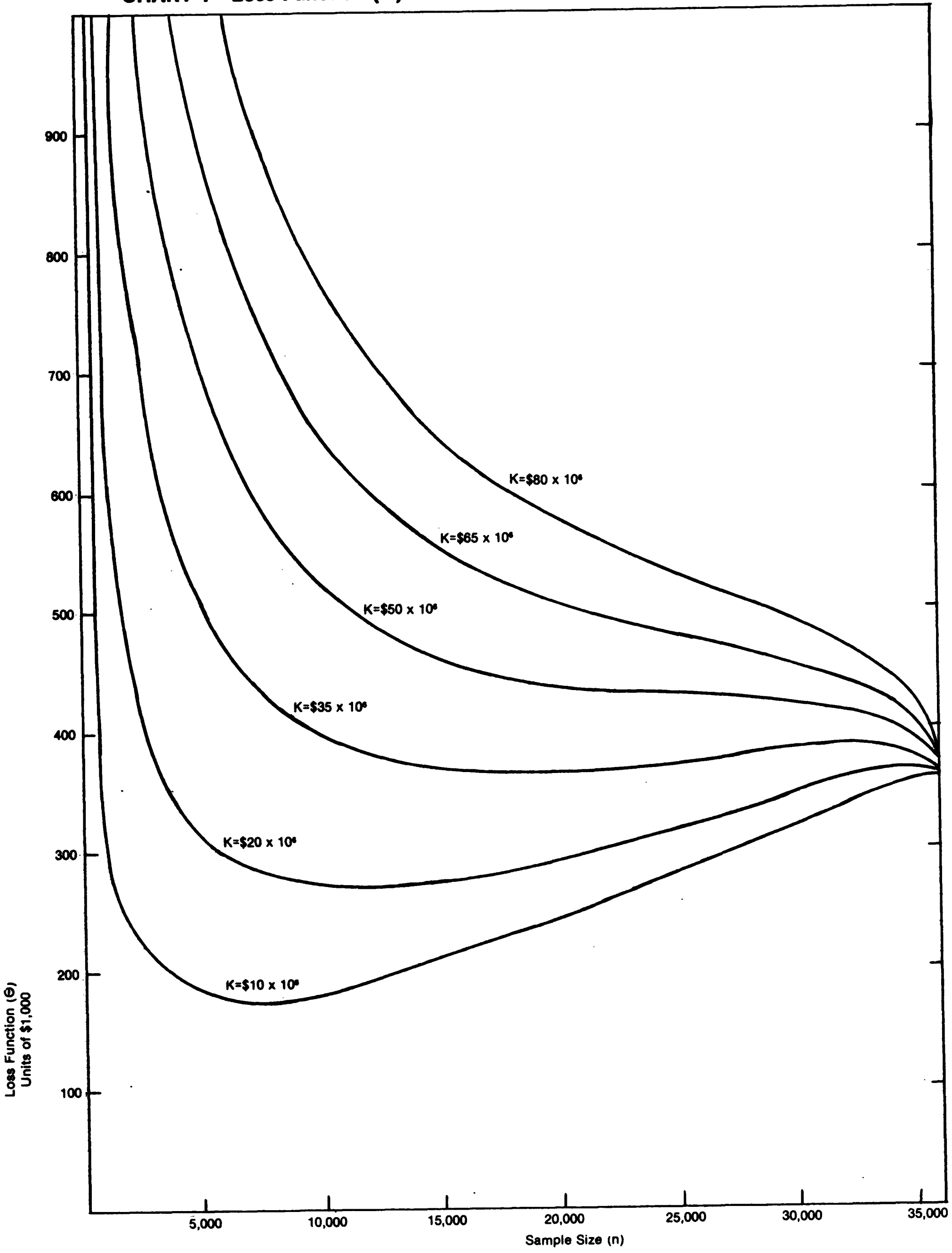

Table 1 - Optimum Sample Size for
Determining Federal Fiscal Liability (FFL)
for Errors in Determination of
State-Funded Supplementary Payments

C = $200

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
State payments subject to FFL	Eligibility error rate	Overpayment error rate	Expected Federal Liability	Optimum sample size	Cost of selecting and reviewing sample	Standard deviation of estimated liability	The chances that the FFL will not be over-estimated by more than the amount shown are approximately	
Y	P_e	P_o	A	n_{opt}	c n	$\sigma_{\hat{A}}$	95 percent	99 percent
$ 1,000,000	.05	.07	24,000	34	6,800	34,403	56,765	80,159
(N~2,500)	.10	.15	96,000	43	8,600	42,697	70,450	99,484
	.15	.21	160,000	48	9,600	47,613	78,561	110,938
$ 10,000,000	.05	.07	240,000	160	32,000	159,622	263,376	371,919
(N~25,000)	.10	.15	960,000	199	39,800	198,105	326,873	461,585
	.15	.21	1,600,000	222	44,400	220,913	364,506	514,727
$ 50,000,000	.05	.07	1,200,000	469	93,800	466,614	769,913	1,087,211
(N~125,000)	.10	.15	4,800,000	581	116,200	579,107	955,527	1,349,319
	.15	.21	8,000,000	648	129,600	645,781	1,065,539	1,504,669
$100,000,000	.05	.07	2,400,000	744	148,800	740,617	1,222,018	1,725,638
(N~250,000)	.10	.15	9,600,000	923	184,600	919,169	1,516,629	2,141,663
	.15	.21	16,000,000	1,029	205,800	1,024,995	1,691,241	2,388,238
$200,000,000	.05	.07	4,800,000	1,181	236,200	1,175,521	1,939,610	2,738,964
(N~500,000)	.10	.15	19,200,000	1,466	293,200	1,458,921	2,407,219	3,399,286
	.15	.21	32,000,000	1,634	326,800	1,626,890	2,684,369	3,790,654