STATISTICAL DECISION AND CONTROL APPROACHES FOR ALLOCATION OF FUNDS

Wray Smith and David Zalkind, Department of Health, Education and Welfare

INTRODUCTION

Many existing and proposed governmental grant-in-aid programs distribute funds annually to each state in accordance with a specific formula or allocation rule that takes into account some presumed measure of "need", as well as of population and, in some cases, of other local factors. Some federal assistance programs and their formula practices are reviewed in Statistical Policy Working Paper 1 (U.S. Department of Commerce 1978). For the purposes of our present analysis we assume the existence of a national program to aid the states in their conduct of some specified social service activity. We assume that it is an open-ended entitlement program providing an annual grant to each state of, say, one dollar per low-income resident. We assume that three types of administrative costs are incurred: fixed program overhead costs, recurring data collection costs, and imputed costs of misallocation of grant funds.

In this paper we consider the first-stage problem facing a statistical program administrator who seeks optimal policies for selecting the timing and size of data collections to update a single measure (such as percentage in poverty) at the national level, assuming for now that very precise annual estimates of population are provided without cost to the program. We plan to treat the second-stage problem of optimal policies for updating subnational estimates by state or by groups of states in a separate paper. Also to be treated elsewhere is the problem presented by multi-item measures (e.g., poverty, unemployment, and housing overcrowding) jointly in a single allocation formula.

THE TIMING AND SAMPLE SIZE PROBLEM

Specifically, we propose to solve the following operational decision problem in survey practice by solving a related optimization problem in inventory control theory: Given (i) that there is a cost for collecting and processing data, (ii) that a cost over time can be imputed to the lack of knowledge of the true value of a socioeconomic parameter used in an allocation formula, and (iii) that the decay in data precision over time can be quantified, determine when to collect new data and how much new data to collect in order to minimize the sum of costs of collection and of lack of knowledge.

In our development below we do we do not start with some predetermined precision or desired coefficient of variation for the parameter estimate to be obtained from each data collection. Rather, we derive optimal pairs (t^o, n^o) for elapsed time since the last previous data collection and for data base size in accordance with a loss function which combines data collection costs and lackof-knowledge costs, with the latter based on an appropriate measure of the dollars misallocated each year owing to lack of perfect knowledge of the allocation parameter value.

Determination of sample size as a statistical decision problem is reviewed briefly by Cochran (1977; section 4.10). Most previous work in this area considers the economic choice of sample size, but not the timing of recurring collections, for estimating a parameter with a fixed but unknown or vaguely known mean value. We consider here the problem of jointly determining timing and sample size when the goal is to estimate a parameter that is assumed to follow approximately a general one-dimensional random walk in discrete time with no adherence to a fixed mean value.

MEASUREMENTS AND ESTIMATES FOR A NONSTATIONARY PARAMETER

Consider a random variable X with stochastic parameter P having a value P_t at time t. For example, X might be a binary random variable taking the value 1 for a person below the poverty line and 0 otherwise, subject to a binomial parameter P_t constrained to the range $0 < P_t < 1$. An underlying model for P_t might be represented by a linear first-order stochastic difference equation,

$$P_t = P_{t-1} + u_t \quad (t = 1,2,\ldots), \qquad (1)$$

where the u_t are normal independent $(0,\sigma_o^2)$ random variables, with P_0 known or precisely estimated, $0 < P_0 < 1$, and $E[|u_t|] << P_0$ (with, for example, $P_0 \doteq 0.1$), so that for periods involving only a few time steps (e.g., a few years in the case of poverty, which is defined in terms of a calendar year accounting period) there is essentially zero chance of P_t wandering to zero or one.

Efforts to measure P_t are assumed to yield

$$\hat{P}_t = P_t + w_t , \qquad (2)$$

where $\hat{P}_t$ is a survey estimate based on a simple random sample of size n_t and w_t is a measurement error term depending on sample size n_t, possibly on the true value P_t of the parameter, and on various interfering factors. A realistic model for survey data on U.S. poverty over the past decade would specifically recognize measurement bias and year-to-year correlations of the w_t, but a simplified model with the w_t treated as if they were

independent zero-mean constant-variance random variables suffices to demonstrate our approach.

For relatively large sample sizes, with $E[P_t] = P_{t-1}$ in a general random walk framework, we claim that old data, properly discounted for loss of precision, is "equivalent" to new data. We may thus use old data and new data together to form an estimate P_t' of P_t,

$$P_t' = a_t \hat{P}_t + (1-a_t) P_{t-1}' , \qquad (3)$$

$0 \leq a_t \leq 1$, where a_t is a function of time and of sample sizes and is determined by a method described below. We use a caret to denote a measurement at time t on a single data set collected at time t and a prime to denote an estimate constructed at time t from two or more data sets collected at different times. One may observe that (3) is analogous to exponential smoothing when a_t is constant.

COST FUNCTIONS

We assume that the cost of collecting and processing new data with sample size n is of the form

$$\begin{aligned} C(n) &= c_o + c\,n , \quad \text{if } n > 0 , \\ &= 0 , \quad \text{if } n = 0 , \end{aligned} \qquad (4)$$

where c_o is a fixed cost we incur each time we undertake a new data collection and c is the unit cost of each observation.

It should be noted that we may include in c_o indirect administrative burden as well as operational start-up cost and include in c not only the direct cost of each interview but also an imputed cost of burden on respondents.

We define lack-of-knowledge cost per unit time as the loss which arises because at time t, $t \geq r$, we possess only the information from a sample of size n_r, collected and instantaneously processed at time r. We ascribe a dollar loss to our lack of knowledge since it causes us to misallocate grant-in-aid funds. This cost may be represented formally by a loss function, namely $L(t-r,n_r,P_r',P_t)$, where the first argument is the time elapsed since the data were collected, the second argument is the size of the simple random sample that was taken, the third argument is our estimate of the parameter at time r, and the fourth argument is the unknown value of the parameter at time t.

We have assumed that P_r is slowly varying, in the sense that it is likely to remain within a few percentage points of its current level for several years. For the purposes of our present analysis, we interpret L as conditional on the most recent P_r estimate, suppressing explicit dependence on P_r' or P_t, writing $L(t-r,n_r)$.

DECAY AND EQUIVALENT SAMPLE SIZE

We characterize a sample of size n_r taken at time r as equivalent to a sample of size n_t' taken at time t, with n_t' given implicitly by

$$L(t-t,n_t') = L(0,n_t') = L(t-r,n_r) . \qquad (5)$$

We will call n_t' the <u>equivalent sample size</u> corresponding to an n_r decayed to time t, since by time t a sample of size n_r taken at time r yields the same lack-of-knowledge cost as would data with sample size n_t' taken at time t. Furthermore, if at time r the equivalent sample size for all the knowledge we still possess from acquisitions up to and including time r is n_r', then we may treat n_r' as if it were the n_r in (5) and write

$$n_t' = n_r' R(t-r,n_r') , \qquad (6)$$

where R is a retention function representing the fraction remaining at time t of the equivalent sample size n_r' that existed at time r. The amount retained is $n_r' R(t-r,n_r')$ while the amount decayed is thus $n_r'[1-R(t-r,n_r')]$.

We will refer to this latter quantity as the demand or depletion due to decay over an interval of length $t-r$. The functional form of the retention function R depends on the lack-of-knowledge cost function L, since from (5) and (6) the key requirements that R must satisfy are

$$L(t-r,n_r) = L(t-s,n_r R(s-r,n_r)) \qquad (7a)$$

and

$$R(t-r,n_r) = R(s-r,n_r)R(t-s,n_r R(s-r,n_r)) \qquad (7b)$$

for all s such that $r \leq s \leq t$.

We select as a representative cost function the lack-of-knowledge cost function

$$L(t-r,n_r) = A_1[\, A/n_r + (t-r)B \,]^{B_1} , \qquad (8)$$

where A, A_1, B, and B_1 are positive constants. A_1 is a scale factor transforming the second factor, which is a measure of current data quality, into a dollar loss. The magnitude of A_1 will depend on the value of the information in the data base in the sense of the seriousness of resulting allocation errors. A_1 will be expressed in different units for different B_1. A and B are weights assigned to the size-dependent and time-dependent components, respectively, of the loss function; for example, A/n_r may be proportional to the variance of the parameter estimate based on a sample of size n_r and $(t-r)B$ may be proportional to the variance of a random walk of duration $t-r$. In that case, if $B_1 = 1$, then L is proportional to the sum of the two variance components and is a quadratic loss function. If

$B_1 = 0.5$, then L is proportional to the corresponding standard error of P_t'.

The explicit R corresponding to the L function of (8) is

$$R(t-r,n_r) = [\ 1 + (n_r/A)(t-r)B\ ]^{-1} \ , \quad (9)$$

which for fixed n_r and discrete t has a decay pattern over time as shown in Figure A. We restrict R by the relation $R \cdot n_r \geq 1$, thus implicitly bounding t and hence L .

We then have directly what we may regard as an updated Bayesian estimate of the parameter P_t ,

$$P_t' = (n_t/n_t')\ \hat{P}_t + (n_r'\ R(t-r,n_r')/n_t')P_r' \ , \quad (10)$$

where n_t' is the sum of the new sample size and the decayed equivalent of the old sample size. P_t' is thus an average of the old P_r' estimate and the new $\hat{P}_t$ measurement weighted by their respective equivalent sample sizes at time t . For a discussion of Bayesian estimation techniques see DeGroot (1970) or Zellner (1971).

APPLICABLE INVENTORY CONTROL PRINCIPLES

A process evolving through time may be treated as an inventory system if (i) there is a stock of items (e.g., equivalent sample size) which is depleted over time by demand for items, perishability, evaporation, decay, etc., (ii) there exists the possibility of ordering additional items (e.g., a new survey) to increase the inventory level, and (iii) costs are incurred based on the amount of inventory in stock or short (e.g., our lack-of-knowledge cost) and on each order placed to replenish inventory. For further elaboration of the inventory control principles set forth in this section see, e.g., Hillier and Lieberman (1974) or Naddor (1966).

The downward segments in the graph of Figure A represent decline in the stock level, which we interpret as demand corresponding to data decay. Demand (or decay) is assumed to occur according to a power law pattern such that the rate of decay is a nonincreasing function of time; cf. (9). Replenishments made every t units of time (called the scheduling-period) for the amount q (called the lot size) are represented by the upward segments of the graph. Another way of interpreting what is happening is that whenever the inventory level falls to s or below (called the reorder-point), the decisionmaker orders the amount q which arrives instantly. A third interpretation is that whenever the amount of inventory falls to level s , or below, a sufficient quantity is ordered so that the inventory level rises to the amount S (called the order-level).

This last interpretation is the most robust in the sense that such (s,S) policies are optimal under a wide variety of circumstances. For the deterministic demand rate treated in this paper the three types of ordering policies described above are equivalent. The cost function K , the expected cost of inventory that may be carried or short during this period, is usually a convex function of y , the number of items on hand at the beginning of the period (after replenishments, if any, have arrived).

As in equation (4), replenishment costs are usually taken to be of the form:

$$C(q) = c_o + cq \ , \quad \text{if } q > 0 \ ,$$

$$= 0 \ , \quad \text{if } q = 0 \ ,$$

where a fixed charge c_o is incurred if an order is placed and a cost of c monetary units per unit ordered is also incurred.

Thus if we are ordering, in order to optimize,

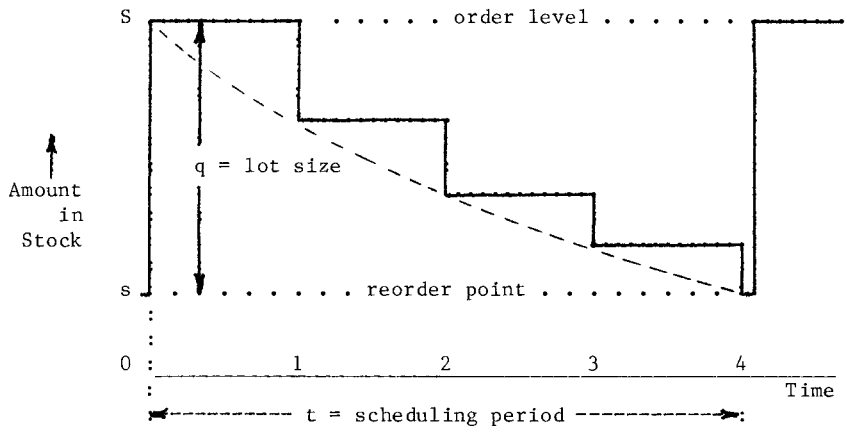


Figure A. Graph of the Behavior of a Simple Deterministic Inventory System with a Power Demand Pattern and No Backordering

we need only find the value of y that minimizes $cy + K(y)$. Call this value S and consider

Figure B. We can see from Figure B that if $x > S$, then

$$c_o + cy + K(y) > cx + K(x) \qquad (11)$$

for all $y > x$. For all $x > s$ subtracting cx from both sides of (11), we have

$$c_o + c(y-x) + K(y) > K(x) \qquad (12)$$

so we should not order. However, for $x \leq s$, the inequality (12) is reversed, that is,

$$c_o + c(S-x) + K(S) \leq K(x) \qquad (13)$$

and it is less expensive to place an order for the amount $S-x$ than to take any other action.

Treating y as a continuous variable and $K(y)$ as a differentiable function, the optimum S is the solution to

$$\frac{dK(y)}{dy} + c = 0 . \qquad (14)$$

Furthermore, the optimum s is the smallest value (actually unique since the function is convex) that satisfies the expression

$$cs + K(s) = c_o + cS + K(S). \qquad (15)$$

We now introduce leadtime. Suppose S were the optimal order-level when leadtime is 0; i.e., for instantaneous delivery of an order. If the leadtime is actually v , then the optimal order-level S^* is $S + D(v)$ where $D(v)$ is demand over an interval v . Thus, if orders were placed every t units of time, starting at time 0, then ordering up to S^* will cause the inventory level to be at S at times v , $t + v$, $t + 2v$, etc. This policy minimizes costs over time intervals of length t , namely $[v + kt, v + (k+1)t)$, $k = 0,1,2,\ldots$. Actions by the decisionmaker cannot affect what happens in the time interval $[0,v)$, so we do not consider this interval in the optimization problem. The effect of leadtime on the cost function to be minimized is shown in the next section.

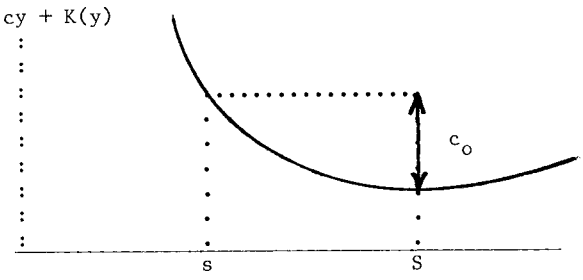


Figure B. Graph of cy + K(y)

OPTIMAL SOLUTION TO THE TIMING AND SAMPLE SIZE PROBLEM

We now make use of the similarity in structure between our statistical policy problem and the inventory control problem of the previous section. First, suppose we have a single-period problem of determining the timing and sample size of one data collection. Assume that the equivalent sample size at the beginning of the period is n' , then for a period of length one time unit the cost, with no ordering, is

$$C(n') = L(1,n') . \qquad (16)$$

However, if we do order a sample of lot size $q = n^o - n'$, assuming instantaneous delivery, the expected cost during the time period is

$$C(n') = c_o + c(n^o - n') + L(1,n^o) , \qquad (17)$$

if cost is incurred only at the end of the unit period. If we treat the cost over time as accruing continuously (with lack of knowledge costing us more at the end of the period than it does at the beginning) then the cost during a unit period is

$$C(n') = \int_0^1 L(r, n')\, dr . \qquad (18)$$

It can be shown that

$$I_{t-r}(n) = (t-r)^{-1} \int_r^t L(s,n)\, ds \qquad (19)$$

is convex in n for any t since $L(s,n)$ is convex in n for every s , $s \geq r \geq 0$.

For simplicity, in the remainder of this section, we let $r = 0$. The pair of cost equations (16) and (17) also provide an analog to the inventory problem, with $I_t(n)$ substituting for $K(n)$, when $t = 1$. Because the decay function (e.g., equation (9)) is deterministic in t (and n), our problem may be solved by finding the optimal scheduling-period t^o (the the time interval between reorders) rather than a reorderpoint s. That is, for any given time interval t between reorders, we can find an optimal order-level $S = n^o$.

Since demand (i.e., decay) is deterministic, finding an optimal s and finding an optimal t are equivalent, and it is easier to find the optimal t and then infer the optimal s . Technically, we are finding the optimal solution for a scheduling-period order-level (t,S) policy rather than a reorder-point order-level (s,S) policy.

For continuously accruing cost, the minimization problem is

$$\min_t (1/t)\, [I_t(n^o(t)) + c_o + cn^o(t)[1-R(t,n^o(t))]] \qquad (20)$$

in order to get the least average cost per unit time, where $n^o(t)$ is the optimal order-level for time t, found from equations (16) and (17) using the theory of the previous section and $n^o(t)[1-R(t,n^o(t))]$ is demand due to decay of equivalent sample size during a time interval of length t. Although we may treat time as continuous with regard to cost, we will only consider integer values of t in the optimization expression (20), since, as a practical matter, the time interval between surveys is an integral number of months or years.

In order to take leadtime into account, we must determine an appropriate v to reflect the typical time delay between the collection of new data and its availability for use. We then replace n^o with n^* where

$$n^o(t) = n^*(t,v)\ R(v,n^*(t,v)) , \qquad (21)$$

and $n^*(t,v)$ is the optimal order-level for a given scheduling period t when leadtime is v. Thus, expression (20) becomes

$$\min_t\ (1/t)\ [I_t(n^*(t,v) + c_o + cn^*(t,v)[1-R(v,n^*(t,v))]] \ . \qquad (22)$$

In the discrete case, the analog to equation (20) is

$$\min_t (1/t)[\sum_{r=1}^{t} L(r,n^*(t,v)R(r,n^*(t,v))) + c_1 + cn^*(t,v)[1-R(t,n^*(t,v))]] \ . \qquad (23)$$

NUMERICAL EXAMPLES

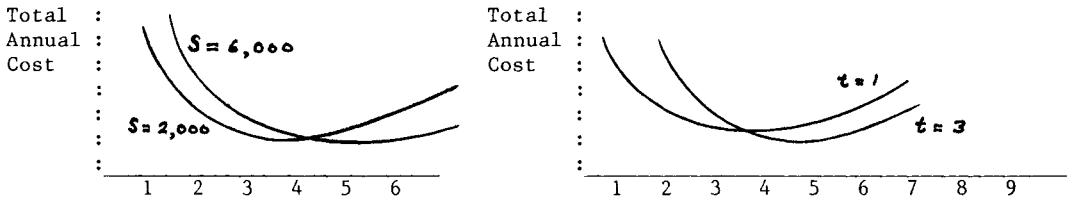
We now give some examples of a specific solution to the timing and sample size problem using (20) as the objective function. We presume that K is unimodal in t when optimal order-levels are used for each t. We thus find the optimal pair $(t^o,n^o(t^o))$ for the scheduling-period and the order-level by this procedure. Our procedure is to find the optimal order-level $n^o = S$ and the associated total cost $K(t,n^o(t))$ when $t = 1$, and then do the same for $t = 2,3,...$ until $K(t,n^o(t))$ starts increasing. Henceforth we will write n^o for $n^o(t^o)$. The sample size each time a survey is commissioned is therefore $n^o[1-R(t,n^o)]$, except that the first survey will need to have sample size n^o, if we start with nothing, or be large enough so that the initial equivalent sample size (possibly using old data) is n^o.

Table 1 contains solutions for cost function (8) with the following values for constants: $A_1 = 1$, $B_1 = 0.5$, $c = \$60$/case and $c_o = \$300{,}000$. A has values of 2.56×10^{15} and 2.56×10^{17}, which correspond to dollar benefits per person of \$1 and \$10. The lack-of-knowledge cost is the expected misallocated dollar amount, which is the dollar amount per person times the mean absolute error (where the latter is 0.8 times the standard deviation) of the number of persons in poverty. The variance of the decay per year ranges from 1×10^{-5} to 9×10^{-5}, which roughly corresponds to existing national poverty data. The constant B, which in conjunction with A determines the decay rate, is found by setting $0.1B/A$ equal to the random walk variance.

Good estimates for the decay rate would depend on observations of temporal changes in the parameter being measured, perhaps based on time series analysis of past data. See, for example, Scott and Smith (1974), who considered stochastic population parameters, although not in relation to the timing and sample size decision problem. We do not treat rate estimation here, but give examples utilizing a range of decay rates with selected values of the other constants in each cost function.

Table 1. Optimal Order-Levels and Sample Sizes for Various Values of Grant Funds Allocated and Decay Rates Assumed. (Ordering cost is \$60 per unit plus a fixed charge of \$300,000. Inconsistencies in the table are due to roundoff errors in the computer program.)

Entitlement per Person in Poverty (\$)	Variance of Random Walk	t^o Optimal Scheduling-Period	S Optimal Order-Level	Optimal Sample Size	Annual Lack-of Knowledge Cost (\$)	Total Annual Cost (\$)
1	1×10^{-5}	4	8104	6193	558,851	726,751
1	3×10^{-5}	2	5160	3900	691,025	958,032
1	5×10^{-5}	2	4691	3867	825,211	1,091,212
1	7×10^{-5}	2	4405	3790	936,746	1,200,457
1	9×10^{-5}	2	4200	3709	1,034,509	1,295,889
10	1×10^{-5}	1	22,315	15,410	4,866,246	6,090,818
10	3×10^{-5}	1	18,136	15,320	7,064,615	8,283,827
10	5×10^{-5}	1	16,466	14,683	8,638,299	9,819,255
10	7×10^{-5}	1	15,423	14,116	9,941,021	11,087,953
10	9×10^{-5}	1	14,671	13,638	11,079,457	12,197,743



(i) Cost as function of scheduling-period t for selected values of S

(ii) Cost as function of order-level S (thousands) for selected values of t

Figure C. Typical Graphs of Total Annual Cost

The table shows the optimal scheduling-period t^o, the corresponding optimal order-level S, the sample size (lot size) S-n needed to maintain that order-level, and the corresponding lack-of-knowledge cost and total cost per year of following that policy.

Figure C depicts the typical variation of total cost per year in relation to elapsed time t between surveys and to the order-level S for selected values of the constants. One can see from the shape of the curves that ordering too frequently is relatively more expensive than ordering too infrequently. It can also be seen that each curve is fairly shallow around its minimum. Therefore, the additional cost of being near but not precisely at the optimal scheduling-period t^o may not be excessive.

CONCLUDING REMARKS

By the above application of elementary optimization methods to a simplified practical problem of periodic data collection we have attempted to demonstrate the potential utility of inventory control and related decision techniques in assessing the tradeoffs between collection costs and lack-of-knowledge costs. We have not explicitly considered stochastic decay in this paper, but instead have treated a multiple time unit deterministic case as a single-period model, possibly revising constant terms for each decision on survey timing and sample size.

An extension in progress of the present results to cases of multi-item and multi-jurisdiction systems with associated loss functions relevant to specific allocation formulas should enhance the practical utility of our approach. Problems associated with a randomly-varying decay process and links with perishable inventory theory will be discussed in a separate study.

REFERENCES

Cochran, W. G. (1977), Sampling Techniques, 3rd ed. New York: John Wiley & Sons.

DeGroot, Morris H. (1970), Optimal Statistical Decisions, New York: McGraw-Hill Book Co.

Hillier, Frederick S. and Lieberman, Gerald J. (1974), Introduction to Operations Research, 2nd ed. San Francisco: Holden-Day.

Naddor, Eliezer (1966), Inventory Systems, New York: John Wiley and Sons.

Scott, A. J. and Smith, T.M.F. (1974), "Analysis of Repeated Surveys Using Time Series Methods" Journal of the American Statistical Association, 69, 674-678.

U.S. Department of Commerce (1978), Statistical Policy Working Paper 1: Report on Statistics for Allocation of Funds, Washington, D.C.: Office of Federal Statistical Policy and Standards.

Zellner, Arnold (1971), An Introduction to Bayesian Inference in Econometrics, New York: John Wiley and Sons.