ADAPTIVE SAMPLING

Steven K. Thompson
University of Alaska - Fairbanks

Adaptive sampling designs are those in which the selection procedure may depend sequentially on observed values of the variable of interest. Gains in efficiency can be achieved through adaptive sampling procedures. In this paper, some results on adaptive sampling are given and examples of adaptive designs are described.

Basu (1969) showed that theoretically the best sampling designs would depend on observed values. Zacks (1969) described the optimal sampling design of fixed sample size from a Bayesian point of view and showed that, except for special cases, such designs would be sequentially adaptive. He gave a sufficient condition for the optimal design to be 'single phase' or nonsequential. Suggestive as this result is, the optiaml design described would be virtually impossible to implement in practice, due to the detailed prior knowledge required of the population and the mathematical complexity involved (Solomon and Zacks, 1970). Subsequently, Zacks (1970) described a much simpler, two phase adaptive design for quality control sampling. Seber (1986) cites the potential importance of adaptive designs for the estimation of animal abundance. Cassel, et al. (1977) review the subject of adaptive designs under the term 'informative' designs.

Examples of Adaptive Sampling Designs

The examples of adaptive sampling designs described in this section are analyzed in detail in Thompson and Ramsey (1982).

The first example is an Alaska shrimp survey in which sampling consists of towing a net across the ocean floor, measuring the amount of shrimp caught and calculating the 'area swept' by the net to estimate average density of the population. When observed abundance is average or above, subsequent nearby tows are made one mile in length, while if observed abundance is below average, nearby tows are made only one-half mile in length. Thus, subsequent sampling intensity depends adaptively on observed abundance. Since locations of tows are selected at random within primary units, the procedure can be shown to be design-unbiased. Because of the schooling or aggregation tendencies of the shrimp, conditional expectation and mean square error are functions of observed abundance at nearby locations. Analysis of the conditional variance strncture showed a gain in efficiency of about 24% for a given amount of sampling effort through use of the adaptive procedure.

The second example is a survey of rare and endangered Hawaiian forest birds in which observers stand at selected sites for a specified time period counting every bird detected of a given species. If unusually high abundance is observed, additional sites nearby are selected. The strategy is not design-unbiased because sites are systematically located rather than selected at random, but is model-unbiased under stationary assumptions. The gain in efficiency with the adaptive procedure is estimated to be about 37% and is due to the aggregation tendencies of the birds.

Advantages of Adaptive Designs

Although the practical adaptive designs described above do produce important gains in efficiency over nonadaptive designs, one is led to suspect that considerably higher gains are possible. To achieve these gains, a combination of theoretical development and evaluation of practical examples is needed.

In the following summary of the theoretical developments, it will first be assumed that the sampling effort is limited to a fixed total sample size n. Suppose we have just observed the first m units selected. Let Y denote the vector of observations of these first-phase units. Let X denote the remaining n-m units to be selected. Should X depend on the initial obsrevations Y?

Let T denote our estimate of some population quantity θ (such as population mean or total). The estimator T will be a function of the total data set of n observations.

The ideal adaptive two-phase design would select the second phase units X which minimize $E\{T-\theta^2|Y,X]$, giving conditional mean square error

$$\inf_X E\{(T-\theta)^2|Y,X\}$$

Under such a design, the unconditional mean square error would be

$$E(T-\theta)^2=E\{\inf_X E\{(T-\theta)^2|Y,X\}\}$$

The set of all possible samples is assumed to be countable and hence we can identify each possible sample with an integer i and write

$$f_i(Y)=E_i\{(T-\theta)^2|Y\}$$

where E_i denotes expectation given sample i.

The mean square error can then be written

$$E(T-\theta)^2=\int\inf_i f_i(y)dP(y)$$

The best nonadaptive design, on the other hand, would give mean square error

$$\inf_i\int f_i(y)dP(y)$$

Such a design would choose the sample i to minimize mean squre error without taking the first phase observations Y into account.

In the above development we have implicitly conditioned on the given selection of first phase sites. The procedure for selecting the first phase units could be simple random sampling or some other probability design. Alternatively, the first phase sample could be chosen to minimize the unconditional mean square error taking into account the design to be used at the second phase. In our comparisons of two-phase adaptive sampling with nonadaptive sampling, we will assume the same selection proceure for each in the first phase, so that our argument will be unaffected by the first phase selection.

By a basic property of integration,

$$\int \inf f_i(y)dP(y) \leq \inf \int f_i(y)dP(y)$$

so that the optimal adaptive procedure will always be as good as or better than the best nonadaptive procedure. The following theorem gives necessary and sufficient conditions for the optimal design to be nonadaptive or single phase, that is, for equality to hold in the above expression.

Theorem 1. Let $\{f_n\}$ be a sequence of Borel measurable functions. Define a sequence $\{h_n\}$ by

$$h_n = f_i \varepsilon \{f_n\} \text{ such that } \int f_i dP = \min_{k \leq n} \int f_k dP.$$

Suppose there exists an integrable function g such that $|h_n| \leq g$ a.e. and suppose $\lim h_n = h$ a.e.

Then $\int \inf f_n dP = \inf \int f_n dP$ if and only if $h = \inf f_n$ a.e.

Before proving the theorem, a few remarks about its meaning for our problem will be given. Essentially, the theorem says that a nonadaptive design, in which the entire sample may be selected ahead of time, will be optimal if and only if there is some possible selection of second phase units which is best for every possible outcome of the first-phase observations.

Proof of Theorem 1. The proof is based on the Dominated Convergence Theorem. First, suppose $h = \inf h_n$ a.e. Then

$$\int \inf f_n = \int h = \int \lim h_n = \lim \int h_n$$

by the D.C.T. By the definition of h_n,

$$\lim \int h_n = \lim_{n \to \infty} \min_{k \leq n} \int f_k = \inf \int f_n,$$

establishing the sufficiency of the condition. Second, suppose $\int \inf f_n = \inf \int f_n$. By the definition of h_n we can write

$$\inf \int f_n = \lim_{n \to \infty} \min_{k \leq n} \int f_k = \lim \int h_n$$

By the D.C.T., $\lim \int h_n = \int \lim h_n = \int h$, i.e., $\int (h - \inf f_n) = 0$. But $h - \inf f_n = 0$ since $h_n \varepsilon \{f_n\}$ so $h = \inf f_n$ a.e., establishing necessity and completing the proof.

We next consider sampling procedures in which the sample size N depends adaptively on observed values. We consider a two-phase procedure in which n_1 units have been selected and observed in the first phase, and total sample size k (equivalently, second phase sample size) may depend on the first stage vector of observations Y. Let k(Y) denote the rule for choosing sample size as a function of the initial observatios Y. Conditional on the initial observations, we may write:

$$F(Y,k(Y)) = E\{(T-\theta)^2 | Y,k(Y)\}$$

so that, unconditionally, $E(T-\theta)^2 = E\{F(Y,k(Y))\}$. The optimal choice of sample size would be given by the function k(Y) which minimizes mean square error $E(T-\theta)^2$ subject to the fixed expected sample size $E(k(Y))=n$. That is, we wish to find a function k(y) which minimizes the integral $\int F(y,k(y))dP(y)$ subject to the constraint $\int k(y)dP(y)=n$.

Finding a function which minimizes an integral subject to a constraint in the form of another integral is the 'isoperimetric' problem of the calculus of variations. The follwoing theorem states the necessary condition for a function k(Y) to be the optimal two-phase sample size choice.

Theorem 2. The function k(y) which minimizes the integral $\int F(y,k(y))dP(y)$ subject to $\int k(y)dP(y)=n$ must satisfy $\partial F/\partial k = \lambda$ for some constant λ.

Proof: By Euler's equation, a necessary condition for the function k(y) to minimize $\int FdP$, suject to fixed $\int kdP$ is that k satisfy $\partial/\partial k\{F(y,k) - \lambda k\} = 0$ where λ is a lagrange multiplier. Hence the necessary condition is $\partial F/\partial k = \lambda$ for all outcomes y.

An important special case is the one in which $E\{(T-\theta)^2 | Y,k\} = c^2(Y)/k + d$ where c is a function of y only and d a constant. This is the case, for example, when units are selected by simple random sampling without replacement. In this case, the optimal sample size, based on the first phase observations, is

$$k(y) = nc(y)/E(c(Y))$$

The result is superficially similar to the well known optimal allocation in stratified sampling, with the difference that the allocation is not over strata but over the sample space of all possible outcomes of the first phase observations.

Acknowledgment: This work was partially supported by National Science Foundation Grant DMS-8705812.

References

Basu, D. (1969). Role of the sufficiency and likelihood principles in sample survey theory. Sankhya A 31, 447-454.

Cassel, C.M., Sarndal, C.E., and Wretman, J.H. (1977). Foundations of Inference in Survey Sampling. New York: Wiley.

Seber, G.A.F. (1986). A review of estimating animal abundance. Biometrics 42(2), 267-292.

Solomon, H. and Zacks, S. (1970). Optimal design of sampling from finite populations: A critical review and indication of new research areas. Journal of the American Statistical Association 65, 653-677.

Thompson, S.K. and Ramsey, F.L. (1983). Adaptive sampling of animal populations. Technical Report No. 82. Department of Statistics, Oregon State University, Corvallis.

Zacks, S. (1969). Bayes sequential designs of fixed size samples from finite populations. Journal of the American Statistical Association 64, 1342-9.

Zacks, S. (1970). Bayesian design of single and double stratified sampling for estimating proportions in finite populations. Technometrics 12, 119-30.