# OPTIMUM STRATIFIED SAMPLING USING PRIOR INFORMATION

Kallappa M. Koti, University of Alabama and Karnataka University, India


## Abstract

Bayesian approach to the problem of well known Neyman's stratified allocation, assuming prior information concerning the unknown coefficient of variation (cv), is presented. Only the case of non-sequential estimation based on one-stage sampling is considered. The solution ensures that the larger the relative variability in a stratum, the larger will be the sample size from that stratum, every stratum being represented by at least a specified number of sampling units in the stratified sample. The "least-cost" solution discussed in this note differs from the existing statistical literature on optimal design of sampling from finite populations in that it is based on the assumption that an "inadequate" representation of any stratum in the stratified sample may result in an irreparable loss of information to the investigator. The logic and operation of the solution for the case of cv being negative exponentially distributed is illustrated.


## 1. Introduction

The allocation of resources in a stratified sampling design aimed at estimating a population parameter is usually carried out, keeping in mind either one of the following two alternatives: (i) to achieve maximum precision for a given total cost of the survey, or (ii) to achieve a given precision at a minimum cost. The well known Neyman optimum allocation is one, based on this type of approach. From the Neyman optimum allocation (c.f., Cochran (1963)), it is apparent that the statistician planning the sample survey needs to have some prior information on the behavior of the character under consideration, other than the size of each stratum. The available literature on optimum stratified sampling using prior information consists of the papers by Aggarwal (1959), DeGroot and Starr (1969), Ericson (1965, 1968), Zacks (1970), and many others. An excellent review and criticism of all these may be found in Solomon and Zacks (1970).

In classical decision theory, the merit of a decision is examined by setting up what is called a loss function which with reference to the allocation problem under consideration, is usually made up of two components, one representing the error in the estimate, and the other representing the cost of observation. Then a Bayesian decision maker chooses a decision that minimizes what is called the "average risk" (see Aggarwal (1979)). Another alternative that seems to be appropriate and popular in statistical inference making is to consider the posterior variance (c.f., Ericson (1965); Zacks (1970)) as the average risk and to minimize it subject to a cost constraint. Whatever the approach may be, the basic question is: what prior information is available or can be assumed about the character under study? However, without entering any debate in this context, reference may be made again to Soloman and Zacks (1970). We assume prior information is available concerning the coefficient of variation (cv) of each stratum.

An important factor that is considered while determining the sample size is the inherent variability in the statistical distribution of measurements. In many applied fields, the coefficient of variation (cv = $\sigma/\mu$ or $\sigma/\mu \times 100$) has been used as a quantitative index of the measured variability. Nevertheless, in statistical inference making wherein normality assumption is involved, the cv may help to decide whether a normal approximation could be reasonable (c.f., Cochran (1963); Searls (1964); Rogowaski (1972)). In this article, it is assumed that "loss" associated with the error in the absence of an "adequate representation" of any stratum in the sample happens to be very large and perhaps unknown. Consequently, an expression for the expected prior risk that ensures an "adequate representation" of all strata in the "total sample" has been derived.

## 2. General Framework of the Problem

At the outset of this section, we state the fundamental objective of the sample survey in a normative way, and then introduce the essential requirements and subscripts definitions of this allocation model.

### 2.1 Statement of the Problem

Consider a set $U = \{u_1, u_2, \ldots, u_N\}$ of N distinguishable elements which are classified into k strata, the size of the ith stratum being $N_i$. Let $\underline{\pi} = (\pi_1, \pi_2, \ldots, \pi_k)$ be a vector of given constants, where $\pi_i = N_i/N$. With each element $u_i$ there is associated a real number $X_i = X(u_i)$. Furthermore, let C be the budget available for sampling, and $\underline{c} = (c_1, c_2, \ldots, c_k)$ be a vector of given constants, $c_i$ being the cost per observation of sampling within the ith stratum. Suppose $\underline{\mu} = (\mu_1, \mu_2, \ldots, \mu_k)$ is a vector of the unknown means of k strata. The fundamental objective of the sample survey is to make inference about, say, the population mean $\mu$ defined by

$$\mu = \underline{\pi}\underline{\mu}' \in R \text{ (the real line)}$$

on the basis of an optimum stratified sample

$$\underline{n} = (n_1, n_2, \ldots, n_k), \; n_i \geq 0$$

and where $n_1 + n_2 + \ldots + n_k = n$, a preassigned integer. Of course in this paper, we define an optimum stratified sample as the one that minimizes the expected prior risk r defined by (2.8).

### 2.2 Basic Assumptions and Definitions

The following preliminary definitions and assumptions may help to understand the formulation of the risk function given by (2.7) and the expected prior risk of (2.8). Let $\theta_i$ denote the cv in the ith stratum. It is assumed throughout this paper that $\theta_i$ (i = 1, 2, ..., k) are mutually independently distributed random variables with prior density $g_i(\theta)$ and that

$$G_i(u) = \Pr[\theta_i \le u]. \tag{2.1}$$

It is further assumed unless otherwise mentioned that

$$P[\theta_i \in (0,\infty)] = 1, \; i = 1, 2, \ldots, k. \tag{2.2}$$

It is important to note that we always assume

$$E(\theta_i) < \infty, \; i = 1, 2, \ldots, k. \tag{2.3}$$

Let $Z = \max_{1 \le j \le k}(\theta_j)$. Define I by $X_I$, and w$\ell$g, let $Z_i$ denote the Z given that I = i. We use $P_i$ to represent the probability that I = i, i = 1, 2, ..., k.

We call the k-dimensional Euclidean space $n^k$ defined by

$$n^k = \left\{\underline{n}: \sum_{i=1}^{k} n_i = n\right\}$$

the "action space" for, ultimately the problem reduces to that of choosing a proper vector $\underline{n}$. Specialty of this paper rests on the assumptions under which the loss function L is defined and on the way the risk function is obtained. Let the function L in (2.4) be a non-negative function defined on R x R, and represent the loss incurred when $\mu$ is estimated by $\hat{\mu}(\underline{X}, \underline{n})$.

$$\begin{aligned} L(\mu,\underline{\theta},\hat{\mu}) &= \sum_{i=1}^{k} \ell_i(\mu,\theta_i,\hat{\mu}) + \sum_{i=1}^{k} C_i(n_i) \\ &= e + \sum_{i=1}^{k} C_i(n_i) \end{aligned} \tag{2.4}$$

wherein for each i = 1, 2, ..., k, $\ell_i$ is non-negative unknown, and assumed to be very large whenever for "large" $\theta_i$ the sample size $n_i$ happens to be "small" and $C_i$'s are defined in (2.6).

Definition 2.1. The point $\underline{n_0} = (n_{01}, n_{02}, \ldots, n_{0k}) \in n^k$ at which e in (2.4) is supposed to reach the unique minimum is said to be the "adequate" stratefied sample.

Let $0 \equiv 0(\underline{n_0}) \subset n^k$ represent the base for the over neighborhood system at $\underline{n_0}$.

Definition 2.2. A point $\underline{n} \in 0$ is said to be a stratified sample "admissible with respect to error."

Thus, $\ell_i$ becomes inadmissibly large faster than the difference $n_{0i} - n_i \to n$ for every i = 1, 2, ..., k. Then the requirement of the stratified sampling design viz, "larger the relative variability in a stratum larger be the size of the sample from that stratum" is met by introducing a non-negative continuous, monotonically increasing auxiliary function,

$$Y = \eta(\theta), \; \theta \in (0,\infty) \tag{2.5}$$

with $\eta(0) = a$, a non-negative number, and confirming that "the sample size from the ith stratum is $\eta(\theta_i)$": w$\ell$g, we sometimes write $\eta_i(\theta)$ for $\eta(\theta_i)$. For instance one may set

$$\text{(i)} \quad \eta_i(\theta) = \alpha + \beta\theta_i, \; \beta > 0$$

or

$$\text{(ii)} \quad \eta_i(\theta) = \exp(\alpha + \beta\theta_i), \; \beta > 0$$

depending upon the total sample size n. The constants $\alpha$ and $\beta$ are to be chosen taking into consideration n, C, etc.

For the sake of simplicity, as mentioned at the beginning of this section, we assume that the cost of observation in the ith stratum is proportional to $\eta(\theta_i)$, i.e., for a given set of positive constants $\underline{c} = (c_1, c_2, \ldots, c_k)$, we let

$$\begin{aligned} C_i &= c_i\eta_i(\theta), \; \theta \in S_i \\ &= 0 \text{ otherwise} \end{aligned} \tag{2.6}$$

where

$$S_i = \{\theta: a \le \eta_i(\theta) \le n_i\}, \; i = 1, 2, \ldots, k,$$

in which $n_i$ is the size of the sample to be drawn from the ith stratum. We now put forth the following heuristic arguments. Since $\ell_i$ are unknown, and with an almost adequate stratified sample, e in (2.4) approaches the admissible minimum, it may be treated as a constant. Then as the second term on r.h.s. of (2.4) is independent of $\underline{X}$, we define the risk function R (ignoring e) as follows:

$$\begin{aligned} R(\mu,\underline{\theta},\hat{\mu}) &= \sum_{i=1}^{k} C_i[\eta(\theta_i)] \\ \text{i.e.,} \quad R &= \sum_{i=1}^{k} C_i\eta_i(\theta) \end{aligned} \tag{2.7}$$

by virtue of (2.6). Then an important task is to ensure an "adequate stratified sample," which is accomplished through the derivation of the expected prior risk r as a function of $\underline{n}$, as given in (2.8) below.

$$r(\underline{n}) = E_{\underline{\theta}}(R|\underline{n} \in 0) \tag{2.8}$$

Thus, from the foregoing discussion, it is obvious that ultimately the problem is to find a non-negative vector $\underline{n} \in 0$, that minimizes r of (2.8) subject to the constraint $\Sigma n_i = n$. It may be pointed out in advance that the average prior risk of (2.8) would be different from the traditional Bayes risk which is usually defined as the unconditional expected value of R in that (i) in its derivation the weightage given to $\theta_i$ differs from its actual prior density, and (ii) it is independent of the error component, that is to say that it happens to be simply a cost-function. Therefore r may be sometimes called the "cost-function" rather than the Bayes risk.

## 3. The Two Strata Case

The derivation of the expression for r needs a special attention as it is nested deep in the mathematics of probability and order statistics. Therefore, before going directly to the general case of k(>2) strata, it is worth describing the case of only two strata. The following lemma 3.1 is found useful in deriving the expression for the average prior risk r.

Lemma 3.1. Let $Z = \max(X_0,X_1)$, where $X_0$ and $X_1$ are two independent r.v.s., and if I is defined by

$Z = X_I$, then the p.d.f. $f_i$ of $Z_i$ is given by

$$P_i f_i = g_i(z) G_{1-i}(z), \quad i = 0, 1 \qquad (3.1)$$

where $P_i = \Pr[I = i]$.

Proof. Proof of the lemma is simple, straightforward, and directly follows from No'das (1970).

In case of two strata, the admissible average prior risk (the average cost) is obtained as follows: When $\theta_i > \theta_j$, we set

$$R_i(\underline{\theta}) = c_i \eta(\theta_i) + c_j \eta(\max\theta_j | \theta_i > \theta_j), \quad i \neq j \qquad (3.2)$$

and evaluate the expected prior risk r as

$$r(n_1) = \sum_{i=1}^{2} P_i E(R_i | I=i) \qquad (3.3)$$

Note that r is shown to be a function of $n_1$ alone as $n_2 = n - n_1$. Thus the problem now reduces to that of finding a $n_1$ that minimizes (3.3). In order to obtain the r.h.s. of (3.3), we make use of the lemma 3.1 and a result on conditional expectation of section 6.5.2 in Ash (1972). Then by virtue of (2.6), the usual technique of transformation of r.v.s. gives

$$r(n_1) = \sum_{\substack{i=1 \\ n_i \geq n_j}}^{2} \{c_i \int_a^{n_i} y \cdot g_i[\eta^{-1}(y)] G_j[\eta^{-1}(y)] \left|\frac{d}{dy}\eta^{-1}(y)\right| \cdot dy$$
$$+ c_j \int_a^{n_j} y [\prod_{i=1}^{2} g_i(\eta^{-1}(y))] \left|\frac{d}{dy}\eta^{-1}(y)\right| \cdot dy\} \qquad (3.4)$$

## 3.1. A Particular Case: Uniform Priors

Assume $\theta_1$ and $\theta_2$ to be two independent random variables uniformly distributed, respectively over $(a_1,b_1)$ and $(a_2,b_2)$. Without loss of generality, let

$$0 \leq a_1 \leq a_2 \leq b_1 \leq b_2 < \infty$$

It may be noted here that we are slightly deviating from the assumption (2.2), and consequently we have to add constraints (3.6) to the expression (3.5) of $\gamma$. Then the expression (3.5) directly follows from (3.1) through (3.4).

$$w_1 w_2 r = \sum_{\substack{i=1 \\ i \neq j}}^{2} \{c_i \int_m^{n_i} y[\eta^{-1}(y) - a_i] \left|\frac{d}{dy}\eta^{-1}(y)\right| \cdot dy$$
$$+ c_j \int_m^{n_j} y \left|\frac{d\eta^{-1}(y)}{dy}\right| \cdot dy\} \qquad (3.5)$$

with $n_i \geq n_j$ for i = 1, 2, and where $w_i = b_i - a_i$, $m = \eta(a_2)$, and where we should have

$$n_i \leq \eta(b_i), \quad n - n_i = n_j \leq \eta(b_j), \quad i \neq j \qquad (3.6)$$

If however

$$\eta(\theta_i) = \alpha + \beta\theta_i, \quad i = 1, 2$$

the expression (3.5) reduces to be

$$w_1 w_2 r = \sum_{\substack{i=1 \\ i \neq j}}^{2} \{\frac{c_i}{\beta^2} \int_m^{n_i} y(y - \alpha - \beta a_i) dy + \frac{c_j}{\beta} \int_m^{n_j} y dy\}$$

with $n_i \geq n_j$, for i = 1, 2. This can be simplified to be

$$w_1 w_2 r = \sum_{\substack{i=1 \\ i \neq j}}^{2} \{(\frac{c_i}{\beta})[\frac{1}{3}(n_i^3 - m^3) - \frac{(\alpha + \beta a_i)}{2}(n_i^2 - m^2)]$$
$$+ \frac{c_j}{2}(n_j^2 - m^2)\} \qquad (3.6)$$

wherein $n_i \geq n_j$, for i = 1, 2. Convexity of (3.6) wrt $n_1$ can easily be verified.

## 4. The General Case of k Strata

The derivation of the cost function for a general k(>2) is slightly complicated and is based on concepts similar to those studied in Koti (1979) and No'das (1970). Suppose the k strata are serially numbered from 1 to k, and let for a given i

$$V_{(k-1,i)} = \max_{i \leq j \leq k} (\theta_j, j \neq i) \qquad (4.1)$$

Denoting again the p.d.f. of $Z_i$ by $f_i$, intuitively we should have that

$$P_i f_i(z) = g_i(z) \Pr[V_{(k-1,i)} \leq z] \qquad (4.2)$$

Generalizing the notation given in (4.1), for a specified set $(i_1, i_2, \ldots, i_r)$, we use the notations

$$V_{(k-1,i_1,\ldots,i_r)} =$$
$$\max_{1 \leq j \leq k} (\theta_j, j \neq i_1, i_2, \ldots, i_r), \qquad (4.3)$$

and $H_{(k-r,i_1,\ldots,i_r)}(\cdot)$ to represent the distribution function of V defined in (4.3). Let $E_{ij_1j_2 \ldots j_{k-1}}$ be the set of points on the orthant I=i, such that

$$\theta_i \geq \theta_{j_i} \cdots \geq \theta_{j_{k-1}}$$

Then on $E_{ij_1,j_2,\ldots,j_{k-1}}$, a possible observation on R would be of the form:

$$R_{ij_1j_2\ldots j_{k-1}}(z) = c_i \eta_i(z) + c_{j_1}(\max\theta_{j_1} | \theta_{j_1} \leq \theta_i)$$
$$+ \sum_{\alpha=2}^{r} c_{j\alpha} \eta(\max\theta_{j\alpha} | \theta_{j\alpha} \leq \theta_{j\alpha-1}) \qquad (4.4)$$

Perhaps, at this stage, it is worth making the following observation viz "If we assume for a moment that, for some $0 < a_j < b_j < \infty$, that

$$\Pr[\theta_j \varepsilon (a_j, b_j)] = 1, \quad j = 1, 2, \ldots, k$$

then on $E_{ij_1j_2\ldots j_{k-1}}$, we should have in (4.4) that

$$\max \theta_{j_1} = \min (z, b_{j_1})$$

---

$$\max \theta_{j_{k-1}} = \min (z, b_{j_1} \ldots b_{j_{k-1}}) ,$$

and we further add to the complexity." That is why we stick to the assumption (2.2). As a result (4.4) may be rewritten as

$$R_{ij_1j_2\cdots j_{k-1}}(z) = c_i\eta_i(z) + \sum_{\ell=1}^{k} c_{j_\ell}\eta_{j_\ell}(z) \qquad (4.5)$$

In order to evaluate $r_i = E(R|I=i)$, $i = 1, 2, \ldots, k$, again we make use of section 6.5.2 of Ash (1972) and (2.6). Then by the usual technique of transformation of r.v.s., we obtain the expression for $r_i$ given in (4.6).

$$P_ir_i = c_i\int_a^{n_i} y\cdot g_i[\eta^{-1}(y)]$$

$$\times H_{(k-1,i)}[\eta^{-1}(y)]\left|\frac{d}{dy}\eta^{-1}(y)\right|\cdot dy$$

$$+ \sum_{r=1}^{k} \sum_{\substack{\ell=1 \\ \ell\neq i}} \sum_{j_1<\ldots<j_r} c_\ell\int_a^{n_\ell} y[\prod_{s=1}^{\gamma} g_{j_s}[\eta^{-1}(y)]$$

$$\times g_\ell[\eta^{-1}(y)]$$

$$\times H_{(k-r-2,i,j_1\ldots j_r,\ell)}[\eta^{-1}(y)]\left|\frac{d\eta^{-1}(y)}{dy}\right|\cdot dy \qquad (4.6)$$

wherein $n_i \geq n_{j_1} \geq \cdots \geq n_{j_{k-1}}$, $\underline{n}$ being the decision vector. Then making use of (4.6) we obtain the expression for the expected prior risk as given in (4.7)

$$r(\underline{n}) = \sum_{i=1}^{k} P_ir_i \qquad (4.7)$$

which is to be minimized w.r.t. $\underline{n}$ such that

$$\sum_{i=1}^{k} n_i = n .$$

## 5. Case of 3 Strata

### An Illustration: Exponential Priors

We consider here the allocation of resources to three strata, wherein cv of each stratum is exponentially distributed.

Definition 4.1. A random variable U is said to have an "exponential distribution" with parameter $\lambda$, whenever for $\lambda>0$,

$$g_U(u) = \lambda e^{-\lambda u},\ u > 0$$
$$= 0 \text{ otherwise} \qquad (5.1)$$

We state the following lemma 4.1 without proof.

Lemma 4.1. If $U_i$, $i = 1, 2, 3$, are three r.v.s., independently, exponentially distributed with parameters $\lambda_i$, $i = 1, 2, 3$, and if

$$Y_{(2,i)} = \max_{j\neq i} (U_j)$$

then

$$H_{(2,i)}(z) = 1 - \bar{e}^{\lambda_j z} - \bar{e}^{\lambda_\ell z} + \bar{e}^{(\lambda_j+\lambda_e)z}, \quad z>0 \quad (5.2)$$

Let us assume that all practical considerations led us to set

$$Y = \eta(\theta) = e^{\alpha+\beta\theta} \qquad (5.3)$$

We use throughout this section the following notations.

$$\lambda = \lambda_1 + \lambda_2 + \lambda_3,\ \zeta = \lambda_1\lambda_2\lambda_3 ,$$

and

$$u_i = (\ln(n_i) - \alpha)/\beta$$

Then the expression (4.7) after proper substitution and simplifications works out to be as follows.

$$\bar{e}^{\alpha}r(\underline{n}) = \sum_{i=1}^{3} \{c_i\lambda_i\int_0^{u_i}[\bar{e}^{(\lambda_i-\beta)t} - \bar{e}^{(\lambda_i+\lambda_j\beta)t}$$

$$-\bar{e}^{(\lambda_i+\lambda_\ell-\beta)t} + \bar{e}^{(\lambda-\beta)t}]dt$$

$$+ \sum_{\substack{j=1 \\ \neq i\neq \ell}}^{3} c_j\lambda_i\lambda_j\int_0^{u_j}[\bar{e}^{(\lambda_i+\lambda_j-\beta)t} - \bar{e}^{(\lambda-\beta)t}]dt$$

$$+ \sum_{\substack{\ell=1 \\ \neq i,\ell}}^{3} \zeta\cdot c_\ell \int_0^{u} \bar{e}^{(\lambda-\beta)t}dt\} \qquad (5.4)$$

which is to be minimized with respect to $\underline{n} = (n_1, n_2, n_3)$ such that $n_1 + n_2 + n_3 = n$.

## 6. Concluding Remarks

The basic assumption underlying the model viz "larger the cv, larger be the sample size" seems to be satisfactory and is the ideal one in many real-life problems. However, as a precaution, it is worth stating here that, though in a different context, Rao et al. (1978) have asserted in their study that a larger sample size was required with decreasing coefficient of variation.

Since we have a least-cost allocation that is "admissible w.r.t. error," the result of this paper may help the statistician still at the planning stage, to set an optimal C, the total cost of the sample survey.

With the approach discussed here, the case of unknown strata size $N_i$ as studied by DeGroot and Starr (1969) does not arise. In other words, the advantage of this model is that, for all practical purposes for the sake of finding an optimal $\underline{n}$, the knowledge of $N_i$ is unnecessary.

We do admit here that this work is incomplete in many respects. If however the practitioners accept the solution of this paper as the feasible one for some sample surveys, further research may be carried on the following lines: (i) at the outset, a multivariate analogue of the results of this model seems to be essential and persued, (ii) though derivation of the posterior distribution of $\theta_i$ certainly is a hard exercise, a two stage or sequential sampling design is worth attempting, (iii) finally application of this model in regression analysis employed to study problems in econometrics (c.f., Ericson (1965)) may be considered.

## Bibliography


Aggarwal, Om. P. (1959). Bayes and minimax procedures in sampling from finite and infinite populations. The Annals of Mathematical Statistics, Vol. 30, pp. 206-218.

Ash, Robert B. (1972). Real Analysis and Probability. Academic Press. New York.

Cochran, W. G. (1963). Sampling Techniques, 2nd Edition, John Wiley & Sons. New York.

DeGroot, M. M. and Starr, N. (1969). "Optimum two stage stratified sampling." The Annals of Mathematical Statistics, Vol. 40, pp. 575-82.

Ericson, W.A. (1965). "Optimum stratified sampling using prior information." Journal of the American Statistical Association, Vol. 60, 750-71.

Ericson, W. A. (1968). "Optimal allocation in stratified and multistage sampling using prior information." Journal of the American Statistical Association, Vol. 63, 964-983.

Koti, Kallappa, M. (1979). "A probabilistic model in inventory distribution management." (Submitted for publication in Naval Research Logistics Quarterly).

No'das, Arthur (1970). "On estimating the distribution of a random vector when only the smallest co-ordinator is observable." Technometrics, 12, 923-924.

Rao, P. V., P. S. C. Rao, J. M. Davidson, and L. C. Hammond (1978). "Use of goodness-of-fit tests for characterizing the spatial variability of soil properties." Technical Report No. 135. Department of Statistics. University of Florida.

Rogowaski, A. S. (1972). "Watershed physics: Soil variability criteria." Water Resources Res. 8, 1015-1023.

Searls, Donald T. (1964). "The utilization of a known coefficient of variation in the estimation procedure." Journal of the American Statistical Association, Vol. 59, 1125-1127.

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

Zacks, S. (1970). "Bayesian design of single and double stratified sampling for estimating proportion in finite population." Technometrics, Vol. 12, 119-130.