BAYES AND MINIMAX PROCEDURES FOR ESTIMATING THE ARITHMETIC MEAN OF A POPULATION WITH TWO-STAGE SAMPLING

By Om P. Aggarwal

Iowa State University

1. Introduction. Two-stage sampling (or sub-sampling) in sample surveys is a procedure under which the population is considered as divided into a number of "clusters" or groups of units of the population and sampling is carried out in two stages. At first a sample of clusters is selected, and thus clusters become the first-stage sampling units or primary sampling units (p.s.u.'s); then a sample is taken from each of the selected p.s.u.'s. The ultimate units of the population are these second-stage sampling units. Stratification and cluster sampling are special cases of two-stage sampling when the rate of sampling is 100% at the first and the second stage, respectively. When the rate of sampling at each stage is less than 100%, one generally uses the term "two-stage sampling" or "subsampling."

In the current practice of choosing a survey design, the statisticians use one of the two principles: (i) to get an estimator of maximum precision for a given total cost of the survey, or (ii) to get an estimator of a given precision for a minimum total cost of the survey. The allocation of the resources for a given survey is usually carried out keeping one or the other of these two principles as the guide. The author [1] considered jointly the losses resulting from the errors in the estimators and from the cost of sampling, and obtained Bayes and minimax procedures for the estimation of mean in the case of an infinite population as well as a finite population. The loss function was taken as the sum of two components, one proportional to the square of the error of the estimator and the other proportional to the cost of obtaining and processing the sample. Both the case of a simple random sample and stratified random samples were discussed and a formula was obtained for the optimum allocation of the resources with a simple cost function.

In this paper we shall discuss two-stage sampling. We shall also use, for simplicity, the term "clusters" for first-stage units. The two cases, infinite and finite populations, shall be treated separately. For the sake of generality, we shall consider the case where the clusters are of unequal sizes and obtain the results for equal size clusters as a special case.

2. Infinite populations. Unequal-sized sampling at the second stage. Consider the situation where a statistician is required to estimate the mean μ of some random variable Y with a known upper bound σ_b^2 (> 0) for variance. He chooses a random sample of some predetermined size, say m ($\geqq 1$), but is not allowed to or is unable to observe the values obtained, say $\mu_1, \mu_2, \cdots, \mu_m$. He is, however.

Received 8 December 1965; revised 20 April 1966.

allowed to observe a predetermined number of observations, say n_i ($\geqq 1$) each from the conditional distribution of some other random variables, X_i ($i = 1, 2, \cdots, m$), corresponding to the values $\mu_1, \mu_2, \cdots, \mu_m$ obtained and unobserved by him, the conditional distributions of the X_i being such that

$$E(X_i \mid \mu_i) = \mu_i, \quad E[(X_i - \mu_i)^2 \mid \mu_i] \leqq \sigma_i^2 \ (> 0); \quad i = 1, 2, \cdots, m. \tag{2.1}$$

He is given the values of $\sigma_b^2, \sigma_1^2, \sigma_2^2, \cdots, \sigma_m^2$, and is required to estimate μ, the loss function L being given by

$$L(\mu, \delta) = \alpha(\delta - \mu)^2 + c_b m + \textstyle\sum_{i=1}^m c_i n_i, \tag{2.2}$$

where δ is a numerical-valued function of the sample observed by the statistician and used to estimate μ, c_b the sampling cost per unit for choosing the sample $(\mu_1, \mu_2, \cdots, \mu_m)$ at the first stage and c_i, for each i, the sampling cost per unit of observations on X_i conditional upon μ_i, and α is a positive constant. Without loss of generality we may take $\alpha = 1$. The problem is: what function δ should the statistician adopt as an estimator of μ, and further, what values of the m and n_i should he decide on for the sample sizes. The loss function with $\alpha = 1$ may now be written as

$$L(\mu, \delta) = (\delta - \mu)^2 + c_b m + \textstyle\sum_{i=1}^m c_i n_i. \tag{2.3}$$

3. Bayes and minimax strategies. Since the number of observations corresponding to the different values of the μ_i may be different, we shall obtain Bayes and minimax strategies corresponding to a fixed choice of the μ_i, i.e., corresponding to a fixed choice of clusters, by using the methods developed in [1]. That these strategies lead to an overall minimax strategy follows from a theorem which we shall state and prove presently. This theorem is similar to Theorem 6.1 of [1] which enables us to obtain minimax estimators for the given n_i in the case of stratified sampling and then choose the optimum n_i against the largest allowable σ_i^2. We state the theorem as follows:

THEOREM 3.1. *Suppose the space of strategies for the statistician is a union of spaces, say* $D = \bigcup_\lambda D_\lambda$. *Suppose* λ *is chosen randomly (i.e., not by the statistician of his free will) and that he must use a strategy from* D_λ *if* λ *is chosen. Suppose* δ_λ *is minimax in* D_λ *against the space of nature's strategies* Ω. *Then the strategy* δ *of using* δ_λ *when* λ *is chosen is minimax against* Ω *if:*

(i) *the risk* $R(\omega, \delta_\lambda) = r_\lambda$ *is constant (independent of* ω*) for each* λ,

(ii) r_λ *is a bounded function of* λ, *and*

(iii) *there exists a sequence of* θ*'s in* Ω *(not depending on* λ*) such that the Bayes risk against* θ *in* $D_\lambda \to r_\lambda$ *for all* λ.

PROOF. Since δ is the strategy which says, "Use δ_λ when λ is chosen," we have

$$R(\omega, \delta) = E_{(\lambda)} R(\omega, \delta_\lambda), \tag{3.1}$$

where the expectation is taken over the random variable λ.

Now,

$$\begin{aligned}(3.2)\qquad \max_\omega R(\omega, \delta) &= \max_\omega E_{(\lambda)} R(\omega, \delta_\lambda)\\ &\leqq E_{(\lambda)} \max_\omega R(\omega, \delta_\lambda) = E_{(\lambda)} r_\lambda\end{aligned}$$

because of the Condition (i).

Further, let δ' be a strategy for the statistician which chooses $\delta_\lambda' \varepsilon D_\lambda$ whenever λ is chosen. Then utilizing Conditions (iii) and (ii) respectively,

$$\begin{aligned}\max_\omega R(\omega, \delta') &= \max_\omega E_{(\lambda)} R(\omega, \delta_\lambda')\\ (3.3)\qquad &\geqq \lim_\theta E_{(\lambda)} R(\theta, \delta_\lambda')\\ &= E_{(\lambda)} \lim_\theta R(\theta, \delta_\lambda')\\ &\geqq E_{(\lambda)} \lim_\theta R(\theta, \delta_\lambda) = E_{(\lambda)} r_\lambda .\end{aligned}$$

From (3.2) and (3.3) it follows that

$$(3.4)\qquad \max_\omega R(\omega, \delta) \leqq E_{(\lambda)} r_\lambda \leqq \max_\omega R(\omega, \delta')$$

for any δ', proving that δ is minimax.

Consider now the problem of obtaining Bayes and minimax estimators corresponding to a fixed choice of clusters. Since we do not expect a least favorable distribution for μ, we shall find a sequence of Bayes estimators corresponding to a sequence $\{\lambda_\theta\}$ of prior distributions of μ, where λ_θ is normal with mean zero and variance θ^2, and will be denoted as $N(0, \theta^2)$.

Regarding the n_i as fixed pro tem, we may omit the terms $c_b m + \sum c_i n_i$ from the loss function (2.3) and note that the loss function is simply the mean square error. By an argument similar to that in Section 5 of [1], the Bayes estimator, δ_θ, is the mean of the conditional distribution of μ given the sample, and the Bayes risk, r_θ, is equal to the expected value of the variance of this conditional distribution.

Let the sample available for observation be denoted by $x = \{X_{ij}; i = 1, 2, \cdots, m; j = 1, 2, \cdots, n_i\}$ where X_{ij} is the jth observation on the conditional distribution of X_i given μ_i, and the sampling is independent for each X_i. Assume that Y is distributed as $N(\mu, \sigma_b^2)$ and that the conditional distributions of the X_i, given μ_i, are $N(\mu_i, \sigma_i^2)$. We notice that each $\bar{X}_i = n_i^{-1} \sum_{j=1}^{n_i} X_{ij}$ is a sufficient statistic for the corresponding μ_i and since the distribution of the μ_i depends upon μ, it follows that the set $\bar{X}^* = (\bar{X}_1, \bar{X}_2, \cdots, \bar{X}_m)$ is sufficient for μ. We may thus replace the sample x by the set $\bar{X}^*$ and obtain

$$(3.5)\qquad \delta_\theta = E(\mu \mid \bar{X}^*),$$

$$(3.6)\qquad r_\theta = E\sigma^2_{\mu\mid\bar{X}^*},$$

where the notation for conditional expectation is quite obvious.

Since $\bar{X}_i$ for given μ_i is distributed as $N(\mu_i, \sigma_i^2/n_i)$ and μ_i for given μ is $N(\mu, \sigma_b^2)$, it follows by the independence of their distributions that $\bar{X}_i$ for given μ is $N(\mu, \sigma_b^2 + \sigma_i^2/n_i)$. As the prior distribution of μ has been assumed to be $N(0, \theta^2)$, the joint distribution of the $\bar{X}_i$ $(i = 1, 2, \cdots, m)$ and μ is $(m + 1)$-

variate normal, from which it can be seen that the conditional distribution of μ, given $\bar{X}^*$, is $N(\alpha, \beta)$, where

$$\alpha = \sum_{i=1}^m w_i\bar{X}_i/(\sum_{i=1}^m w_i + \theta^{-2}), \tag{3.7}$$

$$\beta = [\sum_{i=1}^m w_i + \theta^{-2}]^{-1}, \tag{3.8}$$

and where

$$w_i = n_i/(n_i\sigma_b^2 + \sigma_i^2). \tag{3.9}$$

Thus, we see that, for a fixed choice of clusters, the Bayes estimator $\delta_\theta = \alpha$, where α is given by (3.7), and since the variance β, given by (3.8), is independent of $\bar{X}^*$, the Bayes risk $r_\theta = \beta$.

4. Minimax estimator. We now apply Theorem 2.2 of [1] for obtaining a minimax estimator if one exists. By letting $\theta \to \infty$, it is seen that $r_\theta \to r$ where

$$r = (\sum_{i=1}^m w_i)^{-1} = [\sum_{i=1}^m n_i/(n_i\sigma_b^2 + \sigma_i^2)]^{-1}. \tag{4.1}$$

We must now search for an estimator δ^* corresponding to which the risk does not exceed r. Let us try for such an estimator,

$$\lim_{\theta\to\infty} \delta_\theta(x) = \sum_{i=1}^m w_i\bar{X}_i/\sum_{i=1}^m w_i = \delta^*(x), \quad \text{say}. \tag{4.2}$$

Since the $\bar{X}_i$ are normal and independent with mean μ each and variances $1/w_i$, $\delta^*(x)$, being a linear function of the $\bar{X}_i$, is also normal, with mean μ and variance $(\sum_{i=1}^m w_i)^{-1}$. Hence, the risk corresponding to the estimator δ^* is equal to r, from which it follows that δ^* is a minimax estimator of μ.

5. Removal of the assumption of normality. We shall now drop the assumptions on the distributions of Y and the conditional distributions of the X_i. Under the statement of the problem in the beginning of Section 2, Y is a random variable with mean μ and variance $\leqq \sigma_b^2$, and for each i, the conditional distribution of the X_i, given μ_i, has mean μ_i and variance $\leqq \sigma_i^2$. It is easily verified that the risk corresponding to the estimator δ^* of μ, under these more general conditions, $R(\mu, \delta^*) \leqq r$, and thus δ^* is indeed a minimax estimator under the conditions stated in Section 2.

6. Minimax choice of sample sizes. A minimax sampling scheme for the given m clusters is, by Theorem 6.1 of [1], one which chooses the n_i so as to be "optimum" if the conditional variances of the X_i given μ_i are as large as the assumptions allow, viz. σ_i^2. Thus, the choice of the n_i for the given m clusters can be made by minimizing the risk

$$R(\mu, \delta^*) = [\sum_{i=1}^m n_i/(n_i\sigma_b^2 + \sigma_i^2)]^{-1} + c_bm + \sum_{i=1}^m c_in_i. \tag{6.1}$$

Theoretically speaking, the risk (6.1) should be minimized over the choice of the n_i under the restriction that they be positive integers. However, even without this restriction, the problem of the minimization of (6.1) appears to be a hopeless task to attempt, in general. Under some simplifying assumptions, it may be possible to obtain approximate solutions, and we shall discuss one such case in Section 9, after first considering the case of equal-sized sampling.

7. Infinite populations. Equal-sized sampling at the second stage. For the special case of equal-sized samples from each cluster, we assume that for each i, $n_i = n$, $c_i = c_w$, and $\sigma_i^2 = \sigma_w^2$. The Condition (2.1) and the loss function (2.3) then reduce to

$$E(X_i \mid \mu_i) = \mu_i, \qquad E[(X_i - \mu_i)^2 \mid \mu_i] \leqq \sigma_w^2 (>0), \tag{7.1}$$

and

$$L(\mu, \delta) = (\delta - \mu)^2 + c_b m + c_w mn, \tag{7.2}$$

respectively.

Since the number of observations for each value of the μ_i is the same, we do not need Theorem 3.1 for this special case. We may simply assume that the numbers m and n have been determined somehow and restrict ourselves to the choice of Bayes and minimax estimators corresponding to fixed m and n. The procedure outlined in Sections 3, 4 and 5 will now lead to the following results, obtained by replacing each n_i by n, σ_i^2 by σ_w^2 and c_i by c_w :

$$\text{Bayes estimator } \delta_\theta = n\theta^2 \textstyle\sum_{i=1}^m \bar{X}_i / (mn\theta^2 + n\sigma_b^2 + \sigma_w^2) \tag{7.3}$$

$$\text{Bayes risk } r_\theta = [mn/(n\sigma_b^2 + \sigma_w^2) + \theta^{-2}]^{-1} + c_b m + c_w mn. \tag{7.4}$$

$$\text{Minimax estimator } \delta = \bar{X} \tag{7.5}$$

$$\text{Minimax risk} = (\sigma_b^2/m) + (\sigma_w^2/mn) + c_b m + c_w mn. \tag{7.6}$$

8. Minimax strategy for choosing *m* and *n*. By a repeated application of Theorem 6.1 of [1], a minimax strategy for the statistician to choose m and n is to choose the "optimum" values of m and n against the maximum allowed variances. An exact solution of m and n minimizing the risk $(\sigma_b^2/m) + (\sigma_u^2/mn) + c_b m + c_w mn$ is difficult to obtain. However, an approximate solution is provided by standard calculus methods, ignoring the discreteness of m and n, as

$$m \cong \sigma_b / c_b^{\frac{1}{2}} \quad \text{and} \quad n \cong (c_b/c_w)^{\frac{1}{2}} (\sigma_w/\sigma_b). \tag{8.1}$$

9. A special case in unequal-sized sampling—approximate solution to the minimax choice of sample sizes. Suppose $c_i\sigma_i^2 = c$ for all i, then the minimax risk for the given m clusters, given by (6.1'), can be written as

$$R(\mu, \delta^*) = \{\textstyle\sum_{i=1}^m 1/[\sigma_b^2 + (c/c_i n_i)]\}^{-1} + c_b m + \sum_{i=1}^m c_i n_i, \tag{9.1}$$

and this is minimized, for given $\sum_{i=1}^m c_i n_i$, when $\sum_{i=1}^m (c_b^2 + (c/c_i n_i))^{-1}$ is maximized. It is easily seen that this requires $c_i n_i$ to be the same for all i (neglecting the discreteness problem). It then follows that $\sigma_i^2/n_i = c_i\sigma_i^2/c_i n_i = k$ (say) for all i. Thus the w_i, given by (3.9), are the same for all i, and from

(4.2), the estimator δ^* becomes the simple mean of the (cluster) sample means. The risk (9.1) reduces to

$$R(\mu, \delta^*) = (\sigma_b^2/m) + (k/m) + c_b m + (cm/k), \tag{9.2}$$

where k is at our disposal (through the n_i). The minimum of (9.2) occurs for $k = mc^{\frac{1}{2}}$, which determines the values of the n_i for given m.

In this case we can also say something about the choice of m. Neglecting the discreteness problem, we see that the minimum of (6.1) over $n_1, n_2, \cdots, n_m$, where $c_i\sigma_i^2 = c$ for all i, is given by

$$R(\mu, m, c_1, \cdots, c_m, \sigma_1^2, \cdots, \sigma_m^2) = (\sigma_b^2/m) + c_b m + 2c^{\frac{1}{2}}. \tag{9.3}$$

Taking the expectation over the joint distribution of the c_i and the σ_i^2, we note that, if m "clusters" are taken,

$$ER(\mu, m, c_1, \cdots, c_m, \sigma_1^2, \cdots, \sigma_m^2) = (\sigma_b^2/m) + c_b m + 2c^{\frac{1}{2}}. \tag{9.4}$$

It is easily seen that the value of m which minimizes (9.4) is given by

$$m = \text{integer nearest to } [(\sigma_b^2/c_b) + \tfrac{1}{4}]^{\frac{1}{2}}. \tag{9.5}$$

It may be pointed out here that the overall sample mean will not ordinarily be minimax. From the form (4.2) of the minimax estimator obtained, it is clear that if it were the overall sample mean, the minimax choice of the n_i would have to be one making w_i proportional to n_i which from (3.9) indicates that $n_i\sigma_b^2 + \sigma_i^2$ should be the same for each i. This leads to $n_i = \bar{n} + (\sigma_w^2 - \sigma_i^2)/\sigma_b^2$, where $\bar{n}$ and σ_w^2 are simple averages of the n_i and the σ_i^2, respectively. This would be rather surprising since the "clusters" with larger upper bounds for the variance would then be allocated the smaller number of sampling units.

10. Finite population. Unequal-sized clusters and unequal samples from the sampled clusters. Suppose we have a finite population consisting of M (>1) sub-populations, the ith sub-population $(i = 1, 2, \cdots, M)$ consisting of N_i (>1) units, and that X_{ij} denotes some numerical characteristic of the jth unit $(j = 1, 2, \cdots, N_i)$ in the ith sub-population, the mean for which is denoted by $\mu_i = N_i^{-1} \sum_{j=1}^{N_i} X_{ij}$. In what follows, we shall use the terms clusters or first-stage units or primary sampling units indiscriminately for the sub-populations. It is required to estimate the population mean $\mu = (\sum_{i=1}^{M} N_i)^{-1} \sum_{i=1}^{M} N_i\mu_i = (\sum_{i=1}^{M} N_i)^{-1} \sum_{i=1}^{M} \sum_{j=1}^{N_i} X_{ij}$, with a two-stage sampling plan as follows. An ordered set of positive integers is decided upon, say $(m, n_1, n_2, \cdots, n_M)$. A simple random sample (without replacement) of m clusters is drawn out of the total M clusters in the first stage. In the second stage, whenever the ith cluster of the population, $i = 1, 2, \cdots, M$ is selected in the sample at the first stage, a simple random sample (without replacement) of size n_i units is drawn from the N_i units comprising that cluster. Tlus, a sample consisting of a total number of $\sum n_i$ units is obtained for observation, where the summation $\sum$ is taken over those values of i which are selected in the first stage. If c_b is the cost of sampling per cluster at the first stage and c_i the cost of sub-sampling per unit in the ith cluster, the loss function L is assumed to be given by

$$L(\mu, \delta) = (\delta - \mu)^2 + c_b m + \sum c_i n_i, \tag{10.1}$$

where δ is an estimator for μ.

We assume that this finite population has been obtained by nature (or a conscious being) in the following manner. Nature first chooses a real number μ and then, using some distribution ω_b in the hyperplane $\mu_1 + \mu_2 + \cdots + \mu_M = M\mu$ in the M-dimensional Euclidean space (M-space), and subject to the restriction that

$$E_{\omega_b}\{\textstyle\sum_{i=1}^{M} (\mu_i - \mu)^2 \mid \mu\} \leqq (M - 1)\sigma_b^2, \tag{10.2}$$

where σ_b is a given positive number, obtains $\mu_1, \mu_2, \cdots, \mu_M$. Next, by using an ordered set $\{\omega_i; i = 1, 2, \cdots, M\}$ of distributions ω_i on hyperplanes in the N_i-spaces of the form

$$X_{i1} + X_{i2} + \cdots + X_{iN_i} = N_i\mu_i; \quad (i = 1, 2, \cdots, M), \tag{10.3}$$

and subject to the restrictions that

$$E_{\omega_i}\{\textstyle\sum_{j=1}^{N_i} (X_{ij} - \mu_i)^2 \mid \mu_i\} \leqq (N_i - 1)\sigma_i^2; \quad (i = 1, 2, \cdots, M), \tag{10.4}$$

where the σ_i are given positive numbers, nature obtains the finite population $\Pi = \{X_{ij}; j = 1, \cdots, N_i; i = 1, \cdots, M\}$. The parameter $\omega \varepsilon \Omega$ thus consists of the ordered set $(\mu, \omega_b, \omega_1, \cdots, \omega_M)$ whose elements have been defined above. If the sample obtained for observation is denoted by $x = \{x_{ij}\}$ where j ranges over the n_i values selected from $(1, 2, \cdots, N_i)$ in the second stage (and renumbered as $1, 2, \cdots, n_i$ for the sake of convenience of notation), while i ranges over the m values selected from $(1, 2, \cdots, M)$ at the first stage (and also renumbered as $1, 2, \cdots, m$ for the sake of convenience of notation), p_ω for $\omega \varepsilon \Omega$ is the distribution of x from the population Π distributed according to ω.

By Theorem 6.1 of [1], a minimax strategy for this problem is a minimax strategy for a given m, where m is chosen in an optimum way. By Theorem 3.1, this would be a minimax strategy corresponding to a given choice of m clusters. Accordingly, we shall obtain Bayes and minimax strategies corresponding to some given m clusters. In addition to m, if we also fix the n_i for the time being, we note that the terms $c_b m + \sum c_i n_i$ in the loss function (10.1) may be omitted temporarily, thus reducing the loss function to mean square error. If ξ denotes a prior distribution used by nature for picking $\omega \varepsilon \Omega$, and δ denotes the estimator used by the statistician to estimate $g(\omega) = \mu$, it will be seen, as before [1], that a Bayes estimator δ_ξ and the Bayes risk r_ξ are given by

$$\delta_\xi(x) = E(\mu \mid x), \tag{10.5}$$

and

$$r_\xi = E[E(\mu - \delta_\xi(x))^2 \mid x], \tag{10.6}$$

respectively.

11. Bayes estimators. Let the strategy ξ of nature for picking $\omega \varepsilon \Omega$ be a member of the sequence $\{\lambda_\theta\}$ where λ_θ is defined as follows: Pick μ from $N(0, \theta^2)$, and given μ, let the distribution ω_b, with probability one, be singular M-variate normal with mean μ, variance $\sigma_b^2(M - 1)/M$ for each component and covariance $-\sigma_b^2/M$ for each pair of components, and for each μ_i $(i = 1, 2, \cdots, M)$ so produced, let the distribution ω_i, with probability one, be singular N_i-variate normal distribution with mean μ_i, variance $\sigma_i^2(N_i - 1)/N_i$ for each component and covariance $-\sigma_i^2/N_i$ for each pair of components. Although the distribution of the sample x can be obtained from the joint distribution of the μ_i, $i = 1, \cdots, m$, and the x_{ij}, $i = 1, \cdots, m; j = 1, \cdots, n_i$, which is a $\sum (n_i + 1)$-variate normal distribution, it is not necessary to do so. We note that since the distribution of the μ_i depends upon μ, the set $(\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_m)$ where $\bar{x}_i = n_i^{-1} \sum_{j=1}^{n_i} x_{ij}$, which is sufficient for the set $(\mu_1, \mu_2, \cdots, \mu_m)$ in the conditional distribution of the sample (which depends upon $\mu_1, \mu_2, \cdots, \mu_m$), is also sufficient for μ when the distribution of the sample x depends upon μ. Consequently, the set $(\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_m)$ may replace the sample in Equations (10.5) and (10.6) above. We shall use θ rather than ξ as the subscript in both δ and r.

The distribution of $\bar{x}_i$, given μ_i, can be seen to be $N(\mu_i, v_i)$ where $v_i = (n_i^{-1} - N_i^{-1})\sigma_i^2$, and since the sampling is independent for each cluster, the joint distribution of the $\bar{x}_i$ given the μ_i is the product of m normal distributions. The sampled μ_i are themselves distributed as m-variate normal with mean μ and variance $\sigma_b^2(M - 1)/M$ for each component and covariance $-\sigma_b^2/M$ for each pair of components. Let $\bar{x}^*$, μ^*, and e denote the $m \times 1$ column vectors with components $(\bar{x}_1, \cdots, \bar{x}_m)$, $(\mu_1, \cdots, \mu_m)$, and $(1, \cdots, 1)$ respectively and I an $m \times m$ unit matrix. If $N_k(u, \Sigma)$ denotes the k-variate normal distribution with mean vector u and covariance matrix Σ, then, given μ, μ^* is distributed as $N_m(\mu e, A)$ where $A = \sigma_b^2(I - M^{-1}ee')$, and, given μ^*, the distribution of $\bar{x}^*$ is $N_m(\mu^*, B)$ where B is the diagonal matrix with the v_i as the diagonal elements. It is then easily seen that $\bar{x}^*$ for given μ is distributed as $N_m(\mu e, W)$, where $W = A + B$. Since μ has the prior distribution $N(0, \theta^2)$, the joint distribution of μ and $\bar{x}^*$ is $(m + 1)$-variate normal from which it can be shown that the conditional distribution of μ, given $\bar{x}^*$, is $N(\alpha, \beta)$, where

$$\alpha = e'W^{-1}\bar{x}^*/(e'W^{-1}e + \theta^{-2}), \tag{11.1}$$

and

$$\beta = (e'W^{-1}e + \theta^{-2})^{-1}. \tag{11.2}$$

Thus, for fixed choice of clusters, the Bayes estimator $\delta_\theta = \alpha$, and since the variance β is independent of $\bar{x}^*$, the Bayes risk $r_\theta = \beta$, where α and β are given by (11.1) and (11.2), respectively.

12. Minimax estimator. We again apply Theorem 2.2 of [1] for obtaining a minimax estimator if one exists. When $\theta \rightarrow \infty$, we see that $r_\theta \rightarrow r = (e'W^{-1}e)^{-1}$. We must now search for an estimator δ^* corresponding to which the risk does not exceed r. Let us try

$$\delta^*(x) = \lim_{\theta \to \infty} \delta_\theta(x) = e'W^{-1}\bar{x}^*/e'W^{-1}e. \tag{12.1}$$

The risk corresponding to the estimator δ^*, apart from the sampling costs, is given by

$$(12.2) \qquad E_\omega(\delta^* - \mu)^2 = [1/(e'W^{-1}e)^2]E[e'W^{-1}(\bar{x}^* - \mu e)]^2 = e'W^{-1}W_1W^{-1}e/(e'W^{-1}e)^2,$$

where $W_1 = A_1 + B_1$, $A_1 = \eta_b^2(I - M^{-1}ee')$, and B_1 is the diagonal matrix with diagonal elements $v_i' = (n_i^{-1} - N_i^{-1})\eta_i^2$, and where

$$(12.3) \qquad \eta_b^2 = (M - 1)^{-1}E_{\omega_b}\{\textstyle\sum_{i=1}^M (\mu_i - \mu)^2 \mid \mu\} \leqq \sigma_b^2,$$

and

$$(12.4) \qquad \eta_i^2 = (N_i - 1)^{-1}E_{\omega_i}\{\textstyle\sum_{j=1}^{N_i} (X_{ij} - \mu_i)^2 \mid \mu_i\} \leqq \sigma_i^2; \quad i = 1, \cdots, M.$$

After some algebraic manipulations, it can be shown that $e'W^{-1}W_1W^{-1}e \leqq e'W^{-1}e$, thus leading to the result that $E_\omega(\delta^* - \mu)^2 \leqq (e'W^{-1}e)^{-1} = r$ and proving that $\delta^*(x)$ is a minimax estimator of μ. We note that this estimator is a weighted mean of $\bar{x}_1, \cdots, \bar{x}_m$, and may be written as

$$(12.5) \qquad \delta^*(x) = \textstyle\sum_{i=1}^m w_i\bar{x}_i/\sum_{i=1}^m w_i,$$

where w_i is the ith element of $W^{-1}e$, and is given by

$$(12.6) \qquad w_i = [d_i(1 - (\sigma_b^2/M)\textstyle\sum_{j=1}^m d_j^{-1})]^{-1}$$

and where

$$(12.7) \qquad d_i = (n_i^{-1} - N_i^{-1})\sigma_i^2 + \sigma_b^2.$$

The maximum possible risk, apart from the sampling costs, is given by

$$(12.8) \qquad r = (e'W^{-1}e)^{-1} = (\textstyle\sum_{i=1}^m w_i)^{-1} = (\sum_{i=1}^m d_i^{-1})^{-1} - M^{-1}\sigma_b^2.$$

13. Minimax choice of sample sizes. As discussed in Section 6, a minimax sampling scheme, for the m sampled clusters, is one which chooses the n_i so as to be "optimum" if the variances "within clusters" are as large as allowed by the assumptions, viz., σ_i^2. Thus the choice of the n_i for the given m clusters can be made by minimizing the risk

$$(13.1) \qquad R(\mu, \delta^*) = (\textstyle\sum_{i=1}^m d_i^{-1})^{-1} - M^{-1}\sigma_b^2 + c_b m + \sum_{i=1}^m c_i n_i.$$

As mentioned earlier, the risk (13.1) should be minimized over the choice of the n_i under the restriction that they be positive integers. However, even without this restriction, the problem of minimization of (13.1) appears to be hopeless in general. The problem may be solved, in special cases, with the techniques available for numerical analysis.

14. Finite population. Equal-sized clusters and equal-sized sampling from the sampled clusters. For this special case, suppose that, for each i, $N_i = N$ and

$n_i = n$. We may also assume that $c_i = c_w$ and $\sigma_i^2 = \sigma_w^2$ for each i. Since the number of observations from each sampled cluster is the same, it is not necessary to appeal to Theorem 3.1 for this special case. We may assume, as before, that the numbers m and n have been determined somehow and restrict ourselves to the choice of Bayes and minimax estimators for μ corresponding to fixed m and n. The procedure of Section 11 can be gone through now to obtain Bayes estimators directly or the same result can be obtained by substituting N for N_i, n for n_i and σ_w^2 for σ_i^2 in the results (11.1) and (11.2). It will be noticed that v_i now becomes independent of i, say v, and thus the matrix B reduces to vI. The square matrix W is now of the form in which each diagonal element has the same value, say a, and each off-diagonal element has also the same value, say $b \neq a$. By utilizing an easily derivable result, that the sum of the elements of any row or column of the inverse of such a matrix W is equal to the reciprocal of the sum of the elements of any row or column of W, we obtain, from (11.1) and (11.2) respectively, the following results:

$$\text{(14.1)} \qquad \text{Bayes estimator } \delta_\theta = \theta^2 \bar{x}_{..}/(\theta^2 + V), \qquad \text{and}$$

$$\text{(14.2)} \qquad \text{Bayes risk } r_\theta = (\theta^{-2} + V^{-1})^{-1} + c_b m + c_w mn,$$

where $\bar{x}_{..}$ and V are defined by

$$\text{(14.3)} \qquad \bar{x}_{..} = m^{-1} \textstyle\sum_{i=1}^{m} \bar{x}_i ,$$

the overall sample mean, and

$$\text{(14.4)} \qquad V = (m^{-1} - M^{-1})\sigma_b^2 + m^{-1}(n^{-1} - N^{-1})\sigma_w^2$$

respectively.

By using the limiting process directly as in Section 12, or noticing that the d_i and hence the w_i of that section are independent of i for this special case, and thus the estimator (12.5) becomes a simple average of all the sample means, $\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_m$, it is seen that the overall sample mean, $\bar{x}_{..}$, is a minimax estimator for given m and n. The minimax risk, for given m and n, becomes

$$\text{(14.5)} \qquad R = (m^{-1} - M^{-1})\sigma_b^2 + m^{-1}(n^{-1} - N^{-1})\sigma_w^2 + c_b m + c_w mn.$$

By a repeated application of Theorem 6.1 of [1], a minimax strategy for the statistician is to choose the "optimum" values of m and n against the maximum allowed variances. An approximate solution is provided by standard calculus methods, and we get, ignoring the discreteness of m and n,

$$\text{(14.6)} \qquad m \cong [(\sigma_b^2 - \sigma_w^2/N)/c_b]^{\frac{1}{2}}$$

$$n \cong [(c_b/c_w)(\sigma_w^2/(\sigma_b^2 - \sigma_w^2/N))]^{\frac{1}{2}}.$$

REFERENCE

[1] Aggarwal, Om P. (1959). Bayes and minimax procedures in sampling from finite and infinite populations—I. *Ann. Math. Statist.* **30** 206–218.