The Annals of Mathematical Statistics
1969, Vol. 40, No. 2, 575–582

OPTIMAL TWO-STAGE STRATIFIED SAMPLING[1]

BY M. H. DEGROOT AND N. STARR[2]

Carnegie-Mellon University

1. Introduction and summary. Suppose that a fixed total number N of observations are to be made in a population Π which is composed of two strata Π_1 and Π_2. For $i = 1, 2$, it is assumed that each observation in the stratum Π_i has a normal distribution with unknown mean θ_i and specified precision r_i ($r_i > 0$). It should be kept in mind that the precision of any normal distribution is the reciprocal of the variance.

Let p denote the unknown proportion of the total population Π which is included in the stratum Π_1. Then the mean $\bar{\theta}$ of the population Π is given by the equation $\bar{\theta} = p\theta_1 + q\theta_2$, where $p + q = 1$. In this paper, the problem of estimating the value of $\bar{\theta}$ will be studied.

It will be assumed that the loss which results from any estimate δ is the squared error $(\delta - \bar{\theta})^2$. It is well known that for any prior distribution of θ_1, θ_2, and p, the Bayes estimate of $\bar{\theta}$, after all of the observations have been taken, will be the mean of the posterior distribution of $\bar{\theta}$. Furthermore, the expected loss from this estimate will be the variance of the posterior distribution of $\bar{\theta}$. Therefore, we must find a sampling procedure for which the expected value with respect to the prior distribution of this posterior variance will be minimized.

Throughout this paper, it will be assumed that the joint prior distribution of θ_1, θ_2, and p is as follows: θ_1, θ_2, and p are independent; the distribution of p is a beta distribution with parameters α and β ($\alpha > 0, \beta > 0$); and for $i = 1, 2$, the distribution of θ_i is a normal distribution with mean μ_i and precision h_i. This joint distribution has the following fundamental property: After any number of observations have been taken from Π, Π_1, or Π_2, the posterior joint distribution of θ_1, θ_1, and p will again be of the same form and, in particular, θ_1, θ_1, and p will again be independent under their posterior distribution.

We assume that sampling will be carried out in two stages. At the first stage, a random sample of size m ($0 \leqq m \leqq N$) will be taken from the whole population Π. At the second stage, the remaining observations $N - m$ are to be allocated between the two strata Π_1 and Π_2. Hence, at the second stage, n_i observations are taken from the stratum Π_i, where $n_1 + n_2 = N - m$. The problem is to find an optimal choice of the design constants m, n_1, and n_2. Note that the value of m must be chosen in advance of any sampling, whereas the constants n_1 and n_2 need not be chosen until the values of the first m observations obtained from the whole population have been studied.

Received 24 April 1968.

[1] This research was supported in part by a grant from the R. K. Mellon Foundation and by the National Science Foundation grant GP-6296.

[2] Now at the University of Michigan.

In this paper, we shall develop effective approximations to the optimal sampling procedure for situations in which the total number N of available observations is large and, therefore, the optimal number m of observations which should be obtained at the first stage will also be large. The techniques which will be presented can be extended for studying populations which are composed of k strata ($k \geqq 2$), in each of which the observations have a normal distribution. However, although the theory can be extended without difficulty, the actual computations become somewhat more complex, and we shall not consider these extensions.

The optimal allocation of observations from the Bayesian point of view has also been studied by Ericson [2], [3] and Draper and Guttman [1]. Ericson [2] studied a related optimal one-stage stratified sampling scheme in which the proportion in each stratum is known. Draper and Guttman considered the optimal allocation at the second stage of a two-stage process, extending [2]. Ericson [3] investigated an optimal two-stage design different from ours in a nonresponse context. Here we shall study the basic problem of finding the optimal choice of m at the first stage.

2. The allocation at the second stage. Suppose that a fixed number t of observations are to be allocated between the two strata Π_1 and Π_2, and let n_i be the number of observations which will be taken from Π_i, where $n_i \geqq 0$ and $n_1 + n_2 = t$. In this section we shall find the values of n_1 and n_2 for which the expected loss is minimized. In Section 3 we shall let $t = N - m$ and regard these results as specifying the optimal allocation for the second stage of the two-stage process described earlier. There, the prior distribution of the three parameters θ_1, θ_2, and p for the problem of choosing n_1 and n_2 will actually be the posterior distribution of these three parameters after the first stage of sampling has been completed.

For any distribution such that θ_1, θ_2, and p are independent, it can be shown by a routine computation that

$$(1) \quad \text{Var } (\bar{\theta}) = E^2(p) \text{ Var } (\theta_1) + E^2(q) \text{ Var } (\theta_2) + \text{Var } (p)E[(\theta_1 - \theta_2)^2].$$

After n_i observations have been taken from Π_i for $i = 1, 2$, it follows from Bayes' theorem that the posterior distribution of p will be the same as its prior distribution, but the posterior distribution of θ_i will be normal with mean μ_i' and precision h_i' where

$$(2) \qquad \mu_i' = (h_i\mu_i + n_ir_i\bar{x}_i)/(h_i + h_ir_i), \qquad h_i' = h_i + n_ir_i.$$

Here, $\bar{x}_i$ is the average of the n_i observations from Π_i.

For any given values of n_1 and n_2, and any random variable W, let $E'(W)$ and Var$'$ (W) denote the expectation and variance of W under the posterior distribution after $\bar{x}_1$ and $\bar{x}_2$ have been observed. Furthermore, let $E(W)$ and Var (W) denote the expectation and variance of W under the prior distribution. In these terms, we must find values of n_1 and n_2 which minimize $E[\text{Var}' (\bar{\theta})]$.

It can be shown that

$$(3)\quad E\{E'[(\theta_1 - \theta_2)^2]\} = E[\theta_1 - \theta_2)^2] = h_1^{-1} + h_2^{-1} + (\mu_1 - \mu_2)^2 = A, \text{ say.}$$

It is important to note that the value of A, as specified by equation (3), does not depend on the choice of n_1 and n_2.

It now follows that for any given values of n_1 and n_2, the expected loss is specified by the equation:

$$(4)\quad E[\text{Var}'(\bar{\theta})] = E^2(p)/(h_1 + n_1 r_1) + E^2(q)/(h_2 + n_2 r_2) + A \text{ Var } (p).$$

Let

$$(5)\qquad \rho = (r_2/r_1)^{\frac{1}{2}} E(q)/E(p).$$

Then it can be shown that, subject to the constraints that $n_i \geqq 0$ $(i = 1, 2)$ and $n_1 + n_2 = t$, the value of n_1 which minimizes (4) is specified as follows:

$$(6\text{I})\qquad n_1 = 0 \quad \text{if} \quad t < (\rho h_1 - h_2)/r_2 ;$$

$$(6\text{II})\qquad n_1 = (h_2 - \rho h_1 + t r_2)/(\rho r_1 + r_2) \quad \text{if}$$

$$t \geqq \max\{(\rho h_1 - h_2)/r_2, (h_2 - \rho h_1)/\rho r_1\};$$

$$(6\text{III})\qquad n_1 = t \quad \text{if} \quad t < (h_2 - \rho h_1)/\rho r_1 .$$

One interesting property of the solution (6) is that the optimal values of n_1 and n_2 depend on the distribution of the unknown proportion p only through its expectation $E(p)$. In particular, the optimal allocation when it is known that p has a certain value p_0 will be the same as the optimal allocation when the value of p is unknown but $E(p) = p_0$. This property has also been noted by Draper and Guttman [1] in their problems.

3. The expected loss for two-stage sampling. Suppose now that a total of N observations are to be taken. A certain number m of them are first taken from the whole population Π and the number k_i which lie in stratum Π_i are noted $(k_1 + k_2 = m)$ as well as the values of these observations. The remaining $t = N - m$ observations are then to be allocated between the two strata.

Suppose again that the prior distribution of θ_1, θ_2, and p is as specified in Section 1. After the m observations have been taken in the first stage of sampling, the posterior distribution of p will be a beta distribution with parameters $\alpha' = \alpha + k_1$ and $\beta' = \beta + k_2$. For $i = 1, 2$, the posterior distribution of θ_1 will be normal with mean μ_i' and precision h_i', where μ_i' and h_i' are specified by equation (2) with n_i replaced by k_i. Furthermore, θ_1, θ_2, and p will still be independent under this posterior distribution. Again, for any random variable W, let $E'(W)$ and $\text{Var}'(W)$ denote the expectation and variance of W with respect to this posterior distribution.

Now suppose that the remaining $N - m$ observations will be allocated optimally between the strata Π_1 and Π_2, as prescribed by equation (6). The expected loss from this allocation, as computed at the end of the first stage of

sampling, can be found from equations (4), (5), and (6). Let ρ' denote the value of ρ specified by equation (5) with $E(p)$ and $E(q)$ replaced by $E'(p)$ and $E'(q)$. The expected loss will depend on which of the three conditions given in equation (6) is correct at the end of the first stage of sampling when $t = N - m$, ρ is replaced by ρ', and h_i is replaced by h_i' $(i = 1, 2)$. Let S_{I} denote the event that the inequality in (6I) is correct when the indicated replacements are made. Similarly, let S_{II} denote the event that the inequality in (6II) is correct, and let S_{III} denote the event that the inequality in (6III) is correct. Then the expected loss L' as computed at the end of the first stage of sampling, can be expressed as follows:

(7I) $$L' = [E'(p)^2/(h_1 + k_1r_1) + [E'(q)]^2/(h_2 + (N - k_1)r_2) + A' \operatorname{Var}'(p)$$ if S_{I} occurs;

(7II) $$L' = [r_2^{\frac{1}{2}}E'(p) + r_1^{\frac{1}{2}}E'(q)]^2/(h_1r_2 + h_2r_1 + Nr_1r_2) + A' \operatorname{Var}'(p)$$ if S_{II} occurs;

(7III) $$L' = [E'(p)]^2/(h_1 + (N - k_2)r_1) + [E'(q)]^2/(h_2 + k_2r_2) + A' \operatorname{Var}'(p)$$ if S_{III} occurs.

Here, A' denotes the value of A specified by equation (3) with h_i and μ_i replaced by h_i' and μ_i' for $i = 1, 2$.

We must now find a value of m such that $E(L')$ is minimized, where L' is specified by equation (7) and the expectation is computed under the prior distribution of θ_1, θ_2, and p. The variable A' which appears in equation (7) depends on the actual values of the first m observations and not just on the numbers k_1 and k_2 of these observations which belong to each of the two strata. Therefore, the first step in the computation of $E(L')$ will be the computation of the conditional expectation $E(A' \mid k_1, k_2)$ of A' when the numbers k_1 and k_2 are known $(k_1 + k_2 = m)$ but the values of these observations are not known. The result of this computation is

$$E(A' \mid k_1, k_2) = A, \tag{8}$$

where A is specified by equation (3).

The posterior distribution of p is a beta distribution as specified earlier. Therefore,

(9) $$E'(p) = (\alpha + k_1)/(\alpha + \beta + m) = 1 - E'(q) \text{ and}$$
$$\operatorname{Var}'(p) = (\alpha + k_1)(\beta + k_2)/(\alpha + \beta + m)^2(\alpha + \beta + m + 1).$$

It now follows from equations (8) and (9) that, for any given values of k_1 and k_2 $(k_1 + k_2 = m)$, the expected loss L is as follows:

(10I) $$L = (\alpha + \beta + m)^{-2}\{(\alpha + k_1)^2/(h_1 + k_1r_1) + (\beta + k_2)^2/[h_2 + (N - k_1)r_2] + A(\alpha + k_1)(\beta + k_2)/(\alpha + \beta + m + 1)\};$$

$$(10\text{II}) \quad L = (\alpha + \beta + m)^{-2}\{[r_2^{\frac{1}{2}}(\alpha + k_1) + r_1^{\frac{1}{2}}(\beta + k_2)]^2/(h_1 r_2 + h_2 r_1 + N r_1 r_2) + A(\alpha + k_1)(\beta + k_2)/(\alpha + \beta + m + 1)\};$$

$$(10\text{III}) \quad L = (\alpha + \beta + m)^{-2}\{(\alpha + k_1)^2/[h_1 + (N - k_2)r_1] + (\beta + k_2)^2/(h_2 + k_2 r_2) + A(\alpha + k_1)(\beta + k_2)/(\alpha + \beta + m + 1)\}.$$

The value of L is specified by equation (10I), (10II), or (10III) according as S_{I}, S_{II}, or S_{III} occurs.

We must now find a value of m which minimizes $E(L)$. For any given value of m, this expectation is computed under the distribution of k_1 and k_2 determined from the specified prior distribution of p. Because of the appearance of the random variables k_i in the denominators of some of the ratios in equation (10), and because of the fact that L has a different form on each of the three events S_{I}, S_{II}, and S_{III}, the value of $E(L)$ does not seem to be computable in a form that is suitable for minimization with respect to m.

However, a few simple results should be noted. If $h_1 \to \infty$ and $h_2 \to \infty$ in the prior distribution, then the values of θ_1 and θ_2 are known with high precision, and the only uncertainty about the population Π which remains concerns the value of p. For this reason, all N observations should be taken in the first stage from the whole population Π (i.e., $m = N$).

Also, if $\alpha \to \infty$ and $\beta \to \infty$ in the prior distribution then the value of p is known with high precision, and the statistician should not take any observations from the whole population Π (i.e., $m = 0$). He should allocate all N of the available observations in an optimal way between the two strata.

We shall now develop some approximations which are appropriate when the values of h_1, h_2, α, and β are not too large, and a large number N of observations are to be taken.

4. Approximations for large samples. Suppose now that we let $N \to \infty$. The value of m which minimizes $E(L)$ must also become infinite, for otherwise the posterior variance of p would be bounded away from 0 and, hence, so also would the minimum expected loss. As $m \to \infty$, then with probability 1,

$$(11) \qquad k_1/m \to p, \qquad k_2/m \to q, \qquad \text{and} \qquad \rho' \to cq/p,$$

where $c = (r_2/r_1)^{\frac{1}{2}}$. Furthermore, as $N \to \infty$ and $m \to \infty$ we shall let

$$(12) \qquad s = \lim (N/m)$$

and shall assume that s is well defined ($1 \leqq s \leqq \infty$).

It can now be shown from equations (6) and (10) that the value $L^* = Nr_2L$ can be approximated as follows, when $N \to \infty$ and $m \to \infty$:

$$(13\text{I}) \quad L^* = s[c^2 p + q^2/(s - p) + Bpq] \quad \text{if} \quad s < p + (q/c);$$

$$(13\text{II}) \quad L^* = (cp + q)^2 + Bspq \quad \text{if} \quad s \geqq \max\{p + q/c, cp + q\};$$

$$(13\text{III}) \quad L^* = s[(cp)^2/(s - q) + q + Bpq] \quad \text{if} \quad s < cp + q.$$

Here B is a constant specified by the equation

$$(14) \qquad B = r_2 A = r_2[(1/h_1) + (1/h_2) + (\mu_1 - \mu_2)^2].$$

TABLE

Optimal value s^* as a function of b ($\alpha = \beta = 1$)

$c = 1.25$		$c = 1.50$		$c = 1.75$	
b	s^*	b	s^*	b	s^*
0.000	1.250	0.000	1.500	0.000	1.750
0.010	1.181	0.011	1.380	0.012	1.579
0.050	1.138	0.075	1.288	0.055	1.481
0.101	1.114	1.150	1.242	0.309	1.318
0.506	1.046	0.750	1.114	1.237	1.154
1.012	1.016	1.500	1.056	3.094	1.051
$\geqq \gamma = 1.6875$	1.000	$\geqq \gamma = 3.750$	1.000	$\geqq \gamma = 6.1875$	1.000
$c = 2.00$		$c = 3.00$		$c = 5.00$	
b	s^*	b	s^*	b	s^*
0.000	2.000	0.000	3.000	0.000	5.000
0.018	1.758	0.024	2.521	0.072	3.759
0.090	1.609	0.240	2.076	0.504	2.920
0.450	1.400	1.200	1.644	2.880	2.020
1.800	1.187	4.800	1.268	7.200	1.594
4.500	1.060	12.000	1.079	43.200	1.058
$\geqq \gamma = 9.000$	1.000	$\geqq \gamma = 24.000$	1.000	$\geqq \gamma = 72.000$	1.000

We shall assume, without loss of generality, that $r_1 \leqq r_2$ and, hence, that $c \geqq 1$. With this assumption, the inequality in equation (13I) can never be satisfied since s must be a least 1. Therefore, if $f(p)$ denotes the prior beta pdf of p, then it follows from equation (13) that

$$(15) \quad V(s) = E(L^*) = \int_0^{s'} (cp + q)^2 f(p)\, dp + s\int_{s'}^1 [(cp)^2/(s - q) + q] f(p)\, dp + bs.$$

Here

$$(16) \quad s' = (s - 1)/(c - 1) \quad \text{and} \quad b = BE(pq) = B\alpha\beta/(\alpha + \beta)(\alpha + \beta + 1)$$

We have introduced the notation $V(s)$ in equation (15) to indicate that we shall now be primarily interested in studying the behavior of $E(L^*)$ as a function of s. Among all values of s ($s \geqq 1$), we must find a value s^* which minimizes $V(s) = E(L^*)$. It does not seem to be possible to obtain s^* as an explicit function of the given constants b, c, α, and β. However, various properties of the optimal value s^* can be developed.

5. Properties of the optimal procedure. Let b, c, α, and β be fixed constants such that $b > 0$, $c \geqq 1$, $\alpha > 0$, and $\beta > 0$, and let s^* be a value of s which minimizes $V(s)$, as defined by equation (15), over the set where $s \geqq 1$. Then, in problems with a large number of available observations, a proportion $1/s^*$ of the total number of observations should be taken from the whole population II at the first stage.

It is shown in the next lemma that s^* is never greater than c. Therefore, according to the approximations which we are using, at least the proportion $1/c$ of the available observations should be taken at the first stage from the whole population II. However, we have already remarked that if the value of p is known with high precision then the optimal procedure will specify taking very few observations at the first stage. These remarks serve to emphasize that our approximations are appropriate only when the values of α and β are not very large.

LEMMA. $s^* \leqq c$.

PROOF. For $s \geqq c$, $V(s)$ can be written in the following form:

$$V(s) = \int_0^1 (cp + q)^2 f(p)\, dp + bs. \tag{17}$$

This is an increasing linear function of s. Therefore, $V(c) < V(s)$ for $s > c$.

Since $s^* \geqq 1$, it follows from Lemma 1 that $s^* = 1$ when $c = 1$. This result has a highly interesting interpretation: If $r_1 = r_2$, then, regardless of the prior information about θ_1, θ_2, and p (within the range of the approximations which are being used), *all* of the observations should be taken from the whole population II at the first stage and no observations should remain to be allocated at the second stage.

Next, let γ be a non-negative constant defined as follows:

$$\gamma = (c^2 - 1)E(q) = (c^2 - 1)\beta/(\alpha + \beta). \tag{18}$$

Then we can establish the following result.

THEOREM. *If* $0 < b < \gamma$, *then* $V(s)$ *has a unique minimum at a value* s^* *such that* $1 < s^* < c$. *If* $b \geqq \gamma$, *then* $s^* = 1$.

PROOF. A straightforward computation shows that for $1 < s < c$, the derivative $V'(s)$ can be expressed as follows:

$$V'(s) = b - \int_{s'}^1 [(cp/(s - q))^2 - 1]qf(p)\, dp. \tag{19}$$

It can be seen from equation (19) that $V'(s)$ will be strictly increasing over the interval where $1 \leqq s \leqq c$, and that

$$V'(1) = b - \gamma \quad \text{and} \quad V'(c) = b. \tag{20}$$

The theorem now follows from these facts.

It can be seen from equations (19) and (20) that $s^* \to c$ as $b \to 0$. However, it should be kept in mind that the approximations which have been developed here are appropriate only when h_1 and h_2 are not too large and hence, by equations (14) and (16), they are appropriate only when b is not too small. It also follows from (19) and (20) that the value s^* which minimizes $V(s)$ is a decreasing function of b, for $0 < b < \gamma$. Thus, s^* decreases from c to 1 as b varies between 0 and γ. This observation has a highly interesting interpretation. Since b varies inversely with the precisions h_1 and h_2, then s^* varies directly with h_1 and h_2. In other words, the less we know about θ_1 and θ_2 (as measured by the prior precisions of these values), the smaller will be the proportion of observa-

tions we reserve for allocation at the second stage. This feature of the solution, which appears to be contrary to intuition, can be heuristically explained as being a consequence of the expression (1) for the variance of $\bar{\theta}$. The last term in this expression is Var $(p)E[(\theta_1 - \theta_2)^2]$. We have noted that the value of $E[(\theta_1 - \theta_2)^2]$ does not depend on the choice of m, n_1, and n_2, and that this value varies inversely with h_1 and h_2. Thus, the smaller that h_1 and h_2 are, the greater the factor which multiplies Var (p) will be; hence, the greater will be the importance of reducing this variance by taking a large number of observations from the total population.

6. Numerical results. When $0 < b < \gamma$, the unique solution s^* of the equation $V'(s) = 0$ can typically be computed without difficulty for moderate values of α and β. As a numerical example we take the case $\alpha = \beta = 1$, and compute the optimal value s^* of s as a function of the value b. These computations are prepared for $c = 5/4, 3/2, 7/4, 2, 3, 5$.

REFERENCES

[1] DRAPER, N. R. and GUTTMAN, I. (1967). Some Bayesian stratified two-phase sampling results. *Biometrika* **55** 131–139.

[2] ERICSON, W. A. (1965). Optimum stratified sampling using prior information. *J. Amer. Statist. Assoc.* **60** 750–771.

[3] ERICSON, W. A. (1967). Optimal sample design with nonresponse. *J. Amer. Statist. Assoc.* **62** 63–78.