OPTIMAL ALLOCATION IN STRATIFIED SAMPLING WITH PARTIAL INFORMATION

by

D.V. Lindley
2 Periton Lane
Minehead
Somerset TA24 8AQ
ENGLAND

and

J.J. Deely
Department of Mathematics
University of Canterbury
Christchurch, NEW ZEALAND
and Purdue University

Technical Report #92-01

Department of Statistics
Purdue University

June 1992

*

OPTIMAL ALLOCATION IN STRATIFIED SAMPLING WITH PARTIAL INFORMATION

by

D.V. Lindley[1] **and J.J. Deely**[2]

[1]2 Periton Lane, Minehead, Somerset TA24 8AQ, England. [2]Department of Mathematics, University of Canterbury, Christchurch, New Zealand.

Summary

The classical allocation formula of Neyman is replaced by a new formula which can accomodate prior information on the ratio of means in two strata. To accomplish this, a Bayesian approach is adopted and in this context partial prior information is quantified over the spectrum from absolute certainty to completely unknown.

Key words: Stratified sampling, optimal allocation, prior information, Bayesian methods.

1. Introduction

It is well known that in the presence of appropriate prior information stratified sampling can lead to increased efficiency over simple random sampling (SRS) for various types of estimation problems in survey sampling. In particular consider the problem of estimating the population total when there are two strata for which the total number of units as well as the variances of the measurement variable are known for each stratum. Furthermore suppose each of these variances is smaller than the total population variance and the allocation of a fixed total number of samples is done in the optimal way. Then the variance of the classical stratified estimate of the total is always smaller than the corresponding variance with SRS. The result is true under more general circumstances but the above is quite useful in the practical situation. For a full discussion of the many ramifications possible, see Cochran (1977, Chapter 5). The basic idea is that whatever basis for stratification is used, it should produce variances within the strata smaller than the variance over the total population. Thus the ability to stratify sensibly requires some prior knowledge of the relevant parameters.

Often in the practical situation when this kind of prior information is available, it is also the case that additional information in the form of proportions of the variable of interest within the strata is known approximately. Examples such as the number of stock owned by the large and small farmers, or the proportion of yearly traffic during the weekends and the weekdays, and many other apparent applications illustrate this point. When such information is available it is to be expected that further efficiencies are available. However it appears that such situations have not been treated in the literature. It is the purpose of this paper to deal with this kind of prior information. To do this, we adopt a Bayesian approach which allows the additional information about proportions to be quantified over the spectrum from absolute certainty to completely unknown. Appropriate formulas for computing both the optimal allocation of the sample to the strata, which allocation is required to minimize the posterior variance, as well as the estimates to be used for those allocations are provided as a function of this quantification. There are two extreme cases. If the proportion is unknown, the allocation reduces to that of Neyman (1934). If the proportion is known exactly, then only one stratum is sampled. In Section 2 the model used will be discussed and the new allocation formula given. The appropriate estimator for the total and a numerical example are given in Section 3. Derivations for the aforementioned formulas are provided in Section 4.

Previous work on a Bayesian approach to stratified sampling is contained in papers by Ericson (1965, 1988) and references therein. Whereas Ericson does use a multivariate normal model and obtain a formula for the posterior variance similar to our Equation (6), he does not deal with prior information as we have defined it and consequently does not obtain the explicit formula for the allocation (our Equation (1)). Other work by Soland (1967) deals with a general framework and discusses the resulting linear-programming problem without carrying through the special analysis required in our case where there is partial knowledge of the ratio of the means.

2. Mathematical Notation and Optimal Allocation

Consider two groups, or strata, of units. Associated with each unit is a real number. Within each group these numbers are independently and identically distributed conditional upon parameters. For group i there is a known number N_i of units and the distribution is normal with mean θ_i and variance $\sigma_i^2 (i = 1, 2)$. N_1, N_2, σ_1, σ_2 are known; θ_1, θ_2 are unknown. It is proposed to take random samples of sizes n_1, n_2 from the groups to determine the numbers associated with the sample units, and hence estimate the grand total $N_1\theta_1 + N_2\theta_2$. By making this association to the grand total, we will be effectively ignoring the finite correction factor, or assuming that the N's are large. The specific problem discussed in this paper is that of determining, for a fixed total sample size $n = n_1 + n_2$, the optimum allocation of the sample numbers between the groups; in other words, the choice of n_2, n_1 being determined as $n - n_2$. There are two known results in this field.

If nothing is known about the values of θ_1 and θ_2, then the optimum allocation is to choose the ratio n_1/n_2 to be equal to $N_1\sigma_1/N_2\sigma_2$ (Neyman(1934)). The latter value plays an important role in the calculations. It will be written γ_0 and referred to as the Neyman allocation. The second result applies when the ratio $N_1\theta_1/N_2\theta_2$ of the totals for the two groups is known. This ratio will be denoted by γ and its known value by γ_1. The result says that if γ_1 exceeds (is less than) the Neyman allocation γ_0, then all the observations should be made on the first (second) group so that no sampling of the other group is made. Alternatively, if $\gamma_1 > (<) \gamma_0$, then $n_2(n_1) = 0$. This result, though intuitively acceptable, does not seem to appear in the literature but will appear as a special case of our general result.

Notice that the two results cited represent extreme situations and provide contrasting results. In the first, nothing is known of the ratio $N_1\theta_1/N_2\theta_2$; in the second, it is known precisely. In the first, both groups are sampled with a familiar proportion in each. In the second, only one group is sampled and the known value of the ratio used to make inferences about the other. In the present paper we extend these results, offering intermediate cases where the ratio is only partly known.

The criterion for optimality in both cases is minimization of variance. From the frequentist sampling-theory viewpoint, the variance is that of the estimate of the grand total. In the Bayesian approach, the posterior variance of the grand total is the relevant quantity. Because of the assumed normality, the same result holds in the alternative view provided our interpretation of 'partly known' is suitably interpreted.

There are two reasons for using the Bayesian approach, apart from the general consideration that it alone provides a logically coherent attitude to problems of inference and decision. First, this is a situation in which there is an appreciable amount of prior information expressible in probability terms. Second, it is possible not merely to provide an estimate of the total and its standard error, but a complete probability distribution for it. With the distributions being normal, the distinction is relatively unimportant, but in general it is valuable to have the more complete specification.

Before giving proofs, we give a description and commentary on the results. The partial knowledge of $N_1\theta_1/N_2\theta_2$ is described through a bivariate, normal distribution for θ_1 and θ_2, the unit means of the two groups. Only two features of this distribution are relevant to the calculations. The first is the ratio of the means μ_1/μ_2. This is needed to evaluate the most likely value of $N_1\theta_1/N_2\theta_2$, namely $N_1\mu_1/N_2\mu_2 = \gamma_1$ in the notation used above. The second is the residual variance ω^2 of θ_1, were the value of θ_2 known. In regression language, ω^2 is the variance of θ_1 about the linear regression of θ_1 on θ_2. We imagine that the scientist conducting the investigation would think about the likely scatter of θ_1 for a known value of θ_2, giving limits on θ_1 that might be interpreted as about 4ω apart (using 95% normal limits).

Using ω and γ_1 to express the initial knowledge about $\gamma = N_1\theta_1/N_2\theta_2$, our result says that to achieve least posterior variance of $N_1\theta_1 + N_2\theta_2$, the value of n_2 should be

$$n_2 = \frac{n}{1+\gamma_0} + \frac{\sigma_1^2}{\omega^2}\frac{(1+\gamma_1)(\gamma_0-\gamma_1)}{\gamma_0(1+\gamma_0)} \tag{1}$$

if this lies in $(0, n)$. If the right-hand side exceeds n (is less than 0), then n_2 should be $n(0)$.

In commenting on (1), let us first notice how the two results cited above are special cases. If nothing is known about the ratio, the uncertainty of θ_1 given θ_2 is large and $\omega \to \infty$. The second term in (1) vanishes and $n_2 = n/(1+\gamma_0)$, or $n_1/n_2 = \gamma_0$, the Neyman allocation. If the ratio is known, the uncertainty vanishes and $\omega \to 0$. The second term in (1) therefore tends to $+\infty$ if $\gamma_0 > \gamma_1$, so that $n_2 = n$; and to $-\infty$ if $\gamma_0 < \gamma_1$, so that $n_2 = 0$, again in agreement with the result cited.

The term $n/(1+\gamma_0)$ in (1) provides, as we have just seen, agreement with the Neyman allocation. The last term provides for departure from the Neyman allocation; its sign depending on whether or not the most reasonable value, γ_1, of the ratio is greater or less than γ_0. (If $\gamma_1 = \gamma_0$ we are back to the Neyman allocation.) Note that it can always be supposed that $\gamma_0 \geq 1$ by renumbering the two groups if necessary. The magnitude of this second term depends primarily on σ_1^2/ω^2. This is a ratio of two variances. Since σ_1^2 is the variance of an individual measurement whereas ω^2 is the variance of their mean, (and then conditional upon θ_2), ω^2 will often be much smaller than σ_1^2 and hence the ratio appreciates. The effect of the second term is to increase (decrease) the amount of sampling in the second group if γ_1, the most reasonable value of γ_0, is below (above) the Neyman allocation. This term does not depend on n, the total size of the sample, so that its effect is greatest with small values of n. As $n \to \infty$, its effect is slight and we are effectively back to allocation in proportion to the Neyman allocation. For example, if $\gamma_0 = 2$, $\gamma_1 = 3$, the optimum allocation according to (1) has $n_2 = \frac{1}{3}n - \frac{2}{3}\frac{\sigma_1^2}{\omega_2}$. If $\sigma_1^2/\omega^2 = 30$, this gives $n_2 = \frac{1}{3}n - 20$. Here such a lot is known about the ratio that it is not until $n = 60$ that there is any allocation to the second group. If $\sigma_1^2/\omega^2 = 3$, the value $n_2 = \frac{1}{3}n - 2$ and there are samples taken from both groups as soon as n exceeds 6.

There are two further points before we leave consideration of the optimum allocation of samples between the two groups and pass to consideration of the best estimate of the total $N_1\theta_1 + N_2\theta_2$. The reader may be puzzled by the occurrence of σ_1^2 in the formula (1) for the optimum, but not σ_2. This is because we have chosen to compare σ_1^2 with ω^2, the residual variance of θ_1. An equivalent formula could be provided involving σ_2^2 and the residual variance of θ_2. A second consideration is that the calculations leading to the optimum are based on opinions about θ_1 and θ_2 being expressed through a bivariate, normal distribution. This can only be considered as an approximation when opinion is essentially about the ratio θ_1/θ_2 (or $N_1\theta_1/N_2\theta_2$). This is easily seen by noting that limits on a ratio of the form $\theta_1/\theta_2 \pm d$ imply greater variation in θ_1 for fixed θ_2 when the latter is large than when it is small, whereas our value, dependent on ω, is expressed for a single value of θ_2. We recommend that in thinking about ω, θ_2 be chosen around μ_2, the most reasonable value for θ_2. The normal distribution can also include nonsensical, negative values of θ_1 or θ_2. This can happen if γ_1 is very large or very small, especially when ω is large. Our result may not be appropriate in such circumstances. It should always be regarded as an approximation since the value of the variance near the optimum changes very little with n_2 so that the exact value is not critical.

3. Estimation of the Total and Numerical Example

We now pass from discussion of the optimum allocation of samples to the estimation of the total $N_1\theta_1 + N_2\theta_2$ for all the units in the two groups. The new quantities that enter here are x_1 and x_2, the sample means for the n_1 and n_2 units sampled in the two groups. The mean of the posterior distribution of the total is

$$\frac{\frac{n_1 n_2}{\sigma_1^2 \sigma_2^2}(N_1 x_1 + N_2 x_2) + \left(\frac{x_1 n_1}{\sigma_1^2}\frac{\gamma_1 N_2}{N_1} + \frac{x_2 n_2}{\sigma_2^2}\right)\frac{(\gamma_1+1)N_2}{\omega^2}}{\frac{n_1 n_2}{\sigma_1^2 \sigma_2^2} + \frac{1}{\omega^2}\left(\frac{n_1}{\sigma_1^2}\left(\frac{\gamma_1 N_2}{N_1}\right)^2 + \frac{n_2}{\sigma_2^2}\right)}. \tag{2}$$

This can serve as a point estimate of the total for those who prefer to think outside the Bayesian paradigm. As $\omega \to \infty$, so that there is effectively no knowledge of the ratio $N_1\theta_1/N_2\theta_2$, the mean (or estimate) tends to the obvious one in such circumstances. As $\omega \to 0$, so that the ratio is known precisely as γ_1, it tends to

$$\frac{\left(\frac{x_1 n_1}{\sigma_1^2}\frac{\gamma_1 N_2}{N_1} + \frac{x_2 n_2}{\sigma_2^2}\right)(\gamma_1 + 1)N_2}{\left(\frac{n_1}{\sigma_1^2}\left(\frac{\gamma_1 N_2}{N_1}\right)^2 + \frac{n_2}{\sigma_2^2}\right)} \tag{3}$$

which can be shown to be the appropriate estimate in these cases. In the general case when $0 < \omega < \infty$, (2) provides a weighted average of these two estimates. Notice that (2) holds for all values of n_1 and n_2 and not just for the optimal allocation given by (1). In particular, when γ is known to be γ_1, we saw that all the sampling is from one group and $n_2(n_1) = 0$ when $\gamma_1 > (<)\gamma_0$. Thus when $n_1 = 0$, (3) is simply $N_2 x_2(1 + \gamma_1)$, the estimate $N_2 x_2$ for the group sampled, inflated by the known factor $(1 + \gamma_1)$.

The distribution of the total, given the data, is normal, so the only remaining feature to be described is the variance. The expression for this is

$$\frac{N_1^2}{\gamma_0^2}\frac{\left\{\gamma_0^2 n_2 + n_1 + \frac{\sigma_1^2}{\omega^2}(1+\gamma_1)^2\right\}}{\left\{\frac{n_1 n_2}{\sigma_1^2} + \left(\frac{n_1 \gamma_1^2}{\gamma_0^2} + n_2\right)\frac{1}{\omega^2}\right\}}. \tag{4}$$

Alternatively this may be thought of as the sampling variance of the estimate given by (2). As ω tends to zero or infinity, the previously noted and familiar values of (4) can be obtained.

Numerical Example.

Traffic counting data for 341 days (243 weekdays and 98 weekend days) each with 24 hourly counts from a city street in Auckland, New Zealand, provide the basic data set to be used here to illustrate the proposed method for stratified sampling. One of the reasons these counts were recorded was to see how well a sample of hours could predict the total over the year. For this particular road it is known that the weekday traffic occupies nearly all of the total, since it is a main artery into a business and industrial area which is closed on the weekend and hence would receive very little traffic, relatively speaking, during that time. Also, for our purposes here, the main hours of concern are between 7 a.m. and 6 p.m.. Thus we stratify the total of 3751 such hours into Strata 1 consisting of 2673 weekday hours, and Strata 2 consisting of 1078 weekend hours, with standard deviations 100 and 45 respectively. This gives $\gamma_0 = 5.5$. It was known that the proportion p of weekday traffic accounted for about 95% of the total (within 1%). The most likely value of θ_2, the average hourly traffic for Strata 2, was thought to be about 40 vehicles per hour. Using this value and the range of values on p, upper and lower bounds on θ_1 can be obtained from the relationship

$$N_1\theta_1/p = N_2\theta_2/(1-p).$$

Hence

$$\omega = \left(\theta_1^u - \theta_1^l\right)/4 = \left[\left(\frac{p^u}{1-p^u} - \frac{p^l}{1-p^l}\right)\frac{N_2\theta_2}{N_1}\right] = (96/4 - 94/6)(N_2/4N_1)40 = 33.6$$

and $\sigma_1^2/\omega^2 = 8.86$. Using these values in equation (1) we obtain

$$n_2 = (n/6.5) - 66.$$

Suppose a sample of 500 hours is to be taken to estimate the total traffic between these hours. Then the optimal allocation is to take 490 (10) from Strata 1 (2). This should be compared to the Neyman allocation which is 423 (77). Note that if the size of the sample to be taken is smaller than 435 then all of such samples would be taken in Strata 1. Some measure of the efficiency of this allocation can be seen by comparing the variances for the appropriate estimates of the total traffic. The variance in (4) and the relevant variance for the estimate under Neyman allocation are 13,430 and 14,123 respectively, a reduction of approximately 5%, which is to be compared with a maximum reduction of 11% for perfect knowledge of γ (i.e. p).

4. Derivations

This section is concerned with stating our assumptions about the opinion of $\gamma = N_1\theta_1/N_2\theta_2$ prior to the data and proving the results given above. Under the assumption of normality for the measurements on the units and with the variances σ_1^2 and σ_2^2 known, the sample means, x_1 and x_2, together with n_1 and n_2, provide sufficient statistics. This joint distribution is normal with

$$E\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$$

and dispersion $\begin{pmatrix} \sigma_1^2/n_1 & 0 \\ 0 & \sigma_2^2/n_2 \end{pmatrix}$. These facts provide the likelihood function for θ_1 and θ_2, given the data. The knowledge of (θ_1, θ_2) prior to the data is assumed to be described by a bivariate, normal distribution. The notation used is

$$E\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \text{ with dispersion } \begin{pmatrix} \tau_1^2 & \rho\tau_1\tau_2 \\ \rho\tau_1\tau_2 & \tau_2^2 \end{pmatrix}.$$

The values of these five hyperparameters will later be specialized, but for the moment they are kept quite general.

Lemma 1. The posterior distribution of (θ_1, θ_2) given the data (x_1, x_2) is normal. The dispersion matrix is

$$B = \frac{\begin{pmatrix} \frac{n_2}{\sigma_2^2} + \frac{1}{\tau_2^2(1-\rho^2)} & \frac{\rho}{\tau_1\tau_2(1-\rho^2)} \\ \frac{\rho}{\tau_1\tau_2(1-\rho^2)} & \frac{n_1}{\sigma_1^2} + \frac{1}{\tau_1^2(1-\rho^2)} \end{pmatrix}}{\left[\left(\frac{n_1}{\sigma_1^2} + \frac{1}{\tau_1^2(1-\rho^2)}\right)\left(\frac{n_2}{\sigma_2^2} + \frac{1}{\tau_2^2(1-\rho^2)}\right) - \frac{\rho^2}{\tau_1^2\tau_2^2(1-\rho^2)^2}\right]}$$

and the mean is

$$B \times \begin{pmatrix} \frac{n_1 x_1}{\sigma_1^2} + \frac{\mu_1}{\tau_1^2(1-\rho^2)} - \frac{\mu_2\rho}{\tau_1\tau_2(1-\rho^2)} \\ \frac{n_2 x_2}{\sigma_2^2} - \frac{\mu_1\rho}{\tau_1\tau_2(1-\rho^2)} + \frac{\mu_2}{\tau_2^2(1-\rho^2)} \end{pmatrix}.$$

The proof is immediate from standard, normal theory. The posterior precision (the inverse of the dispersion) is the sum of the prior and likelihood precisions. Easy matrix inversions provide the result. The posterior mean is a weighted average of the prior and data means with weights proportional to their respective precisions.

For simplicity of notation, write

$$a_{ii} = \frac{\sigma_i^2}{\tau_i^2(1-\rho^2)}, \quad a_{12} = \frac{\rho\sigma_1\sigma_2}{\tau_1\tau_2(1-\rho^2)}, \quad M_i^2 = \sigma_i^2 N_i^2. \tag{5}$$

Using the dispersion matrix B in Lemma 1, it easily follows that the posterior variance of the total $N_1\theta_1 + N_2\theta_2$ is

$$\frac{\{M_1^2(n_2+a_{22}) + 2M_1M_2a_{12} + M_2^2(n_1+a_{11})\}}{[(n_1+a_{11})(n_2+a_{22}) - a_{12}^2]}. \tag{6}$$

That allocation will be regarded as optimal which minimizes the variance (omitting the adjective 'posterior') function in (6) with respect to the variable n_2 subject to the total sample size $n = n_1 + n_2$ being fixed. With $n_1 = n - n_2$ and n_2 written temporarily as x, (6) is of the form $(M_2^2 - M_1^2) \times h(x) = (x-a)/(x-r_1)(x-r_2)$ for suitable r_1, r_2 and a. Suppose, without loss of generality, $M_2 > M_1$. The case $M_2 = M_1$ is straightforward. We will require two Lemmas to obtain the main result.

Lemma 2: If $r_2 < 0 < n < r_1 < a$, the least value of $h(x)$ in $[0, n]$ occurs at $x_1 = (r_1 + Sr_2)/(1+S)$ if this lies in $[0, n]$. If $x_1 < 0 (> n)$, the least value is at $0(n)$. Here S is $+[(a-r_1)/(a-r_2)]^{1/2}$.

Proof: Write

$$h(x) = \frac{1}{r_1 - r_2}\left\{\frac{a-r_2}{x-r_2} - \frac{a-r_1}{x-r_1}\right\}.$$

Then $h'(x)$ vanishes where

$$\frac{a-r_1}{(x-r_1)^2} = \frac{a-r_2}{(x-r_2)^2},$$

that is, at $x_1 = \frac{r_1+Sr_2}{1+S}$ and $x_2 = \frac{r_1-Sr_2}{1-S}$. For x in (r_2, r_1), $h(x) > 0$, tending to $+\infty$ at r_1 and r_2. Hence x_1 provides a local minimum. (It is easily shown, though irrelevant to our calculations, that $a < x_2$.) $\square$

It is now shown that the inequalities assumed in Lemma 2 do hold for our $h(x)$. The denominator of (6) is, with $n_1 = n - n_2$, a quadratic in n_2 with its arms down, yet it is positive both when $n_2 = 0$ and $n_2 = n (n_1 = 0)$. Hence $r_2 < 0 < n < r_1$.

The numerator of (6), when multiplied by $(n_1 + a_{11})$, is, on rearrangement,

$$\{M_1^2(n_2+a_{22})(n_1+a_{11}) - M_1^2a_{12}^2\} + \{M_1a_{12} + M_2(n+a_{11})\}^2.$$

If this vanishes (at $x = a$), since the second term is positive, the first, which is M_1^2 times the denominator, must be negative. Hence a lies outside the interval (r_2, r_1) in which the denominator is positive. Since a is positive, it exceeds r_1. $\square$

It remains only to calculate x_1 explicitly. Rather than use the expression obtained in the proof of Lemma 2, it is easier to proceed directly.

Lemma 3: The values of n_2 at the turning points of (6) when subjected to the constraint $n_1 + n_2 = n$ are

$$\frac{a_{12}(\pm M_1 - M_2) \pm nM_2 \pm a_{11}M_2 - a_{22}M_1}{M_1 \pm M_2}.$$

(Where there is a choice of signs the upper (lower) ones are to be taken together.)

Proof: Write

$$z_i = n_i + a_{ii}$$

so that (6) is more simply

$$\frac{M_1^2 z_2 + 2M_1M_2a_{12} + M_2^2 z_1}{z_1 z_2 - a_{12}^2}$$

and the constraint $n_1 + n_2 = n$ is

$$(z_1 - a_{11}) + (z_2 - a_{22}) = n. \tag{7}$$

Use a Lagrange multiplier μ. Differentiate with respect to z_1 and z_2 and equate to zero. The results are

$$\left(z_1z_2 - a_{12}^2\right) M_2^2 - \left(M_1^2 z_2 + 2M_1M_2a_{12} + M_2^2 z_1\right) z_2 + \mu \left(z_1z_2 - a_{12}^2\right)^2 = 0$$

and

$$\left(z_1z_2 - a_{12}^2\right) M_1^2 - \left(M_1^2 z_2 - 2M_1M_2a_{12} - M_2^2 z_1\right) z_1 + \mu \left(z_1z_2 - a_{12}^2\right)^2 = 0.$$

These simplify to yield equivalently

$$(M_1z_2 + M_2a_{12})^2 = \mu(z_1z_2 - a_{12}^2)^2$$

and

$$(M_2z_1 + M_1a_{12})^2 = \mu(z_1z_2 - a_{12}^2)^2.$$

Elimination of μ gives

$$(M_1 z_2 + M_2 a_{12}) = \pm(M_2 z_1 + M_1 a_{12})$$

which, together with the constraint (7) provides two sets (according to whether the plus or minus sign is selected) of linear equations. It is easily verified that the elimination of z_1 yields

$$z_2 = [a_{12}(\pm M_1 - M_2) \pm (n + a_{11} + a_{22})M_2]/(M_1 \pm M_2)$$

from which the stated value of $n_2 = z_2 - a_{22}$ immediately follows.

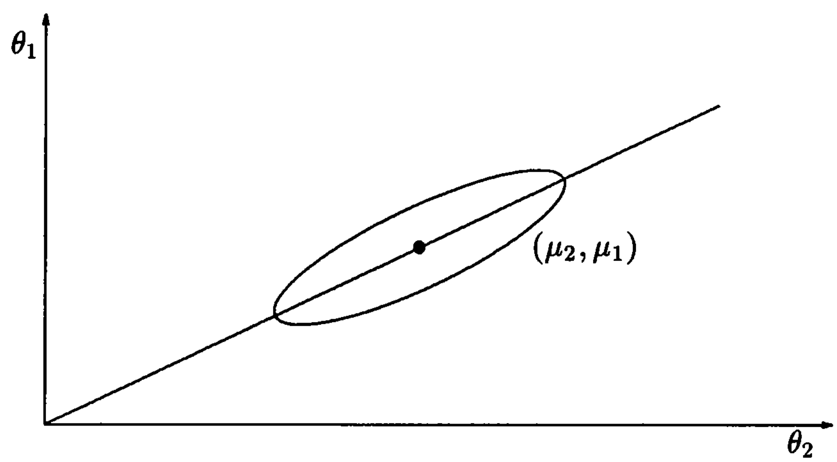
Combining now Lemmas 2 and 3 gives the following main result.

Theorem: The optimum allocation of n samples between the two groups has

$$n_2 = \frac{a_{12}(M_1 - M_2) + nM_2 + a_{11}M_2 - a_{22}M_1}{M_1 + M_2} \tag{8}$$

provided this value lies in the interval $[0, n]$. If it is less than zero (greater than n), the optimum allocation is $n_2 = 0$ ($n_2 = n$), i.e. sample from only one of the groups.

This result holds for any values of the five hyperparameters in the bivariate normal prior. These values are now specialized to reflect knowledge of the ratio $N_1\theta_1/N_2\theta_2 = \gamma$. The next figure



shows the sort of normal ellipse required. Clearly the mean (μ_1, μ_2) must lie on the line $N_1\theta_1/N_2\theta_2 = \gamma_1$, where γ_1 is the most reasonable value of γ. Hence

$$\mu_1/\mu_2 = \gamma_1 N_2/N_1.$$

If the major axis of the ellipse is to lie along the same line, the variances must satisfy

$$\tau_1^2/\tau_2^2 = \gamma_1^2 N_2^2/N_1^2. \tag{9}$$

The residual variance of θ_1, given θ_2, is $\tau_1^2(1-\rho^2)$, denoted by ω^2, and expresses the uncertainty about the exact value of the ratio γ. Next we suppose the uncertainty of θ_2 is large, that is τ_2^2, and therefore by (9), τ_1^2 tend to infinity. To keep ω^2 finite it is necessary to simultaneously let ρ tend to $+1$. Under these conditions, in the limit

$$a_{11} = \frac{\sigma_1^2}{\omega^2},\ a_{22} = \frac{\sigma_1^2}{\omega^2}\frac{M_2^2}{M_1^2}\gamma_1,\ a_{12} = \frac{\sigma_1^2}{\omega^2}\frac{M_2}{M_1}\gamma_1,\ a_{11}a_{22} - a_{12}^2 = 0. \tag{10}$$

Substitution of these results into the expression (8) for the optimum value of n_2 easily gives the form stated earlier in (1), remembering that $M_1/M_2 = \gamma_0$.

To obtain the expressions for the posterior mean (2) and variance (4), it is again only necessary to insert the special values in (10) into the expressions for the mean and variance given in Lemma 1. Notice that neither involve the means μ_1, μ_2. This is expected for the variance but the terms in μ_1 and μ_2 vanish when the algebra for the mean is performed.

5. Possible extensions.

There are several possible extensions of the results presented here. A simple one is to consider a general, linear function $c_1\theta_1 + c_2\theta_2$ instead of the total with $c_i = N_i$. The expression for the posterior variance persists except that M_i is now $\sigma_i c_i$ instead of $\sigma_i N_i$. With this notational change, the optimum allocation (8) is unaltered. The special prior knowledge leading to the results around (9) persists but their substitutions into the expression for a_{ij} equation (10), do not lead to the forms given there since M_i has a different interpretation. Instead we have $a_{12} = \sigma_1^2\gamma_1 N_2\sigma_2/\omega^2 N_1\sigma_1$ etc. Insertion of these values into (8) with $M_i = \sigma_i c_i$ leads to the required result.

A much more difficult extension is from 2 to k strata. Ericson (1965) highlights the unexpected behaviour that can arise, though, unlike us, he does not give general results. To appreciate the reason for the difficulty, take the general expression for the posterior dispersion matrix of the θ's that generalizes the case $k = 2$ (see B in Lemma 1). This matrix is proportional to one in which the i^{th} diagonal entry is linear in n_i, but independent of the other sample sizes, and the off-diagonal elements do not involve the sample sizes. As a result, the determinant of this matrix is of degree k in the n's. Consequently the posterior variance of the total is the ratio of a linear function of the n's to a polynomial in them of degree k. The analytical minimization is therefore appreciably harder and is made even more so by the boundary conditions $n_i \geq 0$.

Another way to proceed with more than 2 strata is iteratively. We illustrate with $k = 3$. Having found, with $n_1 + n_2 = n_{12}$ fixed, the optimum allocation between the first two strata for general n_{12}; a possibility is to find the optimum allocation between n_{12}, n_3, with $n_{12} + n_3$ fixed, using the method for two strata already developed. However, this fails for the following reason. Our task is to minimize the variance of $N_1\theta_1 + N_2\theta_2 + N_3\theta_3$. This is

$$\text{var}\ (N_1\theta_1 + N_2\theta_2) + 2\ \text{cov}\ (N_1\theta_1 + N_2\theta_2, N_3\theta_3) + \ \text{var}\ (N_3\theta_3).$$

The first stage just mentioned attends only to the first term and omits consideration of the second, covariance term. Hence the optimum allocation between the first two strata when considered alone, may not be optimum when the third is introduced. The iteration therefore fails.

A third, possible extension is to multivariate data. This causes no problems provided that (i) when a sample is taken, all the variables are observed, and (ii) the quantity of interest is a linear function. Then, with two strata, there are only two variables, n_1, n_2 as before, and the posterior variance of the required quantity is of a similar nature to that in the present note.

A fourth, and formidable, extension is to unknown variances. This is hard for two reasons. First, because it now makes sense to sample a stratum in order to learn about the variance within it, rather than just for the total. Consequently, when our procedure leads, as we have seen it can, to no observations being taken in a stratum, there will now be a possible reason to take some because of inadequate knowledge of the variance. The second reason for unknown variances presenting additional problems can be seen by looking at the expression for B, the posterior dispersion matrix, in Lemma 1. It has the amazing property that it does not involve the data. Once the strata sizes have been selected, but before the observations taken, it is known for sure what precision will be obtained. With unknown variances, this no longer holds. One can only say what the expected precision is. In other words, an additional operation, of expectation, has to be performed.

The whole field of optimum allocation of resources is very difficult. Sometimes numerical procedures can be developed that will deal with individual cases. The present paper develops an analytic solution which, despite its serious limitations, does provide insight in a way that computational solutions find difficult without many replications.

References

Cochran, W.G. (1977), *Sampling Techniques*, Third Edition, Wiley.

Ericson, W.A. (1965), "Optimum stratified sampling using prior information," *JASA* 60, 750-771.

Ericson, W.A. (1988), "Bayesian inference in finite populations," *Handbook of Statistics* 6, 213-246, ed. P.R. Krishnaiah and C.R. Rao. Elsevier.

Neyman, J. (1934), "On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection," *Journal of the Royal Statistical Society* 97, 558-606.

Soland, R.M. (1967), "Optimal multivariate stratified sampling with prior information," *Skand. Aktuar.* 50, 31-39.

Résumé

On remplace la formule classique de l'allocation de Neyman d'une formule nouvelle qui peut sáccomoder a l'information prior sur le rapport des mogennes dans deux strata. Pour cóchever on choix le method Bayesien et en ce coutexte, l'information prior partielle se quantifie sur l'étendue de certitude absolue à l'inconnue complete.