Werk

Titel: Optimum Allocation in Bayesian Stratified Two Phase Sampling when there are m-Att...

Autor: Khan, M.Z.

Jahr: 1976

PURL: https://resolver.sub.uni-goettingen.de/purl?358794056_0023|log41

Kontakt/Contact

Digizeitschriften e.V.
SUB Göttingen
Platz der Göttinger Sieben 1
37073 Göttingen

info@digizeitschriften.de

Metrika, Band 23, 1976, Seite 211–219. Physica-Verlag, Wien.

Optimum Allocation in Bayesian Stratified Two-Phase Sampling when there are m-Attributes

By *M.Z. Khan,* Aligarh[1])

Abstract: The problem of optimally allocating a sample among k-strata at the second phase of a two-phase sampling procedure when the sampling is for proportion has been discussed by *Newbold* [1971] and then by the author [*Khan*, 1972].

This paper deals with optimum allocation of the sample among k-strata at the second phase of a two-phase sampling procedure, when the sampling is for m-attributes. The problem is to estimate the proportion of each attribute in the population. Here $(m-1)$ dimensional Dirichlet distribution is taken as the prior distribution.

1. Preliminaries

Suppose that there are k-strata. Let the proportion of individuals possessing the j-th attribute in the i-th stratum be P_{ij}, $j = 1, 2, \ldots, m$, $i = 1, 2, \ldots, k$. Suppose two phase sampling procedure is adopted. At the first phase a predetermined number of observations n_i are taken from the i-th stratum. Let the number of individuals in the sample from the i-th stratum possessing the j-th attributes be a_{ij}, $j = 1, 2, \ldots, m$, $i = 1, 2, \ldots, k$. Let N_i be the number of observations to be taken in the i-th stratum in the second phase of sampling and suppose a number b_{ij}, $j = 1, 2, \ldots, m$ of individuals possess the j-th attribute out of the number N_i, $i = 1, 2, \ldots, k$.

Let π_i be the proportions of individuals falling in the i-th stratum, so that $\sum_{i=1}^{k} \pi_i = 1$.

Let C be the total cost to be allowed in the survey. Suppose a proportion α of the total cost is spent in the first phase of sampling. So a portion $(1-\alpha)\,C$ is left for taking observations in the second phase. If C_1 is the cost of observing one individual from the i-the stratum, then

$$\Sigma\, C_i n_i = \alpha C,\ \Sigma\, C_i N_i = (1-\alpha)\, C \tag{1.1}$$

[1]) *Mohd. Zubair Khan*, Dept. of Statistics, Aligarh Muslim University, Aligarh-202001, India.

We have to estimate the overall population proportions (of each attribute)

$$P_{.j} = \pi_1 P_{1j} + \pi_2 P_{2j} + \ldots + \pi_k P_{kj} \tag{1.2}$$

$$\text{for } j = 1, 2, \ldots, m.$$

And the problem is to allocate N_i's optimally to the different strata, subject to the budget restriction (1.1).

Two different approaches are employed, namely preposterior analysis and the posterior or Bayesian analysis.

2. Recall from the Previous Paper

In our earlier paper we derived the allocation formulae for different strata in the second phase of a two phase sampling procedure when the sampling was for proportion. There we had to estimate the overall population proportion $P = \pi_1 P_1 + \pi_2 P_2 + \ldots + \pi_k P_k$, where P_i, $i = 1, 2, \ldots, k$ was the proportion of individuals, possessing the attribute, in the i-th stratum. Assuming the likelihood of the first phase observations as binomial and the prior distribution of $P_1, P_2, \ldots, P_k$ as beta, namely

$$p(P_i) \propto P_i^{r_i'-1} (1-P_i)^{n_i'-r_i'-1}, \; n_i' \geqslant 0, r_i' = 0, 1, \ldots, n_i'. \tag{2.1}$$

We obtained (running closely on the lines of *Newbold* [1971] the various allocation formulae as follows, (The notations being in conformity with those given in the present paper except for the second subscripts which which not called for).

In the case of preposterior analysis the allocation formula was

$$N_i' = \frac{(1-\alpha)\, C}{C_i^{1/2}} \frac{A_i^{1/2} (n_i' + n_i)^{1/2}}{\sum_{i=1}^{k} A_i^{1/2} (n_i' + n_i)^{1/2} C_i^{1/2}} \tag{2.2}$$

$$\text{for } i = 1, 2, \ldots, k$$

where

$$A_i = \frac{\pi_i^2 \hat{P}_i (1-\hat{P}_i)}{(n_i' + n_i + 1)} \tag{2.3}$$

and

$$\hat{P}_i = \frac{a_i + r_i'}{n_i' + n_i} \quad \text{for } i = 1, 2, \ldots, k \tag{2.4}$$

where a_i was the number of individuals in the sample from the i-th stratum possessing the attribute.

In the case of posterior analysis, the allocation formula was

$$N_i' = \frac{(C + \Sigma C_i n_i')}{C_i^{1/2}} \frac{A_i^{1/2} (n_i' + n_i)^{1/2}}{\sum_{i=1}^{k} A_i^{1/2} (n_i' + n_i)^{1/2} C_i^{1/2}} - (n_i' + n_i) \qquad (2.5)$$

$$\text{for } i = 1, 2, \ldots, k.$$

where A_i was as in (2.3).

If the above formula produced negative values, we reallocated the second phase sample according to the following reallocation formula

$$N_i' = \frac{(1-\alpha) C + \Sigma_+ (n_i' + n_i) C_i}{C_i^{1/2}} \frac{A_i^{1/2} (n_i' + n_i)^{1/2}}{\Sigma_+ A_i^{1/2} (n_i' + n_i)^{1/2} C_i^{1/2}} - (n_i' + n_i)$$

$$\text{for } i = 1, 2, \ldots, k \qquad (2.6)$$

where Σ_+ indicated that the summation was only over those strata for which N_i' had been positive. In the case of above again producing negative values, the repetition of the reallocation process was suggested.

The π_i unknown case was also discussed, treatment being the same as in the present paper.

It was shown that the above formulae included as special cases all the results given by *Newbold* [1971] (all $n_i' = 1, r_i' = 1/2$). They also included as special case, the allocation resulting from the uniform prior distribution (all $n_i' = 2, r_i' = 1$) and that arising when we had no prior information, in the sense that we depended solely on the sample outcome of the first phase observations (all $n_i' = 0, r_i' = 0$).

3. Preposterior Analysis

We take the assumption that the sample size in each stratum is small in relation to all the classes, so that distribution of a_{ij} and b_{ij} are approximately multinomial [*Cochran*, 1963, p. 59]. The likelihood of the first phase observations is then

$$p(a_{ij}|P_{ij}) = \prod_{i=1}^{k} \frac{n_i!}{a_{i1}!\, a_{i2}! \ldots a_{im}!} P_{i1}^{a_{i1}} P_{i2}^{a_{i2}} \ldots P_{im}^{a_{im}} \qquad (3.1)$$

We assume that $P_{1j}, P_{2j}, \ldots, P_{kj}$ $(j = 1, 2, \ldots, m)$ are a priori independent and follow the $(m-1)$ dimensional Dirichlet distribution with parameters $\nu_1, \nu_2, \ldots, \nu_{m-1}$ defined by

$$\frac{\Gamma(\sum_{j=1}^{m} \nu_{ij})}{\prod_{j=1}^{m} \Gamma(\nu_{ij})} P_{i1}^{\nu_{i1}-1} P_{i2}^{\nu_{i2}-1} \ldots P_{im}^{\nu_{im}-1} \tag{3.2}$$

where

$$P_{im} = 1 - \sum_{j=1}^{m-1} P_{ij} \text{ and } 0 \leqslant P_{ij} \leqslant 1 \tag{3.3}$$

Here it should be noted that the Dirichlet includes as special cases the locally uniform prior (all $\nu_i = 1$), *Jeffrey*'s [1961, p. 180] invariant prior (all $\nu_i = 1/2$) and the diffused prior (all $\nu_i = 0$). Thus from the results derived below, we can get allocation formulae arising from locally uniform prior, *Jeffrey*'s invariant prior and with the diffused prior merely by taking all $\nu_i = 1$, all $\nu_i = 1/2$ and all $\nu_i = 0$ respectively. It should also be noted that the results hold good even if we have different degree of prior information in different strata.

Combining this prior distribution (3.2) with the likelihood (3.1) we find that the posterior density of the P_{ij} after the first phase is proportional to

$$\prod_{i=1}^{k} \frac{n_i!\, \Gamma(\nu_{i.})}{\prod_{j=1}^{m} a_{ij}!\, \Gamma(\nu_{ij})} P_{i1}^{a_{i1}+\nu_{i1}-1} P_{i2}^{a_{i2}+\nu_{i2}-1} \ldots P_{im}^{a_{im}+\nu_{im}-1}$$

$$\text{for } 0 \leqslant P_{ij} \leqslant 1, j = 1, \ldots, m \tag{3.4}$$

where

$$\nu_{i.} = \sum_{j=1}^{m} \nu_{ij} \text{ and } P_{im} = 1 - \sum_{j=1}^{m-1} P_{ij} \tag{3.5}$$

which is again a Dirichlet distribution with parameters

$$a_{i1} + \nu_{i1}, a_{i2} + \nu_{i2}, \ldots, a_{im-1} + \nu_{im-1}.$$

The likelihood of the second phase observations is given by

$$p(b_{ij} \mid P_{ij}) = \prod_{i=1}^{k} \frac{N_i!}{b_{i1}!\, b_{i2}! \ldots b_{im}!} P_{i1}^{b_{i1}} P_{i2}^{b_{i2}} \ldots P_{im}^{b_{im}} \tag{3.6}$$

Product of (3.4) and (3.6) gives

$$\prod_{i=1}^{k} \frac{n_i!\, N_i!\, \Gamma(\nu_{i.})}{\prod_{j=1}^{m} a_{ij}!\, b_{ij}!\, \Gamma(\nu_{ij})} P_{i1}^{a_{i1}+\nu_{i1}-1} P_{i2}^{a_{i2}+\nu_{i2}-1} \ldots P_{im}^{a_{im}+\nu_{im}-1} P_{i1}^{b_{i1}} P_{i2}^{b_{i2}} \ldots P_{im}^{b_{im}} \tag{3.7}$$

which is the Dirichlet multinomial distribution (*Mosimann* [1972] calls it compound multinomial distribution). The first two moments of (3.7) are given by: [*Mosimann*, 1972].

$$E(b_{ij}) = N_i \hat{P}_{ij} \tag{3.8}$$

$$V(b_{ij}) = N_i (n_i + N_i + \nu_{i.}) \frac{(a_{ij} + \nu_{ij})(n_i + \nu_{i.} - a_{ij} - \nu_{ij})}{(n_i + \nu_{i.})^2 (n_i + \nu_{i.} + 1)} \tag{3.9}$$

where

$$\hat{P}_{ij} = \frac{a_{ij} + \nu_{ij}}{n_i + \nu_{i.}} \tag{3.10}$$

Note that both $\Sigma\, \pi_i P_{ij}$ $(j = 1, 2, \ldots, m)$ a posteriori after the first phase and $\Sigma\, \pi_i b_{ij}/N_i$ $(j = 1, 2, \ldots, m)$ preposteriori before the second phase have the same expectation namely $\Sigma\, \pi_i \hat{P}_{ij}$ $(j = 1, 2, \ldots, m)$, where $\hat{P}_{ij}$ is given by (3.10). Hence $\Sigma\, \pi_i b_{ij}/N_i$ $(j = 1, 2, \ldots, m)$ is a natural preposterior estimator of $P_{.j} = \Sigma\, \pi_i P_{ij}$ for $j = 1, 2, \ldots, m$.

Its variance is given by

$$\begin{aligned} V(\sum_i \pi_i b_{ij}/N_i) &= \sum_i \frac{\pi_i^2}{N_i^2} V(b_{ij}) \\ &= \sum_i \frac{\pi_i^2}{N_i} \frac{(n_i + N_i + \nu_{i.})(a_{ij} + \nu_{ij})(n_i + \nu_{i.} - a_{ij} - \nu_{ij})}{(n_i + \nu_{i.})^2 (n_i + \nu_{i.} + 1)} \\ &= \sum_i \frac{\pi_i^2}{N_i} \frac{(n_i + N_i + \nu_{i.})}{(n_i + \nu_{i.} + 1)} \hat{P}_{ij}(1 - \hat{P}_{ij}) \\ &= \sum_i \left\{ 1 + \frac{(n_i + \nu_{i.})}{N_i} \right\} A_{ij} \text{ for } j = 1, 2, \ldots, m \end{aligned} \tag{3.11}$$

where

$$A_{ij} = \frac{\pi_i^2 \hat{P}_{ij}(1 - \hat{P}_{ij})}{(n_i + \nu_{i.} + 1)} \qquad \text{for } j = 1, 2, \ldots, m \tag{3.12}$$

and $\hat{P}_{ij}$ is same as in (3.10).

Now the weighted sum of the variance of each proportion is given by

$$\sum_{j=1}^{m} l_j V(P_{.j}) = \sum_{j=1}^{m} l_j \sum_{i=1}^{k} \left\{1 + \frac{n_i + \nu_{i.}}{N_i}\right\} A_{ij}$$

$$= \sum_{i=1}^{k} \left\{1 + \frac{n_i + \nu_{i.}}{N_i}\right\} \sum_{j=1}^{m} l_j A_{ij} \tag{3.13}$$

To get the optimum allocation of N_i, we minimize the weighted sum of the variance of each proportion (3.13) subject to the budget restriction (1.1). (Considering each proportion an independent variate. The method has been suggested by *Yates* [1960], reproduced by *Cochran* [1963, p. 120]). Thus the optimum allocation is given by

$$N_i' = \frac{(1-\alpha)\, C}{C_i} \frac{(\sum_{j=1}^{m} l_j A_{ij})^{1/2} (n_i + \nu_{i.})^{1/2} C_i^{1/2}}{\sum_{i=1}^{k} (\sum_{j=1}^{m} l_j A_{ij})^{1/2} (n_i + \nu_{i.})^{1/2} C_i^{1/2}} \tag{3.14}$$

for $i = 1, 2, \ldots, k$

where A_{ij} is given by (3.12).

If we consider unweighted sum (all $l_j = 1$) instead of weighted sum of the variance the formula (3.14) becomes a straight forward generalization of the allocation formula (2.2), as we should have reasonably expected.

4. Bayesian Analysis

In this procedure the posterior density of the P_i after the first phase is used as the prior distribution for the second phase. The likelihood of the second phase observations is given by (3.6). Therefore combining (3.4) and (3.6) we get the posterior density after the second phase as

$$p(P_{ij} \mid a_{ij}, b_{ij}) = \frac{n_i!\, N_i!\, \Gamma(\nu_{i.})}{\prod_{j=1}^{m} a_{ij}!\, b_{ij}!\, \Gamma(\nu_{ij})} P_{i1}^{a_{i1}+b_{i1}+\nu_{i1}-1} P_{i2}^{a_{i2}+b_{i2}+\nu_{i2}-1} \ldots P_{im}^{a_{im}+b_{im}+\nu_{im}-1} \tag{4.1}$$

$$(0 \leqslant P_{ij} \leqslant 1,\ i = 1, 2, \ldots, k)$$

which is again a Dirichlet distribution. Hence we find that

$$E''(P_{ij}) = P_{ij}'' \tag{4.2}$$

$$V''(P_{ij}) = \frac{P''_{ij}(1-P''_{ij})}{(n_i + N_i + \nu_{i.} + 1)} \tag{4.3}$$

where

$$P''_{ij} = \frac{a_{ij} + b_{ij} + \nu_{ij}}{n_i + N_i + \nu_{i.}} \tag{4.4}$$

where E'' and V'' denote the posterior expectation and posterior variance after the second phase respectively.

Note that the variance of $P_{.j} = \Sigma\, \pi_i P_{ij}$ for $j = 1, 2, \ldots, m$ involves the b_{ij}, which will not be known at the end of the first phase. So, we take expectation of $V''(P_{.j})$ over the future distribution of b_{ij} to get the expected variance of $P_{.j}$ for $j = 1, 2, \ldots, m$, which is free from b_{ij}. After some algebraic manipulation we get

$$E\left\{V''(P_{.j})\right\} = E\left\{\sum_{i=1}^{k} \frac{\pi_i^2 P''_{ij}(1-P''_{ij})}{(n_i + N_i + \nu_{i.} + 1)}\right\}$$

$$= \sum_{i=1}^{k} \frac{(n_i + \nu_{i.}) A_{ij}}{(n_i + N_i + \nu_{i.})} \qquad \text{for } j = 1, 2, \ldots, m. \tag{4.5}$$

where A_{ij} is the same as defined in (3.12).

Now the weighted sum of the variances of each proportion (4.5) is given by

$$\sum_{j=1}^{m} l_j \left(E\left\{V''(P_{.j})\right\}\right) = \sum_{j=1}^{m} l_j \left(\sum_{i=1}^{k} \frac{n_i + \nu_{i.}}{(n_i + N_i + \nu_i)} A_{ij}\right)$$

$$= \sum_{i=1}^{k} \frac{(n_i + \nu_{i.})}{(n_i + N_i + \nu_{i.})} \left(\sum_{j=1}^{m} l_j A_{ij}\right) \tag{4.6}$$

As in the preposterior analysis, to get the optimum allocation of N_i we minimize the weighted sum of the variances of each proportion, that is we minimize (4.6) subject to the budget restriction (1.1). Thus the optimum allocation is given by

$$N'_i = \frac{(C + \Sigma\, C_i \nu_{i.})}{C_i} \frac{\left(\sum_{j=1}^{m} l_i A_{ij}\right)^{1/2} (n_i + \nu_{i.})^{1/2} (C_i)^{1/2}}{\sum_{i=1}^{k} \left(\sum_{j=1}^{m} l_j A_{ij}\right)^{1/2} (n_i + \nu_{i.})^{1/2} (C_i)^{1/2}} - (n_i + \nu_{i.})$$

$$\text{for } i = 1, 2, \ldots, k \tag{4.7}$$

where A_{ij} is the same as defined in (3.12).

Again if we consider unweighted sum of the variances instead of weighted one, the allocation formula (4.7) becomes a straight forward generalization of the allocation formula (2.5).

Now it may happen that this procedure produces some N_i' which are negative. In this case it is necessary to reallocate the second phase observations. The procedure is as follows: Let the reallocated values be $\hat{N}_i'$. If N_i' is zero or negative, set $\hat{N}_i'$ equal to zero. The remaining $\hat{N}_i'$ are found by minimizing the variance only over those strata for which N_i' was positive. The remaining $\hat{N}_i'$ are then given by

$$\hat{N}_i' = \frac{(1-\alpha)\,C + \Sigma_+ (n_i + \nu_{i.})\,C_i}{C_i} \; \frac{(\sum_{j=1}^{m} l_j A_{ij})^{1/2} (n_i + \nu_{i.})^{1/2} C_i^{1/2}}{\Sigma_+ (\sum_{j=1}^{m} l_j A_{ij})^{1/2} (n_i + \nu_{i.})^{1/2} C_i^{1/2}} (n_i + \nu_{i.})$$

$$\text{for } i = 1, 2, \ldots, k \tag{4.8}$$

where Σ_+ indicates that the summation is only over those strata for which N_i' was positive.

Again we notice that if the sampling is for proportion and we take all $l_j = 1$ the formula (4.8) reduces to the reallocation formula (2.6).

The formula (4.8) may still produce some N_i' which are negative. If this so happens we repeat the process of reallocation. Eventually the process terminates with some zeros and some positive values of N_i. [*Draper* and *Guttman*, 1968].

5. π_i Unknown

In this case we assume that 'a priori' the distribution of π_i is the Dirichlet distribution defined by

$$p(\pi_1, \pi_2, \ldots, \pi_k) = \frac{\Gamma(\nu_i' + \nu_2' + \ldots + \nu_k')}{\prod_{i=1}^{k} \Gamma(\nu_i')} \; \pi_1^{\nu_1'-1} \pi_2^{\nu_2'-1} \ldots \pi_k^{\nu_k'-1}$$

And thus parallel results are obtained for preposterior and Bayesian allocations except that π_i are now replaced by

$$\hat{\pi}_i = E(\pi_i) = \frac{n_i + \nu_i'}{\sum_i (n_i + \nu_i')} \qquad \text{for } i = 1, 2, \ldots, k$$

where ν_i''s are the parameters of the Dirichlet distribution.

References

Cochran, W.G.: Sampling Technique, 1963.

Draper, N.R., and *I. Guttman*: Some Bayesian stratified two-phase sampling results, Biometrika **55**, 131–139, 1968.

Jeffreys, H.: Theory of Probability, Oxford 1961.

Khan, M.Z.: Bayesian stratisfied two-phase sampling for proportions: A general case. The Aligarh Bulletin of Mathematics **2**, 69–76, 1972.

Mosiman, J.E.: On the compound multinomial distribution the multivariate β distribution and correlation among proportions, Biometrika **49**, 65–82, 1962.

Newbold, P.: Optimum allocation in stratified two-phase sampling for proportions, Biometrika, **58**, 681–683, 1971.

Eingegangen am 22. Juli 1975