

A Compromise Criterion for Sampling Allocation in Multiobjective Surveys

*Un Criterio di Compromesso per la Definizione dell'Allocazione
del Campione in Indagini Multiobiettivo*

Marco Centra

ISFOL - Via Lancisi, 19 - Roma, e-mail: m.centra@isfol.it

Piero Demetrio Falorsi

ISTAT- Via Magenta 4 - Roma, e-mail: falorsi@istat.it

Riassunto: In questo lavoro si affronta il problema in cui l'indagine debba fornire stime per l'intera popolazione e separatamente per ciascuno degli strati in cui la popolazione stessa è suddivisa. In tale ambito, viene illustrato un criterio metodologico che consente di affrontare in modo coerente il problema e viene proposta una soluzione di compromesso che risulta migliore di altre soluzioni suggerite in letteratura. Alcune sperimentazioni empiriche concludono il lavoro.

Keywords: Multipurpose Stratified Survey; Power Allocation; Compromise Allocation

1. Introduction

In this paper we consider the case in which the domains of interest are the whole population and the single strata. This seems to be a relatively simple case, but when large discrepancies occur in strata sizes, this problem becomes quite complex since an optimal solution for the whole population may produce unreliable estimates for relatively small strata. In this paper we propose a solving criterion for this problem and suggest a compromise allocation that in some sampling contexts may perform better than known methods.

In order to illustrate the sampling context, let us consider a population U of size N partitioned in H strata U_h ($h=1, \dots, H$) of size N_h . Let us denote with y_{hk} the value of the target variable y of the k -th element ($k=1, \dots, N_h$) of the population U_h . There are $H+1$ parameters of interest: the total population mean $\bar{Y}$ and the H stratum means $\bar{Y}_h$ ($h=1, \dots, H$), being

$$\bar{Y} = \frac{1}{N} \sum_{h=1}^H \sum_{k=1}^{N_h} y_{hk} = \sum_{h=1}^H \frac{N_h}{N} \bar{Y}_h; \quad \bar{Y}_h = \frac{1}{N_h} \sum_{k=1}^{N_h} y_{hk}. \quad (1)$$

A sample of *fixed* size n is selected for estimating the parameters of interest; n_h units are selected without replacement and with equal probabilities from the generic stratum h . The Horvitz-Thompson estimates of the above parameters, expressed by

$$\tilde{\bar{Y}} = \frac{1}{N} \sum_{h=1}^H \frac{N_h}{n_h} \sum_{k=1}^{n_h} y_{hk} = \sum_{h=1}^H \frac{N_h}{N} \tilde{\bar{Y}}_h; \quad \tilde{\bar{Y}}_h = \frac{1}{n_h} \sum_{k=1}^{n_h} y_{hk}, \quad (2)$$

have relative standard errors given by

$$\varepsilon(\tilde{\bar{Y}}) = \sqrt{V(\tilde{\bar{Y}})} / \tilde{\bar{Y}} ; \quad \varepsilon(\tilde{\bar{Y}}_h) = \sqrt{V(\tilde{\bar{Y}}_h)} / \tilde{\bar{Y}}_h \quad (3)$$

being

$$V(\tilde{\bar{Y}}_h) = \frac{N_h - n_h}{N_h} \frac{S_h^2}{n_h} ; \quad V(\tilde{\bar{Y}}) = \sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 V(\tilde{\bar{Y}}_h) ; \quad S_h^2 = \frac{1}{N_h - 1} \sum_{k=1}^{N_h} (y_{hk} - \bar{Y}_h)^2 .$$

The *Neyman* allocation

$$n_{ey} n_h = n (N_h S_h) / \sum_{h=1}^H N_h S_h , \quad (4)$$

minimises the variance $V(\tilde{\bar{Y}})$ of the total population mean. But when large discrepancies occur in strata sizes, this method may produce unreliable estimates for those strata. Furthermore, it produces large variability in the distribution of $\varepsilon(\tilde{\bar{Y}}_h)$ when $f = n/N$ is relatively small or when the population strata N_h are unevenly distributed.

2. Compromise allocations

The Power allocation (Särndal *et al.*, 1992, formula 12.7.18), (here denoted as *POW*) may be useful in our context

$$n_{pow} n_h = n (X_h^\alpha \frac{N_h S_h}{Y_h}) / \sum_{h=1}^H X_h^\alpha \frac{N_h S_h}{Y_h} = n (X_h^\alpha CV_{Y_h}) / \sum_{h=1}^H X_h^\alpha CV_{Y_h} \quad (5)$$

where Y_h and X_h denote the totals in the stratum h of the variable of interest y and of an auxiliary variable x , and α is a constant varying in $0, 1$ interval. By setting $\alpha = 1$ and $X_h = Y_h$, the Power allocation coincides with the Neyman allocation. By setting $\alpha = 0$, when CV_{Y_h} ($h=1, \dots, H$) are approximately constant among strata, each stratum has an equal allocation, given by $n_h = n / H$; this choice is optimal for small strata but largely inefficient for estimating $\tilde{\bar{Y}}$. A suitable choice of α makes it possible to increase considerably the precision in small strata, with only a slight reduction of precision in the estimates for large strata and for the population as a whole. Practical experience suggests the choice of α value in the interval $0.33 \leq \alpha \leq 0.5$. We observe, however, that when $\alpha = 0$, the $\varepsilon(\tilde{\bar{Y}}_h)$ values are not exactly equal. The differences among $\varepsilon(\tilde{\bar{Y}}_h)$ are strictly connected to the CV_{Y_h} values and to the strata dimensions, N_h .

Another allocation (here denoted as *COMPI*), frequently used for large surveys may be obtained as linear convex combination between the Neyman allocation and an allocation with equal $n_h = n/H$ in each stratum

$$_{comp1}n_h = \gamma_{ney}n_h + (1 - \gamma)(n/H) \quad 0 \leq \gamma \leq 1. \quad (6)$$

In this paper we propose a *compromise allocation* (COMP2), obtained as a convex combination between the Neyman allocation and an allocation assuring an equal value of the squared relative standard error of the stratum estimates. In formula

$$_{comp2}n_h = \lambda_{ney}n_h + (1 - \lambda)_{eq}n_h \quad 0 \leq \lambda \leq 1. \quad (7)$$

In order to derive an explicit value of $_{eq}n_h$, we first obtain the values of $_{eq}n_h$ as function of the constant value, say η , of the squared relative standard error of the stratum estimates

$$_{eq}\mathcal{E}^2(\bar{\bar{Y}}_h) = \frac{V(\bar{\bar{Y}}_h)}{\bar{\bar{Y}}_h^2} = \frac{1}{\bar{\bar{Y}}_h^2} \frac{S_h^2(N_h - _{eq}n_h)}{_{eq}n_h \cdot N_h} = \eta \Rightarrow _{eq}n_h = \frac{S_h^2}{\eta \cdot \bar{\bar{Y}}_h^2 + S_h^2 / N_h}, \quad (8)$$

then, we derive the value of η by considering the constraint on the sample size

$$\sum_{h=1}^H _{eq}n_h = \sum_{h=1}^H \frac{S_h^2}{\eta \cdot \bar{\bar{Y}}_h^2 + S_h^2 / N_h} = n. \quad (9)$$

Formula (9) defines a function $n = f(\eta)$ that is implicit with respect to η . The η value solution of (9) may be found through a numerical method such as the Newton-Raphson one. The $_{eq}n_h$ are then obtained by inserting in (8) the η value given by (9).

In order to explain the properties of COMP2 allocation, we define an index δ of the variability of $\mathcal{E}(\bar{\bar{Y}}_h)$ among strata as

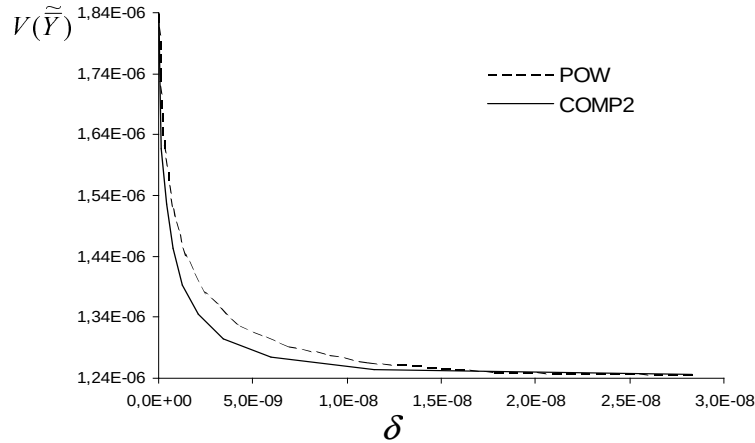
$$\delta = \frac{1}{H} \sum_{h=1}^H \left(\mathcal{E}(\bar{\bar{Y}}_h) - (1/H) \sum_{h=1}^H \mathcal{E}(\bar{\bar{Y}}_h) \right)^2.$$

It is possible to demonstrate that δ is a monotone increasing function of λ , for λ varying in $[0,1]$ interval: (i) $\delta = 0$ for $\lambda = 0$, (ii) $\delta =_{\max} \delta_\lambda > 0$ for $\lambda = 1$. By consequence, a given value of λ , say $\bar{\lambda}$, in $[0, 1]$ λ interval defines an unique value of δ , say $\delta_{\bar{\lambda}}$, in $[0, \delta_{\max}]$ interval. We further note that δ is a function of the α and γ values; δ assumes a minimum value $\delta_{\min} = \delta_{\alpha=\min} = \delta_{\gamma=\min} > 0$ for $\alpha = \gamma = 0$, (ii) $\delta =_{\max} \delta_{\alpha=\max} = \delta_{\gamma=\max} = \delta_{\lambda} > 0$ for $\alpha = \gamma = \lambda = 1$.

An interesting result is that the COMP2 allocation is more efficient than the POW and COMP1 allocations. More in detail, for a fixed value of the variance $V(\bar{\bar{Y}})$, say $\bar{V}$, the COMP2 allocation has a δ value lower than those attained from the POW and COMP1 allocations. Therefore, for a fixed value of δ , say $\bar{\delta}$, the COMP2 allocation variance $V(\bar{\bar{Y}})$ is lower than that obtainable from both the POW and the COMP1 allocations. This COMP2 allocation efficiency improvement may be analytically proved and its demonstration can be provided by the authors. The larger efficiency of COMP2 versus POW allocation may be also illustrated through empirical evidences, which show that COMP1 and POW have roughly the same efficiency level.

The COMP2 allocation efficiency is illustrated in Figure 1, in which the power and the compromise allocations variances $V(\tilde{Y})$ are plotted versus δ . Figure 1 refers to a relevant empirical application on the Italian population. We consider that the sample (with the same size of the Italian labour force survey) should provide estimates at national and regional level; in this application regions are considered as strata.

Figure 1 Plot of $V(\tilde{Y})$ versus δ for POW and COMP2 allocations with α and λ varying in the $[0,1]$ interval.



As noted above, δ is a monotone increasing function of λ . A COMP2 allocation efficiency indicator may be represented by the following

$$eff_{\bar{\lambda}} = {}_{com, \bar{\lambda}} V(\tilde{Y}) / {}_{ney} V(\tilde{Y}) \quad (10)$$

where ${}_{com, \bar{\lambda}} V(\tilde{Y})$ is the variance of $\tilde{Y}$ when using COMP2 allocation with a fixed value $\bar{\lambda}$ of λ , ${}_{ney} V(\tilde{Y})$ is the variance of $\tilde{Y}$ when using Neyman allocation. Thus $eff_1 = 1$; $eff_0 = \max$. It is then possible to define a loss function $\xi(\bar{\lambda})$ with respect to the two conflicting objectives of the survey (minimising $V(\tilde{Y})$ and minimising δ at the same time):

$$\xi(\bar{\lambda}) = 2 - (I_{eff, \bar{\lambda}} + I_{\delta, \bar{\lambda}})$$

$$\text{with } I_{eff, \bar{\lambda}} = 1 - \frac{eff_{\bar{\lambda}} - 1}{eff_0 - 1}; \quad I_{\delta, \bar{\lambda}} = \frac{\delta_{\bar{\lambda}}}{\max_{\lambda} \delta_{\lambda}}.$$

The $\xi(\bar{\lambda})$ is a positive concave function. Therefore, the best choice of $\bar{\lambda}$, is given by the $\bar{\lambda}$ value in which the loss function $\xi(\bar{\lambda})$ attains its minimum value.

References

Särndal, C.E., Swensonn, B., Wretman, J., (1992), *Model Assisted Survey Sampling*, Springer-Verlag New York.