A (BAYESIAN) NOTE ON NON-RANDOM SAMPLES FROM FINITE POPULATIONS

by

M.J. Bayarri
Universitat de València, Spain
and Purdue University

and

B. Font
Universitat de València, Spain

Technical Report #94-32

Department of Statistics
Purdue University

November 1994

*

A (Bayesian) note on non-random samples from finite populations

M.J. Bayarri and B. Font

Universitat de València, Spain

Abstract

When sampling from finite populations in practice, random samples from the whole population are rarely used, due to either the high cost involved or the gaining in information derived from more efficient designs. Bayesian hierarchical models are a natural framework to model the non-randomness in the sample and have been widely and successfully applied towards this end. This paper concentrates on the implications that the design has on the inferences about some characteristics of the finite population, and in a critic comparison among some usual designs.

KEYWORDS: Bayesian prediction; cluster sampling; hierarchical models; stratified sampling.

1. Introduction

A most standard assumption in many statistical inferential processes is that data constitute a random sample of the population of interest. This is not so in the scenario of sampling from finite populations, in which non-random samples are widely used. A very interesting context in which non-random samples arise, that we shall not be treating in this paper, is that of non-randomness due to selection bias (as in Bayarri and DeGroot(1992)[1] and West(1994)[6]). Also, there are instances when the lack of a well defined sampling frame forces the inferences to be based on samples selected somehow haphazardly, and we treated those elsewhere (Bayarri and Font(1994)[2]). In this paper we simply concentrate in the most traditional and familiar alternatives to random samples, that is, in *stratification* and *cluster* sampling.

In a very broad and general sense, stratification is usually sought out, normally with the aim of increasing precision by isolating populations thought to be quite different among themselves; standard stratification is by sex, age, type of habitat, ...etc. On the other hand, clusters are usually encountered, not specially desired, but used because they generally provide a lot more data at a very low additional cost; traditional sampling clusters are households, schools, ...etc. Broadly speaking, the way one thinks about strata is as theirs being some few subgroups of the population, large in size (relative to the size of the finite population), relatively homogeneous within each stratum, and quite heterogeneous accross strata. Clusters brings the idea of small groups, with elements quite similar to each other within each group, the hetereogeneity accross groups being very large. When comparing inferences drawn from a random sample from the finite populations to those based on a stratified or cluster sample of the same size, it is generally perceived that an efficient stratified sample can provide much more information than the random sample, which, in turns, is perceived as being more informative than a cluster sample of the same size (although, of course, a large part of the cluster sample data have been collected basically for free). These intuitive perceptions can in fact be shown to hold when data is analysed from a classical, randomization based (or model assisted), approach to finite population sampling (see, for instance, Cochran(1977)[3]). We have not seen anything similar from the prediction (or model based) approach. This was, in fact, the motivation of this paper. We shall show that the common notions that a stratified sample is "good " and a cluster sample "bad " when compared to random samples, can also be shown to hold in the prediction approach. We shall take a Bayesian point-of-view throught.

This is meant to be only a short note, and we have organised it in three sections, of which this introduction is Section 1. In Section 2 we formulate the problem and remind the results for simple random sampling. In Section 3 a particular two-stage exchangeable model is used to compare random samples with stratified and cluster samples; results are particularized in the usual scenarios for strata and clusters.

2. Formulation

We assume the finite population consisting of N values $y_1, y_2, \ldots, y_N$ of some univariate, continuous characteristic of interest. According to the model-based, or prediction, approach to finite population sampling, $y_1, y_2, \ldots, y_N$ are treated as the realized values of N random variables $Y_1, Y_2, \ldots, Y_N$, whose joint distribution, usually referred to as *superpopulation model*, becomes the natural link between observed and unobserved Y's. (We shall use Y to generically denote any of the random variables in the population.) We assume that, in the superpopulation model

$$E(Y) = \theta_Y, \quad Var(Y) = \sigma_Y^2. \tag{2.1}$$

The natural, and distinctive, goal of statistical analyses of finite populations is to make inferences (predictions) concerning a characteristic, T, of the population, based on a sample of n elements selected from it. We shall (without loss of generality) denote the data by $y_1, y_2, \ldots, y_n$, and assume that the characteristic of interest T is the mean of the population: $T = \overline{Y} = \frac{\sum_{i=1}^N Y_i}{N}$. Since

$$\begin{aligned} E(\overline{Y}|\text{data}) &= \frac{\sum_{i=1}^n y_i + E(\sum_{i=n+1}^N Y_i|\text{data})}{N} = \\ &= f\overline{y}_s + (1-f)E(\overline{Y}_u|\text{data}), \\ Var(\overline{Y}|\text{data}) &= (1-f)^2 Var(\overline{Y}_u|\text{data}), \end{aligned} \tag{2.2}$$

where $f = \frac{n}{N}$ is the *sampling fraction*, and $\overline{Y}_u = \frac{\sum_{i=n+1}^N Y_i}{N-n}$ is the mean of the unobserved Y's, we shall restrict ourselves to consideration of $E(\overline{Y}_u|\text{data})$ and $Var(\overline{Y}_u|\text{data})$. (Notice more generally that $\overline{Y}$ is simply a lineal transformation of $\overline{Y}_u$, so if the distribution of $\overline{Y}$ is desired consideration can also be restricted to deriving the distribution of $\overline{Y}_u$.)

When simple random sampling (SRS from now on) is used, the Y_j's are generally assumed to be i.i.d. with a $N(\theta_Y, \sigma_Y^2)$ distribution. Hence, a superpopulation model frequently used with simple random sampling is

$$Y|\theta_Y, \sigma_Y^2 \sim N_N(\theta_Y 1_N, \sigma_Y^2 I_N), \tag{2.3}$$

where $Y = (Y_1, Y_2, \ldots, Y_N)$, 1_N is the vector of ones, I_N the identity matrix, and N_N reffers to N-variate normal.

In many situations, instead of SRS we may decide to stratify or might have to use cluster sampling. Both situations are naturally modelled by using a two-stage hierarchical model (Scott&Smith(1969)[5]). Accordingly, assume that we have K groups (strata or clusters), with M_i elements in group i, $i = 1, 2, \ldots, K$. We have to change the notation and let Y_{ij} denote observation j in group i. A natural hierarchical model to consider would then be

$$\begin{array}{ll} Y_{ij} \sim N(\theta_i, \sigma_i^2), \quad i = 1, 2, \ldots, K, \quad j = 1, 2, \ldots, M_i, \\ \theta_i \sim N(\theta, \sigma_\theta^2), \qquad \sigma_i^2 \sim \pi(\sigma_i^2), \qquad i = 1, 2, \ldots, K. \end{array} \tag{2.4}$$

But if we are, say, in the design stage, and we are trying to decide whether to use SRS or to stratify, or whether or not to use a cluster sample, then (2.3) and (2.4) should at least be approximately equivalents (they can not be fully matched with this usual formulation). We take this to mean that the marginal first two moments match. Hence, since from (2.4),

$$E(Y) = E[E(Y|\text{group } i)] = E(\theta_i) = \theta, \tag{2.5}$$

it follows that θ in (2.4) has to exactly be θ_Y in (2.3). Also, since

$$\begin{aligned} Var(Y) &= E[Var(Y|\text{group } i)] + Var[E(Y|\text{group } i)] = \\ &= E(\sigma_i^2) + Var(\theta_i) = E(\sigma_i^2) + \sigma_\theta^2, \end{aligned} \tag{2.6}$$

a comparison with (2.4) shows that the following identities have to hold for some $0 < \lambda < 1$;

$$E(\sigma_i^2) = \lambda\sigma_Y^2, \quad \sigma_\theta^2 = (1-\lambda)\sigma_Y^2. \tag{2.7}$$

A similar argument has to also be applied when sampling with covariates. Indeed, in this case, if SRS is used, and X is the $1 \times p$ vector of covariates associated with Y, then the superpopulation model has the Y_i's independent with $E(Y|X) = X\boldsymbol{\beta}$ and some $Var(Y|X)$, where $\boldsymbol{\beta}$ is a $p \times 1$ vector of regressors. (We take the term covariate in a very broad sense allowing also for design constants.) The usual two-stage models have

$$\begin{array}{lll} E(Y|X, \text{group } i) = X\boldsymbol{\beta}_i, & Var(Y|X, \text{group } i) = \sigma_i^2, & i = 1, 2, \ldots, K \\ \boldsymbol{\beta}_i \sim N_p(b_0, V^{-1}), & \sigma_i^2 \sim \pi(\sigma_i^2), & i = 1, 2, \ldots, K. \end{array} \tag{2.8}$$

Again, the model used with random samples should, if not equal at least be a suitable approximation to the superpopulation model derived from the more ellaborated hierarchical models (2.8) and we shall again require the first two moments to match. Since, from (2.8)

$$E(Y|X) = E[E(Y|X, \text{group } i)] = Xb_0, \tag{2.9}$$

it follows that the second-stage prior mean b_0 has to be equal to the vector of regressors $\boldsymbol{\beta}$ of the formulation with SRS. Further insight is gained when comparing the variances. Indeed, we have from (2.8) that

$$\begin{aligned} Var(Y|X) &= E[Var(Y|X, \text{group } i)] + Var[E(Y|X, \text{group } i)] = \\ &= E(\sigma_i^2) + XV^{-1}X^t. \end{aligned} \tag{2.10}$$

It follows that in the SRS formulation, $Var(Y|X)$ has to necessarily depend on X (unless the $\boldsymbol{\beta}_i$'s are known). A possibility that meets the requirement in (2.10) is

$$\begin{aligned} Var(Y|X) &= (XX^t)\sigma^2, \\ E(\sigma_i^2) &= (XX^t)\lambda\sigma^2, \quad V^{-1} = (1-\lambda)\sigma^2 I. \end{aligned} \tag{2.11}$$

This paper aims at pointing to relationships and differences among SRS, strata and cluster sampling as clearly, as possible, and hence will concentrate on easy comparisons in the simplest formulation. We shall therefore only discuss inferences in the exchangeable situations (no covariates) described by (2.3) in the SRS case, and by (2.4) in the strata/clusters case, with the relationship (2.5) and (2.7) holding among them, σ^2 will be assumed known from now on. We next procceed to accordingly derive the posterior predictive distribution of $\overline{Y}_u$, the mean of the unobserved units, for SRS, stratified and cluster samplings. The results for SRS are an easy exercise and will quickly be mentioned next, while the ones for the hierarchical models are given in Section 3.

Assume that data $y_1, y_2, \ldots, y_n$ are obtained by SRS and that superpopulation model (2.3) is thought to be appropiate. Then, if the usual non-informative prior $\pi(\theta_Y) \propto$ constant is used, it is easy to check that the posterior predictive distribution of $\overline{Y}_u$ given the data is normal with

$$E(\overline{Y}_u|\text{data}) = \overline{y}_s, \quad Var(\overline{Y}_u|\text{ data}) = \frac{N}{N-n}\,\frac{\sigma_Y^2}{n}. \tag{2.12}$$

(Notice, by the way, that these results closely match the ones obtained in the classical, randomization-based approach.)

3. A compatible two-stage hierarchical model

As discussed previously, instead of SRS we may wish to stratify or decide to get extra data by exhausting clusters, or have to base inferences in a cluster design for some other reasons. A hierarchical two-stage model (2.4) is then deemed appropriate with the conditions $\theta = \theta_Y$ and (2.7) on its hyperparameters. We take here, for the sake of simplicity of calculus and resulting expressions, a very particular case in which the variances of the Y values around each mean group θ_i are assumed equal: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_K^2$. It then follows from (2.7) that $\sigma_i^2 = \lambda\sigma_Y^2$, for $i = 1, 2, \ldots, K$. If we further take λ to be known the resulting simple model is

$$\begin{aligned} Y_{ij} &\sim N(\theta_i, \lambda\sigma_Y^2), \quad j = 1, 2, \ldots, M_i, \; i = 1, 2, \ldots, K \\ \theta_i &\sim N(\theta_Y, (1-\lambda)\sigma_Y^2), \quad i = 1, 2, \ldots, K \\ \pi(\theta_Y) &\propto \text{constant}. \end{aligned} \tag{3.1}$$

Before deriving the posterior predictive distribution of $\overline{Y}_u$, it is worth pausing and deriving the superpopulation model in this case. By integrating out $\theta_1, \ldots, \theta_K$ from (3.1), it can be seen that the joint distribution of $Y_{11}, \ldots, Y_{KM_K}$ given θ_Y is here a N-variate normal with $E(Y_{ij}) = \theta_Y$ and $Var(Y_{ij}) = \sigma_Y^2$, exactly as in (2.3). The variance-covariance matrix, however, is no longer diagonal.

Instead,

$$Cov(Y_{ij}, Y_{i^*j^*}) = \begin{cases} (1-\lambda)\sigma_Y^2 & \text{if } \; i = i^* \;\; \text{(same group)} \\ 0 & \text{if } \; i \neq i^* \;\; \text{(different groups)} \end{cases} \tag{3.2}$$

Hence, although observations from different groups are independent, observations from the same group are not. In fact, the intra-group correlation coefficient can be seen to precisely be $\rho = 1 - \lambda$. (A more appropriate formulation, specially for some cluster samplings, would take $\sigma_i^2 = \lambda_i \sigma_Y^2$ with $E(\lambda_i) = \lambda$ so as to allow for different intra-cluster correlation coefficients. We shall not pursue it here, however.)

Without loss of generality, we assume that data consist of the first $m_i \leq M_i$ elements of the first $k \leq K$ groups, $i = 1, 2, \ldots, K$, resulting in sample means $\overline{y}_{s_1}, \overline{y}_{s_2}, \ldots, \overline{y}_{s_k}$. The joint posterior predictive distribution of the means of the unobserved elements in the K groups, $\overline{Y}u_1, \ldots, \overline{Y}u_K$ given the data can be shown to be a K-variate normal with mean vector

$$E(\overline{Y}_{u_i}|\text{data}) = \begin{cases} \alpha_i \, \overline{y}_{s_i} + (1-\alpha_i)\, \overline{y}_w & i = 1, 2, \ldots, k \\ \overline{y}_w & i = k+1, \ldots, K \end{cases} \tag{3.3}$$

and covariance matrix

$$\lambda\, \sigma_Y^2 \begin{pmatrix} diag_1^k \, \frac{1}{M_i - m_i} & 0_{k \times (K-k)} \\ 0_{(K-k) \times k} & diag_{k+1}^K \, \frac{1}{M_i} \end{pmatrix} + \sigma_Y^2 \begin{pmatrix} diag_1^k \, \frac{\lambda \alpha_i}{m_i} & 0_{k \times (K-k)} \\ 0_{(K-k)\times k} & (1-\lambda) I_{K-k} \end{pmatrix} +$$
$$+ \frac{(1-\lambda)\sigma_Y^2}{\sum_{i=1}^k \alpha_i} \begin{pmatrix} \frac{\lambda^2}{(1-\lambda)^2} \begin{pmatrix} \frac{\alpha_1}{m_1} \\ \vdots \\ \frac{\alpha_k}{m_k} \end{pmatrix} \left(\frac{\alpha_1}{m_1}, \ldots, \frac{\alpha_k}{m_k}\right) & \frac{\lambda}{1-\lambda} \begin{pmatrix} \frac{\alpha_1}{m_1} \\ \vdots \\ \frac{\alpha_k}{m_k} \end{pmatrix} 1_{K-k}^t \\ \frac{\lambda}{1-\lambda}\, 1_{K-k} \left(\frac{\alpha_1}{m_1}, \ldots, \frac{\alpha_k}{m_k}\right) & 1_{K-k} 1_{K-k}^t \end{pmatrix}, \tag{3.4}$$

where $\overline{y}_w = \sum_{i=1}^k \frac{\alpha_i}{\sum_{i=1}^k \alpha_i}\, \overline{y}_{s_i}$ is a weighted average of the observed sample means and

$$\alpha_i = \frac{m_i}{m_i + (\lambda/(1-\lambda))} = \frac{(1-\lambda) m_i}{(1-\lambda) m_i + \lambda}. \tag{3.5}$$

These expressions admit an easy, intuitive interpretation. Thus, from (3.3) it can be seen that, the means units (groups) that have not been sampled at all, are naturally estimated by a suitable weighted average, $\overline{y}_w$, of all of the sample means. For the observed units, the usual estimate $\overline{y}_{s_i}$ is shrinked towards $\overline{y}_w$ an amount determined by the weights α_i. Also, the covariance between the means of two units that have both been sampled is,

$$Cov(\overline{Y}_{u_i}, \overline{Y}_{u_j}|\text{data}) = \begin{cases} \lambda \sigma_Y^2 \left(\frac{1}{M_i - m_i} + \frac{\alpha_i}{m_i} \right) + \frac{\lambda^2 \sigma_Y^2}{(1-\lambda)\sum_{i=1}^k \alpha_i} \; \frac{alpha_i^2}{m_i^2} & i = j, \; i,j \leq k \\ \frac{\lambda^2 \sigma_Y^2}{(1-\lambda)\sum_{i=1}^k \alpha_i} \; \frac{alpha_i \alpha_j}{m_i m_j} & i \neq j, \; i,j \leq k \end{cases}$$

that of one sampled unit and an unsampled one is,

$$Cov(\overline{Y}_{u_i}, \overline{Y}_{u_{j*}}|\text{data}) = \frac{\lambda\sigma_Y^2}{\sum_{i=1}^k \alpha_i} \frac{\alpha_i}{m_i} \;\; i \leq k, \; j^* > k$$

and the one for two unsampled units is,

$$Cov(\overline{Y}_{u_{i*}}, \overline{Y}_{u_{j*}}|\text{data}) = \begin{cases} \lambda\sigma_Y^2 \left(\frac{1}{M_{i*}} + (1-\lambda)\right) + \frac{(1-\lambda)\sigma_Y^2}{\sum_{i=1}^k \alpha_i} & i^* = j^*, \; i^*, j^* > k \\ \frac{(1-\lambda)\sigma_Y^2}{\sum_{i=1}^k \alpha_i} & i^* \neq j^*, \; i^*, j^* > k \end{cases}$$

The weight that each sample mean $\overline{y}_{s_i}$ receives in the combined $\overline{y}_w$ and the amount of shrinkage of $\overline{y}_{s_i}$ toward $\overline{y}_w$ are determined by α_i, which is a function of both, the number of data taken in unit i and the intra-class correlation ρ. It can be shown that α_i is an increasing function of m_i from 0 (if no data is taken in unit i) to $(1 + \frac{\lambda}{1-\lambda} \frac{1}{M_i})^{-1}$ if the unit (group) is exhaustively sampled. It is also an increasing function of ρ from 0 (when $\rho = 0$) to 1 (when $\rho = 1$). Thus, in (3.3), the relative weight that the sample mean $\overline{y}_{s_i}$ has in estimating $\overline{Y}_{u_i}$ is larger, the larger m_i (which makes $\overline{y}_{s_i}$ a better estimate of $\overline{Y}_{u_i}$) and the larger ρ (if ρ is very large, $\overline{y}_{s_i}$ is an excellent estimate of $\overline{Y}_{u_i}$). Still another interpretation can be given to α_i. By noting that

$$\alpha_i = \frac{Var(\theta_i|\theta_Y)}{Var(\theta_i|\theta_Y) + Var(\overline{Y}_{s_i}|\theta_i)},$$

it follows that α_i can be interpreted as the relative contribution of the variance of θ_i around θ_Y to the total variation of the $\overline{Y}_{u_i}$'s around θ_Y. An easy particular case occurs when the same number of observations is taken in each sampled group, that is, $m_1 = m_2 = \ldots = m_k = \frac{n}{k}$. In this case, all the α_i's are equal $\alpha_1 = \alpha_2 = \ldots = \alpha_k = \alpha$, the relative weight of each sample mean $\frac{\alpha_i}{\sum_{j=1}^k \alpha_j}$ is simply $\frac{1}{k}$, so that the combined $\overline{y}_w$ is the sample mean $\overline{y}_s$.

¿From the joint disribution of $\overline{Y}_{u_1}, \ldots, \overline{Y}_{u_K}$ it is easy to derive the distribution of $\overline{Y}_u = \sum_{i=1}^K \gamma_i \overline{Y}_{u_i}$ where $\gamma_i = \frac{M_i - m_i}{N-n}$, which is normal with mean

$$E(\overline{Y}_u|\text{data}) = \frac{1}{N-n} \left\{ \sum_{i=1}^k (M_i - m_i)[\alpha_i \overline{y}_{s_i} + (1-\alpha_i)\overline{y}_w] + (\; sum_{i=k+1}^K M_i)\overline{y}_w \right\} \quad (3.6)$$

and variance

$$\begin{aligned} Var(\overline{Y}_u|\text{data}) &= \frac{\sigma_Y^2}{(N-n)^2} \{\lambda(N-n) + \\ &+ \lambda \sum_{i=1}^k (M_i - m_i) m_i^{-1} \alpha_i \left[M_i - m_i + \frac{N - n - \sum_{i=1}^k (M_i - m_i)\alpha_i}{\sum_{i=1}^k \alpha_i} \right] + \\ &+ (1-\lambda)(\sum_{k+1}^K M_i \left[M_i + \frac{N - n - \sum_{i=1}^k (M_i - m_i)\alpha_i}{\sum_{i=1}^k pha_i} \right])\}, \quad (3.7) \end{aligned}$$

we shall further discuss the expresions (3.6) and (3.7) above in two specially interesting particular cases of strata and clusters.

3.1. Strata

In the most usual scenario for stratified sampling there are very few strata (that is, K is small) and observations are taken in all of them, so that $k = K$. If the stratification is "good ", most of the variability among the values of Y is due to variability accross strata. When $k = K$, expresion (3.6) can be rewritten as

$$E(\overline{Y}_u|\text{data}) = \sum_{i=1}^{K} q_i[\alpha_i \overline{y}_{s_i} + (1-\alpha_i)\overline{y}_w], \tag{3.8}$$

where $q_i = (M_i - m_i)/(N-n)$ is the percentage of unsampled units which lies on stratum i, $i = 1, 2, \ldots, K$. That is, the estimate of $\overline{Y}_u$ is just a weighted average of the shrinkage estimates of the mean $\overline{Y}_{u_i}$ in each stratum; strata with larger proportions of unsampled units get larger weights in the pooled estimate, as it could be expected. The posterior variance of $\overline{Y}_u$ (3.7) is, for $k = K$,

$$Var(\overline{Y}_u|\text{data}) = \frac{N}{N-n}\,\frac{\sigma_Y^2}{n}\{\lambda f + (1-\lambda) f S_{\text{strata}}\}, \tag{3.9}$$

where $S_{\text{strata}} = \sum_{i=1}^{K} (M_i - m_i)(1-\alpha_i)[q_i + \frac{1-\sum_{j=1}^{k}\alpha_j q_j}{\sum_{j=1}^{k}\alpha_j}]$.

An important particular case is that of proportional allocation , in which $m_i/M_i = n/N$. In this case, $q_i = (M_i - m_i)/(N-n) = M_i/N$ (or stratum weight), and $S_{\text{strata}} = [\lambda/(1-\lambda)][(1-f)/f]$ so that (3.9) now becomes

$$Var(\overline{Y}_u|\text{data}) = \frac{N}{N-n}\,\frac{\sigma_Y^2}{n}\lambda = \lambda Var(\text{SRS}). \tag{3.10}$$

(Var(SRS) reffers to the variance of $\overline{Y}_u$ when predicted from a SRS as given in (2.12).) The mean $E(\overline{Y}_u|\text{data})$, however, is not $\overline{y}_s$ as in SRS, but still the weighted average in (3.8) that can not be substantially reduced, and hence comparison with SRS is not direct. Nevertheless, in the particular case in which all the strata have similar sizes, we can assume as an approximation $M_1 = M_2 = \ldots = M_K$, so that with proportional allocation, $m_1 = m_2 = \ldots = m_K$. In this case the variance (3.9) is given by (3.10) and the mean $E(\overline{Y}_u|\text{data})$ is simply $\overline{y}_s$, the overall sample mean.

Hence, we get, in the last case, the same posterior predictive for $\overline{Y}_u$ here as when using SRS, except that the variance here is smaller (since $0 < \lambda < 1$). Also, recall that $Var(Y) = E(\sigma_i^2) + Var(\theta_i) = \lambda\sigma_Y^2 + (1-\lambda)\sigma_Y^2$; good stratification means that $Var(\theta_i)$ is large, that is, $(1-\lambda)$ is large and we can obtain a substancial reduction on the variance of $\overline{Y}_u$, thus confirming the known results.

3.2. Clusters

A usual scenario in cluster sampling is having many clusters (K is large in comparison with N), of quite small size (the M_i's are small), from which a small sample (k is much smaller than K) is taken and sampled exhaustively ($m_i = M_i$ for the k sampled units, $i = 1, 2, \ldots, k$). Clusters are usually encountered, and they usual consists of elements which are quite similar, with cluster means varying almost as much as the individual values of Y in the worst cases.

The posterior predictive distribution of $\overline{Y}_u$ is normal with mean and variances given by (3.6) and (3.7), where here $m_i = M_i$ for $i = 1, 2, \ldots, k$. That is,

$$\begin{aligned} E(\overline{Y}_u|\text{data}) &= \overline{y}_w, \\ Var(\overline{Y}_u|\text{data}) &= \frac{N}{N-n}\frac{\sigma_Y^2}{n}\{\lambda f + (1-\lambda) f S_{\text{cluster}}\}, \end{aligned} \tag{3.11}$$

where $S_{\text{cluster}} = \frac{\sum_{i=k+1}^{K} M_i^2}{N-n} + \frac{N-n}{\sum_{j=1}^{k}\alpha_j}$.

In the particular case of all the clusters being roughly of the same size, we can as an approximation take $M_1 = M_2 = \ldots = M_K = M$, and then

$$\begin{aligned} E(\overline{Y}_u|\text{data}) &= \overline{y}_s, \\ Var(\overline{Y}_u|\text{data}) &= \frac{N}{N-n}\frac{\sigma_Y^2}{n}[\lambda + M(1-\lambda)] = \\ &= [\lambda + M(1-\lambda)]Var(\text{SRS}). \end{aligned} \tag{3.12}$$

It is immediate to see from (3.12) and (3.10) that when comparing SRS, strata an cluster samples in these particular cases, all the three result in the same estimate for $\overline{Y}_u$, but that the variance when clusters are used is larger than the variance of $\overline{Y}_u$ for both SRS and stratified samples. Recall again that $Var(Y) = E(\sigma_i^2) + Var(\theta_i) = \lambda\sigma_Y^2 + (1-\lambda)\sigma_Y^2$ and that in cluster samples $Var(\theta_i)$ is almost as large as $Var(Y)$, so that $(1-\lambda)$ is close to 1 and we can get a $Var(\overline{Y}_u|\text{data})$ much larger than the one obtained with SRS. In fact, as $\lambda \to 0$ (or $\rho \to 1$), the variance in (3.12) goes to $MVar(\text{SRS})$, a result that could be expected.

We have so far compared the three designs -SRS, stratification and clustering- on the basis of the variance of the posterior predictive distribution. Such a comparison can be made more appealing when interpreted as a comparison among the entropies of those predictive destributions. The entropy of a (continuous) distribution with density $p(x)$ is given by

$$-\int p(x)\log p(x)dx,$$

and it has been long used (see, for instance, Renyi(1961)[4]) as a measure of the information contained in $p(x)$. (Note that a relative measure of information, such as the kullback-Leibler divergence between the prior and posterior predictive distributions can not be used here because the former is an improper distribution.)

The entropy of a normal distribution with mean μ and variance σ^2 can be computed to be

$$H\{N(x|\mu, \sigma^2)\} = \frac{1}{2} + \log \sqrt{2\pi} + \log \sigma. \tag{3.13}$$

Hence, comparing the variances of the posterior predictive distributions is equivalent to comparing their entropies in our examples. In fact, the difference between the entropies of the predictive distributions corresponding to two different designs can be seen to be the logarithm of the ratio of their standard desviation, which from (3.12) and (3.10) can be seen to have very simple expressions.

Aknowledgements

We are indebted to Josemar Rodrigues for his useful suggestions. This work has been supported in part by the Generalitat Valenciana, under grant GV-1081/93.

References

[1] Bayarri, M.J. and DeGroot, M.H. (1992). A "BAD" view of weighted distributions and selection models. In *Bayesian Statistics 4*. (J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith eds.), Oxford: Clarendon Press. 17-33.

[2] Bayarri, M.J. and Font, B. (1994). Random routes. *Technical Report 6-94*. Departament of Statistics and O.R., University of Valencia.

[3] Cochran, W.G. (1977). *Sampling Techniques. Third edition.* John Wiley & Sons. New York.

[4] Renyi, A. (1961). On measures of entropy and information. *Proc. Fourth Berkeley Symp.*. **1** (J. Neyman and E.L. Scott, eds.). Berkeley: Univ. California Press, 547-561.

[5] Scott, A. and Smith, T.M.F. (1969). Estimation in Multi-stage Surveys. *J. Amer. Statist. Association.* **64**, 830-840.

[6] West, M. (1994). Discovery sampling and selection models. In *Statistical Decision Theory and Related Topics V.* (S.S. Gupta and J.O. Berger, eds.), Springer-Verlag. New York. 221-235.