BAYESIAN STATISTICS 3, pp. 437–451
J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, (Eds.)

To Weight or not to Weight, That is the Question

(Whether 'tis nobler in the mind to suffer
The slings and arrows of outrageous fortune,
Or to take arms against a sea of troubles,
And by opposing end them?) Hamlet, Act 3, Scene 1.

T. M. F. SMITH
University of Southampton

SUMMARY

Weighting by the inverse unit selection probabilities is the basis of randomization inference. In a model-based framework probability designs are ignorable and so probability weights have no obvious role. This issue of whether to weight or not is examined by following Rubin (1983) and conditioning on the selection probabilities. Using results from size biased sampling it is shown that randomization estimators can be justified.

Keywords: RANDOMIZATION; WEIGHTING; CONDITIONAL INFERENCE; IGNORABLE DESIGNS, SIZE-BIASED SAMPLING; REGRESSION; ROBUST ESTIMATION.

1. INTRODUCTION

Statisticians frequently seek to protect themselves against outrageous fortune by an act of randomization. In sample surveys this may involve the use of different selection probabilities for different population units and the inverse selection probabilities may then be used as weights in forming estimates of population totals. These weights are basic to randomization inference and any method of estimation which fails to use them is treated with great suspicion. An alternative to randomization inference is to assume that the distribution of population values can be represented by a probability model. A sample selection mechanism using randomization can be ignored for model-based inferences and then there is no apparent role for probability weights. The problem addressed in this paper is whether probability weights have a role in model-based inference for sample surveys.

2. RANDOMIZATION INFERENCE

Let I_i be an indicator variable for unit $i, i = 1, \ldots, N$, in a finite population, such that

$$I_i = \begin{cases} 1 & \text{if } i \in s \\ 0 & \text{otherwise,} \end{cases}$$

where $s = (i_1, \ldots, i_n)$ is the set of labels selected by a sampling mechanism. For samples of fixed size n we have $\sum_{i=1}^N I_i = n$ and

$$\Pr(I_i = 1) = \pi_i, \tag{2.1}$$

which is the inclusion probability for the ith unit when randomization is employed. We assume that $0 < \pi_i < 1$ for all i.

A sampling mechanism is a rule for selecting s, a subset of the population units. Let $\boldsymbol{X}$ denote the prior knowledge available to a statistician before drawing the sample and let $\boldsymbol{Y}$ denote the $N \times p$ matrix of values of the survey variables of interest. A sampling mechanism of the form

$$p(s|\boldsymbol{X}) \tag{2.2}$$

for which $0 < \pi_i = \sum_{s \ni i} p(s|\boldsymbol{X}) < 1$ is called strongly ignorable, Rosenbaum and Rubin (1983). Random sampling schemes satisfy this condition, but quota sampling schemes may not, see Smith (1983). In practice the observed sample may also be determined by a non-response selection mechanism which is not under the statistician's control and may depend on the survey variables $\boldsymbol{Y}$. Such a mechanism would not be ignorable, see Little (1982). In this paper we assume throughout that the selection mechanism is strongly ignorable.

Let $\mathcal{S}$ denote the $\binom{N}{n}$ possible samples which might be drawn. The probability distribution on $\mathcal{S}$ determined by $p(s|\boldsymbol{X})$ is the randomization distribution. From the statistical point of view it has the interesting property of being completely known; it is not indexed by any unknown parameters, nor is it directly related to the survey variables $\boldsymbol{Y}$. If T is some function of $\boldsymbol{Y}$ of interest and $\hat{T}_s$ is an estimator of T then the only statistical operation of any content is to take expectations with respect to $p(s|\boldsymbol{X})$, that is to form

$$E_p(\hat{T}_s) = \sum_{s \in \mathcal{S}} \hat{T}_s p(s|\boldsymbol{X}). \tag{2.3}$$

Since $\boldsymbol{Y}$ can take any values the only useful general constraint is to require that

$$E_p(\hat{T}_s) = T \qquad \text{for all possible } \boldsymbol{Y}, \tag{2.4}$$

that is to require that estimators be p-unbiased. When T is a total and the estimators are linear in the indicator variables I_i, so that

$$\hat{T}_s = \sum_{i=1}^{N} w_i g(\boldsymbol{Y}_i) I_i, \quad \text{and} \quad T = \sum_{i=1}^{N} g(\boldsymbol{Y}_i), \tag{2.5}$$

p-unbiasedness leads to

$$E_p(\hat{T}_s) = \sum_{i=1}^{N} w_i g(\boldsymbol{Y}_i)\pi_i = \sum_{i=1}^{N} g(\boldsymbol{Y}_i) \qquad \text{for all } \boldsymbol{Y},$$

so that

$$w_i = \pi_i^{-1},$$

the inverse probability weight.

Example 1. The population mean

Let $T = \bar{Y} = \Sigma_1 Y_i/N$. An unbiased estimator is

$$\hat{T}_{1s} = \frac{1}{N}\sum_{i \in s} w_i Y_i, \quad \text{with } w_i = \pi_i^{-1}. \tag{2.6}$$

Which is the well known Horvitz-Thompson estimator. If a computer package is used for data analysis then the weighted estimator will be

$$\hat{T}_{2s} = \sum_{i \in s} w_i Y_i / \sum_{i \in s} w_i, \tag{2.7}$$

which is now a ratio and is not unbiased. However, $\hat{T}_{2s}$ is component-wise unbiased in the sense that $\Sigma_{i \in s} w_i Y_i$ is an unbiased estimator of $\Sigma_1^N Y_i$ and $\Sigma_{i \in s} w_i$ is an unbiased estimator of N. $\hat{T}_{2s}$ was suggested by Hajek (1971) as a possible solution to the Basu elephant problem and has the desirable property of being location invariant, which is not true for $\hat{T}_{1s}$.

Example 2. A regression coefficient.

Let

$$B = \sum_{i=1}^{N} (Y_{1i} - \bar{Y}_1)Y_{2i} - \bar{Y}_2) / \sum_{i=1}^{N} (Y_{2i} - \bar{Y}_2)^2 \tag{2.8}$$

be the finite population regression coefficient between Y_1 and Y_2. Apparently this is sometimes of interest. There are many alternative estimators of B, all of which are biased. Applying the weights w_i to each unit $i \in s$ gives

$$\hat{B}_w = \frac{\sum_s w_i \sum_s Y_{1i} Y_{2i} w_i - \sum_s Y_{1i} w_i \sum_s Y_{2i} w_i}{\sum_s w_i \sum_s Y_{2i}^2 w_i - \left(\sum_s Y_{2i} w_i\right)^2}, \tag{2.9}$$

which is the analogue of (2.7). In (2.9) a term like $\Sigma_s Y_{1i} Y_{2i} w_i$ is the unbiased estimator of $\Sigma_{i=1}^N Y_{1i} Y_{2i}$, and so (2.9) can be viewed as a function of unbiased estimators of totals T_j. So if $h(\boldsymbol{T})$ is the function of interest $h(\hat{\boldsymbol{T}}_s)$ is the component-wise unbiased estimator. All standard sample survey estimators are in this class so randomization inference for sample surveys is closely tied to p-unbiasedness. Taylor series expansions give the conditions under which this is a reasonable approach.

For a general multiple regression problem with Y_{1i} regressed on $\boldsymbol{Y}_{2i}, i = l, \ldots, N$, then the weighted estimator is

$$\hat{\boldsymbol{B}}_w = \left(\boldsymbol{Y}_{2s}^T \boldsymbol{w}_s \boldsymbol{Y}_{2s}\right)^{-1} \boldsymbol{Y}_{2s}^T \boldsymbol{w}_s \boldsymbol{Y}_{1s}, \tag{2.10}$$

where $\boldsymbol{Y}_{1s}$, is the $n \times 1$ vector of dependent variables in s, $\boldsymbol{Y}_{2s}$ is the $n \times p$ matrix of explanatory variables in s, and $\boldsymbol{w}_s = \text{diag}(w_i, i \in s)$ is the $n \times n$ matrix with sample weights down the diagonal. This is the solution obtained by using the weighting option in a standard package of statistical programs. It should be noted however, that the variance associated with (2.10) in packages is usually the weighted least squares variance

$$\hat{V}(\hat{\boldsymbol{B}}_w) = \left(\boldsymbol{Y}_{2s}^T \boldsymbol{w}_s \boldsymbol{Y}_{2s}\right)^{-1} \hat{\sigma}^2, \tag{2.11}$$

and this is not the randomization variance derived from $p(s|\boldsymbol{X})$, see Rao (1975).

3. ALTERNATIVES TO RANDOMIZATION INFERENCE

3.1. Ordinary Least Squares (OLS)

If Rubin (1976), Rosenbaum and Rubin (1983), Sugden and Smith (1984) etc., say that random sampling is ignorable for inference then why not ignore it? Ignoring sample weights surely implies using equal weights which in turn implies ordinary least squares as a criterion for regression analysis.

The OLS estimator is

$$\hat{B}_0 = \left(Y_{2s}^T Y_{2s}\right)^{-1} Y_{2s} Y_{1s}, \tag{3.1}$$

which is an unweighted alternative to $\hat{B}_w$ in (2.10). This is frequently chosen for the analysis of data from a complex survey as a default option. As we shall see in Section 3.2 this approach takes the word ignorable at face value and fails to read the small print.

Social surveys are usually designed to be self-weighting, in which case the OLS estimator is also the component-wise unbiased estimator. However, as stated above, the least squares variance is not the correct p-variance and clustering in the design can lead to considerable inflation of the true p-variance relative to the OLS variance, see Kish and Frankel (1974).

3.2. Adjusted Least Squares (ALS)

A full model of a survey requires the joint distribution of the survey variables Y, the prior variables X and the sample selection variable s. Formally we can write

$$f(s, Y, X; \lambda) = p(s|X) f(Y|X; \theta) g(X; \phi), \tag{3.2}$$

where $\lambda = (\theta, \phi)$ is a vector of parameters. The sample data comprise $d_s = (s, Y_s, X)$, and then

$$f(d_s; \lambda) = p(s|X) g(X; \phi) f_s(Y_s|X; \theta), \tag{3.3}$$

where $f_s(Y_s|X;\theta) = \int f(Y|X;\theta) dy_{\bar{s}}$, and $Y_{\bar{s}}$ denotes $\{Y_i : i \notin s\}$. If X is known then predictive inferences about $Y_i; i \in \bar{s}$, can be made via the conditional distribution $f(Y|X;\theta)$ ignoring the design $p(s|X)$. If X_i, $i \in \bar{s}$ is not known then the design $p(s|X)$ contains potentially usefull information that will help the satistician to predict X_i, $i \in \bar{s}$ and hence Y_i, $i \in \bar{s}$, see Scott (1977), Sugden and Smith (1984).

From the sample data the parameters θ can be estimated from the conditional distribution $f_s(Y_s|X;\theta)$, and the parameters ϕ from the marginal distribution $g(X;\phi)$. In the regression problem in Section 2 the parameter of interest was a regression coefficient between Y variables and is thus defined in the marginal distribution of Y, which is not directly observable from the sample data. The problem is how to use the sample data to estimate parameters in the marginal distribution of Y?

If (X, Y) has a multivariate normal distribution, or equivalently if $E(Y|X)$ is linear in X with a constant covariance matrix, where $E(\cdot)$ denotes expectation with respect to the model, then the adjusted least squares estimators of μ_y, Σ_{yy}, the mean vector and covariance matrix of the marginal distribution of Y are

$$\hat{\mu}_Y = m_y + b_{yx}(M_x - m_x),$$

and

$$\hat{\Sigma}_{YY} = s_{yy} + b_{yx}(S_{xx} - s_{xx}) b_{yx}^T, \tag{3.4}$$

where $S = \begin{pmatrix} s_{xx} & s_{xy} \\ s_{yx} & s_{yy} \end{pmatrix}$ is the *unweighted* sample covariance matrix of (X_s, Y_s), $b_{yx} = s_{yx}^{-1} s_{xx}$, S_{xx} is the finite population (known) covariance matrix of X, and m_y, m_x are the unweighted sample mean vectors of (Y, X), and M_x is the finite population mean vector of X. If $\hat{\Sigma}_{YY}$ is partitioned according to (Y_1, Y_2), then the adjusted regression coefficient of Y_1 on Y_2 becomes

$$\hat{B}_A = \hat{\Sigma}_{Y_1Y_2} \hat{\Sigma}_{Y_2Y_2}^{-1} \tag{3.5}$$

The properties of (3.5), (3.1) and (2.10) have been compared in a simulation study by Holt, Smith and Winter (1980), Smith (1981), under various sampling schemes. For unequal probability selection schemes the OLS estimator B_0, is badly biased, while both $\hat{B}_w$ and $\hat{B}_A$ remain approximately unbiased provided the population satisfies the linearity and homoscedasticity assumption. Under this assumption $\hat{B}_A$ is generally more efficient than $\hat{B}_w$. These results are given in Table 1 for a multivariate normal model.

	Means				S.D.			
Design	$\hat{B}_0$	$\hat{B}_W$	$\hat{B}_A$	$\hat{B}_{AW}$	$\hat{B}_0$	$\hat{B}_W$	$\hat{B}_A$	$\hat{B}_{AW}$
D_1	.721	.721	.721	.721	.041	.041	.041	.041
D_2	.721	.721	.721	.721	.041	.041	.041	.041
D_3	.725	.719	.722	.719	.041	.043	.041	.043
D_4	.735	.722	.725	.722	.041	.054	.041	.054
D_5	.737	.722	.724	.722	.042	.063	.042	.062
D_6	.746	.720	.721	.719	.041	.010	.044	.109
D_7	.702	.723	.719	.723	.039	.043	.039	.043
D_8	.677	.716	.711	.716	.036	.085	.036	.085
D_9	.673	.719	.710	.719	.035	.123	.037	.123

Table 1. *Biases and standard deviations of estimated regressions.*
Simulated population, $N = 7,027$.
Y_1 = log(expenditure on food)
Y_2 = log(total expenditure)
X = log(expenditure in housing)
Mean and covariance matrix from Family Expenditure Survey.
Population regression $Y_1 = 1.74 + 0.71Y_2$
Finite population regression $Y_1 = 1.63 + 0.72Y_2$

What happens if the regressions are not linear or the variances are hecteroscedastic? Pfeffermann and Holmes (1985), Holmes (1987) show that $\hat{B}_A$ is not robust to these changes, whereas $\hat{B}_w$ remains approximately unbiased. In Table 2 some results are shown for repeated samples from a real finite population, the data being the U. K. Family Expenditure Survey for 1977.

3.3. A Compromise Estimator

Nathan and Holt (1980) show that the OLS estimator $\hat{B}_0$ is biased in the conditional distributions given (X, s) while the adjusted estimator $\hat{B}_A$ is approximately unbiased provided that the model is true. The empirical results in Table 2 suggest that $\hat{B}_A$ is not robust to departures from the model assumption but that the p-weighted estimator does have robustness properties. Can we get the best of both worlds by using a p-weighted version of $\hat{B}_A$? Nathan and Holt propose such an estimator and this is the estimator $\hat{B}_{AW}$ in Tables 1 and 2.

From the simulations it appears that $\hat{B}_{AW}$ shares the robustness properties of $\hat{B}_0$. When the simulation results are plotted in bands according to the value of X then it appears that

Design	Means $\hat{B}_0$	$\hat{B}_W$	$\hat{B}_A$	$\hat{B}_{AW}$	S.D. $\hat{B}_0$	$\hat{B}_W$	$\hat{B}_A$	$\hat{B}_{AW}$
D_1	.714	.714	.713	.713	.047	.047	.047	.047
D_2	.714	.714	.713	.713	.048	.048	.047	.047
D_3	.701	.712	.694	.711	.051	.049	.050	.049
D_4	.693	.711	.668	.708	.056	.063	.055	.063
D_5	.669	.706	.645	.703	.058	.065	.056	.066
D_6	.656	.699	.608	.691	.063	.111	.063	.116
D_7	.660	.701	.698	.700	.042	.088	.044	.087
D_8	.677	.716	.711	.716	.036	.085	.036	.085
D_9	.658	.712	.701	.712	.040	.123	.043	.0123

Table 2. *Biases and standard deviations of estimated regressions.*
Real population, $N = 7,027$.
Details as above
Finite population regression $Y_1 = 1.74 + 0.71Y_2$

$\hat{B}_{AW}$ has better properties than $\hat{B}_0$ in the conditional distribution given (X, s). It really does seem to benefit from both approaches!

Faced with these empirical results which favour p-weighting how should somebody who believes in models proceed? Brewer (1979), Little (1983), both advocate estimators based on models which are then protected against model misspecification by choosing the sampling scheme and estimator to make the estimator approximately p-unbised. The estimator $\hat{B}_{AW}$ is chosen in this spirit. DuMouchel and Duncan (1983) have examined the issue of weighting for regression analysis in a wider context. They have considered cases where weighting should not be used, for example when certain models are strongly believed to be true, and have suggested that weighting might be used when B in (2.8) is the parameter of interest. In this latter case they advocate testing the difference between $\hat{B}_0$ and $\hat{B}_W$ and if there is no difference using $\hat{B}_0$. If a difference is found then they adovocate introducing extra variables into the model to explain the difference. In our context they widen the regression to include variables in the design set X. In their example this strategy works and conditional on the X variables an unweighted regression explains the data adequately.

4. A MODEL-BASED JUSTIFICATION FOR WEIGHTING

The proposals in the previous section for including probability weights into estimation are to some extent ad hoc. In Brewer's approach the design must be shown to be consistent with the model while in Little's approach the selection probabilities are stratified after selection to make the model consistent with the desing. DuMouchel and Duncan's final proposal is to condition on the X variable, thus extending the model beyond the marginal distribution of Y.

In this section we show that probability weights can feature naturally in model-based inference by following Rubin (1983) and conditioning on the vector $\pi = (\pi_1(x), \ldots, \pi_N(x))$ of inclusion probabilities rather than the whole design set X. The target for inference is still some property of the marginal distribution of Y such as a predictive inference about $\bar{Y}$, the finite population mean. Rubin showed that the vector π, the propensity score, is frequently an adequate summary of the prior information X in the sense that

$$p(s|X) = p(s|\pi), \tag{4.1}$$

and that this still enables $p(s|X)$, or $p(s|\pi)$, to be ignored for model-based inference. He then suggests using the joint distribution of (Y, π) rather than that of (Y, X) for constructing the model. Let the data be $d_s = (s, Y_s, \pi)$ then

$$\begin{aligned} f(d_s;\lambda) &= p(s|X)g(\pi;\phi)f_s(Y_s|\pi;\theta) \\ &= p(s|\pi)g(\pi;\phi)f_s(Y_s|\pi;\theta); \end{aligned} \tag{4.2}$$

and this has the same form as (3.3) and so $p(s|\pi)$ can be ignored for inference about θ, ϕ, or Y. Rubin argues further that frequently it will be simpler to construct $f(Y|\pi;\theta)$ than $f(Y|X;\theta)$.

Unfortunately in social surveys most designs are self-weighting which means that $\pi_i(x)$ is constant for all $i = 1, \ldots, N$. In this case π contains no useful information. However, by expanding π to $\pi^* = \binom{\pi}{L}$, where L is the set of higher level labels denoting the stratification and clustering in the design, and then conditioning on π^* leads to stratification models and multi-level models (components of variance) which are adequate for modelling Y.

When the weights in π are not all equal then we can distinguish two cases:

(i) stratification, with $\pi_h = \frac{n_h}{N_h}$ in stratum h, and not all π_h equal;
(ii) variable probability sampling with $\pi_i \neq \pi_j$, for some $j \neq i$.

With stratification a predictive inference about $N\bar{Y} = \Sigma_h N_h \bar{Y}_h$ leads naturally to weights involving N_h/n_h where n_h is the sample size in stratum h. But inferences about $\bar{Y}_h$ or S_h^2 do not require these weights, nor do inferences about linear combinations $\Sigma_h w_h \bar{Y}_h$, when W_h are known. These latter inferences are made when data from a survey in one area are used to predict the results for a different population in either time or space or both. For example surveys on the annoyance due to aircraft noise around London Heathrow have been used to predict possible annoyance at potential sites for a third London airport and the probability weights N_h/n_h for Heathrow have no role for such inferences.

We consider the case where the weights are a measure of size of a sampling unit. The prior data X may contain many variables and the resulting summary into π is at best very crude. However, for inferences about Y all that is required is π. Before sampling, the joint distribution of (Y, π) is

$$f(Y, \pi) = f(Y|\pi)g(\pi), \tag{4.3}$$

and after sampling on π the superpopulation distribution is modified to

$$f_s(Y, \pi) = f(Y|\pi)g_s(\pi), \tag{4.4}$$

where strong ignorability, with $0 < \pi_i \dot{<} 1$, imples that all units have a chance of inclusion the sample, so that $f(Y|\pi)$ can be estimated from the sample data for all Y. Now since unit i is selected with probability proportional to size $\pi_i, g_s(\pi_i)$ is the size biased distribution

$$g_s(\pi_i) = \frac{\pi_i g(\pi_i)}{\mu_\pi}, \tag{4.5}$$

where $\mu_\pi = \int \pi g(\pi) d\pi$. In a finite population

$$\mu_\pi = \frac{1}{N}\sum_{i=1}^{N} \pi_i = \frac{n}{N}, \tag{4.6}$$

for a fixed sample size design and then

$$f_s(Y_i, \pi_i) = f(Y_i|\pi_i)N\pi_i g(\pi_i)/n. \tag{4.7}$$

We assume that the sample data comprise independent observations from the size biased distribution (4.5) or (4.7).

Size-biased samples have been studied by many authors, for example, Cox (1969), Patil and Rao (1978). The moments of the sampled distribution are simply related to those of the original distribution, for example,

$$\begin{aligned} E_s\left(\frac{\mu_\pi}{\pi}Y^r\right) &= \int Y^r \frac{\mu_\pi}{\pi} f(Y|\pi) g_s(\pi) dY d\pi \\ &= \int Y^r f(Y|\pi) g(\pi) dY d\pi \\ &= E(Y^r). \end{aligned} \quad (4.8)$$

Now since the sample data can be considered as a random sample from $f_s(Y, \pi)$

$$m_s(r) = \frac{1}{n}\sum_{i\in s} \frac{\mu_\pi}{\pi_i} Y_i^r = \frac{1}{N}\sum_{i\in s} \frac{Y_i^r}{\pi_i} \quad (4.9)$$

is an unbiased estimator of $E(Y^r)$. In particular

$$m_s(1) = \frac{1}{N}\sum_{i\in s}\frac{Y_i}{\pi_i}$$

is an unbiased estimator of $\mu_y = E(Y)$. This is the well known Horvitz-Thompson estimator given by (2.6).

For more complex functions of moments such as ratios or regression coefficients component-wise unbiased estimation leads to probability weighted estimators similar to (2.9). Thus conditioning on π and using results from size-biased sampling leads to distribution free methods of moments estimators identical to the classical p-weighted estimators. Clearly if the distributions in (4.3) can be specified accurately then more efficient methods estimation can be employed. In sample surveys the populations are very complex and highly multivariate and can rarely be specified accurately. In such cases a robust estimation procedure is highly desirable and the method of moments estimator leading to the p-weighted estimator for size-biased sampling must be a serious contender.

5. THE ADJUSTED LEAST SQUARES ESTIMATOR

In Section 3.2 the ALS estimator was introduced. From the form of the estimator (3.4) it can be seen that it adjusts the unweighted estimator m_y or s_{yy} for lack of balance in the sample on the prior variables X. Thus $\hat{\mu}_y$ is adjusted for the difference between M_x and m_x and $\hat{\Sigma}_{yy}$ for the difference between S_{xx} and s_{xx}. These adjustments are exact if the regressions are linear and homoscedastic, but as we saw in the simulation study in Table 2 the results do not appear to be robust to departures from these assumptions. How can the model-based estimators be adjusted to take into account lack of balance in the sample on the known auxiliary variables X?

If the sample selection probabilities $\pi(x)$ are related to the size of a particular variable X_1, say, then the sample points will mainly occur for large values of X_1. In Figure 1 we show a non-linear regression between Y and X_1 and the OLS regression and p-weighted regression fitted to a pps sample. The OLS regression line fits the data points and gives a good approximation to the true regression curve $E(Y|X_1)$ for large values of X_1. The ALS curve gives large weight to the points with small values of X_1 and gives a regression line which approximates the entire curve of $E(Y|X_1)$. Clearly it is the latter regression which is required if $E(Y|X_1)$ is to be approximated by a linear regression in the sense of Mouchart and Simar (1980). This approximate population regression can then be used for predictive inferences about unobserved Y_i.

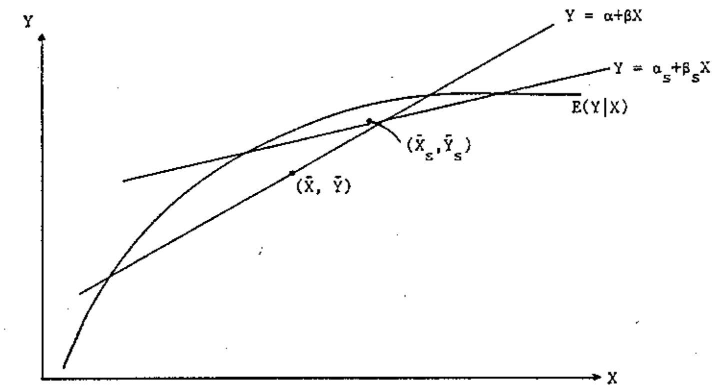

Figure 1. *Population and sample regressions*

$Y = \alpha + \beta X$ is the population linear regression.
$Y = \alpha_s + \beta_s$ is the sample linear regression.
$Y = E(Y|X)$ is the true (non-linear) population regression.

The relevant population model for regression adjustment is the joint distribution of Y and X_1, given by

$$f(Y, X_1) = f(Y|X_1)g(X_1). \quad (5.1)$$

The sampling scheme $p(s|X)$ is based in principle on the complete set of prior variables X, but if in fact the size measure component is a function only of X_1 then after sampling we have

$$f_s(Y, X_1) = f(Y|X_1)\pi(X_1)g(X_1)/\mu_\pi^*, \quad (5.2)$$

where $\mu_\pi^* = \int \pi(x_1)g(X_1)dx_1$. The linear ALS estimator of μ_Y is then

$$\hat{\mu}_Y = m_y + b_{yx_1}^*(M_{x_1} - m_{x_1}), \quad (5.3)$$

where $b_{yx_1}^* = \hat{\Sigma}_{yx_1}\hat{\Sigma}_{x_1x_1}^{-1}$.

Now the components in $\hat{\Sigma}_{yx_1}$ are the component-wise estimators of Σ_{yx_1} and using the results for size-biased sampling

$$\begin{aligned} E_s\left(YX_1\frac{\mu_\pi^*}{\pi(X_1)}\right) &= \int yx_1 \frac{\mu_\pi^*}{\pi(X_1)}\frac{\pi(X_1)}{\mu_\pi^*} f(y|x_1)g(x_1)dydx, \\ &= E(YX_1). \end{aligned}$$

Thus as before $\frac{1}{N}\Sigma_{i \in s} \frac{y_i x_{1i}}{\pi_i}$ is an unbiased estimator of $E(YX_1)$, with similar expressions for the other components.

These results suggest that the adjusted least squares estimator is not the compromise estimator $\hat{B}_{AW}$ proposed by Nathan and Holt (1980) but the modified version given by (5.3) in which only the slope is subject to p-weighting. The properties of $\hat{\mu}_Y$ in (5.3) are currently under investigation.

The overall conclusion is to agree with Rubin (1983) that the selection probabilities, π, can play a useful role in a model-based approach to finite population inference and moreover if a robust approach to inference is employed then the p-weighted estimators which are so fundamental in randomization inference appear as natural moment estimators using the ideas of size-biased sampling.

REFERENCES

Brewer, K. R. W. (1979). A class of robust sampling designs for large-scale surveys. *J. Amer. Statist. Assoc.* **74**, 911–914.

Cox, D. R. (1969). Some sampling problems in technology. *New Developments in Survey Sampling.* (N. L. Johnson and H. Smith Jr., eds.). New York: Wiley.

DuMouchel, W. H. and Duncan, G. J. (1983). Using sample weights in multiple regression analysis of stratified samples. *J. Amer. Statist. Assoc.* **78**, 535–543.

Hajek, J. (1973). Discussion of Basu, D.: "An essay on the logical foundations of survey sampling", Part I. *Foundations of Statistical Inference.* Holt, Rinehart and Winston of Canada Ltd.

Holmes D. J. (1987). *Ph. D. Thesis.* Southampton, U.K.: University of Southampton.

Holt, D. and Smith, T. M. F. (1976). The design of surveys for planning purposes. *The Australian J. of Statistics.* **18**, 37–44.

Holt, D., Smith, T. M. F. and Winter, P. D. (1980). Regression analysis of data from complex surveys. *J. Roy. Statist. Soc. A* **143**, 474–87.

Kish, L. and Frankel, M. R. (1974). Inference from complex samples (with Discussion). *J. Roy. Statist. Soc. B* **36**, 1–17.

Mouchart, M. and Simar, L. (1980). Least squares approximation in Bayesian analysis. *Bayesian Statistics,* Proceedings of the First Int'l. Meeting in Valencia, University Press, Valencia.

Little, R. J. A. (1982). Models for non-response in sample surveys. *J. Amer. Statist. Assoc.* **77**, 237–250.

Little, R. J. A. (1983). Estimating a finite population mean from unequal probability samples. *J. Amer. Statist. Assoc.* **78**, 596–604.

Nathan, G. and Holt, D. (1978). The effect of survey design on regression analysis. *J. Roy. Statist. Soc. B* **42**, 377–386.

Patil, G. P. and Rao, C. R. (1978). Weighted distributions and size biased sampling with applications, etc. *Biometrics* **34**, 179–190.

Pfeffermann, D. and Holmes, D. J. (1985). Robustness considerations in the choice of a method of inference for regression analysis of survey data. *J. Roy. Statist. Soc. A* **148**, 268–278.

Rao, J. N. K. (1975). Analytic studies of sample survey data. *Survey Methodology* **1**, supplementary issue, Statistics Canada.

Rosenbaum, P. R. and Rubin, D. R. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Rubin, D. B. (1985). The use of propensity scores in applied Bayesian inference. *Bayesian Statistics* **2**. (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds). Amsterdam: North-Holland.

Scott, A. J. (1977). On the problem of randomization in survey sampling. *Sankhya C* **39**, 1–9.

Smith, T. M. F. (1981). Regression analysis for complex surveys. *Current Topics in Survey Sampling.* Academic Press, 267–92.

Smith, T. M. F. (1983). On the validity of inferences from non-random samples. *J. Roy. Statist. Soc. A* **146**, 394–403.

Sugden, R. and Smith, T. M. F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika* **71**, 495–506.

DISCUSSION

M. J. BAYARRI (*University of Valencia*)

For a long time, it has been commonly argued that the inclusion probabilities, π, had no role to play in the Bayesian approach to sample surveys. After the last two Valencia meetings the situation seems to be changing. As a matter of fact, Professor Rubin in Valencia 2 (Rubin, 1985) showed how the π's can be useful as a coarse summary of the information provided by the covariates, easing the task of modelling as well. Now, in Valencia 3, Prof. T. M. F. Smith uses Rubin's proposal for modelling and carries the argument one step further: He shows how the largely condemned (by Bayesian audiences) classical weighted estimators can also arise from a model-based approach to sample surveys. Those applied Bayesian statisticians who quietly use randomization estimators in their applications owe a debt of gratitude to both of them.

The question raised in the title of the paper, however, is not whether to use or not to use weighted estimators, but whether to weight or not to weight. This question got me interested in whether the π's could be not just useful or justifiable but even interesting. It might very well turn out that Bayesians would ask for the units to be selected with probability proportional to size, for instance, if that selection provided greater information than simple random sampling.

The following discussion is restricted to the "weighted" part of the model, as presented by Professor Smith, that is, to the size-biased version of $g(\pi)$,

$$g^b(\pi) = \frac{\pi g(\pi)}{\mu_\pi}. \tag{1}$$

This size-biased distribution is just a particular case of what Rao (1965) called *weighted distributions*, in which the original density is multiplied by some general weight function and renormalized. Professor DeGroot and myself are currently working on this topic and have already obtained some preliminary results showing that, in some situations, the experiment that selects a random sample from the weighted distribution is *sufficient*, in the Blacwkell sense (Blackwell, 1951,1953), for the experiment selecting a random sample from the original or unweighted distribution. Then, in these situations, given the choice, a Bayesian would always select the "weighted" experiment because for every decision problem involving the parameter indexing the distribution, and every prior distribution for it, the expected Bayes risk would be smaller with the weighted experiment than with the unweighted one.

Size-biased distributions being particular cases of weighted distributions, it is natural to ask what would be the case in this scenario. Needless to say, different answers will be obtained depending on the model $g(\pi)$ we have in mind. In what follows, we will study Fisher information for different models $g(\pi)$ and their size-biased versions $g^b(\pi)$. We will denote by $\mathcal{E}_{\text{original}}$ and $\mathcal{E}_{\text{size-biased}}$ the experiments in which a random sample is obtained from the original density $g(\pi)$ and its size-biased version $g^b(\pi)$, respectively. Also, $\mathcal{E}_1 \succ_F \mathcal{E}_2$ will mean that $\mathcal{E}_1$ provides greater Fisher information than $\mathcal{E}_2$ for every value of the parameter considered ($\approx_F$ will mean equal Fisher information). $I(\cdot)$ and $I_b(\cdot)$ will denote Fisher information in one observation from $g(\pi)$ and $g^b(\pi)$ respectively.

One difficulty with both Rubin's paper and Smith's paper is that they provide no hints about what a sensible model $g(\pi)$ could be, but π being a probability, the natural guess would be a beta distribution. Also, we don't expect π to be too big, particularly if N (the size of the finite population) is large, so that, as a first simple model we will consider

$$g_1(\pi|\theta) = Be(\theta, \theta + k) \tag{2}$$

where k is a constant (presumably related to N). It is easily found that the size-biased version of (2) is

$$g_1^b(\pi|\theta) = Be(\theta+1, \theta+k) \tag{3}$$

and also that one observation from (3) provides greater Fisher information than one observation from (2) so that, in this case,

$$\mathcal{E}_{\text{size-biased}} \succ_F \mathcal{E}_{\text{original}}; \tag{4}$$

that is, it would be convenient for us to select the π's with probability proportional to size.

When thinking about selection probabilities π, we somehow feel that the value $1/N$ has some special meaning that should be reflected in $g(\pi)$. The model we will consider next is a mixture of Pareto related distributions and has the advantage over (2) of being more spiked around $1/N$ and of being far more easy to handle from a Bayesian point of view. Thus, let's consider now

$$\begin{aligned} g_2(\pi|\theta) &= \theta\pi^{\theta-1}N^{\theta-1} && \text{for } 0 \le \pi \le \frac{1}{N} \\ &= \theta(1-\pi)^{\theta-1}\left(\frac{N}{N-1}\right)^{\theta-1} && \text{for } \frac{1}{N} \le \pi \le 1 \end{aligned} \tag{5}$$

which is a density for $\theta > 0$, but values of $\theta \ge 1$ seem more sensible in this context (when $\theta < 1$, (5) is U shaped). Figure 1 shows the shape adopted by $g_2(\pi|\theta)$ for selected values of θ. Notice that the mode of this distribution (for $\theta > 1$) is precisely $1/N$ and that greater values of θ correspond to distributions which are more and more spiked around their modes. More general mixtures of this type of Pareto related distribution were studied in Bayarri (1984), and a particular mixture, which is a two parameter generalization of (5), was used in Bayarri (1985) in a Bayesian goodness-of-fit context; there it was called the alpha distribution, a name due to Bernardo (1982) who apparently first introduced it.

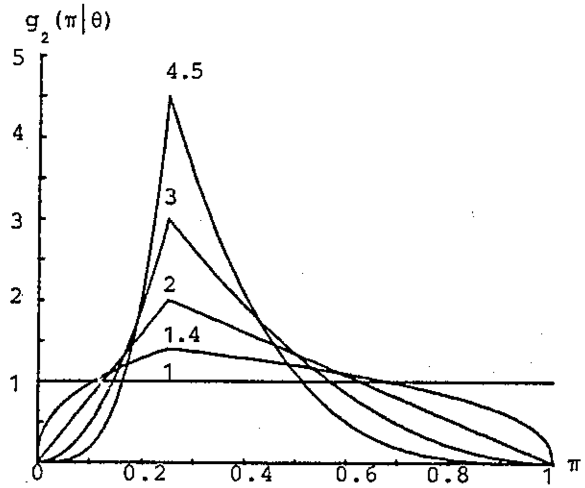

Figure 1. *The density $g_2(\pi|\theta)$ for $N = 4$ and $\theta = 1, 1.4, 2, 3, 4.5$*

The size-biased version of (5) is found to be

$$\begin{aligned} g_2^b &= \frac{1+\theta}{N+\theta-1}(N\pi)^{\theta} && \text{for } 0 \le \pi \le \frac{1}{N} \\ &= \frac{1+\theta}{N+\theta-1}N^{\theta}\pi\left(\frac{1-\pi}{N-1}\right)^{\theta-1} && \text{for } \frac{1}{N} \le \pi \le 1. \end{aligned} \tag{6}$$

It can be shown that, in this case, $I(\theta) \ge I_b(\theta)$ for all values of θ, so that

$$\mathcal{E}_{\text{original}} \succ_F \mathcal{E}_{\text{size-biased}}$$

and the situation is just the opposite to the one encountered before.

In the two examples just presented, we have selected some distributions $g(\pi)$ to explain the behavior of π and assume that the data is a random sample from their size biased versions $g^b(\pi)$. But really we are not very used to thinking about models for the probabilities of selection π, so that we will deduce the last model to be studied directly from the distribution of the covariates X.

Assume X is a univariate positive random variable with density $f_X(x)$. As usual, we consider $X_1, \ldots, X_N$, the finite population, to be a random sample from f_X. We are assuming that the data we have is a sample from $X_1, \ldots, X_N$ selected with probability proportional to size. Thus, associated with $X_1, \ldots, X_N$ there is the corresponding finite population of π's : $\pi_1, \ldots, \pi_N$, where $\pi_i = X_i/(\sum_{i=1,\ldots,N} X_i)$ and we select X_i with probability π_i (notice that there is an slight variation here with respect to the paper: these π_i's add to one, while the ones in the paper add to n).

If we want to model the behavior of π instead of the behavior of X, then we assume that a sample is going to be drawn from $\pi_1, \ldots, \pi_N$ with probability proportional to size, that is, π_i is selected with probability π_i. In this process, the distribution of data, $g_s(\pi)$ in the paper, is given by:

$$g_s(\pi) = N\pi \int t f_X(\pi t) g_Y[(1-\pi)t]dt, \tag{7}$$

where Y represents the sum of $N-1$ i.i.d. random variables from f_X, and g_Y its density.

Let's take an example. Again, for a positive random variable it would be natural to try a gamma distribution, that is, $f_X(x) = Ga(\alpha, \beta)$. Then it is found that

$$g_s(\pi) = Be\{\alpha+1, (N-1)\alpha\}. \tag{8}$$

One interesting fact about (8) is that this beta distribution is just the size-biased version of the $Be\{\alpha, (N-1)\alpha\}$ distribution. So that in general, let's assume that

$$g_3(\pi) = Be\{\alpha, k\alpha\}, \tag{9}$$

where k is a constant. Notice that, if $k = N-1$ as in the gamma example, $E(\pi) = 1/N$ so that $1/N$ is again regarded as a special value in the distribution of π. As we have already said, $g_3^b(\pi) = Be(\alpha+1, \alpha k)$ and in this case it is found that $I(\alpha) = I_b(\alpha)$ for all α, so that

$$\mathcal{E}_{\text{size-biased}} \approx_F \mathcal{E}_{\text{original}}$$

and both experiments are totally equivalent with regard to this criterion.

We have thus encountered three situations in which the behaviour of Fisher information is different. Of course, all the examples refer solely to the X part of the model, without referring to the possible effects of weighting in the marginal distribution of Y. However, in our opinion the general conclusion to be drawn is that, even if Professor Smith has shown us how the π's can enter the picture, there is not yet a clear answer to the question posed.

Indeed, it is not surprising for a Bayesian to conclude that whether to weight or not to weight will depend on the particular decision problem at hand.

I wouldn't like to finish without asking Professor Smith a couple of questions: We have just seen the weighted estimators appearing as a result of both modelling π (instead of X) and using the method of moments for estimation. Rubin (1985) already cautioned us that π can be "too coarse a summary" of the information provided by X, so that my first question refers to whether this type of modelling is the only way to make the π's play a role in a model-based approach to sample surveys. Also, it is well known that the method of moments can exhibit a number of undesirable features; Has Professor Smith tried a Bayesian or at least a likelihood approach to estimation?. I am looking forward to seeing some results in this direction, but in turn, it would imply selecting suitable $f(Y|\pi)$ and $g(\pi)$. How would Professor Smith select these models?.

The second question relates to robustness: Has Professor Smith studied the issue of whether modelling $f(Y|\pi)$ and $g(\pi)$ produces more robust inferential results than modelling $f(Y|X)$ and $g(X)$?. If so, it could be another reason for using the former alternative and for weighting, thus helping to answer the question raised in the title of the paper.

R. A. SUGDEN (*Goldsmiths' College, London*)

Smith gives some answers and suggests new approaches to the vexed question for a Bayesian: What is the role of the design in survey sampling inference?

Contrary to the impression possibly given in Section 3.1., it is important to realise that inferences can depend on the design even in the ignorable case. For example under a normal error regression through the origin on a single "size" variable with error variance proportional to squared size and a probability proportional to size design, it is easy to show, through the likelihood (4.1), that the Bayes posterior mean of the population total is just the Horvitz-Thompson design-unbiased estimator (4.9) but with an additional term representing a sum of "residuals".

As shown in Sugden and Smith (1984), the design is no longer ignorable when not all the inclusion probabilities are observed. However, some aspects of ignorable inference may be preserved e.g. in the above the posterior mean depends only on the inclusion probabilities of sample units so is unaltered.

All statements about ignorability (or not) have been made by authors assuming the model is correct. A Bayesian who lacks confidence in his model must seek model elaboration —see Royall and Pfeffermann (1982)— or adopt a distribution free approach such as the method of moments that Smith suggests. In the former case ignorability may no longer hold unless sufficient design information is available. An alternative to the latter is some form of non-parametric maximum likelihood estimation of the finite population distribution function, see Vardi (1982). A problem with the method of moments here is that it essentially amounts to imposing (component-wise) design unbiasedness.

REPLY TO THE DISCUSSION

Dr. Bayarri makes many interesting points which take the discussion far beyond my modest aims. I was concerned only with the problem of inference *after* a sample has been selected using a randomized design with unequal selection probabilities. The dilemma for a Bayesian is that if the design variable $\boldsymbol{X}$ is known for all units in the population then any design of the form $p(s|\boldsymbol{x})$ contains less information than $\boldsymbol{X}$ itself and so can be ignored for inference. But sample designs are constructed by knowledgeable statisticians so surely they must contain useful information and as such should not be ignored. The resolution of the dilemma is found by constructing an appropriate conditional inference. Rubin's contribution is to show that frequently there will exist a reduction of the design (prior) information $\boldsymbol{X}$ which is an

adequate summary of $\boldsymbol{X}$ for inference on Y. He shows further that under certain conditions the vector π of inclusion probabilities will provide such an adequate summary of $\boldsymbol{X}$. Inferences can then be made conditional on π and as such they will depend on π. Dr. Sugden makes this point in his discussion but his phrasing is misleading since the inference does not depend on the sampling mechanism $p(s|\boldsymbol{x})$ but only on the units in the sample and their inclusion probabilities π.

My contribution was to consider how the information in π might be used for inference about $\boldsymbol{Y}$. The complexity of most survey populations means that precise models are difficult to specify and even harder to justify and so I did not attempt to model $g(\pi)$ along the lines suggested by Dr. Bayarri. Instead I adopted one form of model-free estimation, namely the methods of moments estimators, because it gave the traditional π-weighted estimators.

As both Dr. Sugden and Dr. Bayarri point out other methods of estimation could have been considered. These would lead to different estimators and to comparisons of efficiency. Which one to choose will depend on the strenght of one's prior belief about the underlying data structure. In some further sets of simulation results based on real data we have found cases where the π-weighted estimator is inefficient unconditionally (over all samples) and is appalling conditionally (given the sample). Thus as Dr. Bayarri concludes from her models there are some cases when π-weights are good and some when they are not. There is still no simple answer to the question of whether to use π-weights or not. So the Bayesian who says that using π-weights is always wrong is wrong and the traditional statistician who says that π-weights should always be used is equally wrong.

In the absence of precise models we still need a robust procedure. My own belief is that stratification after selection on $\boldsymbol{X}$(or π) is the best general purpose robust procedure for survey inference. This employs the π-weights indirectly through the stratification rather than directly through the weighting.

REFERENCES IN THE DISCUSSION

Bayarri, M. J. (1985). A Bayesian test for goodness-of-fit. *Tech. Rep.* Departamento de Estadística e I.O. University of Valencia. Presented at the 1985 Joint Statistical Meetings (ASA, Biometrics, IMS).

Bayarri, M. J. (1984). *Contraste Bayesiano de Modelos Probabilísticos*. Ph. D. Thesis. University of Valencia.

Bernardo, J. M. (1982). Contraste de modelos probabilísticos desde una perspectiva Bayesiana. *Trabajos de Estadística* **32**, 16–30.

Blackwell, D. (1951). Comparison of experiments. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 93–102. Berkeley, CA: University of California Press.

Blackwell, D. (1953). Equivalent comparison of experiments. *Ann. Math. Statist.* **24**, 265–272.

Rao, C. R. (1965). On discrete distributions arising out of methods of ascertainment. *Classical and Contagious Discrete Distributions*, (G. P. Patil, ed.), 320–333. Calcutta: Statistical Publishing Society.

Royall, R. M. and Pfeffermann, D. (1982). Balanced samples and robust Bayesian inference in finite population sampling. *Biometrika* **69**, 401–410.

Vardi, Y. (1982). Nonparametric estimation in the presence of length bias. *Ann. Statist.* **10**, 616–620.