STATISTICA, anno LXXIII, n. 3, 2013

A DESIGN-BASED APPROXIMATION TO THE BAYES INFORMATION CRITERION IN FINITE POPULATION SAMPLING

Enrico Fabrizi
Dipartimento di Scienze Economiche e Sociali, Università Cattolica del S. Cuore, Piacenza, Italia
Parthasarathi Lahiri
Joint Program in Survey Methodology, University of Maryland, College Park, MD, United States.

1. INTRODUCTION

Survey researchers frequently use statistical models. However, such models, known as superpopulation models (Deming and Stephen, 1941), are generally used to describe finite populations of interest and have been used earlier for evaluation, sampling design development and making inferences on either the relevant superpopulation or the finite population parameters. In an analytic use of survey data (Deming, 1953) where the main goal is to address various scientific questions, inferences for the superpopulation parameters are more important than those for the finite population parameters. In an excellent review article, Graubard and Korn (2002) discussed the importance of inferences for superpopulation parameters using survey data and cited a number of practical examples such as the estimation of superpopulation means, linear regression and logistic regression coefficients using complex survey data from the U.S. National Health Interview Survey, the third National Health and Nutrition Examination Survey and the 1986 National Hospital Discharge Survey.

Model selection among different plausible models has received considerable attention in statistical literature. The Institute of Mathematical Statistics (IMS) monograph on model selection edited by Lahiri (2001) contains four long review articles that critically examine various classical and Bayesian approaches to model selection. For further important developments in the Bayesian literature on the subject see Spiegelhalter et al. (2002). The impact of the superpopulation model misspecification has been studied in the literature. See Holt et al. (1980), Hansen et al. (1983), and others. However, to our knowledge the related issue of model selection, especially the well known likelihood-based methods such as the Bayesian Information Criterion (BIC), has received little attention in survey research literature.

One important feature that distinguishes a superpopulation model selection for a finite population from that for a hypothetical infinite population is that we must derive the superpopulation model selection criterion from the knowledge of the observed sample. Finite populations studied in social and economic surveys are generally very complex and heterogeneous. Moreover, if we are interested in a regression model for a

response variable, possible covariates are limited to those measured in other questions of the same questionnaire or those available from administrative records that can be linked to survey responses on a statistical unit basis. Moreover, if prediction of finite population means or totals is one of the goals, as it is often the case, the choice is further restricted to those for which population means/totals are known. For this reason some survey researchers feel that 'most statistical models in finite population inference are either wrong or (at best) incomplete' (Kott, 1989).

The main goal of this paper is to find a model selection criterion that is capable of discriminating among models even though some features of the models being considered are misspecified. In particular, we discuss a simple approximation to the BIC for the analysis of complex survey data that avoids specification of the full superpopulation likelihood. If the full specification of the superpopulation likelihood for the finite population is possible and the sampling design is ignorable, there is conceptually no problem in deriving the likelihood for the complex sample and so an extension of the BIC to finite population sampling in such a situation is quite straightforward. However, this may not be always the case. The sample and the population likelihood functions may be different because of informative sampling (Pfeffermann, 2009) and their relationship may be complex. Moreover, the effect of informative sampling may be present when sampling design features (e.g., stratification, clustering and size variables) are not accurately known to the data analyst.

In Section 2, we review the Bayes factor and its relation to the BIC for hypothetical infinite population. In Section 3, we critically examine two possible ways to adapt the BIC in the context of the finite population sampling. The first approach consists in finding a formula for the BIC based on the superpopulation likelihood for the finite population and then estimating this finite population BIC. We argue that this model selection criterion does not even work for a simple hypothesis testing problem, a special case of model selection, with data collected by a simple random sampling with replacement. This approach makes the disagreement between the data and the null hypothesis look more than it really is. The second approach is the BIC based on the sample likelihood. This certainly provides us a meaningful model selection criterion. However, the basic requirement for this approach is the specification of a full superpopulation likelihood for the finite population. In Section 3, we discuss the impact of superpopulation model misspecification on the BIC based on the sample likelihood.

In Section 4, we propose a new model selection criterion that is essentially the Wald statistics based on a survey-weighted estimator of the superpopulation parameter of interest and its randomization-based variance estimator. Our model selection criterion is robust and can be used, for example, to test the significance of a regression coefficient with unspecified distribution for the error term using complex survey data. We show that under certain regularity conditions, the new model selection criterion is indeed an approximation to the BIC for a large sample. In Section 5, we verify the regularity conditions for two commonly used sampling designs. We provide results from a Monte Carlo simulation study in Section 6. Our simulation results demonstrate good performance of the new criterion in a complex situation involving clustered binary data with unknown intra-cluster correlation.

2. THE BAYES FACTOR AND THE *BIC*

The Bayesians frequently use the Bayes factor (BF) in hypothesis testing and model selection problems. To illustrate the BF, let $y_s = (y_1, ..., y_n)$ be an independent and identically distributed sample from a distribution belonging to a family of probability distributions parameterized by (β, θ) with $dim(\beta, \theta) = m$ and $dim(\beta) = m_0$. Consider the following hypothesis testing problem:

$$M_0 : \theta = \theta_0 \qquad \text{versus} \qquad M_a : \theta \in \mathbb{R}^{m-m_0}. \tag{1}$$

The BF is defined as the ratio of the aposteriori and the apriori odds in favour of the larger model M:

$$BF = \frac{prob\,(M|y_s)}{prob\,(M_0|y_s)} \Big/ \frac{prob\,(M)}{prob\,(M_0)} = \frac{\int p\,(y_s|\beta,\theta)\,\pi\,(\beta,\theta)\,d\beta d\theta}{\int p\,(y_s|\beta,\theta_0)\,\pi_0\,(\beta)\,d\beta}, \tag{2}$$

where $\pi(\beta, \theta)$ and $\pi_0(\beta)$ are the joint prior distribution of β and θ and marginal prior distribution of β, respectively. The calculation of the BF requires a full specification of the prior distributions for the parameters in both M_0 and M. In many applications rules for 'objectively' selecting priors have been proposed (see Berger and Pericchi, 2001). Alternatively, one can use a suitable approximation to the logarithm of the BF. One popular approximation is the Bayes Information Criterion (Schwartz, 1978) given by:

$$S = \lambda - \frac{m - m_0}{2} \log n,$$

where $\lambda = \ell(\hat{\beta}, \hat{\theta}) - \ell_0(\hat{\beta}_0, \theta_0)$ is the logarithm of the likelihood ratio; $\ell(\hat{\beta}, \hat{\theta})$ is the log-likelihood evaluated at the maximum likelihood estimator $(\hat{\beta}, \hat{\theta})$ of (β, θ) under model M; $\ell(\hat{\beta}_0, \theta_0)$ is the log-likelihood evaluated at the maximum likelihood estimator $\hat{\beta}_0$ of β under model M_0.

The statistic S is based on the Laplace approximation to the integrals appearing in the numerator and denominator of Eq. (2). See Kass and Wassermann (1995) for details. The quality of the approximation S to the logarithm of the Bayes Factor depends on the prior distributions of the unknown parameters under M_0 and M. In general, it is rather crude since it neglects terms up to a constant order. Nonetheless, Kass and Wassermann (1995) showed that for a suitable choice of the prior distributions (e.g., unit information prior)

$$S = \log BF + O_p(n^{-1/2}).$$

Moreover, the Bayes Information Criterion is popular among frequentists because it incorporates a penalized deviance criterion.

In a hypothesis testing problem, S can be compared against the scale of evidence introduced by Jeffreys (1961) as an alternative to the frequentist scale of evidence introduced by R.A. Fisher in the 1920s. For a discussion and the comparison between the Jeffreys' and Fisher's scales of evidence, see Efron and Gous (2001). It should be stressed that S is a consistent model selection method, i.e., if one of the hypotheses (models) being tested is true, the BIC selects the true hypothesis with probability 1 as the sample size tends to infinity. For the problem (1), S goes to $+\infty$ ($-\infty$) with probability 1 if M (M_0) is true. This property is enjoyed by all penalized deviance criteria with penalty factors of the order $o(n)$.

3. Two possible approaches to adapt BIC to the finte population sampling

For this section and the rest of the paper, we need a few notations. Let $U = \{1,\ldots,N\}$ denote the units of a finite population of known size N. Let $y_U = (y_1,...,y_N)$, where y_i is the value of a characteristic of interest for the ith unit of the finite population ($i = 1,\ldots,N$). Let $p(s)$ be the probability of drawing a particular sample s from the universe of all possible samples $\mathbb{S}$. Thus, $p(s) \geq 0$ and $\sum_{s\in\mathbb{S}} p(s) = 1$. Let $d_s = \{d_i : i \in s\}$, where d_i contains all possible design and other auxiliary information on the unit $i \in s$. For example, d_i may contain information on the label and sampling weight w_i for the unit $i \in s$. The sampling weight w_i is defined as the inverse of the inclusion probability for the unit i and represents a certain number of units in the finite population. Define $y_s = \{y_i : i \in s\}$ and $z_s = [d_s, y_s]$. In the following two subsections, we discuss two possible approaches to extend the BIC to select model for the superpopulation for y_U.

3.1. *An estimator of the finite population BIC*

Let the observations y_i ($i = 1,\ldots,N$) of the finite population be generated randomly from $N(\theta, 1)$. Consider the following hypothesis testing problem, a special case of model selection:

$$M_0 : \theta = 0 \qquad M_a : \theta \neq 0. \tag{3}$$

If all units of the finite population were observed, then it is easy to see that the BIC based on all the observations in the finite population is given by

$$S_{POP}(y_U) = \frac{N}{2}\bar{y}_U^2 - \frac{1}{2}\log N. \tag{4}$$

We call $S_{POP}(y_U)$ the finite population BIC. Of course, we cannot use $S_{POP}(y_U)$ since $\bar{y}_U$ is unknown. Let $\hat{\bar{y}}_U$ be a design-consistent estimator of $\bar{y}_U$. An estimator is design-consistent estimator for the corresponding finite population parameter if it converges to the true finite population parameter as $n \to \infty$, where the convergence is with respect to probability induced by the sample design. We observe that, since $n \leq N$, the limit $n \to \infty$ makes sense only in a setting in which the population size N is also allowed to increase. We assume a mathematical definition of the limit for $n \to \infty$ that is consistent with most literature on inference in finite population sampling. A description of this framework may be found in Isaki and Fuller (1982). Replacing $\bar{y}_U$ by a design-consistent estimator $\hat{\bar{y}}_U$, the following naïve model selection criterion is obtained:

$$S_{Plugin}(z_s) = \frac{N}{2}\hat{\bar{y}}_U^2 - \frac{1}{2}\log N. \tag{5}$$

We note that the simple plug-in approach as described above does not work even for a simple random sampling with replacement. Under this sampling design, when N is very large compared to n (the sample size), one would expect a reasonable finite population sampling implementation of S to be very close to the following standard BIC S_{IID} obtained under the assumption of independently and identically distributed

observations from a normal population:

$$S_{IID}(y_s) = \frac{n\bar{y}_s^2}{2} - \frac{1}{2}\log n. \tag{6}$$

This is a reasonable expectation since in this case simple random sampling from a finite population can be regarded as a random sample from the assumed hypothetical superpopulation. But, if we replace $\bar{y}_U$ in (5) by the usual design-consistent estimator $\bar{y}_s$, we obtain:

$$S_{Plugin}(z_s) - S_{IID}(y_s) = \frac{(N-n)}{2}\bar{y}_s^2 - \frac{1}{2}\ln\left(\frac{N}{n}\right). \tag{7}$$

This difference tends to 0 when $n \to N$ but, for N fixed, it diverges to infinity as $N \to \infty$ and not to 0 as we would like. This implies that for N large enough, Eq. (5) provides stronger evidence against M_0 than Eq. (6). The reason is that Eq. (5) approximates S we would have obtained if all the units in the finite population were observed and thereby making the disagreement between the data and the null hypothesis look more than it really is.

3.2. *The BIC based on the exact likelihood for the sample*

Like in the standard BIC calculation for a hypothetical infinite population, this approach is also based on the sample likelihood. However, we must obtain the sample likelihood using the superpopulation model for the finite population and the sampling design used. Survey populations usually have complex structures and misspecification of the assumed model is quite likely (see Kott, 1991). This is a serious issue in large scale sample survey. Since the BIC is based on the sample likelihood it may be subject to model misspecification. We now illustrate this point through a simple example.

Let the observations in the finite population be normally distributed with common mean θ. We assume that the observations within the same cluster are equally correlated, the common intra-cluster correlation being τ. Furthermore, observations from two different clusters are assumed to be uncorrelated. We consider the same testing problem on the overall population mean θ as in Eq.(3).

For the finite population described in the previous paragraph, a cluster sampling is often employed. Suppose we have a finite population of size N divided into M clusters each of size N_c. A sample of m clusters is selected by simple random sampling (with replacement) and all the units of the sampled cluster are selected. Thus, $n = mN_c$. In this case a suitable model for y_s is given by

$$y_{ij} = \theta + \alpha_j + e_{ij},$$

where α_j and e_{ij}'s are all uncorrelated with $V(\alpha_j) = \tau$ and $V(e_{ij}) = 1-\tau$ for $j = 1,\ldots,m$, $i = 1,\ldots,N_c$. Note that marginally $V(y_i) = 1$ so we are consistent with the model assumed in the *iid* case. This leads to $\bar{y}_s$ as the maximum likelihood estimator of θ and to the following BIC:

$$S(z_s) = \frac{1}{2}\frac{n\bar{y}_s^2}{\{1+(N_c-1)\tau\}} - \frac{1}{2}\log n. \tag{8}$$

We note that

$$S_{IID}(y_s) - S(z_s) = \frac{n\bar{y}_s^2}{2}\left\{\frac{(N_c - 1)\tau}{1 + (N_c - 1)\tau}\right\},$$

where $S_{IID}(y_s)$, given in Eq. (6), is the appropriate BIC when there is no clustering of the population units. The error increases with $\tau > 0$. In other words, if we neglect the clustering of the population units and $\tau > 0$, we reject the null hypothesis more often than we really should.

Unfortunately, unlike the previous example the likelihood for the sample may be very complicated. To this end, reconsider the same hypothesis testing problem of Eq. (3) based on a probability proportional to size with replacement sampling in which the size variable X is positively correlated with the target variable Y. One can consider a model for $f(y_s|d_s = x_s)$. However, we are interested in testing a hypothesis for the superpopulation mean θ that characterizes the marginal distribution of Y and not the mean conditional on X. Since the sampling design is not simple random sampling and larger values of Y are more likely to be observed, we need to obtain a marginal likelihood for y_s by integrating out x_s:

$$f(y_s) = \int f(y_s|x_s) f(x_s)\, dx_s.$$

This is certainly not as simple as the previous example. Actually, when analyzing data from complex surveys we observe a sample from $Y|d_s$. A researcher may not be interested in $f(y|d_s)$ but may be interested in an appropriate marginal model – one that averages out some of the population features incorporated in the sampling design. For instance, one may be interested in testing hypothesis about the overall population mean, ignoring possible differences among the means of different subgroups of the population. In general, some degree of aggregation in modelling may be necessary in complex surveys from a finite population (see Holt , 1989).

In any case, the calculation of the BIC based on the sample likelihood requires that we use all information needed to specify a suitable model for $f(y_s|d_s)$. This may not be the case in many applications. It is typical that the analyst may not be provided with all the information about the sample design but only with the sampling weights, defined as the inverse of the inclusion probabilities and adjusted for post-stratification and non-response.

4. A ROBUST DESIGN-BASED APPROXIMATION TO THE BIC

Let y_U be a realization from an underlying superpopulation distribution characterized by a parameter θ. We are interested in testing $M_0 : \theta = \theta_0$ vs $M_a : \theta \neq \theta_0$. In this case the BIC is given by $S = \lambda - \frac{1}{2}\log n$, where $\lambda = \ell(\hat{\theta}) - \ell(\theta_0)$ is the logarithm of the likelihood ratio.

As noted in the previous section, it is often difficult or even impossible to obtain an exact expression for the sample likelihood due to a complex superpopulation model as explained in section 3.2 and informative sampling. In this section, we consider a design-based approximation to S that essentially involves an estimator of θ using the following method and its design-consistent variance estimator.

Let $f(y_U,\theta)=0$ be an estimating equation for θ. The solution $T(y_U)$ of the equation $f(y_U,T(y_U))=0$ is known as the corresponding descriptive population quantity (CDPQ) of θ. We can estimate $T(y_U)$ by a design-based estimator $\hat{T}(z_s)$. For example, $\hat{T}(z_s)$ could be obtained using the pseudo-maximum likelihood approach. A thorough discussion of this class of methods can be found in Särndal et al. (1992).

We propose the following model selection criterion:

$$S_{DB}=\frac{1}{2}W_{DB}-\frac{1}{2}\log n, \tag{9}$$

where $W_{DB}=\left\{\hat{V}_D\left(\hat{T}(z_s)\right)\right\}^{-1}\left(\hat{T}(z_s)-\theta_0\right)^2$ and $\hat{V}_D\left(\hat{T}(z_s)\right)$ is a consistent estimator of $V_D\left(\hat{T}(z_s)\right)$, the variance of $\hat{T}(z_s)$ under the randomization distribution. It may be noted that the use of the total sample size n in Eq.(9) may be misleading in some cases. The 'effective' sample size is often a more relevant measure of sample information. Actually, we can obtain a different model selection criterion if we replace n by $n^*=n/\text{Deff}$ in Eq. (9), where

$$\text{Deff}=\frac{V_D\left(\hat{T}(z_s)\right)}{V_{SRS}\left(\hat{T}(y_s)\right)}$$

and $V_{SRS}\left(\hat{T}(y_s)\right)$ is the randomization variance of the un-weighted estimator $\hat{T}(y_s)$ of $T(y_U)$ under a simple random sampling of size n. However, we note that the order of log(Deff) is often small compared to $\ell(\hat{\theta})-\ell(\theta_0)$. Thus, asymptotically $S_{DB}(n^*)\cong S_{DB}(n)$ in most cases since $\log(n^*)=\log(n)-\log(\text{Deff})$.

The following theorem shows that S_{DB} approximates S well, the error of approximation is of lower order than the error of approximating $\log BF$ by S.

THEOREM 1. *Under the following regularity conditions:*
i) $\hat{\theta}-\theta_0=O_\xi\left(n^{-1/2}\right)$, *under model* M_a, *where* $O_\xi\left(n^{-1/2}\right)$ *denotes a stochastic order with respect to the superpopulation distribution* ξ;
ii) $\ell(\theta)$ *is twice differentiable with* $-\ell''(\hat{\theta})=I(\theta_0)+O_\xi\left(n^{-1/2}\right)$, *where* $I(\theta_0)=-E\left\{\frac{\partial^2 l(z,\theta)}{\partial\theta^2}\right\}_{\theta=\theta_0}$ *is the Fisher information matrix evaluated at* θ_0;
iii) $\hat{T}(z_s)=\hat{\theta}+o_{D\xi}\left(n^{-1/2}\right)$ *where* $o_{D\xi}\left(n^{-1/2}\right)$ *denotes a stochastic order with respect to the compound model/randomization distribution* $D\xi$;
iv) $\hat{V}_D\left(\hat{T}(z_s)\right)=\{I(\theta_0)\}^{-1}+o_{D\xi}\left(n^{-1}\right)$.
We have

$$S-S_{DB}=o_{D\xi}(n^{-1/2}).$$

PROOF. Using the Taylor series expansion of $\ell(\theta)$ around $\hat{\theta}$, and evaluating it at θ_0, we have

$$\lambda=\ell(\hat{\theta})-\ell(\theta_0)=-\frac{1}{2}\ell''(\hat{\theta})(\theta_0-\hat{\theta})^2+o_\xi\left[(\theta_0-\hat{\theta})^2\right]$$

so that regularity conditions (*i*) and (*ii*) imply

$$-\frac{1}{2}\ell''(\hat{\theta})(\hat{\theta}-\theta_0)^2=\frac{1}{2}I(\theta_0)(\hat{\theta}-\theta_0)^2+o_\xi(n^{-1/2}).$$

Now using regularity conditions (*iii*) and (*iv*), we have

$$W_{DB} = I(\theta_0)(\hat{\theta} - \theta_0)^2 + o_{D\xi}(n^{-1}).$$

The theorem now follows from the fact that $S_{DB} = \frac{1}{2} W_{DB} - \frac{1}{2} \log n$ and $S = \lambda - \frac{1}{2} \log n$. □

We note that the regularity conditions of Kass and Wassermann (1995), given in their section 2, are analogous to our assumptions i) and ii). Thus, we can conclude that under (*i*) and (*ii*) and unit information priors

$$\log BF = S_{DB} + O_{D\xi}(n^{-1/2}).$$

5. TWO EXAMPLES

In this section we verify the regularity conditions needed to prove Theorem 1 for two well-known sampling designs and the associated superpopulation models.

5.1. *One-stage cluster sampling and the associated one-way random effects model (as in Skinner, 1989, p. 37)*

Consider a clustered finite population described by the following superpopulation model

$$y_{ij} = \theta + \alpha_j + e_{ij},$$

where α_j and e_{ij} are uncorrelated with $V(\alpha_j) = \tau\sigma_0^2$ and $V(e_{ij}) = (1-\tau)\sigma_0^2$, $j = 1,\ldots,M$, $i = 1,\ldots,N_c$. Note that τ can be interpreted as the intra-cluster correlation coefficient. Suppose we are interested in testing $M_0 : \theta = \theta_0$ vs $M_a : \theta \neq \theta_0$ based on a one-stage cluster sample in which m clusters are selected by simple random sample without replacement.

For the one-way random effects model, we have $\hat{\theta} = \bar{y}_s$. Condition (*i*) is a standard property of the maximum likelihood estimator in regular problems. In order to verify condition (*ii*), note that $I(\theta_0) = \frac{n}{[1+(N_c-1)\tau]\sigma_0^2}$, (see Searle et al., 1996, p. 80) and the fact that the log-likelihood function is a quadratic form with $-\ell(\hat{\theta})$ free from θ and y_s. Under the sampling design, $\hat{T}(z_s) = \hat{\theta}$ so condition (*iii*) is trivially verified. Turning to condition (iv), we note that under the cluster sampling design: $T(y_U) = \bar{y}_U$, $\hat{T}(z_s) = \bar{y}_s$. Thus, $\hat{\theta} = T(y_U)$ and

$$\hat{V}_D(\bar{y}_s) = \frac{N-n}{N} \frac{[1+(N_c-1)\tau]}{n} s_y^2,$$

where $s_y^2 = (n-1)^{-1} \sum_{i \in s} (y_i - \bar{y}_s)^2$. Condition (*iv*) can now be verified by showing that $E\left\{\hat{V}_D(\bar{y}_s)\right\} = \{I(\theta_0)\}^{-1} + o\left(n^{-1}\right)$ and $V\left\{\hat{V}_D(\bar{y}_s)\right\} = o\left(n^{-2}\right)$, where E and V denote the expectation and the variance with respect to both the sampling design and the model.

5.2. *Two-stage sampling and the associated one-way random effects model*

We consider the same one-way random effects model and the same testing problem for a two-stage sampling where m clusters are selected by simple random sample without replacement and n_c second-tage units are randomly selected from each sampled cluster. In this case, it can be shown that $\hat{\theta} = \bar{y}_s$ and $I(\theta_0) = \frac{n}{[1+(n_c-1)\tau]\sigma_0^2}$. Verification of conditions (*i*)-(*iii*) is similar to that of one-stage cluster sampling case. To verify condition (iv), we first note that $T(y_U) = \bar{y}_U$, $\hat{T}(z) = \bar{y}_s$, and

$$\hat{V}_D(\bar{y}_s) = \frac{N-n}{Nn} s_{yt}^2 + \frac{1}{N}\left(\frac{N_0}{n_0} - 1\right) s_{ye}^2,$$

where $s_{yt}^2 = \frac{n_c}{m-1}\sum_{j=1}^{m}\left(\bar{y}_j - \bar{y}_s\right)^2$ and $s_{ye}^2 = \frac{n_c}{m(n_0-1)}\sum_{j=1}^{m}\sum_{i=1}^{n_c}\left(y_{ij} - \bar{y}_j\right)$ (see Cochran, 1977, Theorem 10.2).

Noting that $s_{ty}^2 = \frac{n-1}{m-1}\frac{1}{n_c}[1+(n_c-1)\tau]s_y^2$, we have

$$\hat{V}_D(\bar{y}_s) \simeq \frac{N-n}{Nn}[1+(n_c-1)\tau]s_y^2 + \frac{1}{N}\left(\frac{N_c}{n_c} - 1\right) s_{ye}^2.$$

Verification of condition (*iv*) is now similar to that of Example 1.

6. Monte Carlo simulation

As mentioned in the introduction, the main advantage of our proposed model selection criterion S_{DB} is that it can be applied even when the exact BIC (S_E) cannot be obtained because of the unavailability of the exact sample likelihood. However, it is important to understand its performance when the sample likelihood can be fully specified so we can compare with the exact BIC, the gold standard. In this section, we achieve this goal using a Monte Carlo simulation. In our simulation study, we include a naïve BIC (S_N), a BIC that ignores the sampling design, to understand the role of the sampling design. Consider an artificial finite population that consists of $M = 200$ clusters each of size $N_c = 10$. Thus, the size of the finite population is $N = 2000$. Suppose we are interested in a binary variable Y. We assume that for $i = 1,\ldots,N_c$, $j = 1,\ldots,M$:

$$\begin{aligned} y_{ij} &\overset{ind}{\sim} Ber(\pi_j), \\ \pi_j &\overset{ind}{\sim} Beta\left(\frac{\mu}{\gamma}, \frac{1-\mu}{\gamma}\right). \end{aligned} \tag{10}$$

Note that the above model implies that the common marginal proportion and the common intra-cluster correlation are μ and $\gamma(\gamma+1)^{-1}$, respectively. For this simulation study, γ is assumed to be positive and thus the higher the value of γ the higher the intra-cluster correlation.

Let us consider the following hypothesis testing problem:

$$M_0 : \mu = 0.25 \qquad M_1 : \mu \neq 0.25.$$

TABLE 1
Different settings for model parameters

Setting	Population mean	Mixing parameter	Sample size
1	$\mu = 0.25$	$\gamma = 1$	$n = 30 \quad (m = 3)$
2	$\mu = 0.25$	$\gamma = 1$	$n = 60 \quad (m = 6)$
3	$\mu = 0.25$	$\gamma = 0.\bar{3}$	$n = 30 \quad (m = 3)$
4	$\mu = 0.25$	$\gamma = 0.\bar{3}$	$n = 60 \quad (m = 6)$

The choice $\mu = 0.25$ suggested by the null hypothesis allows fairly skewed sampling distribution of the number of 'ones'. In generating finite populations we consider two different values of γ, $\gamma = 1$ and $\gamma = 0.\bar{3}$, which correspond to intra-cluster correlation coefficients of 0.5 and 0.25, respectively. As far as the sampling design is concerned, we assume simple random sampling (with replacement) of clusters and consider two different sample sizes: $n = 30$ and $n = 60$ (i.e., a sample of 3 and 6 clusters). In summary, we consider four different settings characterized by the values summarized in Table 1.

If we completely ignore the clustering of the observations, we can specify a binomial likelihood and compute our maximum likelihood estimate of μ as $\hat{\mu} = n^{-1}y$, where y is the number of ones observed in the sample. We refer to this solution as S_N. On the contrary, if we consider the clustered population model given by Eq. (10), we can specify the exact Beta-Binomial likelihood for the parameter vector (μ, γ). In this case, the maximum likelihood estimate $(\hat{\mu}, \hat{\gamma})$ cannot be obtained in a closed form, but can be computed using a numerical method (see Griffiths, 1973, for details). The S statistic based on this exact likelihood at the sample clustering level is referred to as S_E. The performances of S_N and S_{DB} are compared with S_E.

As in section 6.1, in order to summarize the evidence provided by various statistics in favour or against the null hypothesis, we consider the logarithm of the scale of evidence proposed by Jeffreys (1961) and the same cut-off point of 1.1. Values larger than 1.1 are supposed to provide 'positive' evidence against the model suggested by the null hypothesis.

The entries in Table 2 represent the percentage of samples with statistics lower than 1.1 over 1000 simulated samples, each drawn independently according to the sampling design described above. Clearly, the effect of clustering on S_N is very severe for all the three cases, the acceptance rates being considerably lower than those using our gold standard S_E. The difference between the S_N and S_E increases as the intra-cluster correlation increases. The increase in the sample size contributes very little in resolving the difference. Our approximation S_{DB} is tracking S_E very well even for this non-normal situation and for a moderate sample size. Needless to say, both S_{DB} and S_E are not affected by the variation of the intra-cluster correlation.

In Table 3 we compare the behaviour of the three procedures under a few selected alternatives: $\mu_{ALT1} = 0.5$, $\mu_{ALT2} = 0.6$, $\mu_{ALT3} = 0.75$, $\mu_{ALT4} = 0.9$. The entries of Table 2 and Table 3 have similar interpretations. Under all null hypotheses considered, S_{DB} and S_E perform quite closely. They both seem to be rather conservative in rejecting the null hypothesis compared to S_N. We stress that S_N underestimates the evidence against the model suggested by the null hypothesis since it is derived from a model that does not incorporate intra-cluster homogeneity. Settings 1-3 correspond to high intra-cluster cor-

TABLE 2
Percentage of S statistics lower than 1.1 under M_0

Setting	S_N	S_E	S_{DB}
1	57	97	87
2	51	96	93
3	75	99	86
4	77	99	93

TABLE 3
Percentages of various S statistics lower than 1.1 under different alternative hypotheses

Hypothesis	Setting	S_N	S_E	S_{DB}
$\mu_{ALT1} = 0.5$	1	64	75	71
	2	54	63	67
	3	37	64	69
	4	23	49	58
$\mu_{ALT2} = 0.6$	1	51	59	65
	2	35	49	49
	3	23	51	51
	4	7	39	41
$\mu_{ALT3} = 0.75$	1	18	35	45
	2	8	31	27
	3	4	27	4
	4	0	9	12
$\mu_{ALT4} = 0.9$	1	4	7	26
	2	0	1	5
	3	0	6	17
	4	0	0	2

relation coefficient that reduces the effective sample size substantially. For this reason both S_{DB} and S_E have problems in finding positive evidence against the wrong model when it is quite close to the true one (e.g., null hypothesis 1). This effect is somewhat weaker in settings 4 and 5, which correspond to a lower level of the intra-cluster correlation coefficient.

7. CONCLUDING REMARKS

We have presented a robust approximation to the BIC that can be used with complex survey data. Our method is expected to be useful in situations where it is *not* possible to obtain the *exact* likelihood for the sample since our proposed method merely requires an estimator of the superpopulation parameter with good design-based properties (e.g., psuedo-maximum likelihood) and its design-consistent variance estimator. Thus, this paper fills in an important research gap in the analytic use of survey data. In the future we plan to extend our proposed method to a general variable selection problem and compare with an alternative informative sampling approach when a working superpopulation model can be specified.

ACKNOWLEDGEMENTS

The second author's research was supported by the National Science Foundation SES-085100.

REFERENCES

J. O. BERGER (2001). *Objective Bayesian Methods for Model Selection: Introduction and Comparison*. In P. LAHIRI (ed.), *Model Selection*, Institute of Mathematical Statistics, Lecture Notes - Monograph series, Vol. 38.

W. G. COCHRAN (1977). *Sampling Techniques*. John Wiley & Sons, New York.

W. E. DEMING, F. F. STEPHEN (1941). *On the interpretation of censuses as samples*. Journal of the American Statistical Association, 36, pp. 45–49.

W. E. DEMING (1953). *On the distinction between enumerative and analytic surveys*. Journal of the American Statistical Association, 48, pp. 244–255.

B. EFRON, A. GOUS (2001). *Scales of Evidence for Model Selection: Fisher versus Jeffreys*. In P. LAHIRI (ed.), *Model Selection*, Institute of Mathematical Statistics, Lecture Notes - Monograph series, Vol. 38.

B. I. GRAUBARD, E. L. KORN (2002). *Inference for superpopulation parameters using sample surveys*. Statistical Science, 17, pp. 73–96.

W. E. DEMING (1953). *Maximum Likelihood Estimation for the Beta-Binomial Distribution and an Application to the Household Distribution of the Total Number of Cases of a Disease*. Biometrics, 29, pp. 637–648.

M. H. HANSEN, W. G. MADOW, B. J. TEPPING (1983). *An Evaluation of Model-Dependent and Probability-Sampling Inference in Sample Surveys*. Journal of the American Statistical Association, 78, pp. 776–793.

D. HOLT, T. M. F. SMITH, P. D. WINTER (1980). *Regression Analysis of Data from Complex Surveys*. Journal of the Royal Statistical Society, Ser. A, 143, pp. 474–487.

D. HOLT (1989). *Introduction to Part C*. In C. J. SKINNER, D. HOLT, T. M. F. SMITH (eds.), *Analysis of Complex Surveys*, John Wiley & Sons, Chicester, pp. 209–220.

C. T. ISAKI, W. A. FULLER (1982). *Survey Design under the Regression Superpopulation Model*. Journal of the American Statistical Association, 77, pp. 89–96.

H. JEFFREYS (1961). *Theory of Probability*. Oxford University Press, Oxford.

P. LAHIRI (2001). *Model Selection*. Institute of Mathematical Statistics, Lecture Notes - Monograph series, Vol. 38.

R. E. KASS, L. WASSERMANN (1995). *A Reference Test for Nested Hypotheses and Its Relationship to the Schwartz Criterion*. Journal of the American Statistical Association, 90, pp. 928–934.

P. S. KOTT (1989). *Robust Small Domain Estimation using Random Effects Modelling*. Survey Methodology, 15, pp. 3–12.

P. S. KOTT (1991). *A Model-Based Look at Linear Regression with Survey Data*. The American Statistician, 45, pp. 107–112.

D. PFEFFERMANN (2009). *Inference under informative sampling*. In D. PFEFFERMANN, C. R. RAO (eds.), *Handbook of Statistics 29: Sample Surveys: Inference and Analysis*, Elsevier, Amsterdam, pp. 455–487.

C. E. SÄRNDAL, B. SWENSSON, J. WRETMAN (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

S. R. SEARLE, G. CASELLA, C. E. MCCULLOGH (1996). *Variance Components*. John Wiley & Sons, New York.

G. SCHWARTZ (1978). *Estimating the dimension of a model*. The Annals of Statistics, 6, pp. 461–464.

C. J. SKINNER (1989). *Introduction to Part A*. In C. J. SKINNER, D. HOLT, T. M. F. SMITH (eds.), *Analysis of Complex Surveys*, John Wiley & Sons, Chicester, pp. 23–28.

D. J. SPIEGELHALTER, N. G. BEST, B. P. CARLIN, A. VAN DER LINDE (2002). *Bayesian measures of model complexity and fit*. Journal of the Royal Statistical Society, Ser. B, 64, pp. 583–639.

SUMMARY

A design-based approximation to the Bayes Information Criterion in finite population sampling

In this article, various issues related to the implementation of the usual Bayesian Information Criterion (BIC) are critically examined in the context of modelling a finite population. A suitable design-based approximation to the BIC is proposed in order to avoid the derivation of the exact likelihood of the sample which is often very complex in a finite population sampling. The approximation is justified using a theoretical argument and a Monte Carlo simulation study.

Keywords: Bayes factor; Hypothesis testing; Model selection; Pseudo-maximum likelihood; Cluster sampling