Bayesian Sample Surveys: some reflections

Brunero Liseo

June 2023

Sapienza Università di Roma
brunero.liseo@uniroma1.it

with M.G. Ranalli (Università di Perugia) & Marco di Zio (ISTAT, Roma)

XIII Giornata della ricerca MEMOTEF

Intro

Main goal:

To explore and discuss the potential role of Bayesian ideas and techniques in modern survey sampling.

Summary:

- the theoretical conflict between design-based methods and the likelihood principle;
- the ineluctability of a shift towards model-based techniques in modern survey statistics;
- review of the most prominent and promising ideas for a Bayesian theory of inference for finite populations.

Peculiarity of Survey Sampling

Survey Sampling is different from standard inference because:

- In standard inference we observe units and assume that what we observe is somehow close to an average value. We actually do not care about individual values
- In survey sampling, we are interested in the entire vector of the values of some variables. However, some of them are perfectly detected, but most of them are unknown.

The conflict in a simple context

- Simple random sampling w/o replacement drawn from a population P with N identified units.
- N is known and units identified through their labels $\{1, 2, \ldots, N\}$.
- Draw a sample s of size n and assume that the randomization scheme assigns a probability $p(s)$ to this specific sample.
- Quantity of interest is the vector of values of a variable $\boldsymbol{Y}$ on the entire population, $\boldsymbol{Y}_P = (y_1, y_2, \ldots, y_N)$, or maybe $\tau = f(\boldsymbol{Y}_P)$, after observing $\boldsymbol{Y}_{P \setminus s}$.

The conflict in a simple context

- Design-based approach based on the H-T strategy: take an empirical version of τ, with the n observations weighted with the inverse of their probabilities to be included in the sample.
- Unbiasedness is referred to the randomization scheme (the sampling design).
- Basu formally showed that the likelihood function for Y_P (or τ) is flat , i.e.

$$L(\boldsymbol{Y}_{P\setminus s}; \boldsymbol{Y}_s) = \begin{cases} k & \text{for all } \boldsymbol{Y}_{P\setminus s} \text{ compatible with } \boldsymbol{Y}_s \\ 0 & \text{otherwise} \end{cases}.$$

Conflict III

- Many scientists have interpreted this result as evidence of a general inadequacy of the $\boldsymbol{L}$ - and the likelihood principle itself - as the main inferential instrument in this framework.
- Basu and other Bayesians believe that this context offers an example where $\boldsymbol{L}$ provides obvious but correct results, and this clarifies the insufficient level of modeling of the design-based methods.
- There is no explicit association or link between what we observe and what we don't . . .

Conflict IV

This is precisely the core of the discussion as reported in Ghosh and Meeden [1997]:

> *As we remarked before, the basic problem of finite population sampling is deciding what one learns about the unobserved units from the observed sampled units.*

Different views

- From a design perspective, the link is given by the representativeness of the sample.
- From a model-based perspective, the $\boldsymbol{Y}_s$ and $\boldsymbol{Y}_{P\setminus s}$ are conditionally independent with a common distribution
- from a Bayesian perspective the link is explicitly provided by the notion of exchangeability
- the Polya Posterior approach (aka *Bayesian Bootstrap* in finite population) automatically assumes *similarities* among $\boldsymbol{Y}_s$ and $\boldsymbol{Y}_{P\setminus s}$.
- A common view: The unseen should be similar to the seen ...

How to justify procedure

1. Design-based procedures are justified in terms of unbiasedness . It works on average, but may dramatically fail (recall the Jumbo & Sambo example by Basu [1971])[1]

2. Model-based procedures are likelihood slaves and they work as long as the model is sensible and provides a good fit

3. A Bayesian procedure is justifiable when the implicit prior used is a sensible prior.

   Here sensible means, at least, *with full support* and *not so concentrated* . . .

[1] The Circus Example, D. Basu (1971)
https://www.umass.edu/cluster/ed/unpublication/yr2000/c00ed72.PDF

Two Bayesian proposals

Two (or maybe more ...) Bayesian proposals.

- Calibrated Bayesian strategy [Little, 2022]
- Polya Posterior [Ghosh and Meeden, 1997]
- Bayesian nonparametrics [Mendoza et al., 2021]

Calibrated Bayes I

- Little [2011] (or Little [2022]) has strongly advocated the use of Bayesian methods in survey sampling and, more generally, in Official Statistics.
- Main point: a compromise between various approaches.
- While inference procedures should follow a Bayesian road, design features like clustering and stratification should be explicitly incorporated into the model in order to avoid the sensitivity of inference to model misspecification.

Calibrated Bayes II

Quoting Little [2022],

a purely design-based approach to finite population inference is no longer able to

> *adequately address many of the problems of modern sample survey*

and a model-based approach is deemed necessary. In this respect, developments in SAE represented the Trojan horse.

- However, the model-based approach should be dressed in a Bayesian suit in order to incorporate in a more natural way survey sample design features.

Calibrated Bayes III

The model-based framework expressed by the equation

$$p_{s,y;z}(s,y;z,\theta,\psi) = p_{y;z}(y;z,\theta)p_{s|y;z}(s|y;z,\psi),$$

- Here θ is a vector of parameters directly related to the variable of interest y and ψ only refers to the mechanism of inclusion.
- Ignore the non-response, issue: then $S|Z,Y$ does not depend on Y.
- The likelihood contribution to inference is restricted to $p_{y|z}(y;z,\theta)$.
- Combined with a prior on θ, it produces a posterior predictive distribution

$$P(Y_{P\setminus s}|Y_s,z) = \int p(Y_{P\setminus s}|Y_s;z,\theta)p(\theta|Y_s;z)d\theta$$

Calibrated Bayes IV

- The above consideration simply rules out any chance that the Bayesian answers could be efficient from a frequentist perspective if the word "frequentist " is meant in terms of the sampling mechanism.
- We believe that the frequentist properties should be considered either with respect to the conditional model induced by the family of distributions $p_{y;z}(y;z,\theta)$, or to the joint distribution $p_{y,s;z}(y,s;z,\psi,\theta)$.
- It is known [see Berger et al., 2009, Consonni et al., 2018] that a correct frequentist coverage of Bayesian procedures can be obtained only through the use of formal "noninformative" priors [see Berger et al., 2009, Consonni et al., 2018].

Priors

- The use of extra-data information represents one of the distinguished features of Bayesian inference.
- The term extra-data is often interpreted as subjective
- There are many instances where previous knowledge can be adequately used to train our model, for example, periodical releases of indexes.

Polya Posterior

- It can be considered the finite population adaptation of the Bayesian Bootstrap [Rubin, 1981].
- In the simplest scenario: Population of N units; we draw a SRS of size n, say Y_s.
- goal: to estimate the mean θ of some function $h(Y_P)$.

PP: how it works

1. Put the n observed units in another urn U_2 and let the other $N - n$ units in the original urn U_1 .
2. draw a unit from U_2, and record its value y ;
3. draw a unit from U_1, attach to it the y value and replace both units in U_2;
4. repeat steps 2-3 until U_1 is empty.

The Pseudo-Posterior

- This way we simulate a single realization of the entire population.
- Repeat the procedure a huge number M of times, to get a pseudo-posterior distribution of Y_P and θ
- This is only a pseudo-posterior since "no prior" has been introduced.

However ...

A result from Lo [1988]

Assume to observe in Y_s, k distinct values; let n_j be their frequencies ($j = 1; \dots, k$). Let m_j^* be the random frequencies of the k distinct values in a single Polya experiment. Then the following statements hold:

- The random vector $(m_1^*, \dots, m_k^*)|Y_s$ is Dirichlet-Multinomial [Mosimann, 1962] with parameters $(N - n; n_1, n_2, \dots, n_k)$
- As $N \to \infty$,

$$\left(\frac{m_1^*}{N-n}, \dots, \frac{m_k^*}{N-n}\right)|Y_s \xrightarrow{d} \mathrm{Dirichlet}(n_1, n_2, \dots, n_k),$$

- This also implies that the implicit prior for the Polya Posterior is a Dirichlet$(0, \dots, 0)$ on any set of observed (in the sample) frequencies.
- Not so easy to handle complex schemes although several extensions have been proposed (Strief and Meeden [2013], Lazar et al. [2008] etc.)

Bayesian nonparametrics

- In the last 20 yrs we experienced an explosion, both in theoretical and applied terms, of Bayesian nonparametric methods.
- Survey sampling has not yet been hit by this wave although the seminal papers by Lo [1986, 1988] seem to have paved the way.
- Some recent exceptions are Mendoza et al. [2021] and Savitsky and Toth [2016], Paddock [2002].
- Model considered until now are very simple. However, we believe there is much to do along this path ...

Conclusions

- FPS is an important and peculiar chapter of Statistics.
- it deserves particular attention and a specifically suited methodology.
- Bayesian inference is a solid, prescriptive, and coherent mathematical theory, sometimes difficult to combine with the practical difficulties of FPS.

Basu himself, the leader of the anti-design party wrote

> *The Bayesian as a surveyor must make all kinds of compromises... He may even agree to introduce an element of randomization into his plan... I can not put this enormous speculative process into a jacket of a theory. I happen to believe that data analysis is more than a scientific method...*

References

D. Basu. An essay on the logical foundations of survey sampling. I. In *Foundations of statistical inference (Proc. Sympos., Univ. Waterloo, Waterloo, Ont., 1970)*, pages 203–242. Holt, Rinehart and Winston of Canada, Toronto, Ont., 1971.

J.O. Berger, J.M. Bernardo, and D. Sun. The formal definition of reference priors. *Annals of Statistics*, 37: 905–938, 2009.

G. Consonni, D. Fouskakis, B. Liseo, and I. Ntzoufras. Prior Distributions for Objective Bayesian Analysis. *Bayesian Analysis*, 13(2):627 – 679, 2018.

M. Ghosh and G. Meeden. *Bayesian methods for finite population sampling*. Chapman & Hall, London, 1997.

R. Lazar, G. Meeden, and D. Nelson. A noninformative Bayesian approach to finite population sampling using auxiliary variables. *Survey Methodology*, 34:51–64, 2008.

R.J. Little. Calibrated Bayes, for statistics in general, and missing data in particular. *Statistical Science*, 26(2): 162–174, 2011.

R.J. Little. Bayes, buttressed by design-based ideas, is the best overarching paradigm for sample survey inference. *Survey Methodology*, 48:257–281, 2022.

A.Y. Lo. Bayesian Statistical Inference for Sampling a Finite Population. *Annals of Statistics*, 14(3):1226–1233, 1986.

A.Y. Lo. A Bayesian bootstrap for a finite population. *Annals of Statistics*, 16:1684–1695, 1988.

M. Mendoza, A. Contreras-Cristán, and Gutièrrez-Pena. Bayesian Analysis of Finite Populations under Simple Random Sampling. *Entropy*, 23:318, 2021.

J.E. Mosimann. On the compound multinomail distribution, the multivariate β-distribution and correlations among proportions. *Biometrika*, 49:65–77, 1962.

S. Paddock. Bayesian nonparametric multiple imputation of partially observed data with ignorable nonresponse. *Biometrika*, 89(3):529–538, 2002.

D. B. Rubin. The Bayesian bootstrap. *Annals of Statistics*, 9:130–134, 1981.

T.D. Savitsky and D. Toth. Bayesian estimation under informative sampling. *Electronic Journal of Statistics*, 10: