Bayesian survey design to guide the end of AIDS

Adam Howes (ath19@ic.ac.uk)

Contents

Abstract

As the HIV epidemic enters its fifth decade, meeting the need for timely, cost-effective, robust surveillance requires advancement of a new survey archetype. Although household surveys provide estimates which are nationally representative, they are prohibitively expensive to carry out, and so only typically occur every four or five years. Furthermore, household surveys are statistically inefficient, failing to leverage prior responses or other sources of data to guide sampling effort. We propose the use of an adaptive, model-based, sampling approach called Bayesian survey design (BSD). We show that our approach enhances the efficiency and responsiveness of HIV surveys, whilst reducing the cost and logistical complexity compared to existing practise. To do so, we use the Manicaland cohort study to simulate survey designs and empirically evaluate their statistical properties.

1 Introduction

There exists a fundamental conflict between Bayesian decision theory and randomisation (O'Hagan 1987). Gelman (2008) writes from the perspective of a critic of Bayesian statistics that:

> The mathematics of Bayesian decision theory lead inexorably to the idea that random sampling and random treatment allocation are inefficient, that the best designs are deterministic. I have no quarrel with the mathematics here, the mistake lies deeper in the philosophical foundations, the idea that the goal of statistics is to make an optimal decision.

In the field of disease surveillance, there is a good argument to be made that the reason we collect data is to make an optimal decision, to invest the limited resources available to limit the burden of disease as much as possible.

There is widespread precedent for the use of prior information to guide sampling effort and improve efficiency the basis for techniques such as stratification, including the use of explicit models to determine strata (Godfrey, Roshwalb, and Wright 1984).

Our proposed sampling method could be carried out continuously, rather than once every four or five years. This would allow hiring permanent staff, saving much of the costs associated to training.

Other relevant points to discuss include covariate adjustment, matching, adaptive clinical trials, randomised control trials, active case finding algorithms (Choun et al. 2019), data defect (Meng 2018; Bradley et al. 2021).

The remainder of this paper is organised as follows. In Section 2 we review the existing approach of national household surveys, before describing ideas from probabilistic numerics which we apply in Section 3 to create a Bayesian approach to survey design. In Section 4 we demonstrate our approach via simulations using population cohort data from the Manicaland Project. We discuss our conclusions and recommendations for future work in Section 5.

2 Background

2.1 Survey design and household surveys

In survey sampling (Lohr 2009) a subset of sampling units from a sampled population are selected. Ideally, the sampled population corresponds to the target population of interest. For example, the population of interest might be all individuals aged 15-49 in a particular country. The subset of sampling units are selected from a list of all possible sampling units called a sampling frame. Sampling units may not correspond to the unit upon which measurements are made, called the observational unit and multiple stages of sampling units may be used. To quantify uncertainty based on sampling variability it is important to use a probability sample, where each observational unit has a known, non-zero probability $\pi_k = \mathbb{P}(\text{unit } k \text{ in sample}) > 0$ of being included in the sample. Keeping track of these probabilities by using survey weights $w_k = 1/\pi_k$ is one way to take the representativeness of the sample into account.

The Demographic and Health Survey (MEASURE DHS 2012) typically employs two-stage household-based sampling design in which, at the first stage, an enumeration area (EA) is selected, before, in the second stage, households are chosen from within each chosen EA. This design is also known as a two-stage cluster sample, where the clusters or primary sampling units (PSUs) correspond to EAs and the secondary sampling units (SSUs) correspond to households. Cluster sampling reduces sample efficiency as members within a cluster tend to be more similar than those between clusters. However, in practise cluster sampling convenient and, when clusters are defined geographically, less costly than the alternatives. In the first stage of the DHS, a fixed number of urban and rural EAs from each region are selected by probability proportional to size (number of households) sampling. Unlike clustering, stratification tends to increase sample efficiency.

Sample efficiency and cost are two central considerations for running any survey. Rather than managing these considerations with techniques like clustering and stratification, we will see that BSD explicitly uses an acquisition function to balance the potential information gain and cost of each new sample.

2.2 Probabilistic numerics

BSD takes inspiration from probabilistic numerical methods (Cockayne et al. 2019) like Bayesian optimisation (Močkus 1975; Snoek, Larochelle, and Adams 2012) and Bayesian numerical integration (Diaconis 1988; Briol et al. 2019; Zhu et al. 2020) which use statistical models to approach numerical analysis problems.

For example, in numerical integration design points $q_{1:J} \in \mathcal{Q}^n$ may be used to estimate integrals

$$\mathbb{E}_p[f] = \int_{\mathcal{Q}} f(q)p(q)\mathrm{d}q \approx \frac{1}{J}\sum_{j=1}^{J} f(q_j) = \widehat{\mathbb{E}}_p[f], \tag{1}$$

where f is the objective function and p is a probability density. In adaptive Bayesian numerical integration, design points q_{j+1} are sequentially chosen based on a model $\hat{f}$ for the objective function trained on all previous data $f(q_{1:j})$. This is in contrast to classical quadrature approaches which typically select all J design points at the outset. The model $\hat{f}$ is combined with an acquisition function $a : \mathcal{Q} \to \mathbb{R}$ whose role is to encourage selection of points which result in favourable estimator properties. An example of such a property is minimising the variance. Given data $q_{1:n}$, the next point selected to be included in the design q_{i+1} can be found by maximising the acquisition function

$$q_{j+1} \in \arg\max_{q \in \mathcal{Q}} a(q) \tag{2}$$

This typically results in selecting design points with good coverage of the space $\mathcal{Q}$ whilst being concentrated in regions which contribute most to the integral. In many ways the goals of numerical integration and survey design are analogous. Both look to make statements about an objective function f on a broader domain $\mathcal{Q}$ based upon a limited number of samples $f(q_{1:J})$.

A combination of model and acquisition function can also be used to solve optimisation problems. In optimisation the aim is to find a minimiser $q^\star \in \arg\min_{q\in\mathcal{Q}} f(q)$ of the objective function f or failing that some value q such that $|f(q) - f(q^\star)|$ is small. As with Bayesian quadrature, a model and acquisition function may be used to approach this problem. Bayesian optimisation algorithms often initially explore the space, before only later spending samples to refine the most promising minima.

BSD applies the same principle of probabilistic numerics, namely the use of a statistical model and acquisition function, to survey design. Depending on the goals of the survey, the acquisition function in BSD may look more similar to those used in either quadrature or optimisation tasks. This particular issue, as well as a broader outline of our methodology, is discussed in the following section.

3 Methodology

3.1 Outline

BSD is an active learning approach to survey design where new elements of the design are chosen based upon existing information. Like the above probabilistic numerical methods, BSD is based upon combination of a statistical model and acquisition function.

In analogy to the numerical integration example in Section 2, the design space $\mathcal{Q}$ of a survey corresponds to the sampled population, that is a collection of all units $\mathcal{U} = \{1, \ldots, N\}$. Each unit may be associated to covariates $x_i \in \mathcal{X}$[1] and outcomes $y_i \in \mathcal{Y}$. The cost of sampling unit i could be determined by its covariates such that $c_i = c(x_i)$.

Suppose that we wish to use BSD to perform a household prevalence survey in a particular country. The objective function f is the true underlying HIV prevalence and $\hat{f}$ is a model of HIV prevalence. The model $\hat{f}$ must be at least of the resolution of the sampling unit. Though in multi-stage sampling designs, it is possible to combine BSD with other sampling approaches. For example, in analogy to the DHS approach, an area-level model $\hat{f}$ may be used to select EAs using BSD, before selecting households with another approach,

[1]We suppose that the covariates may be categorised as either spatial s, temporal t, or other z such that $\mathcal{X} = \mathcal{S} \times \mathcal{T} \times \mathcal{Z}$.

such that a household-level model is not required. Note that this type of approach may have something in common with existing online stratified sampling approaches such as Bennett and Carvalho (2010).

One possible goal for a national household prevalence survey is to precisely estimate national prevalence. This is equivalent to accurately estimating the integral of f over the sampled population. As such, a suitable acquisition function for this task would aim to minimise the posterior variance of this integral. More broadly, appropriate acquisition function choice depends upon the aims of the survey, which must first be elicited before being represented mathematically. This process might be challenging as the aims of any survey may be numerous, vague, and difficult for multiple actors to precisely agree upon. Examples of other survey goals include determining if a quantity is above or below a threshold, such as would be required for the 90-90-90 and 95-95-95 HIV goals. We expect that complex acquisition functions may be composed by combination of simpler acquisition functions.

3.2 Acquisition functions

Distance costs Suppose that the cost of taking sample $j+1$ is proportional to its spatial distance from sample j such that $c_{j+1} = c(|s_{j+1} - s_j|)$.

Composition of acquisition functions Separate acquisition functions $a_1, \ldots, a_A$ may be composed together to obtain a joint acquisition function a.

Effects of batch acquisition In a realistic survey setting, refitting the emulator after each new data point is impractical. Rather than selecting a single new point to be sampled, batch acquisition functions select many.

3.3 Concerns

Incorrect model BSD is a model-based method and will be inefficient when the model used is incorrect. For example, if the model claims with high certainty that an area has low prevalence, then BSD may not chose to sample units in that area, never learning that it was wrong. The cost of using more information to guide sampling is that if that information is misleading then the method will perform poorly. However, it is unclear that choosing not to use relevant information is preferable, particularly in resource constrained settings where we do not have the luxury of disregarding our prior in exchange for robustness.

Incorrect acquisition function The results of a DHS survey are used for a wide variety of purposes. If the acquisition function is not properly designed, then BSD may over optimise for a single, narrowly defined goal such that the results are less broadly useful.

Over-complication BSD may face higher levels of implementation, logistics and data processing challenges than other, simpler approaches. The DHS surveys use a relatively simple sampling design for this reason (MEASURE DHS 2012), noting that "in large-scale surveys non-sampling errors (coverage errors, errors commited in survey implementation and data processing, etc.) are usually the most important sources of error".

Wrong application HIV prevalence may prove not be the best outcome to demonstrate BSD because it changes very slowly over time. Other application areas such as tracking food security, or other more rapidly evolving diseases such as coronavirus, may be more well suited.

Bad incentives Randomisation reduces the risk of any experimenter adversarially choosing among designs to favour one hypothesis.

3.4 Expected behavior

Spatio-temporal We expect to observe that in spatio-temporal models, as uncertainty increases over over time, spatial locations are returned to for sampling.

Evidence synthesis There may be more complex behaviours for models which integrate multiple sources of evidence. For example, a model which integrates household survey prevalence data and ANC prevalence data might have ANC random effects which could be informed by surveying close to ANC sites. As such, although

some behaviours of BSD, such as achieving spatial balance, may closely resemble other, simpler, approaches, in simple cases for more complex evidence synthesis models BSD may display types behaviour which may not be anticipated as easily.

3.5 Multilevel regression and poststratification

Incorporating survey weights into model-based SAE estimators (Chen, Wakefield, and Lumely 2014) is challenging. Gelman et al. (2007) puts it simply by stating that "survey weighting is a mess". Rather than survey weighting, BSD could use poststratification, a technique which, like weighting, accounts for sample non-representativeness.

Poststratification is most well known as a part of "multilevel regression and poststratification" (MRP) (Gelman and Little 1997), a technique popular predominantly in the political science literature. MRP combines a SAE (multilevel) model for sparse data together with a poststratification step. Poststratification involves collecting demographic information about the target population (e.g. age, sex, ethnicity – these are known as poststratification variables), which can also be collected during the survey for each respondent. The multilevel model is used to estimate the quantity of interest for each combination of the poststratification variables, called cells. Population representative estimates can then be calculated by aggregating cell estimates in proportion to target population demographics.

Further work in BSD may look to incorporate additional poststratification variables such as age into the acquisition function by modelling them, further informing the selection process.

4 Experiments

4.1 Population cohort indicator data

As a part of the Manicaland general population cohort study (Gregson et al. 2017) six censuses of all households have been taken across twelve sites in Manicaland province, Zimbabwe between 1998 and 2013. Manicaland province is an ideal setting:

- Manicaland province is at the "leading edge" of what high HIV burden epidemics will look like in the years to come. There is a very high prevalence HIV epidemic, but HIV incidence has declined rapidly and ART coverage is high.
- The province has a varied HIV epidemiology, including a mix of urban and rural populations, mix of industries, and is traversed by key transportation routes from the capital Harare to other countries.

The Manicaland data provides a unique opportunity to evaluate survey designs in a realistic setting where the truth is known. We use simulations to carry out two experiments:

1. **Empirical simulations** We use historical population cohort data from the Manicaland Project population HIV cohort to simulate survey samples and compare results to the full cohort data to evaluate alternative designs and biases.
2. **Synthetic population simulations** We use the PopART-IBM [@pickles2021popart] applied to the Manicaland cohort population to simulate future HIV epidemic scenarios and sample designs within the population. Simulated survey results are compared to "true" epidemic indicators produced by model simulations. Model simulations represent HIV epidemic dynamics and health system engagements among population risk groups, demographic groups, and districts within Manicaland province.

4.2 Empirical simulations

We consider the survey sampling design strategies given by Table 1. We simulate S designs indexed by s from each strategy. For each design strategy d and round t of the survey we compute direct estimates $\hat{p}^s_{dt}$ for the proportion of individuals with value one for the indicator under consideration at round t. We compare the

Sampling design strategy	Details
D1. DHS	Two-stage cluster design: probability-proportional-to-size sampling of EAs stratified within each district by urban and rural status, followed by systematic sampling of households.
D2. EA-BSD	Two-stage cluster design: BSD of EAs with posterior variance acquisition function and distance costs, followed by systematic sampling of households. For the first survey round we use D1.
D3. HH-BSD	BSD at the household level with posterior variance acquisition function and distance costs. For the first survey round we use D1.

Table 1: All sampling designs considered.

collection of direct survey estimates $\{\hat{p}_{dt}^s\}_{s=1}^S$ to the true proportion p_t using the mean squared error (MSE), which may be decomposed as a sum of variance and bias terms as follows

$$\text{MSE}_{dt} = \frac{1}{S}\sum_s (\hat{p}_t - p_{dt}^s)^2 = \text{Var}_{dt}(\hat{p}_{dt}^s) + \text{Bias}(\hat{p}_{dt}^s, p_t)^2. \tag{3}$$

We compare each design according to $\text{MSE}_d = \sum_t \text{MSE}_{dt}$.

4.3 Synthetic population simulations

5 Discussion

Acknowledgements

AH was supported by the EPSRC Centre for Doctoral Training in Modern Statistics and Statistical Machine Learning (EP/S023151/1).

Bibliography

Bennett, Paul N, and Vitor R Carvalho. 2010. "Online Stratified Sampling: Evaluating Classifiers at Web-Scale." In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 1581–84.

Bradley, VC, S Kuriwaki, M Isakov, D Sejdinovic, X Meng, and S Flaxman. 2021. "Unrepresentative Big Surveys Significantly Overestimate US Vaccine Uptake."

Briol, François-Xavier, Chris J Oates, Mark Girolami, Michael A Osborne, Dino Sejdinovic, et al. 2019. "Probabilistic integration: A role in statistical computation?" *Statistical Science* 34 (1): 1–22.

Chen, Cici, Jon Wakefield, and Thomas Lumely. 2014. "The use of sampling weights in Bayesian hierarchical models for small area estimation." *Spatial and Spatio-Temporal Epidemiology* 11: 33–43.

Choun, Kimcheng, Tom Decroo, Tan Eang Mao, Natalie Lorent, Lisanne Gerstel, Jacob Creswell, Andrew J Codlin, Lutgarde Lynen, and Sopheak Thai. 2019. "Performance of Algorithms for Tuberculosis Active Case Finding in Underserved High-Prevalence Settings in Cambodia: A Cross-Sectional Study." *Global Health Action* 12 (1): 1646024.

Cockayne, Jon, Chris J Oates, Timothy John Sullivan, and Mark Girolami. 2019. "Bayesian Probabilistic Numerical Methods." *SIAM Review* 61 (4): 756–89.

Diaconis, Persi. 1988. "Bayesian Numerical Analysis." *Statistical Decision Theory and Related Topics IV* 1: 163–75.

Gelman, Andrew et al. 2007. "Struggles with Survey Weighting and Regression Modeling." *Statistical Science* 22 (2): 153–64.

Gelman, Andrew. 2008. "Objections to Bayesian statistics." *Bayesian Analysis* 3 (3): 445–49.

Gelman, Andrew, and Thomas C Little. 1997. "Poststratification into Many Categories Using Hierarchical Logistic Regression."

Godfrey, James, Alan Roshwalb, and Roger L Wright. 1984. "Model-Based Stratification in Inventory Cost Estimation." *Journal of Business & Economic Statistics* 2 (1): 01–09.

Gregson, Simon, Owen Mugurungi, Jeffrey Eaton, Albert Takaruza, Rebecca Rhead, Rufurwokuda Maswera, Junior Mutsvangwa, et al. 2017. "Documenting and Explaining the HIV Decline in East Zimbabwe: The Manicaland General Population Cohort." *BMJ Open* 7 (10). https://doi.org/10.1136/bmjopen-2017-015898.

Lohr, Sharon L. 2009. *Sampling: Design and Analysis*. Nelson Education.

MEASURE DHS. 2012. "Sampling and Household Listing Manual: Demographic and Health Surveys Methodology."

Meng, Xiao-Li. 2018. "Statistical Paradises and Paradoxes in Big Data (i): Law of Large Populations, Big Data Paradox, and the 2016 US Presidential Election." *The Annals of Applied Statistics* 12 (2): 685–726.

Močkus, Jonas. 1975. "On Bayesian methods for seeking the extremum." In *Optimization Techniques IFIP Technical Conference*, 400–404. Springer.

O'Hagan, Anthony. 1987. "Monte Carlo Is Fundamentally Unsound." *The Statistician*, 247–49.

Snoek, Jasper, Hugo Larochelle, and Ryan P Adams. 2012. "Practical Bayesian optimization of machine learning algorithms." In *Advances in Neural Information Processing Systems*, 2951–59.

Zhu, Harrison, Xing Liu, Ruya Kang, Zhichao Shen, Seth Flaxman, and François-Xavier Briol. 2020. "Bayesian Probabilistic Numerical Integration with Tree-Based Models." *arXiv Preprint arXiv:2006.05371*.