Haphazard Intentional Sampling:
Extension to multiple groups and empirical evaluation.

Student: Rafael Peçanha Waissman

Advisor: Marcelo de Souza Lauretto

Co-advisor: Júlio Michael Stern

Contributions: Miguel Gabriel Ribeiro Miguel

Introduction:

Large-scale sampling and experimental design problems usually demand large staff and infrastructure and expensive field operations to cover a representative group of the population of interest. Even then, pure (or stratified) randomized experiments do not guarantee efficient control over specific sets of covariates, and there may be large divergences between sample and population statistics. To address this problem, Lauretto et al. [1,2] and Fossaluza et al. [3] developed the haphazard intentional sampling method, an approach that combines intentional sampling, using methods of numerical optimization for an appropriate objective function, with random disturbances ensuring good decoupling properties. For a fixed sample size, this technique aims at diminishing the distance between sample and population regarding specific covariates of interest or, the other way around, minimizing the sample size needed to achieve good enough expected agreement between sample and population regarding specific covariates of interest.

This method can be applied in several contexts, such as allocations of treatment and control groups in medical trials [2] or in statistical sampling problems [4]. In this work, the performance of the haphazard intentional sampling method is compared to pure random sampling and to the rerandomization methods proposed by Morgan and Rubin [5]. The performance of the methods is compared in two case studies in a batch of numerical simulations regarding the proximity of generated samples to covariate means of the total population and the precision of ensuing statistical estimators.

Objectives:

Our first objective was to evaluate the performance of haphazard intentional sampling method, comparing to pure random sampling and to the rerandomization methods proposed by Morgan and Rubin [5]. The first case study concerns the prevalence of SARS-CoV-2, using covariates from public data sets generated by the 2010 Census of the Brazilian Institute of Geography and Statistics (IBGE) and is inspired in EPICOVID 19 [6], a population-based survey conducted in 133 Brasilian municipalities.

Another goal was to formulate a generalization of Haphazard Sampling [1, 2] for multiple groups and to evaluate the use of three aforementioned methods in a multiple-group allocation problem. This goal motivates the second case study, which is based on S-project [7], a study involving the entire population of a Brasilian city, Serrana/SP to access Coronavac vaccine efficacy.

Methods:

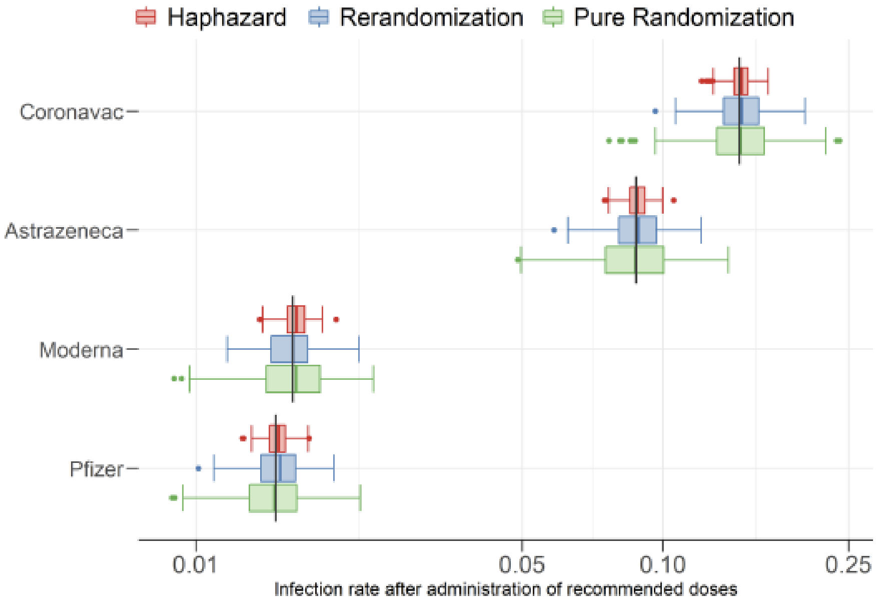
Haphazard method requires the empirical calibration of an auxiliary parameter that regulates the amount of noise added to the deterministic loss function modeling the unbalance between sample groups. This auxiliary parameter has to be large enough to achieve good decoupling (what corresponds to the haphazard character of the method) and, at the same time, small enough not to disrupt the loss function goal of achieving well-balanced groups (which corresponds to the intentional character of the method). Empirical pre-experiments were performed to define a proper calibration of this parameter and a suitable CPU time, which is sufficiently large to the solver to find suboptimal solutions but is also feasible for a large number of repetitions.

Computational experiments were performed with 300 repetitions of the sampling / allocation procedure, in order to access balance distributions for the three methods and the quality of each method in provide good estimations for SARS-CoV-2 prevalence (first case) and vaccine efficacy (second case). In the first case study, our performance experiments used a subset of 10 municipalities of the 133 in the original Epicovid19 study, covering a wide range of population size and characteristics. Following the original Epicovid19 protocol, a sample of 25 census sectors was selected at each municipality. In the second study case, we consider the scenario in which the population of each census sector receives the same vaccine and the 45 census sectors are allocated in four groups, of sizes (12, 11, 11, 11), using the three methods under study.

Results:

When comparing means of the 15 socio-demographic covariates used in the allocation procedures, differences are remarkably smaller for the haphazard allocations than for the rerandomization allocations, which, in turn, are remarkably smaller than for the pure randomization allocations. The same pattern is verified in all 10 municipalities.

On second study case, the results remained similar for the multiple groups formulation. Pure random allocation showed severe unbalanced that reached 6 square roots for some variables. Also, the vaccines distribution according to Haphazard Intentional Allocation provides the smallest variation around the real efficacies.



Discussion and conclusion:

Both the haphazard and rerandomization methods proved to be reliable and robust, outperforming the standard randomization method. Moreover, the haphazard method consistently outperformed the rerandomization method. This increased performance has a direct impact in the design and implementation of clinical trials allowing a target precision of experiment to be achieved with a reduced sample size. Reduced sample sizes immediately imply reduced costs of implementation as well as a faster conclusion of the experiment. Even more, reduced sample sizes help to mitigate ethical concerns related to potential side effects and other uncertain dangers that are inherent to any clinical trial. The haphazard intentional sampling method requires the formulation of Mixed-Integer Programming optimization problems and the use of numerical optimization software.

The aforementioned results motivate some topics for further research. In recent years, the performance of Mixed-Integer Quadratic Programming solvers has improved considerably. Hence, in subsequent articles, we shall explore the viability of direct use of the Mahalanobis loss function without recourse to the surrogate linear hybrid loss function. Moreover, we shall conduct a detailed comparative power analysis between all methods at hand and their variations.

This work was published in Entropy Journal [8]. The publication includes inference methods based on this sampling paradigm, which are part of Miguel Gabriel Ribeiro Miguel's Thesis.

References:

1. Lauretto, M.S.; Nakano, F.; Pereira, C.A.B.; Stern, J.M. Intentional Sampling by goal optimization with decoupling by stochastic perturbation. AIP Conf. Proc. 2012, 1490, 189–201.

2. Lauretto, M.S.; Stern, R.B.; Morgan, K.L.; Clark, M.H.; Stern, J.M. Haphazard intentional allocation and rerandomization to improve covariate balance in experiments. AIP Conf. Proc 2017, 1853, 050003.

3. Fossaluza, V.; Lauretto, M.S.; Pereira, C.A.B.; Stern, J.M. Combining optimization and randomization approaches for the design of clinical trials. In Interdisciplinary Bayesian Statistics; Springer: New York, NY, USA, 2015; pp. 173–184.

4. Lauretto, M.S.; Stern, R.B.; Ribeiro, C.O.; Stern, J.M. Haphazard intentional sampling techniques in network design of monitoring stations. Proceedings 2019, 33, 12.

5. Morgan, K.L.; Rubin, D.B. Rerandomization to balance tiers of covariates. J. Am. Stat. Assoc. 2015, 110, 1412–1421.

6. EPICOVID19. Available online: http://www.epicovid19brasil.org/?page_id=472 (accessed on 21 August 2020).

7. Instituto Butantan. S Project. Available online: https://projeto-s.butantan.gov.br/ (accessed on 14 December 2021).

8. Miguel, M.G.R.; Waissman, R.P.; Lauretto, M.S.; Stern, J.M. Haphazard Intentional Sampling in Survey and Allocation Studies on COVID-19 Prevalence and Vaccine Efficacy. *Entropy* 2022, *24*, 225.