The Journal of Open Source Software

samplics: a Python Package for selecting, weighting and analyzing data from complex sampling designs.

Mamadou S. Diallo[*1]

1 UNICEF, The United Nations Children's Fund

DOI: 10.21105/joss.03376

Software

- Review
- Repository
- Archive

Editor: Arfon Smith

Reviewers:

- @rchew
- @soodoku

Submitted: 13 December 2020
Published: 08 December 2021

Summary

Survey sampling techniques are used in various fields to obtain information about a large population by studying a fraction of its elements. As a result, it helps produce a significant portion of the official statistics by national governments and international organizations. For example, the Demographic and Health Survey (DHS) and the Multiple Cluster Indicator Survey (MICS) have been collecting demographic and health indicators for more than 35 years and 25 years respectively in over 100 countries. DHS and MICS are two of the primary sources of data for tracking the progress towards achieving the Sustainable Development Goals (SDGs). Similarly, numerous political and socio-economic branches of society rely on survey sampling to estimate the characteristics of populations of interest.

Until the initiation of `samplics`, Python did not have a library for analyzing complex survey samples similar to the R survey package (Lumley, 2004) and several commercial software such as SAS, SPSS, and Stata. `samplics` is a Python package developed to provide a comprehensive set of APIs to select random samples, adjust sample weights, produce design-based survey estimates, and predict small area parameters.

Statement of Need

`samplics` aims at providing a comprehensive statistical package for analyzing survey sample data. The primary target audiences are survey statisticians and other data analysts working with sample data obtained from complex survey designs. Data specialists can use this package to produce, analyze, and use official statistics. It also can help teach statistical concepts given the wide use of Python in Education and the simplicity of the `samplics` APIs.

When designing a survey, `samplics` can calculate sample sizes by stratum based on expected proportions and level of precision for the indicator of interest as well as measures of the complexity of the design such as survey design effects. After sample sizes are determined, `samplics` can calculate selection probabilities according to the sample selection strategy. To ensure the representativeness of the random sample, `samplics` will compute design weights and adjust them for non-response, post-stratification, and calibration. `samplics` provides Taylor-based and replication-based techniques for calculating population parameter estimates and associated measures of uncertainty. Finally, `samplics` has a small area estimation subpackage that can predict small area parameters. Note that `samplics` can be used independently for any of the steps described in the paragraph.

The sections below provide more details on the survey sampling techniques implemented in `samplics`.

*Corresponding author.

The Journal of Open Source Software

Survey Sampling Techniques

In large-scale surveys, often complex random mechanisms are used to select samples. Estimations obtained from such samples must reflect the complexity of the random mechanism to ensure correct approximations of the population parameters by sample estimates (Cochran, 1977), (Kish, 1965), and (Lohr, 2010). For Python users, `samplics` provides a comprehensive ecosystem of survey sampling techniques.

Sample Selection

The sample selection mechanism, a fundamental aspect of survey sampling, guides the statistical techniques employed to ensure the representativeness of the sample. In `samplics`, the focus is on random sampling techniques where units in the target population have a known probability of inclusion in the sample. let us assume that the target population has N units, and let us note π_i, the probability of unit i to be included in the sample. That is $P(I_i = 1) = \pi_i$, where I_i indicates whether unit i was included or not in the sample. The sample selection techniques implemented in `samplics` can be viewed as the combination of three key concepts: selection probability, stratification, and clustering. SRS results in an equal probability of selection for all sampling units, $P(I_i = 1) = \pi$. Stratification is a technique that consists of dividing the target population into m partitions, and sample selection is performed independently in each partition called stratum. Stratification is commonly used to divide the population, hence the sample, into homogeneous groups, e.g., income class, gender, ethnic group, etc. But it can also be used to control sample sizes by stratum; for example, governments often use stratification to ensure proper coverage of geographical administrative entities in the sample. Clustering is useful when a sample frame is unavailable for the units of interest or the operational cost of directly selecting the units and collecting the data is too high. In cluster sampling design, units of interest are grouped into clusters, and a sample of clusters is selected first (one-stage cluster sampling). Clustering can be done at multiple levels resulting in two-stage (or more) cluster sampling designs. Probability proportional to size (PPS) methods, e.g., Systematic, Brewer's method, Hanurav-Vijayan method, Murphy's method, and Rao-Sampford's method, are commonly used to select the clusters (Brewer & Hanif, 1983). Generally, cluster sampling leads to unequal probabilities of inclusion of sample units.

Sample Weighting

Sample weighting is the main mechanism used in surveys to formalize the representativeness of the sample. In complex surveys, sample weighting is composed of two main steps. First, the design (or base) weights are calculated as the inverse of the selection probabilities. Let us us assume that π_i is the final selection probability for unit i in the sample. Hence, $d_i = \frac{1}{\pi_i}$, where d_i is the design weight associated with unit i and can be interpreted as the average number of units in the target population represented by i including itself. Second, the design weights are adjusted to compensate for distortions due to shortcomings of the sample design implementation. Often, the initial weight adjustment corrects for nonresponse. This adjustment consists of defining response classes, then inflating the sample weights within response classes to compensate for the loss of sampled units due to nonresponse. In complex surveys, it is common to perform multiple sample weight adjustments. Hence, within a response class, the adjusted sample weights can be obtained as follows:

$$w_i = d_i * \prod_{k=1}^{K} a_k,$$

where a_k is the adjustment factor for step k. When reliable auxiliary information is available at the population level, poststratification and calibration can be used to adjust the sample weights. `samplics` also computes replicate weights, i.e., balanced repeated replication (BRR), bootstrap, and jackknife, often used to estimate complex parameters such as quantiles. (Valliant & Dever, 2018) provides a step-by-step guide for calculating sample weights for complex sampling designs.

Sample Estimation

As mentioned above, estimation of population parameters e.g., total, mean, median, coefficient of correlation, regression coefficients, etc., is one of the main objectives of surveys sampling. The sample weight is the primary mechanism for generalizing the sample estimate to approximate the equivalent population parameter. let us us us us consider the population parameter, total, defined as $Y = \sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{j=1}^{M_{hi}} w_{hij}y_{hij}$, where H is the number of strata, N_h is the number of primary sampling units (PSUs) in the population from stratum h and M_{hj} is the number of units from PSU i in stratum h. It follows that the sample estimate of the total is defined as

$$\hat{Y} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}y_{hij}I_{hij},$$

where n_h is the number of PSUs in the sample from stratum h and m_{hj} is the number of units in the sample from PSU i in stratum h. I_{hij} denotes the inclusion status of unit hij to the sample i.e., $I_{hij} = 1$ if unit hij is included in the sample otherwise $I_{hij} = 0$. The uncertainty estimation of the sample estimate must reflect the sampling mechanism and the weight adjustments. `samplics` provides two main frameworks for computing uncertainties, linearization (Taylor series) and replication.

Using the Taylor series method, the variance of the total is estimated as

$$\hat{V}(\hat{Y}) = \sum_{h=1}^{H}\frac{n_h(1-f_h)}{n_h-1}\sum_{i=1}^{n_h}(y_{hi\cdot} - \bar{y}_{h\cdot\cdot})^2,$$

where $y_{hi\cdot} = \sum_{j=1}^{m_{hi}} w_{hij}y_{hij}$, $\bar{y}_{h\cdot\cdot} = \sum_{i=1}^{n_h} y_{hi\cdot}/n_h$, and f_h is the sampling rate for the first stage from stratum h. The formula can be extended to the two-stage sampling design where second stage clusters or secondary sampling units (SSUs) are randomly selected from the PSUs prior to the selection of final sample units within selected SSUs. Under the two-stage sampling design, the Taylor series variance estimate of the total is

$$\hat{V}(\hat{Y}) = \sum_{h=1}^{H}\frac{n_h(1-f_h)}{n_h-1}\sum_{i=1}^{n_h}(y_{hi\cdot} - \bar{y}_{h\cdot\cdot})^2 + \sum_{h=1}^{H} f_h \sum_{i=1}^{n_h}(1-f_{hi})\frac{m_{hi}}{m_{hi}-1}\sum_{j=1}^{m_{hi}}(y_{hij\cdot} - \bar{y}_{hi\cdot\cdot})^2,$$

where $\hat{Y} = \sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{j=1}^{M_{hi}}\sum_{k=1} w_{hijk}y_{hij}I_{hijk}$, $y_{hij\cdot} = \sum_{k=1}^{m_{hij}} w_{hijk}y_{hijk}$, $\bar{y}_{hi\cdot\cdot} = \sum_{j=1}^{m_{hi}} y_{hij\cdot}/m_{hi}$, and f_{hi} is the sampling rate for the second stage of sampling from PSU i in stratum h. The variance estimation of the total can be extended to other population parameters that are functions of the sample weight. For example, the variance estimates of the mean and ratio are obtained by replacing y_{hijk} by $(y_{hijk} - \hat{\bar{Y}})/\hat{W}$ and $(y_{hijk} - \hat{R}x_{hijk})/\hat{X}$, respectively, where $\hat{\bar{Y}} = \hat{Y}/\hat{W}$, $\hat{W} = \sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{j=1}^{M_{hi}}\sum_{k=1} w_{hijk}$, $\hat{X} = \sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{j=1}^{M_{hi}}\sum_{k=1} x_{hijk}$ and $\hat{R} = \hat{Y}/\hat{X}$. Furthermore, the variance estimators in this section are extensible to domain analysis.

Suppose that θ is the population parameter of interest. Under the replication framework, multiple replicates, say R, of the sample are drawn following a given selection scheme (Efron

& Tibshirani, 1994) and (Wolter, 2007). For each replicate, a set of replicate weights is constructed by multiplying the sample weights by an adjustement factor a_{hi}. The resulting weights, called the replicate weights, are then used to obtain the R replicate estimates of the population parameter i.e. $\hat{\theta}_{(r)}, r = 1, ..., R$. The estimate of the variance of $\hat{\theta}$ is then given by

$$\hat{V}(\hat{\theta}) = \sum_{r=1}^{R} c_r (\hat{\theta}_{(r)} - \bar{\theta}_{(.)})^2,$$

where $\bar{\theta}_{(.)} = \frac{1}{R} \sum_{r=1}^{R} \hat{\theta}_{(r)}$. Both c_r and a_{hi} are specific to the replication method.

For **Bootstrap**, we have $c_r = 1/R$ and $a_{hi} = \frac{n_h}{n_h - 1} m^*_{hi}$, where m^*_{hi} is the number of times PSU hi was resampled. The replication factor c_r is the same across the strata, however the weight adjustment factors a_{hi} are stratum specific.

For **balanced repeated replication (BRR)** with Fay, we have

$$c_r = \frac{1}{R(1-f^2)} \text{ and } a_{hi} = \begin{cases} f & \text{if } Hd(hi) = -1 \\ 2 - f & \text{if } Hd(hi) = 1 \end{cases},$$

where Hd is the Hadarmard matrix. $f = 0$ corresponds to the default BRR method without the Fay adjustment. A Hadamard matrix is a square matrix whose entries are either $+1$ or -1 and whose rows are mutually orthogonal. In the case of BRR-Fay, both the replication factor c_r and the weight adjustment factor a_{hi} are constant across the strata.

For **Jackknife (delete-one)**, we have

$$c_r = \frac{n_{h'} - 1}{n_{h'}} \text{ and } a_{hi} = \begin{cases} \frac{n_{h'}}{n_{h'} - 1} & \text{if } h' = h \text{ and } i \text{ not dropped} \\ 0 & \text{if } h' = h \text{ and } i \text{ dropped} \\ 1 & \text{if } h' \neq h \end{cases}.$$

This formula is easily generalizable to the non stratified design ($H = 1$) by replacing $n_{h'}$ by n and dropping the case $h' \neq h$. The replication factor c_r is stratum specific in the case of Jackknife, which allows a finite-population correction by stratum.

Small Area Estimation (SAE)

When the sample size is insufficient to produce reliable/stable domain level estimates, SAE techniques model the output variable of interest and produce domain level estimates. These domains are referred to as small areas. For the most part, the SAE models are applications of mixed models, see (McCulloch et al., 2008) and (Rao & Molina, 2015) for more details on mixed models. Mixed models allow accounting for the between-area variations by using random area-specific effects and the auxiliary variables contribution through the fixed effects. Small Area Estimation models are generally classified into two classes: the Area-level and the Unit-level models (Rao & Molina, 2015).

Area-level Model

As mentioned above, the Areal-level approach models the variables of interest using known auxiliary information at some aggregated level(s). A common representation of the basic Area-level model is

$$\hat{\theta}\_d = \mathbf{x}_d^T u_d + e_d, \quad d = 1, ..., m,$$

where $u_d \overset{iid}{\sim} N(0, \sigma_u^2)$ and $e_d \overset{iid}{\sim} N(0, \psi_d)$ are independent. The sampling variance ψ_d is assumed to be known; in a real survey this quantity is unknown and must be estimated, then

treated as known for the purpose of deriving the estimates. Under the basic Area-level model, the best predictor (best in the sense of minimizing the mean squared error)) of θ is

$$\hat{\theta}_d^B = (1 - B_d)\hat{\theta}\_d + B_d \mathbf{x}_d^T \tilde{\beta} \quad d = 1, ..., m,$$

where $B_d = \psi_d/(\sigma_u^2 + \psi_d)$ and $\tilde{\beta}$ is the best linear unbiased estimator of β. The empirical best (EB), or empirical Bayes, predictor $\hat{\theta}_d^{EB}$ is obtained by replacing the unknown parameter in the expression of $\hat{\theta}_d^B$ by their estimators. The parameters of the model, β and σ_u^2 are estimated using method of moment (MOM), maximum likelihood (ML), restricted maximum likelihood (REML), or other suitable techniques. The EB estimator is a weighted average of the survey (direct) estimator $\hat{\theta}_d$ and the regression predictor $\mathbf{x}_d^T \tilde{\beta}$ where the weight is $\hat{B}_d = \psi_d/(\hat{\sigma}_u^2 + \psi_d)$.

Unit-level model

The Unit-level models fit the data at the atomic individual unit level. The basic Unit-level model can be formally defined as follows:

$$\mathbf{Y}_{dj} = \mathbf{x}_{dj}^T \beta + u_d + e_{dj}, \quad j = 1, ..., N_d, \quad d = 1, ..., m,$$

where $u_d \overset{iid}{\sim} N(0, \sigma_u^2)$ and $e_{dj} \overset{iid}{\sim} N(0, \sigma_e^2)$ are independent random normal variables, $\mathbf{x}_{dj}$ is the vector of auxiliary variables, d designates the small area and j designates the the unit within the small-area d. The best linear unbiased predictor (BLUP) estimator of the small area mean $\theta_d = \hat{\mathbf{X}}_d^T \beta + u_d$ is

$$\hat{\theta}_d^B = \bar{\mathbf{X}}_d^T \tilde{\beta} + \gamma_d(\bar{y}_d - \bar{\mathbf{x}}_d^T \tilde{\beta})$$

where $\gamma_d = \frac{\sigma_u^2}{\sigma_e^2 + n_d \sigma_u^2}$, the estimator $\tilde{\beta}$ is the best linear unbiased estimator of β, and n_d is the sample size for small area d. The empirical best linear predictor, $\hat{\theta}_d^{EB}$, is obtained by replacing the model parameters by their estimators in the expression of $\hat{\theta}_d^B$. (Elbers et al., 2003) extends the basic Unit-level model by relaxing the normal distribution of the errors with an empirical semi-parametric model. This model has been used by the World Bank to estimate small area poverty indices. Furthermore, (Molina & Rao, 2010) provide a parametric approach for estimating complex small area parameters such as poverty indices.

Acknowledgments

I would like to acknowledge Dr. Kathleen Wannemuehler, Mr. John Wagai, and Mr. Chibwe Lwamba for their comments on an earlier version of the package that significantly improved `samplics` and its documentation. My Ph.D. research advisor and mentor, Professor Emeritus J.N.K. Rao has continuously kept me engaged in the research community well after I obtained my Ph.D. and moved to the corporation, which allowed me to dig deep into many of the topics implemented into this Python survey library, `samplics`.

References

Brewer, K. R. W., & Hanif, M. (1983). *Sampling with unequal probabilities*. Springer-Verlag New York, Inc. https://doi.org/10.1007/978-1-4684-9407-5

Cochran, W. G. (1977). *Sampling techniques, 3rd edn*. John Wiley & Sons, Inc. ISBN: 978-0-471-16240-7

Efron, B., & Tibshirani, R. J. (1994). *An Introduction to the Boostrap*. Chapman & Hall/CRC. ISBN: 978-0-412-04231-7

Elbers, C., Lanjouw, J. O., & Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. *Econometrica*, *71*, 355–364. https://doi.org/10.1111/1468-0262.00399

Kish, L. (1965). *Survey sampling*. John Wiley & Sons, Inc. ISBN: 978-0-471-10949-5

Lohr, S. L. (2010). *Sampling: Design and analysis, 2nd edn*. Cengage Learning, Inc. ISBN: 978-0-367-27341-5

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, *9*(1), 1–19. https://doi.org/10.18637/jss.v009.i08

McCulloch, C. E., Searle, S. R., & Neuhaus, J. M. (2008). *Generalized, linear, and mixed models*. John Wiley; Sons. ISBN: 978-0-470-07371-1

Molina, I., & Rao, J. N. K. (2010). Small area estimation of poverty indicators. *Canad. J. Statist.*, *38*, 369–385. https://doi.org/10.1002/cjs.10051

Rao, J. N. K., & Molina, I. (2015). *Small area estimation, 2nd edn*. John Wiley & Sons, Hoboken, New Jersey. ISBN: 978-1-118-73578-7

Valliant, R., & Dever, J. A. (2018). *Survey weights: A step-by-step guide to calculation*. Stata Press. ISBN: 978-1-59718-260-7

Wolter, K. M. (2007). *Introduction to variance estimation, 2nd edn*. Springer-Verlag New York, Inc. ISBN: 978-0-387-35099-8