Summary

Survey sampling techniques are used in various fields to obtain information about a large population by studying a fraction its elements. A significant portion of the official statistics by national goverments and international organizations is obtained from sample survey. For example, the Demographic and Health Survey (DHS) and the Multiple Cluster Indicator Survey (MICS) have be 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 main 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 sample surveys to estimate characteristics of populations of interest.

Until now, Python did not have a library for analyzing complex survey samples; similar to R survey package and several commercial software such as SAS, SPSS, Stata, and more. *samplics* is a Python package developed to provide a comprehensive set of python code to select random samples, adjust sample weights, produce design-based survey estimates, and predict small area parameters.

Survey Sampling Techniques

In large scale surveys, often complex random mechanisms are used to select samples. Estimations obtained from such samples must reflect the random mechanism to ensure accurate calculations. Samplics implements a set of sampling techniques for complex survey designs.

Sample Selection

The sample selection mechanism is fundamental aspect of survey sampling that 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 has a know probability of inclusion in the sample. Let assume that the target population has N units and let's note π_i, the probability of unit i to be included in the sample. That is $P(I_i = 1) = \pi_i$, where $I_i = 1$ is unit i is selected (included in the sample) and $I_i = 0$ otherwise. Note that

$P(I_i = 0) = 1 - \pi_i$. The sample selection techniques implemented in *samplics* can be viewed as the result of the implementation of three key concepts: simple random sample, stratification, and clustering. Sample random selection (SRS) is the simplest type of probability sample in which all the samples of same size, say n, have the same probability of realization. SRS results in an equal probability of selection for all sampling units, $P(I_i = 1) = \pi_i$. 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 homogenious groups e.g. by 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 not available for the units of interest or the operational cost of directly selecting the units and collection data is too high. In a cluster sample, 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 higher) cluster sampling. Suppose you want to interview hospital patients on the general quality of the care they received. One sampling option could be to first select a random sample of hospitals; then from the selected hospital list of patients randomly include a subset in the sample for data collection. 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. Generally, cluster sampling leads to unequal probabilities of inclusion of sample units.

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 base (or design) weights are calculated as the inverse of the probabilities of selection. Let assume that π_i is the final probability of selection of 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 base weights are adjusted to compensate for distortions due to shortcomings of the the sample design implementation. The most common adjustment is the weight adjustment due to nonresponse. This adjustment consists of defining response classes, then inflate 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 adjustment steps. Hence, within 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 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.

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 unknown population parameter. Let's consider the population parameter, total, defined as $Y = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hj}} w_{hij} y_{hij}$, 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

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

where n_h is the number of PSUs in stratum h and m_{hj} is the number of units 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

$$\hat{V}(\hat{Y}) = \sum *h = 1_H \frac{n_h(1 - f_h)}{n_h - 1} \sum *i = 1_{n*h}(y * hi. - \bar{y}_{h..})_2,$$

where $y_{hi.} = \sum *j = 1_{m*hi} w * hijy * hij$, $\bar{y}_{h..} = \sum_{i=1}^{n_h} y_{hi.}/n_h$, and f_h is the sampling rate for the first stage of sampling 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 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

$$\hat{V}(\hat{Y}) = \sum *h = 1_H \frac{n_h(1 - f_h)}{n_h - 1} \sum *i = 1_{n*h}(y * hi. - \bar{y}_{h..})_2 + \sum_{h=1}^{H} f * h \sum *i = 1_{n*h}(1 -$$

where $\hat{Y} = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} \sum_{k=1} w_{hijk} y_{hij} I_{hijk}$, $y_{hij.} = \sum_{k=1}^{m_{hij}} w_{hijk} y_{hijk}$, $\bar{y}_{hi..} = \sum_{j=1}^{m_{hi}} y_{hij.}/m_{hi}$, 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. 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 and defined as follows

- 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} stratum specific.
- For balanced repeated replication (BRR) with Fay, we have $c_r = \frac{1}{R(1-f)^2}$ 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. $f = 0$ correspond 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 $c_r = \frac{n_{h'} - 1}{n_{h'}}$ 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 not large enough to produce reliable / stable domain level estimates, SAE techniques can be used to model the output variable of interest and produce domain level estimaetes. These domains are referred to as small areas. The SAE models are for the most part part applications of mixed models, see \citet{mcculloch08} and Chapter 5 of \citet{rao15} for more details on mixed models. Mixed models allow to account 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.

Area-level Model

As mentioned above, the Areal-level approach models the variables of interest using known auxiliary information at some aggregated level(s), see discussion about auxiliary information in Section \ref{aux_vars}. The first use of the Area-level model is found in \citet{fay79} in the context of estimating per capita income (PCI) for small places in the United States with population less than 1,000. This model is referred to as the \textit{Fay-Herriot} model or the basic Area-level model. A common representation of the \textit{Fay-Herriot} model is

$$\hat{\theta}^{\_}d = \boldsymbol{x}^{T}_{d} u_{d} + e_{d}, \quad d = 1, ..., m,$$

where where $u_{d} \sim N(0, \sigma^{2}_{u})$ and $e_{d} \sim N(0, \psi_{d})$ 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 small estimates. The other parameters of the model, β and σ^{2}_{u} are estimated using method of moment (MOM), maximum likelihood (ML), restricted maximum likelihood (REML), or other suitable techniques.

Under the \textit{Fay-Herriot} model, the best predictor (best in the sense of minimizing the mean squared error)) of θ is

$$\hat{\theta}^{B}_{d} = (1 - B_{d})\hat{\theta}^{\_}d + B_{d} \boldsymbol{x}^{T}_{d} \tilde{\boldsymbol{\beta}} \quad d = 1, ..., m,$$

where $B_{d} = \psi_{d} / (\sigma^{2}_{u} + \psi_{d})$ and $\tilde{\boldsymbol{\beta}}$ is the best linear unbiased estimator of $\boldsymbol{\beta}$. 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 EB estimator is a weighted average of the survey (direct) estimator $\hat{\theta}^d$ and the regression predictor $\boldsymbol{x}_d^T\tilde{\boldsymbol{\beta}}$ where the weight is $\hat{B}^d = \psi_d/(\hat{\sigma}_u^2 + \psi_d)$.

Unit-level model

The Unit-level framework models the data at the atomic individual unit level. Hence, both the variable of interest and the auxiliary variables should be available at the unit level which can be an additional implementation challenge. The Unit-level model was first introduced by \citet{battese88} for the prediction of corn and soybean crop areas for 12 counties in north-central Iowa. The basic Unit-level model can be formally defined as follows:

$$\boldsymbol{Y}_{dj} = \boldsymbol{x}_{dj}^T \beta + u * d + e * dj, \quad j = 1, ..., N * d, \quad d = 1, ..., m,$$

where $u_d \sim N(0, \sigma_{2_u})$ and $e * dj \sim N(0, \sigma_2 * e)$ are independent random normal variables, $\boldsymbol{x} * dj$ 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 $\theta_d = \hat{\boldsymbol{X}}_{\boldsymbol{d}}^{\boldsymbol{T}}\boldsymbol{\beta} + u_d$ is

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

where $\gamma_d = \frac{\sigma_2}{\sigma_{2_e} + n_d\sigma_{2_u}}$, the estimator $\tilde{\boldsymbol{\beta}}$ is the best linear unbiased estimator of $\boldsymbol{\beta}$, 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 $\hat{\theta}_d^B$. \citet{elbers03} 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, \citet{molina10} 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 greatly 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 corporation which allowed me to dig deep into many of the topics implemented into this Python survey library, *samplics*.

References