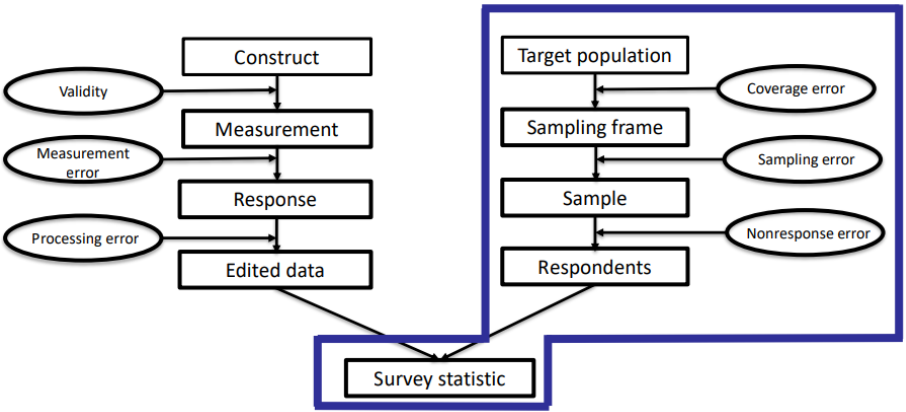
Survey weighting and estimation with R package survey

Total Survey Error Framework



Missing data

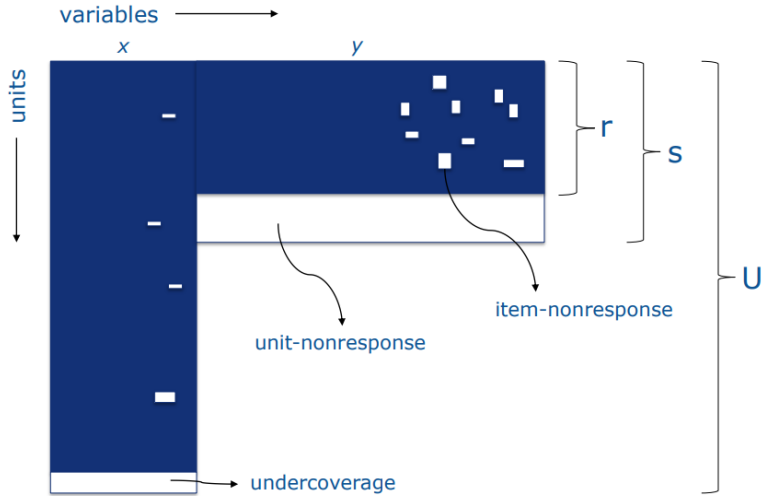
Possible reasons for missing data:

- unit not in sampling frame
- unit not sampled
- unit or item-nonresponse
- imperfect linkage

Adverse effects:

- reduced precision of survey statistics
- biased results, if insufficiently accounted for missing data mechanism (especially challenging for unknown mechanisms such as nonresponse)

Missing data patterns



Imputation and weighting

Imputation to deal with item-nonresponse

- Fill in item-nonresponse to complete the data for all respondents
- Typically missing for a small fraction of respondents

Weighting to deal with unit-nonresponse and generalize to the population

- Assign weights to all respondents
- More practical and often safer than (mass-)imputation

Missing data mechanisms

MCAR (Missing Completely At Random)

Example: simple random sample with 100% response. Unweighted sample means can be used to estimate population means (at least for sufficiently large sample sizes)

MAR (Missing At Random)

Example: stratified sample with 100% response. Include stratum indicators in the analysis to account for the missingness.

NMAR (Not Missing At Random)

Usual case due to unknown mechanisms such as nonresponse

Imputation and weighting models

need to account for

- auxiliary variables that render the missing data mechanism as MAR as possible (design variables, known demographic variables explaining non-response)
- intended analyses/estimations to be conducted with completed/weighted data (domain indicators, regression variables)
- auxiliary variables related to target variables for variance reduction

Survey weighting and estimation, preconditions

Assume that

- we have a completed dataset for respondents, with target variables and auxiliary variables
- auxiliary variables are available for the complete target population, or at least for all respondents together with good estimates of their population totals

Calibration weighting

Let d_i be design weights, i.e. inverse sampling probabilities.

For a set of auxiliary variables x with (accurately) known population totals t_x find the final weights w_i that minimize

$$\sum_{i \in r} \mathrm{dist}(w_i, d_i)$$

subject to

$$\sum_{i \in r} w_i x_i = t_x$$

Generalized Regression Estimator

GREG distance function: $\mathrm{dist}(w_i, d_i) = (w_i - d_i)^2/d_i$

This yields GREG weights w_i (closed form expression)

Resulting GREG estimator for a population total:

$$\hat{t}_y^{\mathrm{GREG}} = \sum_{i \in r} w_i y_i = \hat{t}_y^{\mathrm{HT}} + \hat{\beta}'(t_x - \hat{t}_x^{\mathrm{HT}})$$

where $\hat{t}_z^{\mathrm{HT}} \equiv \sum_{i \in r} d_i z_i$ is the Horvitz-Thompson estimator for the population total of variable z

Survey weighting software

- Sudaan
- Stata (svy, svyset with options rake, regress)
- SAS macros
- CRAN R packages survey, sampling
- Other R packages: ReGenesees (ISTAT),
- brascula (CBS, under development)

Most of them also support variance estimation

R package survey: minimal workflow

Specify the sampling design:

```
des <- svydesign(ids = ~ 1, data = LFSdat)
```

Calibrate to adjust design weights for nonresponse:

```
cal <- calibrate(des, ~ sex*ageclass + province + income,
         population = poptotals)
```

Compute population estimates and corresponding standard errors:

```
est <- svymean(~ unemployed, cal)
```

Survey weighting with R

Demonstration of R package survey

- GREG weighting to reduce non-response bias
- estimation of population totals and means
- variance estimation

Limitations of survey weighting

- Small sample sizes for subpopulations of interest. Here the main issue is variance, not bias. Robust estimation.
- Measurement error
- NMAR missingness

More sophisticated modelling needed to solve such issues. A single set of weights usually no longer suffices.

Further information

- Survey package documentation and website
  http://r-survey.r-forge.r-project.org/survey/
- Lumley, Complex Surveys: A Guide to Analysis Using R (Wiley, 2011)
- Lohr, Sampling - Design and Analysis, (CRC, 2022)