The R Project - The Use of R in Official Statistics

UROS 2019

Just beat it

Bethel Extended Allocation

for Two-stage Sampling Package

Alessio Guandalini

S. Falorsi, A. Fasulo, D. Pagliuca, M.D. Terribili

ISTAT, Italy

beat

stands for

Bethel Extended Allocation for Two-stage

beat

stands for

Bethel Extended Allocation for Two-stage

- implements the extension of the Neyman (1934) – Tschuprow (1923) allocation method to the case of several variables, **adopting a generalization of the Bethel's proposal** (1989)

beat

stands for

Bethel Extended Allocation for Two-stage

- implements the extension of the Neyman (1934) – Tschuprow (1923) allocation method to the case of several variables, **adopting a generalization of the Bethel's proposal** (1989)
- determines the **sample allocation** in the **multivariate** and **multi-domains** case of estimates for **two-stage stratified samples**

Functions in beat

- beth
- beth2st
- bethcv

Methodological background

- **Neyman-Tschuprow** (optimal allocation)

$$ {}_{opt}n_h = n \frac{w_h \ {\sigma_{y_h}}^2}{\sum_{h=1}^{L} w_h \ {\sigma_{y_h}}^2} $$

- **Neyman-Tschuprow** (optimal allocation) with cost constraints

$$ {}_{opt}n_h = C \cdot \frac{w_h \ \sigma_{y_h} / \sqrt{c_h}}{\sum_{h=1}^{L} w_h \ \sigma_{y_h} \ \sqrt{c_h}} $$

where $C = C_0 + \sum_{h=1}^{L} n_h \ c_h$, usually $C_0 = 0$ and $w_h = N_h / N$

Methodological background

Multivariate optimal allocation

- more than one relevant variable for one type of domain [Bethel, 1989]
- more than one relevant variable for many types of domain [Falorsi *et al.*, 1998]

$$\begin{cases} C = min & \\ \sigma\left(\hat{Y}_{j,d}\right) \leq \delta\left(\hat{Y}_{j,d}\right) & \begin{array}{l} j = 1, \ldots, J \\ d = 1, \ldots, D \end{array} \end{cases}$$

Methodological background

Multivariate optimal allocation

- more than one relevant variable for one type of domain [Bethel, 1989]
- more than one relevant variable for many types of domain [Falorsi *et al.*, 1998]

$$\begin{cases} C = min & \\ \sigma\left(\hat{Y}_{j,d}\right) \leq \delta\left(\hat{Y}_{j,d}\right) & \begin{array}{r} j = 1, \ldots, J \\ d = 1, \ldots, D \end{array} \end{cases}$$

Generalised software used in ISTAT (`MAUSS-R`)

Context

Complex sampling design

Context

Complex sampling design

- **PSU** (Primary Stage Units) [*usually stratified*]
  SSU (Secondary Stage Units)

Context

Complex sampling design

- **PSU** (Primary Stage Units) [*usually stratified*]
  SSU (Secondary Stage Units)
- **PSU** can be
  - **SR** (Self-Representative)
  - **NSR** (Non Self-Representative)

Context

Complex sampling design

- **PSU** (Primary Stage Units) [*usually stratified*]
  SSU (Secondary Stage Units)
- **PSU** can be
  - **SR** (Self-Representative)
  - **NSR** (Non Self-Representative)
- two sampling designs coexist toghether
  - cluster $\longrightarrow$ SR
  - two-stage $\longrightarrow$ NSR

Methodological background

Design effect

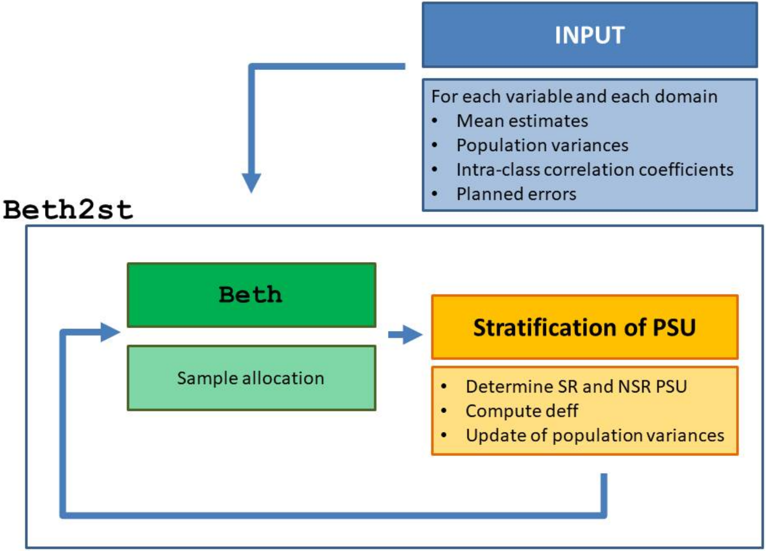
How much the sampling variance under the adopted sampling design is inflated with respect to SRS, on equal sample size

$$
\begin{aligned}
\mathit{deff}(\hat{Y}_{\cdot}) &= \frac{\mathrm{var}(\hat{Y}_{\cdot})}{\mathrm{var}(\hat{Y}_{\cdot,SRS})} \\
&= \frac{N_{SR}^{2}}{n_{SR}}(1 + (\rho_{\cdot,SR}\ (b_{SR} - 1)) + \frac{N_{NSR}^{2}}{n_{NSR}}(1 + (\rho_{\cdot,NSR}\ (b_{NSR} - 1))
\end{aligned}
$$

where

$$
\begin{aligned}
\rho_{\cdot} &= \text{intra-class correlation coefficient} \\
b_{\cdot} &= \text{average size of clusters in the domain}
\end{aligned}
$$

beat



INPUT

For each variable and each domain
- Mean estimates
- Population variances
- Intra-class correlation coefficients
- Planned errors

Beth2st

Beth

Sample allocation

Stratification of PSU

- Determine SR and NSR PSU
- Compute deff
- Update of population variances

beth2st

```
beth2st (strata, errors, psufile, rho, effst=NULL, ...)
```

beth2st

```
beth2st (strata, errors, psufile, rho, effst=NULL, ...)
```

- `strata` - `data.frame` with information on each strata
  - mean estimates (previous survey or other source)
  - population variance (previous survey or other source)
  - unitary cost per interview
  - census strata (1=yes, 0=no)
  - minimum number of interviews in PSU
  - minimun number of PSU
  - size of SSU (Δ)
- `errors` - `data.frame` with planned errors for each variable and each domain

beth2st - input

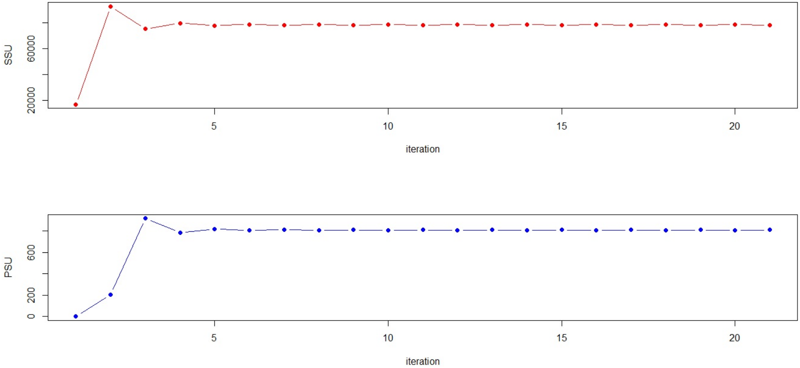
```
beth2st (strata, errors, psufile, rho, effst=NULL, ...)
```

- psufile - data.frame with information on each PSU on
  - strata
  - population size
- effst - data.frame with the "effect of the estimator" for each variable and each domain
- ...

beth2st - output

```
iteraction  PSU-SR  PSU-NSR  PSU-Total    SSU
         0       0        0          0  17027
         1      12      191        203  92446
         2     163      755        918  75040
         3     113      672        785  79555
         ⋮       ⋮        ⋮          ⋮      ⋮
        18     130      680        810  77874
        19     124      682        806  78356
        20     130      680        810  77875
```

beat - output


SSU
20000
60000
5
10
15
20
iteration
PSU
0
200
600
5
10
15
20
iteration

beth2st - output

- iteration - data.frame with the information printed on the screen by beth2st
- alloc - data.frame with sample size with proportional, uniform and optimal allocation (can be used as input for FS4)
- expected - data.frame with the expected error for each variable in each domain
- sensivity - data.frame that can help the evaluation of the allocation
- deft - data.frame with the square root of deff for each variabile in each domain

Conclusions

- beat is a very specific package, useful for the allocation in complex sampling designs
- main users could be, of course, NSIs but in general everyone who need to implement complex sampling design

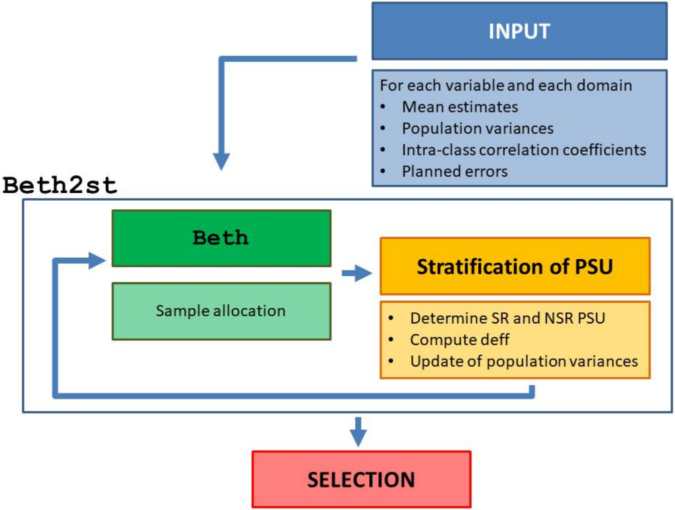
Conclusions

- beat is a very specific package, useful for the allocation in complex sampling designs
- main users could be, of course, NSIs but in general everyone who need to implement complex sampling design

Further perspectives

- **make the package available on** CRAN
- take into account also no-response
- integrate beat (allocation for complex sampling design) with FS4 (selection of PSU for complex sampling designs)

beat



INPUT

For each variable and each domain
- Mean estimates
- Population variances
- Intra-class correlation coefficients
- Planned errors

Beth2st

Beth

Sample allocation

Stratification of PSU

- Determine SR and NSR PSU
- Compute deff
- Update of population variances

SELECTION

References

- BETHEL, James. Sample allocation in multivariate surveys. *Survey methodology*, 1989, 15.1: 47-57.
- COCHRAN, William G. *Sampling techniques*. John Wiley Sons, 2007.
- ISTAT. *MAUSS-R Multivariate Allocation of Units in Sampling Surveys* (User's Manual). https://www.istat.it/it/files/2011/02/user_and_methodological_manual.pdf (Web, 22 March 2019). 2013.
- NEYMAN, Jerzy. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 1934, 97.4: 558-625.
- TSCHUPROW, Al A. On the mathematical expectation of the moments of frequency distributions in the case of correlated observations (Chapters 4-6). *Metron*, 1923, 2: 646-683.

- Stefano Falorsi, stfalors(at)istat.it
- Andrea Fasulo, fasulo(at)istat.it
- Alessio Guandalini, alessio.guandalini(at)istat.it
- Daniela Pagliuca, pagliuca(at)istat.it
- Marco D. Terribili, terribili(at)istat.it