Package ‘UPSvarApprox’

January 20, 2025

Type Package

Title Approximate the Variance of the Horvitz-Thompson Total Estimator

Version 0.1.4

Date 2023-08-26

Description Variance approximations for the
Horvitz-Thompson total estimator in Unequal Probability Sampling
using only first-order inclusion probabilities.
See Matei and Tillé (2005) and Haziza, Mecatti and Rao (2008) for details.

License GPL-3

Encoding UTF-8

RoxygenNote 7.2.3

BugReports https://github.com/rhobis/UPSvarApprox/issues

NeedsCompilation no

Author Roberto Sichera [aut, cre]

Maintainer Roberto Sichera <rob.sichera@gmail.com>

Repository CRAN

Date/Publication 2023-08-27 07:00:07 UTC

Contents

UPSvarApprox-package 2
approx_var_est 2
Var_approx 7

Index **10**

UPSvarApprox-package *UPSvarApprox: Approximate the variance of the Horvitz-Thompson estimator*

Description

Variance approximations for the Horvitz-Thompson total estimator in Unequal Probability Sampling using only first-order inclusion probabilities. See Matei and Tillé (2005) and Haziza, Mecatti and Rao (2008) for details.

Variance approximation

The package provides function `Var_approx` for the approximation of the Horvitz-Thompson variance, and function `approx_var_est` for the computation of approximate variance estimators. For both functions, different estimators are implemented, see their documentation for details.

Author(s)

Maintainer: Roberto Sichera <rob.sichera@gmail.com>

References

Matei, A.; Tillé, Y., 2005. Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. Journal of Official Statistics 21 (4), 543-570.

Haziza, D.; Mecatti, F.; Rao, J.N.K. 2008. Evaluation of some approximate variance estimators under the Rao-Sampford unequal probability sampling design. Metron LXVI (1), 91-108.

See Also

Useful links:

- Report bugs at https://github.com/rhobis/UPSvarApprox/issues

approx_var_est *Approximated Variance Estimators*

Description

Approximated variance estimators which use only first-order inclusion probabilities

Usage

```
approx_var_est(y, pik, method, sample = NULL, ...)
```

Arguments

`y`	numeric vector of sample observations
`pik`	numeric vector of first-order inclusion probabilities of length N, the population size, or n, the sample size depending on the chosen method (see Details for more information)
`method`	string indicating the desired approximate variance estimator. One of "Deville1", "Deville2", "Deville3", "Hajek", "Rosen", "FixedPoint", "Brewer1", "HartleyRao", "Berger", "Tille", "MateiTille1", "MateiTille2", "MateiTille3", "MateiTille4", "MateiTille5", "Brewer2", "Brewer3", "Brewer4".
`sample`	Either a numeric vector of length equal to the sample size with the indices of sample units, or a boolean vector of the same length of `pik`, indicating which units belong to the sample (`TRUE` if the unit is in the sample, `FALSE` otherwise. Only used with estimators of the third class (see Details for more information).
`...`	two optional parameters can be modified to control the iterative procedures in methods `"MateiTille5"`, `"Tille"` and `"FixedPoint"`: `maxIter` sets the maximum number of iterations to perform and `eps` controls the convergence error

Details

The choice of the estimator to be used is made through the argument `method`, the list of methods and their respective equations is presented below.

Matei and Tillé (2005) divides the approximated variance estimators into three classes, depending on the quantities they require:

1. First and second-order inclusion probabilities: The first class is composed of the Horvitz-Thompson estimator (Horvitz and Thompson 1952) and the Sen-Yates-Grundy estimator (Yates and Grundy 1953; Sen 1953), which are available through function `varHT` in package `sampling`;
2. Only first-order inclusion probabilities and only for sample units;
3. Only first-order inclusion probabilities, for the entire population.

Haziza, Mecatti and Rao (2008) provide a common form to express most of the estimators in class 2 and 3:

$$\widehat{var}(\hat{t}_{HT}) = \sum_{i \in s} c_i e_i^2$$

where $e_i = \frac{y_i}{\pi_i} - \hat{B}$, with

$$\hat{B} = \frac{\sum_{i \in s} a_i (y_i/\pi_i)}{\sum_{i \in s} a_i}$$

and a_i and c_i are parameters that define the different estimators:

- `method="Hajek"` [Class 2]

$$c_i = \frac{n}{n-1}(1-\pi_i); \quad a_i = c_i$$

- `method="Deville2"` [Class 2]

$$c_i = (1-\pi_i)\left\{1 - \sum_{j\in s}\Big[\frac{1-\pi_j}{\sum_{k\in s}(1-\pi_k)}\Big]^2\right\}^{-1}; \quad a_i = c_i$$

- `method="Deville3"` [Class 2]

$$c_i = (1-\pi_i)\left\{1 - \sum_{j\in s}\Big[\frac{1-\pi_j}{\sum_{k\in s}(1-\pi_k)}\Big]^2\right\}^{-1}; \quad a_i = 1$$

- `method="Rosen"` [Class 2]

$$c_i = \frac{n}{n-1}(1-\pi_i); \quad a_i = (1-\pi_i)log(1-\pi_i)/\pi_i$$

- `method="Brewer1"` [Class 2]

$$c_i = \frac{n}{n-1}(1-\pi_i); \quad a_i = 1$$

- `method="Brewer2"` [Class 3]

$$c_i = \frac{n}{n-1}\Big(1-\pi_i+\frac{\pi_i}{n} - n^{-2}\sum_{j\in U}\pi_j^2\Big); \quad a_i = 1$$

- `method="Brewer3"` [Class 3]

$$c_i = \frac{n}{n-1}\Big(1-\pi_i-\frac{\pi_i}{n} - n^{-2}\sum_{j\in U}\pi_j^2\Big); \quad a_i = 1$$

- `method="Brewer4"` [Class 3]

$$c_i = \frac{n}{n-1}\Big(1-\pi_i-\frac{\pi_i}{n-1} + n^{-1}(n-1)^{-1}\sum_{j\in U}\pi_j^2\Big); \quad a_i = 1$$

- `method="Berger"` [Class 3]

$$c_i = \frac{n}{n-1}(1-\pi_i)\left[\frac{\sum_{j\in s}(1-\pi_j)}{\sum_{j\in U}(1-\pi_j)}\right]; \quad a_i = c_i$$

- `method="HartleyRao"` [Class 3]

$$c_i = \frac{n}{n-1}\Big(1-\pi_i - n^{-1}\sum_{j\in s}\pi_i + n^{-1}\sum_{j\in U}\pi_j^2\Big); \quad a_i = 1$$

Some additional estimators are defined in Matei and Tillé (2005):

- `method="Deville1"` [Class 2]

$$\widehat{var}(\hat{t}_{HT}) = \sum_{i\in s}\frac{c_i}{\pi_i^2}(y_i - y_i^*)^2$$

  where

$$y_i^* = \pi_i\frac{\sum_{j\in s} c_j y_j/\pi_j}{\sum_{j\in s} c_j}$$

  and $c_i = (1-\pi_i)\frac{n}{n-1}$

- method="Tille" [Class 3]

$$\widehat{var}(\hat{t}_{HT}) = \left(\sum_{i \in s} \omega_i\right) \sum_{i \in s} \omega_i (\tilde{y}_i - \bar{\tilde{y}}_\omega)^2 - n \sum_{i \in s} \left(\tilde{y}_i - \frac{\hat{t}_{HT}}{n}\right)^2$$

where $\tilde{y}_i = y_i/\pi_i, \omega_i = \pi_i/\beta_i$ and $\bar{\tilde{y}}_\omega = \left(\sum_{i \in s} \omega_i\right)^{-1} \sum_{i \in s} \omega_i \tilde{y}_i$

The coefficients β_i are computed iteratively through the following procedure:

1. $\beta_i^{(0)} = \pi_i, \ \forall i \in U$
2. $\beta_i^{(2k-1)} = \frac{(n-1)\pi_i}{\beta^{(2k-2)} - \beta_i^{(2k-2)}}$
3. $\beta_i^{2k} = \beta_i^{(2k-1)} \left(\frac{n(n-1)}{(\beta^(2k-1))^2 - \sum_{i \in U} (\beta_k^{(2k-1)})^2} \right)^{(1/2)}$

with $\beta^{(k)} = \sum_{i \in U} \beta_i^i, \ k = 1, 2, 3, \dots$

- method="MateiTille1" [Class 3]

$$\widehat{var}(\hat{t}_{HT}) = \frac{n(N-1)}{N(n-1)} \sum_{i \in s} \frac{b_i}{\pi_i^3} (y_i - \hat{y}_i^*)^2$$

where

$$\hat{y}_i^* = \pi_i \frac{\sum_{i \in s} b_i y_i / \pi_i^2}{\sum_{i \in s} b_i / \pi_i}$$

and the coefficients b_i are computed iteratively by the algorithm:

1. $$b_i^{(0)} = \pi_i (1 - \pi_i) \frac{N}{N-1}, \ \forall i \in U$$
2. $$b_i^{(k)} = \frac{(b_i^{(k-1)})^2}{\sum_{j \in U} b_j^{(k-1)}} + \pi_i (1 - \pi_i)$$

a necessary condition for convergence is checked and, if not satisfied, the function returns an alternative solution that uses only one iteration:

$$b_i = \pi_i (1 - \pi_i) \left(\frac{N \pi_i (1 - \pi_i)}{(N-1) \sum_{j \in U} \pi_j (1 - \pi_j)} + 1 \right)$$

- method="MateiTille2" [Class 3]

$$\widehat{var}(\hat{t}_{HT}) = \frac{1}{1 - \sum_{i \in U} \frac{d_i^2}{\pi_i}} \sum_{i \in s} (1 - \pi_i) \left(\frac{y_i}{\pi_i} - \frac{\hat{t}_{HT}}{n} \right)^2$$

where

$$d_i = \frac{\pi_i (1 - \pi_i)}{\sum_{j \in U} \pi_j (1 - \pi_j)}$$

- `method="MateiTille3"` [Class 3]

$$\widehat{var}(\hat{t}_{HT}) = \frac{1}{1 - \sum_{i \in U} \frac{d_i^2}{\pi_i}} \sum_{i \in s} (1 - \pi_i) \left(\frac{y_i}{\pi_i} - \frac{\sum_{j \in s} (1 - \pi_j) \frac{y_j}{\pi_j}}{\sum_{j \in s} (1 - \pi_j)} \right)^2$$

  where d_i is defined as in `method="MateiTille2"`.

- `method="MateiTille4"` [Class 3]

$$\widehat{var}(\hat{t}_{HT}) = \frac{1}{1 - \sum_{i \in U} b_i / n^2} \sum_{i \in s} \frac{b_i}{\pi_i^3} (y_i - y_i^*)^2$$

  where

$$y_i^* = \pi_i \frac{\sum_{j \in s} b_j y_j / \pi_j^2}{\sum_{j \in s} b_j / \pi_j}$$

  and

$$b_i = \frac{\pi_i (1 - \pi_i) N}{N - 1}$$

- `method="MateiTille5"` [Class 3] This estimator is defined as in `method="MateiTille4"`, and the b_i values are defined as in `method="MateiTille1"`

Value

a scalar, the estimated variance

References

Matei, A.; Tillé, Y., 2005. Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. Journal of Official Statistics 21 (4), 543-570.

Haziza, D.; Mecatti, F.; Rao, J.N.K. 2008. Evaluation of some approximate variance estimators under the Rao-Sampford unequal probability sampling design. Metron LXVI (1), 91-108.

Examples

```
### Generate population data ---
N <- 500; n <- 50

set.seed(0)
x <- rgamma(500, scale=10, shape=5)
y <- abs( 2*x + 3.7*sqrt(x) * rnorm(N) )

pik <- n * x/sum(x)
s   <- sample(N, n)

ys <- y[s]
piks <- pik[s]

### Estimators of class 2 ---
approx_var_est(ys, piks, method="Deville1")
```

```
approx_var_est(ys, piks, method="Deville2")
approx_var_est(ys, piks, method="Deville3")
approx_var_est(ys, piks, method="Hajek")
approx_var_est(ys, piks, method="Rosen")
approx_var_est(ys, piks, method="FixedPoint")
approx_var_est(ys, piks, method="Brewer1")

### Estimators of class 3 ---
approx_var_est(ys, pik, method="HartleyRao", sample=s)
approx_var_est(ys, pik, method="Berger", sample=s)
approx_var_est(ys, pik, method="Tille", sample=s)
approx_var_est(ys, pik, method="MateiTille1", sample=s)
approx_var_est(ys, pik, method="MateiTille2", sample=s)
approx_var_est(ys, pik, method="MateiTille3", sample=s)
approx_var_est(ys, pik, method="MateiTille4", sample=s)
approx_var_est(ys, pik, method="MateiTille5", sample=s)
approx_var_est(ys, pik, method="Brewer2", sample=s)
approx_var_est(ys, pik, method="Brewer3", sample=s)
approx_var_est(ys, pik, method="Brewer4", sample=s)
```

`Var_approx` *Approximate the Variance of the Horvitz-Thompson estimator*

Description

Approximations of the Horvitz-Thompson variance for High-Entropy sampling designs. Such methods use only first-order inclusion probabilities.

Usage

```
Var_approx(y, pik, n, method, ...)
```

Arguments

`y`	numeric vector containing the values of the variable of interest for all population units
`pik`	numeric vector of first-order inclusion probabilities, of length equal to population size
`n`	a scalar indicating the sample size
`method`	string indicating the approximation that should be used. One of "Hajek1", "Hajek2", "HartleyRao1", "HartleyRao2", "FixedPoint".
`...`	two optional parameters can be modified to control the iterative procedure in `method="FixedPoint"`: `maxIter` sets the maximum number of iterations and `eps` controls the convergence error

Details

The variance approximations available in this function are described below, the notation used is that of Matei and Tillé (2005).

- Hájek variance approximation (`method="Hajek1"`):

$$\tilde{Var} = \sum_{i \in U} \frac{b_i}{\pi_i^2}(y_i - y_i^*)^2$$

where

$$y_i^* = \pi_i \frac{\sum_{j \in U} b_j y_j / \pi_j}{\sum_{j \in U} b_j}$$

and

$$b_i = \frac{\pi_i(1-\pi_i)N}{N-1}$$

- Starting from Hajék (1964), Brewer (2002) defined the following estimator (`method="Hajek2"`):

$$\tilde{Var} = \sum_{i \in U} \pi_i(1-\pi_i)\Big(\frac{y_i}{\pi_i} - \frac{\tilde{Y}}{n}\Big)^2$$

where $\tilde{Y} = \sum_{i \in U} a_i y_i$ and $a_i = n(1-\pi_i)/\sum_{j \in U} \pi_j(1-\pi_j)$

- Hartley and Rao (1962) variance approximation (`method="HartleyRao1"`):

$$\tilde{Var} = \sum_{i \in U} \pi_i \Big(1 - \frac{n-1}{n}\pi_i\Big)\left(\frac{y_i}{\pi_i} - \frac{Y}{n}\right)^2$$

$$-\frac{n-1}{n^2}\sum_{i \in U}\left(2\pi_i^3 - \frac{\pi_i^2}{2}\sum_{j \in U}\pi_j^2\right)\left(\frac{y_i}{\pi_i} - \frac{Y}{n}\right)^2$$

$$+\frac{2(n-1)}{n^3}\left(\sum_{i \in U}\pi_i y_i - \frac{Y}{n}\sum_{i \in U}\pi_i^2\right)^2$$

- Hartley and Rao (1962) provide a simplified version of the variance above (`method="HartleyRao2"`):

$$\tilde{Var} = \sum_{i \in U} \pi_i \Big(1 - \frac{n-1}{n}\pi_i\Big)\left(\frac{y_i}{\pi_i} - \frac{Y}{n}\right)^2$$

- `method="FixedPoint"` computes the Fixed-Point variance approximation proposed by Deville and Tillé (2005). The variance can be expressed in the same form as in `method="Hajek1"`, and the coefficients b_i are computed iteratively by the algorithm:
  1. $$b_i^{(0)} = \pi_i(1-\pi_i)\frac{N}{N-1}, \ \forall i \in U$$
  2. $$b_i^{(k)} = \frac{(b_i^{(k-1)})^2}{\sum_{j \in U} b_j^{(k-1)}} + \pi_i(1-\pi_i)$$

a necessary condition for convergence is checked and, if not satisfied, the function returns an alternative solution that uses only one iteration:

$$b_i = \pi_i(1-\pi_i)\left(\frac{N\pi_i(1-\pi_i)}{(N-1)\sum_{j\in U}\pi_j(1-\pi_j)} + 1\right)$$

Value

a scalar, the approximated variance.

References

Matei, A.; Tillé, Y., 2005. Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. Journal of Official Statistics 21 (4), 543-570.

Examples

```
N <- 500; n <- 50

set.seed(0)
x <- rgamma(n=N, scale=10, shape=5)
y <- abs( 2*x + 3.7*sqrt(x) * rnorm(N) )

pik  <- n * x/sum(x)
pikl <- outer(pik, pik, '*'); diag(pikl) <- pik

### Variance approximations ---
Var_approx(y, pik, n, method = "Hajek1")
Var_approx(y, pik, n, method = "Hajek2")
Var_approx(y, pik, n, method = "HartleyRao1")
Var_approx(y, pik, n, method = "HartleyRao2")
Var_approx(y, pik, n, method = "FixedPoint")
```

Index

approx_var_est, *2*, 2

UPSvarApprox *(*UPSvarApprox-package*)*, 2
UPSvarApprox-package, 2

Var_approx, *2*, 7