Package ‘samplingVarEst’

June 12, 2016

Version 1.0-2

Date 2016-06-12

Title Sampling Variance Estimation

Author Emilio Lopez Escobar [aut, cre, cph] <emilio@numerika.mx>,
Ernesto Barrios Zamudio [ctb] <ebarrios@itam.mx>,
Juan Francisco Munoz Rosas [ctb] <jfmunoz@go.ugr.es>

Maintainer Emilio Lopez Escobar <emilio@numerika.mx>

Description Functions to calculate some point estimators and estimating their variance under unequal probability sampling without replacement. Single and two stage sampling designs are considered. Some approximations for the second order inclusion probabilities are also available (sample and population based). A variety of Jackknife variance estimators are implemented.

Classification/MSC 62D05, 62F40, 62G09, 62H12

Classification/JEL C13, C15, C42, C83

Classification/ACM G.3

Depends R (>= 3.0.0)

License GPL (>= 2)

URL http://www.numerika.mx, http://www.itam.mx

NeedsCompilation yes

Repository CRAN

Date/Publication 2016-06-12 09:41:23

R topics documented:

---

samplingVarEst-package *Sampling Variance Estimation package*

Description

The package contains functions to calculate some point estimators and estimating their variance under unequal probability sampling without replacement. Uni-stage and two-stage sampling designs are considered. The package further contains some approximations for the joint-inclusion probabilities (population and sample based formulae).

Emphasis has been put on the speed of routines as the package mostly uses C compiled code. Below there is a list of available functions. These are grouped in *purpose-lists*, aiming to clarify their usage.

The user should pick a suitable combination of: a population parameter of interest, a choice of point estimator, and a choice of variance estimator.

For these population parameters:	**The available point estimators are:**
total:	`Est.Total.NHT`
	`Est.Total.Hajek`
mean:	`Est.Mean.NHT`
	`Est.Mean.Hajek`
empirical cumulative distribution function:	`Est.EmpDistFunc.NHT`
	`Est.EmpDistFunc.Hajek`
ratio:	`Est.Ratio`
correlation coefficient:	`Est.Corr.NHT`
	`Est.Corr.Hajek`
regression coefficients:	`Est.RegCoI.Hajek`
	`Est.RegCo.Hajek`

For these point estimators:	**The available variance estimators** for *uni-stage samples* are:
`Est.Total.NHT`:	`VE.HT.Total.NHT`
	`VE.SYG.Total.NHT`
	`VE.Hajek.Total.NHT`
`Est.Total.Hajek`:	`VE.Jk.Tukey.Total.Hajek`
	`VE.Jk.CBS.HT.Total.Hajek`
	`VE.Jk.CBS.SYG.Total.Hajek`
	`VE.Jk.B.Total.Hajek`

	VE.EB.HT.Total.Hajek
	VE.EB.SYG.Total.Hajek
Est.Mean.NHT:	VE.HT.Mean.NHT
	VE.SYG.Mean.NHT
	VE.Hajek.Mean.NHT
Est.Mean.Hajek:	VE.Jk.Tukey.Mean.Hajek
	VE.Jk.CBS.HT.Mean.Hajek
	VE.Jk.CBS.SYG.Mean.Hajek
	VE.Jk.B.Mean.Hajek
	VE.EB.HT.Mean.Hajek
	VE.EB.SYG.Mean.Hajek
Est.Ratio:	VE.Lin.HT.Ratio
	VE.Lin.SYG.Ratio
	VE.Jk.Tukey.Ratio
	VE.Jk.CBS.HT.Ratio
	VE.Jk.CBS.SYG.Ratio
	VE.Jk.B.Ratio
	VE.EB.HT.Ratio
	VE.EB.SYG.Ratio
Est.Corr.NHT:	VE.Jk.Tukey.Corr.NHT
Est.Corr.Hajek:	VE.Jk.Tukey.Corr.Hajek
	VE.Jk.CBS.HT.Corr.Hajek
	VE.Jk.CBS.SYG.Corr.Hajek
	VE.Jk.B.Corr.Hajek
Est.RegCoI.Hajek:	VE.Jk.Tukey.RegCoI.Hajek
	VE.Jk.CBS.HT.RegCoI.Hajek
	VE.Jk.CBS.SYG.RegCoI.Hajek
	VE.Jk.B.RegCoI.Hajek
Est.RegCo.Hajek:	VE.Jk.Tukey.RegCo.Hajek
	VE.Jk.CBS.HT.RegCo.Hajek
	VE.Jk.CBS.SYG.RegCo.Hajek
	VE.Jk.B.RegCo.Hajek

For these point estimators:	**The available variance estimators** for *self-weighted two-stage samples* are:
Est.Total.Hajek:	VE.Jk.EB.SW2.Total.Hajek
Est.Mean.Hajek:	VE.Jk.EB.SW2.Mean.Hajek
Est.Ratio:	VE.Jk.EB.SW2.Ratio
Est.Corr.Hajek:	VE.Jk.EB.SW2.Corr.Hajek
Est.RegCoI.Hajek:	VE.Jk.EB.SW2.RegCoI.Hajek
Est.RegCo.Hajek:	VE.Jk.EB.SW2.RegCo.Hajek

For the inclusion probabilities:	**The available functions** are:
1st order inclusion probabilities:	Pk.PropNorm.U
2nd order (joint) inclusion probabilities:	Pkl.Hajek.s
	Pkl.Hajek.U

datasets
oaxaca

Details

To return to this description type:
`help(samplingVarEst)`
or type:
`?samplingVarEst`
To cite, use:
`citation("samplingVarEst")`

`Est.Corr.Hajek` *Estimator of a correlation coefficient using the Hajek point estimator*

Description

Estimates a population correlation coefficient of two variables using the Hajek (1971) point estimator.

Usage

```
Est.Corr.Hajek(VecY.s, VecX.s, VecPk.s)
```

Arguments

`VecY.s`	vector of the variable of interest Y; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecX.s`. There must not be missing values.
`VecX.s`	vector of the variable of interest X; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecY.s`. There must not be missing values.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.

Details

For the population correlation coefficient of two variables y and x:

$$C = \frac{\sum_{k\in U}(y_k - \bar{y})(x_k - \bar{x})}{\sqrt{\sum_{k\in U}(y_k - \bar{y})^2}\sqrt{\sum_{k\in U}(x_k - \bar{x})^2}}$$

the point estimator of C, assuming that N is unknown (see Sarndal et al., 1992, Sec. 5.9) (implemented by the current function), is:

$$\hat{C}_{Hajek} = \frac{\sum_{k\in s} w_k(y_k - \hat{\bar{y}}_{Hajek})(x_k - \hat{\bar{x}}_{Hajek})}{\sqrt{\sum_{k\in s} w_k(y_k - \hat{\bar{y}}_{Hajek})^2}\sqrt{\sum_{k\in s} w_k(x_k - \hat{\bar{x}}_{Hajek})^2}}$$

where $\hat{\bar{y}}_{Hajek}$ is the Hajek (1971) point estimator of the population mean $\bar{y} = N^{-1} \sum_{k \in U} y_k$,

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k \in s} w_k y_k}{\sum_{k \in s} w_k}$$

and $w_k = 1/\pi_k$ with π_k denoting the inclusion probability of the k-th element in the sample s.

Value

The function returns a value for the correlation coefficient point estimator.

Author(s)

Emilio Lopez Escobar.

References

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Sarndal, C.-E. and Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, Inc.

See Also

Est.Corr.NHT
VE.Jk.Tukey.Corr.Hajek
VE.Jk.CBS.HT.Corr.Hajek
VE.Jk.CBS.SYG.Corr.Hajek
VE.Jk.B.Corr.Hajek
VE.Jk.EB.SW2.Corr.Hajek

Examples

```
data(oaxaca)                                 #Loads the Oaxaca municipalities dataset
pik.U <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s     <- oaxaca$sHOMES00                     #Defines the sample to be used
y1    <- oaxaca$POP10                        #Defines the variable of interest y1
y2    <- oaxaca$POPMAL10                     #Defines the variable of interest y2
x     <- oaxaca$HOMES10                      #Defines the variable of interest x
#Computes the correlation coefficient estimator for y1 and x
Est.Corr.Hajek(y1[s==1], x[s==1], pik.U[s==1])
#Computes the correlation coefficient estimator for y2 and x
Est.Corr.Hajek(y2[s==1], x[s==1], pik.U[s==1])
```

`Est.Corr.NHT` *Estimator of a correlation coefficient using the Narain-Horvitz-Thompson point estimator*

Description

Estimates a population correlation coefficient of two variables using the Narain (1951); Horvitz-Thompson (1952) point estimator.

Usage

```
Est.Corr.NHT(VecY.s, VecX.s, VecPk.s, N)
```

Arguments

`VecY.s`	vector of the variable of interest Y; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecX.s`. There must not be missing values.
`VecX.s`	vector of the variable of interest X; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecY.s`. There must not be missing values.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.
`N`	the population size. It must be an integer or a double-precision scalar with zero-valued fractional part.

Details

For the population correlation coefficient of two variables y and x:

$$C = \frac{\sum_{k\in U}(y_k - \bar{y})(x_k - \bar{x})}{\sqrt{\sum_{k\in U}(y_k - \bar{y})^2}\sqrt{\sum_{k\in U}(x_k - \bar{x})^2}}$$

the point estimator of C (implemented by the current function) is given by:

$$\hat{C} = \frac{\sum_{k\in s} w_k(y_k - \hat{\bar{y}}_{NHT})(x_k - \hat{\bar{x}}_{NHT})}{\sqrt{\sum_{k\in s} w_k(y_k - \hat{\bar{y}}_{NHT})^2}\sqrt{\sum_{k\in s} w_k(x_k - \hat{\bar{x}}_{NHT})^2}}$$

where $\hat{\bar{y}}_{NHT}$ is the Narain (1951); Horvitz-Thompson (1952) estimator for the population mean $\bar{y} = N^{-1}\sum_{k\in U} y_k$,

$$\hat{\bar{y}}_{NHT} = \frac{1}{N}\sum_{k\in s} w_k y_k$$

and $w_k = 1/\pi_k$ with π_k denoting the inclusion probability of the k-th element in the sample s.

Value

The function returns a value for the correlation coefficient point estimator.

Author(s)

Emilio Lopez Escobar.

References

Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Narain, R. D. (1951) On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **3**, 169–175.

See Also

Est.Corr.Hajek
VE.Jk.Tukey.Corr.NHT

Examples

```
data(oaxaca)                                 #Loads the Oaxaca municipalities dataset
pik.U <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s     <- oaxaca$sHOMES00                     #Defines the sample to be used
N     <- dim(oaxaca)[1]                      #Defines the population size
y1    <- oaxaca$POP10                        #Defines the variable of interest y1
y2    <- oaxaca$POPMAL10                     #Defines the variable of interest y2
x     <- oaxaca$HOMES10                      #Defines the variable of interest x
#Computes the correlation coefficient estimator for y1 and x
Est.Corr.NHT(y1[s==1], x[s==1], pik.U[s==1], N)
#Computes the correlation coefficient estimator for y2 and x
Est.Corr.NHT(y2[s==1], x[s==1], pik.U[s==1], N)
```

Est.EmpDistFunc.Hajek *The Hajek estimator for the empirical cumulative distribution function*

Description

Computes the Hajek (1971) estimator for the empirical cumulative distribution function (ECDF).

Usage

```
Est.EmpDistFunc.Hajek(VecY.s, VecPk.s, t)
```

Arguments

`VecY.s`	vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s`. There must not be missing values.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.
`t`	value to be evaluated for the empirical cumulative distribution function. It must be an integer or a double-precision scalar.

Details

For the population empirical cumulative distribution function (ECDF) of the variable y at the value t:

$$Fn(t) = \frac{\#(k \in U : y_k \leq t)}{N} = \frac{1}{N}\sum_{k \in U} I(y_k \leq t)$$

the approximately unbiased Hajek (1971) estimator of $Fn(t)$ (implemented by the current function) is given by:

$$\hat{F}n_{Hajek}(t) = \frac{\sum_{k \in s} w_k I(y_k \leq t)}{\sum_{k \in s} w_k}$$

where $I(y_k \leq t)$ denotes the indicator function that takes the value 1 if $y_k \leq t$ and the value 0 otherwise, and where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s.

Value

The function returns a value for the empirical cumulative distribution function evaluated at t.

Author(s)

Emilio Lopez Escobar [aut, cre], Juan Francisco Munoz Rosas [ctb].

References

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

See Also

`Est.EmpDistFunc.NHT`

Examples

```
data(oaxaca)                                         #Loads Oaxaca municipalities dataset
pik.U <- Pk.PropNorm.U(373, oaxaca$HOMES00)          #Reconstructs the inclusion probs.
s     <- oaxaca$sHOMES00                             #Defines the sample to be used
y1    <- oaxaca$POP10                                #Defines the variable of interest y1
Est.EmpDistFunc.Hajek(y1[s==1], pik.U[s==1], 950) #Hajek est. of ECDF for y1 at t=950
```

`Est.EmpDistFunc.NHT` *The Narain-Horvitz-Thompson estimator for the empirical cumulative distribution function*

Description

Computes the Narain (1951); Horvitz-Thompson (1952) estimator for the empirical cumulative distribution function (ECDF).

Usage

```
Est.EmpDistFunc.NHT(VecY.s, VecPk.s, N, t)
```

Arguments

`VecY.s`	vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s`. There must not be missing values.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.
`N`	the population size. It must be an integer or a double-precision scalar with zero-valued fractional part.
`t`	value to be evaluated for the empirical cumulative distribution function. It must be an integer or a double-precision scalar.

Details

For the population empirical cumulative distribution function (ECDF) of the variable y at the value t:

$$Fn(t) = \frac{\#(k \in U : y_k \leq t)}{N} = \frac{1}{N} \sum_{k \in U} I(y_k \leq t)$$

the unbiased Narain (1951); Horvitz-Thompson (1952) estimator of $Fn(t)$ (implemented by the current function) is given by:

$$\hat{F}n_{NHT}(t) = \frac{1}{N} \sum_{k \in s} \frac{I(y_k \leq t)}{\pi_k}$$

where $I(y_k \leq t)$ denotes the indicator function that takes the value 1 if $y_k \leq t$ and the value 0 otherwise, and where π_k denotes the inclusion probability of the k-th element in the sample s.

Value

The function returns a value for the empirical cumulative distribution function evaluated at t.

Author(s)

Emilio Lopez Escobar [aut, cre], Juan Francisco Munoz Rosas [ctb].

References

Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Narain, R. D. (1951) On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **3**, 169–175.

See Also

Est.EmpDistFunc.Hajek

Examples

```
data(oaxaca)                                     #Loads Oaxaca municipalities dataset
pik.U <- Pk.PropNorm.U(373, oaxaca$HOMES00)      #Reconstructs the inclusion probs.
s     <- oaxaca$sHOMES00                         #Defines the sample to be used
N     <- dim(oaxaca)[1]                          #Defines the population size
y1    <- oaxaca$POP10                            #Defines the variable of interest y1
Est.EmpDistFunc.NHT(y1[s==1], pik.U[s==1], N, 950) #NHT est. of ECDF for y1 at t=950
```

Est.Mean.Hajek — *The Hajek estimator for a mean*

Description

Computes the Hajek (1971) estimator for a population mean.

Usage

```
Est.Mean.Hajek(VecY.s, VecPk.s)
```

Arguments

VecY.s	vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s. There must not be missing values.
VecPk.s	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be missing values.

Details

For the population mean of the variable y:

$$\bar{y} = \frac{1}{N} \sum_{k \in U} y_k$$

the approximately unbiased Hajek (1971) estimator of $\bar{y}$ (implemented by the current function) is given by:

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k \in s} w_k y_k}{\sum_{k \in s} w_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s.

Value

The function returns a value for the mean point estimator.

Author(s)

Emilio Lopez Escobar.

References

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

See Also

Est.Mean.NHT
VE.Jk.Tukey.Mean.Hajek
VE.Jk.CBS.HT.Mean.Hajek
VE.Jk.CBS.SYG.Mean.Hajek
VE.Jk.B.Mean.Hajek
VE.Jk.EB.SW2.Mean.Hajek

Examples

```
data(oaxaca)                                #Loads the Oaxaca municipalities dataset
pik.U <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s     <- oaxaca$sHOMES00                    #Defines the sample to be used
y1    <- oaxaca$POP10                       #Defines the variable of interest y1
y2    <- oaxaca$HOMES10                     #Defines the variable of interest y2
Est.Mean.Hajek(y1[s==1], pik.U[s==1])       #Computes the Hajek est. for y1
Est.Mean.Hajek(y2[s==1], pik.U[s==1])       #Computes the Hajek est. for y2
```

`Est.Mean.NHT` *The Narain-Horvitz-Thompson estimator for a mean*

Description

Computes the Narain (1951); Horvitz-Thompson (1952) estimator for a population mean.

Usage

```
Est.Mean.NHT(VecY.s, VecPk.s, N)
```

Arguments

`VecY.s`	vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s`. There must not be missing values.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.
`N`	the population size. It must be an integer or a double-precision scalar with zero-valued fractional part.

Details

For the population mean of the variable y:

$$\bar{y} = \frac{1}{N} \sum_{k \in U} y_k$$

the unbiased Narain (1951); Horvitz-Thompson (1952) estimator of $\bar{y}$ (implemented by the current function) is given by:

$$\hat{\bar{y}}_{NHT} = \frac{1}{N} \sum_{k \in s} \frac{y_k}{\pi_k}$$

where π_k denotes the inclusion probability of the k-th element in the sample s.

Value

The function returns a value for the mean point estimator.

Author(s)

Emilio Lopez Escobar.

References

Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Narain, R. D. (1951) On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **3**, 169–175.

See Also

Est.Mean.Hajek
VE.HT.Mean.NHT
VE.SYG.Mean.NHT
VE.Hajek.Mean.NHT

Examples

```
data(oaxaca)                                 #Loads the Oaxaca municipalities dataset
pik.U <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s     <- oaxaca$sHOMES00                     #Defines the sample to be used
N     <- dim(oaxaca)[1]                      #Defines the population size
y1    <- oaxaca$POP10                        #Defines the variable of interest y1
y2    <- oaxaca$HOMES10                      #Defines the variable of interest y2
Est.Mean.NHT(y1[s==1], pik.U[s==1], N)      #The NHT estimator for y1
Est.Mean.NHT(y2[s==1], pik.U[s==1], N)      #The NHT estimator for y2
```

Est.Ratio *Estimator of a ratio*

Description

Estimates a population ratio of two totals/means.

Usage

```
Est.Ratio(VecY.s, VecX.s, VecPk.s)
```

Arguments

VecY.s	vector of the numerator variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecX.s. There must not be missing values.
VecX.s	vector of the denominator variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecY.s. There must not be missing values. All values of VecX.s should be greater than zero. A warning is displayed if this does not hold and computations continue if mathematical expressions allow this kind of values for the denominator variable.
VecPk.s	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be missing values.

Details

For the population ratio of two totals/means of the variables y and x:

$$R = \frac{\sum_{k\in U} y_k/N}{\sum_{k\in U} x_k/N} = \frac{\sum_{k\in U} y_k}{\sum_{k\in U} x_k}$$

the ratio estimator of R (implemented by the current function) is given by:

$$\hat{R} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k x_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s.

Value

The function returns a value for the ratio point estimator.

Author(s)

Emilio Lopez Escobar.

References

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Narain, R. D. (1951) On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **3**, 169–175.

See Also

VE.Jk.Tukey.Ratio
VE.Jk.CBS.HT.Ratio
VE.Jk.CBS.SYG.Ratio
VE.Jk.B.Ratio
VE.Jk.EB.SW2.Ratio

Examples

```
data(oaxaca)                                #Loads the Oaxaca municipalities dataset
pik.U <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s     <- oaxaca$sHOMES00                    #Defines the sample to be used
y1    <- oaxaca$POP10                       #Defines the numerator variable y1
y2    <- oaxaca$POPMAL10                    #Defines the numerator variable y2
x     <- oaxaca$HOMES10                     #Defines the denominator variable x
Est.Ratio(y1[s==1], x[s==1], pik.U[s==1])   #Ratio estimator for y1 and x
Est.Ratio(y2[s==1], x[s==1], pik.U[s==1])   #Ratio estimator for y2 and x
```

`Est.RegCo.Hajek` *Estimator of the regression coefficient using the Hajek point estimator*

Description

Estimates the population regression coefficient using the Hajek (1971) point estimator.

Usage

```
Est.RegCo.Hajek(VecY.s, VecX.s, VecPk.s)
```

Arguments

`VecY.s`	vector of the variable of interest Y; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecX.s`. There must not be missing values.
`VecX.s`	vector of the variable of interest X; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecY.s`. There must not be missing values.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.

Details

From Linear Regression Analysis, for an imposed population model

$$y = \alpha + \beta x$$

the population regression coefficient β, assuming that the population size N is unknown (see Sarndal et al., 1992, Sec. 5.10), can be estimated by:

$$\hat{\beta}_{Hajek} = \frac{\sum_{k\in s} w_k (y_k - \hat{\bar{y}}_{Hajek})(x_k - \hat{\bar{x}}_{Hajek})}{\sum_{k\in s} w_k (x_k - \hat{\bar{x}}_{Hajek})^2}$$

where $\hat{\bar{y}}_{Hajek}$ and $\hat{\bar{x}}_{Hajek}$ are the Hajek (1971) point estimators of the population means $\bar{y} = N^{-1}\sum_{k\in U} y_k$ and $\bar{x} = N^{-1}\sum_{k\in U} x_k$, respectively,

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

$$\hat{\bar{x}}_{Hajek} = \frac{\sum_{k\in s} w_k x_k}{\sum_{k\in s} w_k}$$

and $w_k = 1/\pi_k$ with π_k denoting the inclusion probability of the k-th element in the sample s.

Value

The function returns a value for the regression coefficient point estimator.

Author(s)

Emilio Lopez Escobar.

References

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Sarndal, C.-E. and Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, Inc.

See Also

Est.RegCoI.Hajek
VE.Jk.Tukey.RegCo.Hajek
VE.Jk.CBS.HT.RegCo.Hajek
VE.Jk.CBS.SYG.RegCo.Hajek
VE.Jk.B.RegCo.Hajek
VE.Jk.EB.SW2.RegCo.Hajek

Examples

```
data(oaxaca)                                #Loads the Oaxaca municipalities dataset
pik.U <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s     <- oaxaca$sHOMES00                    #Defines the sample to be used
y1    <- oaxaca$POP10                       #Defines the variable of interest y1
y2    <- oaxaca$POPMAL10                    #Defines the variable of interest y2
x     <- oaxaca$HOMES10                     #Defines the variable of interest x
#Computes the regression coefficient estimator for y1 and x
Est.RegCo.Hajek(y1[s==1], x[s==1], pik.U[s==1])
#Computes the regression coefficient estimator for y2 and x
Est.RegCo.Hajek(y2[s==1], x[s==1], pik.U[s==1])
```

Est.RegCoI.Hajek *Estimator of the intercept regression coefficient using the Hajek point estimator*

Description

Estimates the population intercept regression coefficient using the Hajek (1971) point estimator.

Usage

```
Est.RegCoI.Hajek(VecY.s, VecX.s, VecPk.s)
```

Arguments

`VecY.s` vector of the variable of interest Y; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecX.s`. There must not be missing values.

`VecX.s` vector of the variable of interest X; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecY.s`. There must not be missing values.

`VecPk.s` vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.

Details

From Linear Regression Analysis, for an imposed population model

$$y = \alpha + \beta x$$

the population intercept regression coefficient α, assuming that the population size N is unknown (see Sarndal et al., 1992, Sec. 5.10), can be estimated by:

$$\hat{\alpha}_{Hajek} = \hat{\bar{y}}_{Hajek} - \frac{\sum_{k\in s} w_k(y_k - \hat{\bar{y}}_{Hajek})(x_k - \hat{\bar{x}}_{Hajek})}{\sum_{k\in s} w_k(x_k - \hat{\bar{x}}_{Hajek})^2}\hat{\bar{x}}_{Hajek}$$

where $\hat{\bar{y}}_{Hajek}$ and $\hat{\bar{x}}_{Hajek}$ are the Hajek (1971) point estimators of the population means $\bar{y} = N^{-1}\sum_{k\in U} y_k$ and $\bar{x} = N^{-1}\sum_{k\in U} x_k$, respectively,

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

$$\hat{\bar{x}}_{Hajek} = \frac{\sum_{k\in s} w_k x_k}{\sum_{k\in s} w_k}$$

and $w_k = 1/\pi_k$ with π_k denoting the inclusion probability of the k-th element in the sample s.

Value

The function returns a value for the intercept regression coefficient point estimator.

Author(s)

Emilio Lopez Escobar.

References

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Sarndal, C.-E. and Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, Inc.

See Also

Est.RegCo.Hajek
VE.Jk.Tukey.RegCoI.Hajek
VE.Jk.CBS.HT.RegCoI.Hajek
VE.Jk.CBS.SYG.RegCoI.Hajek
VE.Jk.B.RegCoI.Hajek
VE.Jk.EB.SW2.RegCoI.Hajek

Examples

```
data(oaxaca)                                #Loads the Oaxaca municipalities dataset
pik.U <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s     <- oaxaca$sHOMES00                    #Defines the sample to be used
y1    <- oaxaca$POP10                       #Defines the variable of interest y1
y2    <- oaxaca$POPMAL10                    #Defines the variable of interest y2
x     <- oaxaca$HOMES10                     #Defines the variable of interest x
#Computes the intercept regression coefficient estimator for y1 and x
Est.RegCoI.Hajek(y1[s==1], x[s==1], pik.U[s==1])
#Computes the intercept regression coefficient estimator for y2 and x
Est.RegCoI.Hajek(y2[s==1], x[s==1], pik.U[s==1])
```

Est.Total.Hajek *The Hajek estimator for a total*

Description

Computes the Hajek (1971) estimator for a population total.

Usage

```
Est.Total.Hajek(VecY.s, VecPk.s, N)
```

Arguments

VecY.s	vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s. There must not be missing values.
VecPk.s	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be missing values.
N	the population size. It must be an integer or a double-precision scalar with zero-valued fractional part.

Details

For the population total of the variable y:

$$t = \sum_{k \in U} y_k$$

the approximately unbiased Hajek (1971) estimator of t (implemented by the current function) is given by:

$$\hat{t}_{Hajek} = N \frac{\sum_{k \in s} w_k y_k}{\sum_{k \in s} w_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s.

Value

The function returns a value for the total point estimator.

Author(s)

Emilio Lopez Escobar.

References

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

See Also

Est.Total.NHT
VE.Jk.Tukey.Total.Hajek
VE.Jk.CBS.HT.Total.Hajek
VE.Jk.CBS.SYG.Total.Hajek
VE.Jk.B.Total.Hajek
VE.Jk.EB.SW2.Total.Hajek

Examples

```
data(oaxaca)                                 #Loads the Oaxaca municipalities dataset
pik.U <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s     <- oaxaca$sHOMES00                     #Defines the sample to be used
N     <- dim(oaxaca)[1]                      #Defines the population size
y1    <- oaxaca$POP10                        #Defines the variable y1
y2    <- oaxaca$HOMES10                      #Defines the variable y2
Est.Total.Hajek(y1[s==1], pik.U[s==1], N)    #The Hajek estimator for y1
Est.Total.Hajek(y2[s==1], pik.U[s==1], N)    #The Hajek estimator for y2
```

`Est.Total.NHT` *The Narain-Horvitz-Thompson estimator for a total*

Description

Computes the Narain (1951); Horvitz-Thompson (1952) estimator for a population total.

Usage

```
Est.Total.NHT(VecY.s, VecPk.s)
```

Arguments

`VecY.s`	vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s`. There must not be missing values.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.

Details

For the population total of the variable y:

$$t = \sum_{k \in U} y_k$$

the unbiased Narain (1951); Horvitz-Thompson (1952) estimator of t (implemented by the current function) is given by:

$$\hat{t}_{NHT} = \sum_{k \in s} \frac{y_k}{\pi_k}$$

where π_k denotes the inclusion probability of the k-th element in the sample s.

Value

The function returns a value for the total point estimator.

Author(s)

Emilio Lopez Escobar.

References

Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Narain, R. D. (1951) On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **3**, 169–175.

See Also

Est.Total.Hajek
VE.HT.Total.NHT
VE.SYG.Total.NHT
VE.Hajek.Total.NHT

Examples

```
data(oaxaca)                               #Loads the Oaxaca municipalities dataset
pik.U <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s     <- oaxaca$sHOMES00                   #Defines the sample to be used
y1    <- oaxaca$POP10                      #Defines the variable of interest y1
y2    <- oaxaca$HOMES10                    #Defines the variable of interest y2
Est.Total.NHT(y1[s==1], pik.U[s==1])       #Computes the NHT estimator for y1
Est.Total.NHT(y2[s==1], pik.U[s==1])       #Computes the NHT estimator for y2
```

---

oaxaca *Municipalities of the state of Oaxaca in Mexico*

Description

Dataset with information about the free and sovereign state of Oaxaca which is located in the south part of Mexico. The dataset contains information of population, surface, indigenous language, agriculture and income from years ranging from 2000 to 2010. The information was originally collected and processed by the Mexico's National Institute of Statistics and Geography (INEGI by its name in Spanish, 'Instituto Nacional de Estadistica y Geografia', http://www.inegi.org.mx/).

Usage

```
data(oaxaca)
```

Format

A data frame with 570 observations on the following 41 variables:

IDREGION region INEGI code.

LBREGION region name (without accents and Spanish language characters).

IDDISTRI district INEGI code.

LBDISTRI district name (without accents and Spanish language characters).

IDMUNICI municipality INEGI code.

LBMUNICI municipality name (without accents and Spanish language characters).

SURFAC05 surface in squared kilometres 2005.

POP00 population 2000.

POP10 population 2010.

HOMES00 number of homes 2000.

HOMES10 number of homes 2010.

POPMAL00 male population 2000.

POPMAL10 male population 2010.

POPFEM00 female population 2000.

POPFEM10 female population 2010.

INLANG00 5 or more years old population which speaks indigenous language 2000.

INLANG10 5 or more years old population which speaks indigenous language 2010.

INCOME00 gross income in thousands of Mexican pesos 2000.

INCOME01 gross income in thousands of Mexican pesos 2001.

INCOME02 gross income in thousands of Mexican pesos 2002.

INCOME03 gross income in thousands of Mexican pesos 2003.

PTREES00 planted trees 2000.

PTREES01 planted trees 2001.

PTREES02 planted trees 2002.

PTREES03 planted trees 2003.

MARRIA07 marriages 2007.

MARRIA08 marriages 2008.

MARRIA09 marriages 2009.

HARVBE07 harvested bean surface in hectares 2007.

HARVBE08 harvested bean surface in hectares 2008.

HARVBE09 harvested bean surface in hectares 2009.

VALUBE07 value of bean production in thousands of Mexican pesos 2007.

VALUBE08 value of bean production in thousands of Mexican pesos 2008.

VALUBE09 value of bean production in thousands of Mexican pesos 2009.

VOLUBE07 volume of bean production in tons 2007.

VOLUBE08 volume of bean production in tons 2008.

VOLUBE09 volume of bean production in tons 2009.

sHOMES00 a sample (column vector of ones and zeros; 1 = selected, 0 = otherwise) of 373 municipalities drawn using the Hajek (1964) maximum-entropy sampling design with inclusion probabilities proportional to the variable HOMES00.

sSURFAC a sample (column vector of ones and zeros; 1 = selected, 0 = otherwise) of 373 municipalities drawn using the Hajek (1964) maximum-entropy sampling design with inclusion probabilities proportional to the variable SURFAC05.

SIZEDIST the size of the district, i.e. the number of municipalities in each district.

sSW_10_3 a sample (column vector of ones and zeros; 1 = selected, 0 = otherwise) of 30 municipalities drawn using a self-weighted two-stage sampling design. The first stage draws 10 districts using the Hajek (1964) maximum-entropy sampling design with clusters' inclusion probabilities proportional to the size of the clusters (variable SIZEDIST). The second stage draws 3 municipalities within the selected districts at the first stage, using equal-probability without-replacement sampling.

Source

Mexico's National Institute of Statistics and Geography (INEGI), 'Instituto Nacional de Estadistica y Geografia' http://www3.inegi.org.mx/sistemas/descarga/

Examples

```
data(oaxaca)                         #Loads the Oaxaca municipalities dataset
mean(oaxaca$INCOME00, na.rm= TRUE)   #Computes INCOME00 mean (note it has NA's)
median(oaxaca$INCOME00, na.rm= TRUE) #Computes INCOME00 median (note it has NA's)
```

`Pk.PropNorm.U` *Inclusion probabilities proportional to a specified variable.*

Description

Creates and normalises the 1st order inclusion probabilities proportional to a specified variable. In the current context, normalisation means that the inclusion probabilities are less than or equal to 1. Ideally, they should sum up to n, the sample size.

Usage

```
Pk.PropNorm.U(n, VecMOS.U)
```

Arguments

`n`	the sample size. It must be an integer or a double-precision scalar with zero-valued fractional part.
`VecMOS.U`	vector of the variable called measure of size (MOS) to which the first-order inclusion probabilities are to be proportional; its length is equal to the population size. Values in VecMOS.U should be greater than zero (a warning message appears if this does not hold). There must not be missing values.

Details

Although the normalisation procedure is well-known in the survey sampling literature, we follow the procedure described in Chao (1982, p. 654). Hence, we obtain a unique set of inclusion probabilities that are proportional to the MOS variable.

Value

The function returns a vector of length n with the inclusion probabilities.

Author(s)

Emilio Lopez Escobar.

References

Chao, M. T. (1982) A general purpose unequal probability sampling plan. *Biometrika* **69**, 653–656.

See Also

Pkl.Hajek.s
Pkl.Hajek.U

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
             #Creates the normalised 1st order incl. probs. proportional
             #to the variable oaxaca$HOMES00 and with sample size 373
pik.U <- Pk.PropNorm.U(373, oaxaca$HOMES00)
sum(pik.U)   #Shows the sum is equal to the sample size 373
any(pik.U>1) #Shows there isn't any probability greater than 1
any(pik.U<0) #Shows there isn't any probability less than 0
```

Pkl.Hajek.s — *The Hajek approximation for the 2nd order (joint) inclusion probabilities (sample based)*

Description

Computes the Hajek (1964) approximation for the 2nd order (joint) inclusion probabilities utilising only sample-based quantities.

Usage

```
Pkl.Hajek.s(VecPk.s)
```

Arguments

VecPk.s	vector of the first-order inclusion probabilities; its length is equal to the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be missing values.

Details

Let π_k denote the inclusion probability of the k-th element in the sample s, and let π_{kl} denote the joint-inclusion probabilities of the k-th and l-th elements in the sample s. If the joint-inclusion probabilities π_{kl} are not available, the Hajek (1964) approximation can be used. Note that this approximation is designed for large-entropy sampling designs, large samples and large populations, i.e. care should be taken with highly-stratified samples, e.g. Berger (2005).

The sample based version of the Hajek (1964) approximation for the joint-inclusion probabilities π_{kl} (implemented by the current function) is:

$$\pi_{kl} \doteq \pi_k \pi_l \{1 - \hat{d}^{-1}(1-\pi_k)(1-\pi_l)\}$$

where $\hat{d} = \sum_{k \in s}(1 - \pi_k)$.

The approximation was originally developed for $d \to \infty$, under the maximum-entropy sampling design (see Hajek 1981, Theorem 3.3, Ch. 3 and 6), the Rejective Sampling design. It requires that the utilised sampling design be of large entropy. An overview can be found in Berger and Tille (2009). An account of different sampling designs, π_{kl} approximations, and approximate variances under large-entropy designs can be found in Tille (2006), Brewer and Donadio (2003), and Haziza, Mecatti, and Rao (2008). Recently, Berger (2011) gave sufficient conditions under which Hajek's results still hold for large-entropy sampling designs that are not the maximum-entropy one.

Value

The function returns a (n by n) square matrix with the estimated joint inclusion probabilities, where n is the sample size.

Author(s)

Emilio Lopez Escobar.

References

Berger, Y. G. (2005) Variance estimation with highly stratified sampling designs with unequal probabilities. *Australian & New Zealand Journal of Statistics*, **47**, 365–373.

Berger, Y. G. (2011) Asymptotic consistency under large entropy sampling designs with unequal probabilities. *Pakistan Journal of Statististics*, **27**, 407–426.

Berger, Y. G. and Tille, Y. (2009) Sampling with unequal probabilities. In *Sample Surveys: Design, Methods and Applications* (eds. D. Pfeffermann and C. R. Rao), 39–54. Elsevier, Amsterdam.

Brewer, K. R. W. and Donadio, M. E. (2003) The large entropy variance of the Horvitz-Thompson estimator. *Survey Methodology* **29**, 189–196.

Hajek, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35**, 4, 1491–1523.

Hajek, J. (1981) *Sampling From a Finite Population.* Dekker, New York.

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

Tille, Y. (2006) *Sampling Algorithms.* Springer, New York.

See Also

Pkl.Hajek.U
Pk.PropNorm.U

Examples

```
data(oaxaca)                                    #Loads the Oaxaca municipalities dataset
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s      <- oaxaca$sHOMES00                     #Defines the sample to be used
#This approximation is only suitable for large-entropy sampling designs
pikl.s <- Pkl.Hajek.s(pik.U[s==1])           #Approx. 2nd order incl. probs. from s
```

```
#First 5 rows/cols of (sample based) 2nd order incl. probs. matrix
pikl.s[1:5,1:5]
```

`Pkl.Hajek.U` *The Hajek approximation for the 2nd order (joint) inclusion probabilities (population based)*

Description

Computes the Hajek (1964) approximation for the 2nd order (joint) inclusion probabilities utilising population-based quantities.

Usage

```
Pkl.Hajek.U(VecPk.U)
```

Arguments

`VecPk.U`	vector of the first-order inclusion probabilities; its length is equal to the population size. Values in `VecPk.U` must be greater than zero and less than or equal to one. There must not be missing values.

Details

Let π_k denote the inclusion probability of the k-th element in the sample s, and let π_{kl} denote the joint-inclusion probabilities of the k-th and l-th elements in the sample s. If the joint-inclusion probabilities π_{kl} are not available, the Hajek (1964) approximation can be used. Note that this approximation is designed for large-entropy sampling designs, large samples and large populations, i.e. care should be taken with highly-stratified samples, e.g. Berger (2005).

The population based version of the Hajek (1964) approximation for the joint-inclusion probabilities π_{kl} (implemented by the current function) is:

$$\pi_{kl} \doteq \pi_k \pi_l \{1 - d^{-1}(1-\pi_k)(1-\pi_l)\}$$

where $d = \sum_{k \in U} \pi_k (1-\pi_k)$.

The approximation was originally developed for $d \to \infty$, under the maximum-entropy sampling design (see Hajek 1981, Theorem 3.3, Ch. 3 and 6), the Rejective Sampling design. It requires that the utilised sampling design be of large entropy. An overview can be found in Berger and Tille (2009). An account of different sampling designs, π_{kl} approximations, and approximate variances under large-entropy designs can be found in Tille (2006), Brewer and Donadio (2003), and Haziza, Mecatti, and Rao (2008). Recently, Berger (2011) gave sufficient conditions under which Hajek's results still hold for large-entropy sampling designs that are not the maximum-entropy one.

Value

The function returns a (N by N) square matrix with the estimated joint inclusion probabilities, where N is the population size.

Author(s)

Emilio Lopez Escobar.

References

Berger, Y. G. (2005) Variance estimation with highly stratified sampling designs with unequal probabilities. *Australian & New Zealand Journal of Statistics*, **47**, 365–373.

Berger, Y. G. (2011) Asymptotic consistency under large entropy sampling designs with unequal probabilities. *Pakistan Journal of Statistics*, **27**, 407–426.

Berger, Y. G. and Tille, Y. (2009) Sampling with unequal probabilities. In *Sample Surveys: Design, Methods and Applications* (eds. D. Pfeffermann and C. R. Rao), 39–54. Elsevier, Amsterdam.

Brewer, K. R. W. and Donadio, M. E. (2003) The large entropy variance of the Horvitz-Thompson estimator. *Survey Methodology* **29**, 189–196.

Hajek, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35**, 4, 1491–1523.

Hajek, J. (1981) *Sampling From a Finite Population.* Dekker, New York.

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

Tille, Y. (2006) *Sampling Algorithms.* Springer, New York.

See Also

Pkl.Hajek.s
Pk.PropNorm.U

Examples

```
data(oaxaca)                                 #Loads the Oaxaca municipalities dataset
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
#(This approximation is only suitable for large-entropy sampling designs)
pikl.U <- Pkl.Hajek.U(pik.U)                 #Approximates 2nd order incl. probs. from U
#First 5 rows/cols of (population based) 2nd order incl. probs. matrix
pikl.U[1:5,1:5]
```

VE.EB.HT.Mean.Hajek *The Escobar-Berger unequal probability replicate variance estimator for the Hajek (1971) estimator of a mean (Horvitz-Thompson form)*

Description

Computes the Escobar-Berger (2013) unequal probability replicate variance estimator for the Hajek estimator of a mean. It uses the Horvitz-Thompson (1952) variance form.

Usage

```
VE.EB.HT.Mean.Hajek(VecY.s, VecPk.s, MatPkl.s,
                    VecAlpha.s = rep(1, times=length(VecPk.s)))
```

Arguments

`VecY.s`	vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s`. There must not be missing values.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.
`MatPkl.s`	matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in `MatPkl.s` must be greater than zero and less than or equal to one. There must not be missing values.
`VecAlpha.s`	vector of the α_k values; its length is equal to n, the sample size. Values in `VecAlpha.s` can be different for each unit and they must be greater or equal to zero. Escobar-Berger (2013) showed that this replicate variance estimator is valid for $\alpha_k \geq 0$. In particular, they suggest using $\alpha_k = 1$ for all units in the sample (the default for `VecAlpha.s` if omitted in the function call). Using $\alpha_k > 1$ results in approximating the Demnati-Rao (2004) linearisation variance estimators. There must not be missing values.

Details

For the population mean of the variable y:

$$\bar{y} = \frac{1}{N} \sum_{k \in U} y_k$$

the approximately unbiased Hajek (1971) estimator of $\bar{y}$ is given by:

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k \in s} w_k y_k}{\sum_{k \in s} w_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. The variance of $\hat{\bar{y}}_{Hajek}$ can be estimated by the Escobar-Berger (2013) unequal probability replicate variance estimator (implemented by the current function):

$$\hat{V}(\hat{\bar{y}}_{Hajek}) = \sum_{k \in s} \sum_{l \in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \breve{\nu}_k \breve{\nu}_l$$

where

$$\breve{\nu}_k = w_k^{\alpha_k} \left(\hat{\bar{y}}_{Hajek} - \hat{\bar{y}}^*_{Hajek,k} \right)$$

for some $\alpha_k \geq 0$ (suggested to be 1, see below comments) and with

$$\hat{\bar{y}}^*_{Hajek,k} = \frac{\sum_{l \in s} w_l y_l - w_k^{1-\alpha_k} y_k}{\sum_{l \in s} w_l - w_k^{1-\alpha_k}}$$

Regarding the value of α_k, Escobar-Berger (2013) show that $\hat{V}(\hat{\bar{y}}_{Hajek})$ is valid for $\alpha_k \geq 0$ but conclude that $\alpha_k > 0$ should be used as $\alpha_k = 0$ corresponds to a naive biased and unstable jackknife. They recommend $\alpha_k = 1$ or $\alpha_k > 1$. If $\alpha_k = 1$, $\hat{V}(\hat{\bar{y}}_{Hajek})$ reduces to the Escobar-Berger (2011) jackknife. Using $\alpha_k > 1$ results in approximating the empirical influence function, i.e. the Gateaux (1919) derivative, or Demnati-Rao (2004) linearisation variance estimators. The larger the α_k, the closer the approximation. Further, Escobar-Berger (2013) give an intuitive explanation of the replication method from a jackknife and bootstrap perspective.

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Demnati, A. and Rao, J. N. K. (2004) Linearization variance estimators for survey data. *Survey Methodology*, **30**, 17–26.

Escobar, E. L. and Berger, Y. G. (2011) Jackknife variance estimation for functions of Horvitz-Thompson estimators under unequal probability sampling without replacement. In Proceeding of the *58th World Statistics Congress*. Dublin, Ireland: International Statistical Institute.

Escobar, E. L. and Berger, Y. G. (2013) A new replicate variance estimator for unequal probability sampling without replacement. *Canadian Journal of Statistics* **41**, 3, 508–524.

Gateaux, R. (1919) Fonctions d'une infinite de variables indeependantes. *Bulletin de la Societe Mathematique de France*, **47**, 70–96.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

See Also

VE.Jk.Tukey.Mean.Hajek
VE.Jk.CBS.SYG.Mean.Hajek
VE.Jk.B.Mean.Hajek
VE.Jk.EB.SW2.Mean.Hajek
VE.EB.SYG.Mean.Hajek

Examples

```
data(oaxaca)                                  #Loads the Oaxaca municipalities dataset
pik.U   <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s       <- oaxaca$sHOMES00                    #Defines the sample to be used
y1      <- oaxaca$POP10                       #Defines the variable y1
y2      <- oaxaca$POPMAL10                    #Defines the variable y2
```

```
Alpha.s <- rep(2, times=373)                 #Defines the vector with Alpha values
#This approximation is only suitable for large-entropy sampling designs
pikl.s  <- Pkl.Hajek.s(pik.U[s==1])          #Approx. 2nd order incl. probs. from s
#Computes the var. est. of the Hajek mean point estimator using y1
VE.EB.HT.Mean.Hajek(y1[s==1], pik.U[s==1], pikl.s)
#Computes the var. est. of the Hajek mean point estimator using y2
VE.EB.HT.Mean.Hajek(y2[s==1], pik.U[s==1], pikl.s, Alpha.s)
```

VE.EB.HT.Ratio — *The Escobar-Berger unequal probability replicate variance estimator for the estimator of a ratio (Horvitz-Thompson form)*

Description

Computes the Escobar-Berger (2013) unequal probability replicate variance estimator for the estimator of a ratio of two totals/means. It uses the Horvitz-Thompson (1952) variance form.

Usage

```
VE.EB.HT.Ratio(VecY.s, VecX.s, VecPk.s, MatPkl.s,
                VecAlpha.s = rep(1, times=length(VecPk.s)))
```

Arguments

`VecY.s`	vector of the numerator variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecX.s`. There must not be missing values.
`VecX.s`	vector of the denominator variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecY.s`. There must not be missing values. All values of `VecX.s` should be greater than zero. A warning is displayed if this does not hold and computations continue if mathematical expressions allow this kind of values for the denominator variable.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.
`MatPkl.s`	matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in `MatPkl.s` must be greater than zero and less than or equal to one. There must not be missing values.
`VecAlpha.s`	vector of the α_k values; its length is equal to n, the sample size. Values in `VecAlpha.s` can be different for each unit and they must be greater or equal to zero. Escobar-Berger (2013) showed that this replicate variance estimator is valid for $\alpha_k \geq 0$. In particular, they suggest using $\alpha_k = 1$ for all units in the sample (the default for `VecAlpha.s` if omitted in the function call). Using $\alpha_k > 1$ results in approximating the Demnati-Rao (2004) linearisation variance estimators. There must not be missing values.

Details

For the population ratio of two totals/means of the variables y and x:

$$R = \frac{\sum_{k\in U} y_k/N}{\sum_{k\in U} x_k/N} = \frac{\sum_{k\in U} y_k}{\sum_{k\in U} x_k}$$

the ratio estimator of R is given by:

$$\hat{R} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k x_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. The variance of $\hat{R}$ can be estimated by the Escobar-Berger (2013) unequal probability replicate variance estimator (implemented by the current function):

$$\hat{V}(\hat{R}) = \sum_{k\in s}\sum_{l\in s} \frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}} \breve{\nu}_k \breve{\nu}_l$$

where

$$\breve{\nu}_k = w_k^{\alpha_k}\left(\hat{R} - \hat{R}_k^*\right)$$

for some $\alpha_k \geq 0$ (suggested to be 1, see below comments) and with

$$\hat{R}_k^* = \frac{\left(\sum_{l\in s} w_l y_l - w_k^{1-\alpha_k} y_k\right) / \left(\sum_{l\in s} w_l - w_k^{1-\alpha_k}\right)}{\left(\sum_{l\in s} w_l x_l - w_k^{1-\alpha_k} x_k\right) / \left(\sum_{l\in s} w_l - w_k^{1-\alpha_k}\right)} = \frac{\sum_{l\in s} w_l y_l - w_k^{1-\alpha_k} y_k}{\sum_{l\in s} w_l x_l - w_k^{1-\alpha_k} x_k}$$

Regarding the value of α_k, Escobar-Berger (2013) show that $\hat{V}(\hat{R})$ is valid for $\alpha_k \geq 0$ but conclude that $\alpha_k > 0$ should be used as $\alpha_k = 0$ corresponds to a naive biased and unstable jackknife. They recommend $\alpha_k = 1$ or $\alpha_k > 1$. If $\alpha_k = 1$, $\hat{V}(\hat{R})$ reduces to the Escobar-Berger (2011) jackknife. Using $\alpha_k > 1$ results in approximating the empirical influence function, i.e. the Gateaux (1919) derivative, or Demnati-Rao (2004) linearisation variance estimators. The larger the α_k, the closer the approximation. Further, Escobar-Berger (2013) give an intuitive explanation of the replication method from a jackknife and bootstrap perspective.

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Demnati, A. and Rao, J. N. K. (2004) Linearization variance estimators for survey data. *Survey Methodology*, **30**, 17–26.

Escobar, E. L. and Berger, Y. G. (2011) Jackknife variance estimation for functions of Horvitz-Thompson estimators under unequal probability sampling without replacement. In Proceeding of the *58th World Statistics Congress*. Dublin, Ireland: International Statistical Institute.

Escobar, E. L. and Berger, Y. G. (2013) A new replicate variance estimator for unequal probability sampling without replacement. *Canadian Journal of Statistics* **41**, 3, 508–524.

Gateaux, R. (1919) Fonctions d'une infinite de variables indeependantes. *Bulletin de la Societe Mathematique de France*, **47**, 70–96.

Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

See Also

VE.Lin.HT.Ratio
VE.Lin.SYG.Ratio
VE.Jk.Tukey.Ratio
VE.Jk.CBS.HT.Ratio
VE.Jk.CBS.SYG.Ratio
VE.Jk.B.Ratio
VE.Jk.EB.SW2.Ratio
VE.EB.SYG.Ratio

Examples

```
data(oaxaca)                                 #Loads the Oaxaca municipalities dataset
pik.U   <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s       <- oaxaca$sHOMES00                   #Defines the sample to be used
y1      <- oaxaca$POP10                      #Defines the numerator variable y1
y2      <- oaxaca$POPMAL10                   #Defines the numerator variable y2
x       <- oaxaca$HOMES10                    #Defines the denominator variable x
Alpha.s <- rep(2, times=373)                 #Defines the vector with Alpha values
#This approximation is only suitable for large-entropy sampling designs
pikl.s  <- Pkl.Hajek.s(pik.U[s==1])          #Approx. 2nd order incl. probs. from s
#Computes the var. est. of the ratio point estimator using y1
VE.EB.HT.Ratio(y1[s==1], x[s==1], pik.U[s==1], pikl.s) #Using default VecAlpha.s
#Computes the var. est. of the ratio point estimator using y2
VE.EB.HT.Ratio(y2[s==1], x[s==1], pik.U[s==1], pikl.s, Alpha.s)
```

VE.EB.HT.Total.Hajek *The Escobar-Berger unequal probability replicate variance estimator for the Hajek (1971) estimator of a total (Horvitz-Thompson form)*

Description

Computes the Escobar-Berger (2013) unequal probability replicate variance estimator for the Hajek estimator of a total. It uses the Horvitz-Thompson (1952) variance form.

Usage

```
VE.EB.HT.Total.Hajek(VecY.s, VecPk.s, MatPkl.s, N,
                     VecAlpha.s = rep(1, times=length(VecPk.s)))
```

Arguments

- `VecY.s` — vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s`. There must not be missing values.
- `VecPk.s` — vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.
- `MatPkl.s` — matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in `MatPkl.s` must be greater than zero and less than or equal to one. There must not be missing values.
- `N` — the population size. It must be an integer or a double-precision scalar with zero-valued fractional part.
- `VecAlpha.s` — vector of the α_k values; its length is equal to n, the sample size. Values in `VecAlpha.s` can be different for each unit and they must be greater or equal to zero. Escobar-Berger (2013) showed that this replicate variance estimator is valid for $\alpha_k \geq 0$. In particular, they suggest using $\alpha_k = 1$ for all units in the sample (the default for `VecAlpha.s` if omitted in the function call). Using $\alpha_k > 1$ results in approximating the Demnati-Rao (2004) linearisation variance estimators. There must not be missing values.

Details

For the population total of the variable y:

$$t = \sum_{k \in U} y_k$$

the approximately unbiased Hajek (1971) estimator of t is given by:

$$\hat{t}_{Hajek} = N \frac{\sum_{k \in s} w_k y_k}{\sum_{k \in s} w_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. The variance of $\hat{t}_{Hajek}$ can be estimated by the Escobar-Berger (2013) unequal probability replicate variance estimator (implemented by the current function):

$$\hat{V}(\hat{t}_{Hajek}) = \sum_{k \in s} \sum_{l \in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \breve{\nu}_k \breve{\nu}_l$$

where

$$\breve{\nu}_k = w_k^{\alpha_k} \left(\hat{t}_{Hajek} - \hat{t}^*_{Hajek,k} \right)$$

for some $\alpha_k \geq 0$ (suggested to be 1, see below comments) and with

$$\hat{t}^*_{Hajek,k} = N \frac{\sum_{l \in s} w_l y_l - w_k^{1-\alpha_k} y_k}{\sum_{l \in s} w_l - w_k^{1-\alpha_k}}$$

Regarding the value of α_k, Escobar-Berger (2013) show that $\hat{V}(\hat{t}_{Hajek})$ is valid for $\alpha_k \geq 0$ but conclude that $\alpha_k > 0$ should be used as $\alpha_k = 0$ corresponds to a naive biased and unstable jackknife. They recommend $\alpha_k = 1$ or $\alpha_k > 1$. If $\alpha_k = 1$, $\hat{V}(\hat{t}_{Hajek})$ reduces to the Escobar-Berger

(2011) jackknife. Using $\alpha_k > 1$ results in approximating the empirical influence function, i.e. the Gateaux (1919) derivative, or Demnati-Rao (2004) linearisation variance estimators. The larger the α_k, the closer the approximation. Further, Escobar-Berger (2013) give an intuitive explanation of the replication method from a jackknife and bootstrap perspective.

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Demnati, A. and Rao, J. N. K. (2004) Linearization variance estimators for survey data. *Survey Methodology*, **30**, 17–26.

Escobar, E. L. and Berger, Y. G. (2011) Jackknife variance estimation for functions of Horvitz-Thompson estimators under unequal probability sampling without replacement. In Proceeding of the *58th World Statistics Congress*. Dublin, Ireland: International Statistical Institute.

Escobar, E. L. and Berger, Y. G. (2013) A new replicate variance estimator for unequal probability sampling without replacement. *Canadian Journal of Statistics* **41**, 3, 508–524.

Gateaux, R. (1919) Fonctions d'une infinite de variables indeependantes. *Bulletin de la Societe Mathematique de France*, **47**, 70–96.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

See Also

VE.Jk.Tukey.Total.Hajek
VE.Jk.CBS.SYG.Total.Hajek
VE.Jk.B.Total.Hajek
VE.Jk.EB.SW2.Total.Hajek
VE.EB.SYG.Total.Hajek

Examples

```
data(oaxaca)                                 #Loads the Oaxaca municipalities dataset
pik.U   <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s       <- oaxaca$sHOMES00                   #Defines the sample to be used
N       <- dim(oaxaca)[1]                    #Defines the population size
y1      <- oaxaca$POP10                      #Defines the variable of interest y1
y2      <- oaxaca$POPMAL10                   #Defines the variable of interest y2
Alpha.s <- rep(2, times=373)                 #Defines the vector with Alpha values
#This approximation is only suitable for large-entropy sampling designs
pikl.s  <- Pkl.Hajek.s(pik.U[s==1])          #Approx. 2nd order incl. probs. from s
```

```
#Computes the var. est. of the Hajek total point estimator using y1
VE.EB.HT.Total.Hajek(y1[s==1], pik.U[s==1], pikl.s, N)
#Computes the var. est. of the Hajek total point estimator using y2
VE.EB.HT.Total.Hajek(y2[s==1], pik.U[s==1], pikl.s, N, Alpha.s)
```

VE.EB.SYG.Mean.Hajek *The Escobar-Berger unequal probability replicate variance estimator for the Hajek (1971) estimator of a mean (Sen-Yates-Grundy form)*

Description

Computes the Escobar-Berger (2013) unequal probability replicate variance estimator for the Hajek estimator of a mean. It uses the Sen (1953); Yates-Grundy(1953) variance form.

Usage

```
VE.EB.SYG.Mean.Hajek(VecY.s, VecPk.s, MatPkl.s,
                     VecAlpha.s = rep(1, times=length(VecPk.s)))
```

Arguments

VecY.s	vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s. There must not be missing values.
VecPk.s	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be missing values.
MatPkl.s	matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in MatPkl.s must be greater than zero and less than or equal to one. There must not be missing values.
VecAlpha.s	vector of the α_k values; its length is equal to n, the sample size. Values in VecAlpha.s can be different for each unit and they must be greater or equal to zero. Escobar-Berger (2013) showed that this replicate variance estimator is valid for $\alpha_k \geq 0$. In particular, they suggest using $\alpha_k = 1$ for all units in the sample (the default for VecAlpha.s if omitted in the function call). Using $\alpha_k > 1$ results in approximating the Demnati-Rao (2004) linearisation variance estimators. There must not be missing values.

Details

For the population mean of the variable y:

$$\bar{y} = \frac{1}{N} \sum_{k \in U} y_k$$

the approximately unbiased Hajek (1971) estimator of $\bar{y}$ is given by:

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k \in s} w_k y_k}{\sum_{k \in s} w_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. The variance of $\hat{\bar{y}}_{Hajek}$ can be estimated by the Escobar-Berger (2013) unequal probability replicate variance estimator (implemented by the current function):

$$\hat{V}(\hat{\bar{y}}_{Hajek}) = \frac{-1}{2}\sum_{k\in s}\sum_{l\in s}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}}(\breve{\nu}_k - \breve{\nu}_l)^2$$

where

$$\breve{\nu}_k = w_k^{\alpha_k}\left(\hat{\bar{y}}_{Hajek} - \hat{\bar{y}}^*_{Hajek,k}\right)$$

for some $\alpha_k \geq 0$ (suggested to be 1, see below comments) and with

$$\hat{\bar{y}}^*_{Hajek,k} = \frac{\sum_{l\in s} w_l y_l - w_k^{1-\alpha_k} y_k}{\sum_{l\in s} w_l - w_k^{1-\alpha_k}}$$

Regarding the value of α_k, Escobar-Berger (2013) show that $\hat{V}(\hat{\bar{y}}_{Hajek})$ is valid for $\alpha_k \geq 0$ but conclude that $\alpha_k > 0$ should be used as $\alpha_k = 0$ corresponds to a naive biased and unstable jackknife. They recommend $\alpha_k = 1$ or $\alpha_k > 1$. If $\alpha_k = 1$, $\hat{V}(\hat{\bar{y}}_{Hajek})$ reduces to the Escobar-Berger (2011) jackknife. Using $\alpha_k > 1$ results in approximating the empirical influence function, i.e. the Gateaux (1919) derivative, or Demnati-Rao (2004) linearisation variance estimators. The larger the α_k, the closer the approximation. Further, Escobar-Berger (2013) give an intuitive explanation of the replication method from a jackknife and bootstrap perspective.

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Demnati, A. and Rao, J. N. K. (2004) Linearization variance estimators for survey data. *Survey Methodology*, **30**, 17–26.

Escobar, E. L. and Berger, Y. G. (2011) Jackknife variance estimation for functions of Horvitz-Thompson estimators under unequal probability sampling without replacement. In Proceeding of the *58th World Statistics Congress*. Dublin, Ireland: International Statistical Institute.

Escobar, E. L. and Berger, Y. G. (2013) A new replicate variance estimator for unequal probability sampling without replacement. *Canadian Journal of Statistics* **41**, 3, 508–524.

Gateaux, R. (1919) Fonctions d'une infinite de variables indeependantes. *Bulletin de la Societe Mathematique de France*, **47**, 70–96.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Sen, A. R. (1953) On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **5**, 119–127.

Yates, F. and Grundy, P. M. (1953) Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B*, **15**, 253–261.

See Also

VE.Jk.Tukey.Mean.Hajek
VE.Jk.CBS.HT.Mean.Hajek
VE.Jk.B.Mean.Hajek
VE.Jk.EB.SW2.Mean.Hajek
VE.EB.HT.Mean.Hajek

Examples

```
data(oaxaca)                                 #Loads the Oaxaca municipalities dataset
pik.U   <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s       <- oaxaca$sHOMES00                   #Defines the sample to be used
y1      <- oaxaca$POP10                      #Defines the variable of interest y1
y2      <- oaxaca$POPMAL10                   #Defines the variable of interest y2
Alpha.s <- rep(2, times=373)                 #Defines the vector with Alpha values
#This approximation is only suitable for large-entropy sampling designs
pikl.s  <- Pkl.Hajek.s(pik.U[s==1])          #Approx. 2nd order incl. probs. from s
#Computes the var. est. of the Hajek mean point estimator using y1
VE.EB.SYG.Mean.Hajek(y1[s==1], pik.U[s==1], pikl.s)
#Computes the var. est. of the Hajek mean point estimator using y2
VE.EB.SYG.Mean.Hajek(y2[s==1], pik.U[s==1], pikl.s, Alpha.s)
```

VE.EB.SYG.Ratio — *The Escobar-Berger unequal probability replicate variance estimator for the estimator of a ratio (Sen-Yates-Grundy form)*

Description

Computes the Escobar-Berger (2013) unequal probability replicate variance estimator for the estimator of a ratio of two totals/means. It uses the Sen (1953); Yates-Grundy(1953) variance form.

Usage

```
VE.EB.SYG.Ratio(VecY.s, VecX.s, VecPk.s, MatPkl.s,
                VecAlpha.s = rep(1, times=length(VecPk.s)))
```

Arguments

VecY.s	vector of the numerator variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecX.s. There must not be missing values.

`VecX.s` vector of the denominator variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecY.s`. There must not be missing values. All values of `VecX.s` should be greater than zero. A warning is displayed if this does not hold and computations continue if mathematical expressions allow this kind of values for the denominator variable.

`VecPk.s` vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.

`MatPkl.s` matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in `MatPkl.s` must be greater than zero and less than or equal to one. There must not be missing values.

`VecAlpha.s` vector of the α_k values; its length is equal to n, the sample size. Values in `VecAlpha.s` can be different for each unit and they must be greater or equal to zero. Escobar-Berger (2013) showed that this replicate variance estimator is valid for $\alpha_k \geq 0$. In particular, they suggest using $\alpha_k = 1$ for all units in the sample (the default for `VecAlpha.s` if omitted in the function call). Using $\alpha_k > 1$ results in approximating the Demnati-Rao (2004) linearisation variance estimators. There must not be missing values.

Details

For the population ratio of two totals/means of the variables y and x:

$$R = \frac{\sum_{k\in U} y_k/N}{\sum_{k\in U} x_k/N} = \frac{\sum_{k\in U} y_k}{\sum_{k\in U} x_k}$$

the ratio estimator of R is given by:

$$\hat{R} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k x_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. The variance of $\hat{R}$ can be estimated by the Escobar-Berger (2013) unequal probability replicate variance estimator (implemented by the current function):

$$\hat{V}(\hat{R}) = \frac{-1}{2} \sum_{k\in s} \sum_{l\in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} (\breve{\nu}_k - \breve{\nu}_l)^2$$

where

$$\breve{\nu}_k = w_k^{\alpha_k} \left(\hat{R} - \hat{R}_k^*\right)$$

for some $\alpha_k \geq 0$ (suggested to be 1, see below comments) and with

$$\hat{R}_k^* = \frac{\left(\sum_{l\in s} w_l y_l - w_k^{1-\alpha_k} y_k\right) / \left(\sum_{l\in s} w_l - w_k^{1-\alpha_k}\right)}{\left(\sum_{l\in s} w_l x_l - w_k^{1-\alpha_k} x_k\right) / \left(\sum_{l\in s} w_l - w_k^{1-\alpha_k}\right)} = \frac{\sum_{l\in s} w_l y_l - w_k^{1-\alpha_k} y_k}{\sum_{l\in s} w_l x_l - w_k^{1-\alpha_k} x_k}$$

Regarding the value of α_k, Escobar-Berger (2013) show that $\hat{V}(\hat{R})$ is valid for $\alpha_k \geq 0$ but conclude that $\alpha_k > 0$ should be used as $\alpha_k = 0$ corresponds to a naive biased and unstable jackknife. They recommend $\alpha_k = 1$ or $\alpha_k > 1$. If $\alpha_k = 1$, $\hat{V}(\hat{R})$ reduces to the Escobar-Berger (2011) jackknife.

Using $\alpha_k > 1$ results in approximating the empirical influence function, i.e. the Gateaux (1919) derivative, or Demnati-Rao (2004) linearisation variance estimators. The larger the α_k, the closer the approximation. Further, Escobar-Berger (2013) give an intuitive explanation of the replication method from a jackknife and bootstrap perspective.

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Demnati, A. and Rao, J. N. K. (2004) Linearization variance estimators for survey data. *Survey Methodology*, **30**, 17–26.

Escobar, E. L. and Berger, Y. G. (2011) Jackknife variance estimation for functions of Horvitz-Thompson estimators under unequal probability sampling without replacement. In Proceeding of the *58th World Statistics Congress*. Dublin, Ireland: International Statistical Institute.

Escobar, E. L. and Berger, Y. G. (2013) A new replicate variance estimator for unequal probability sampling without replacement. *Canadian Journal of Statistics* **41**, 3, 508–524.

Gateaux, R. (1919) Fonctions d'une infinite de variables indeependantes. *Bulletin de la Societe Mathematique de France*, **47**, 70–96.

Sen, A. R. (1953) On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **5**, 119–127.

Yates, F. and Grundy, P. M. (1953) Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B*, **15**, 253–261.

See Also

VE.Lin.HT.Ratio
VE.Lin.SYG.Ratio
VE.Jk.Tukey.Ratio
VE.Jk.CBS.HT.Ratio
VE.Jk.CBS.SYG.Ratio
VE.Jk.B.Ratio
VE.Jk.EB.SW2.Ratio
VE.EB.HT.Ratio

Examples

```
data(oaxaca)                                 #Loads the Oaxaca municipalities dataset
pik.U   <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s       <- oaxaca$sHOMES00                   #Defines the sample to be used
y1      <- oaxaca$POP10                      #Defines the numerator variable y1
y2      <- oaxaca$POPMAL10                   #Defines the numerator variable y2
x       <- oaxaca$HOMES10                    #Defines the denominator variable x
Alpha.s <- rep(2, times=373)                 #Defines the vector with Alpha values
```

```
#This approximation is only suitable for large-entropy sampling designs
pikl.s  <- Pkl.Hajek.s(pik.U[s==1])           #Approx. 2nd order incl. probs. from s
#Computes the var. est. of the ratio point estimator using y1
VE.EB.SYG.Ratio(y1[s==1], x[s==1], pik.U[s==1], pikl.s) #Using default VecAlpha.s
#Computes the var. est. of the ratio point estimator using y2
VE.EB.SYG.Ratio(y2[s==1], x[s==1], pik.U[s==1], pikl.s, Alpha.s)
```

`VE.EB.SYG.Total.Hajek` *The Escobar-Berger unequal probability replicate variance estimator for the Hajek (1971) estimator of a total (Sen-Yates-Grundy form)*

Description

Computes the Escobar-Berger (2013) unequal probability replicate variance estimator for the Hajek estimator of a total. It uses the Sen (1953); Yates-Grundy(1953) variance form.

Usage

```
VE.EB.SYG.Total.Hajek(VecY.s, VecPk.s, MatPkl.s, N,
                      VecAlpha.s = rep(1, times=length(VecPk.s)))
```

Arguments

`VecY.s` vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s`. There must not be missing values.

`VecPk.s` vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.

`MatPkl.s` matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in `MatPkl.s` must be greater than zero and less than or equal to one. There must not be missing values.

`N` the population size. It must be an integer or a double-precision scalar with zero-valued fractional part.

`VecAlpha.s` vector of the α_k values; its length is equal to n, the sample size. Values in `VecAlpha.s` can be different for each unit and they must be greater or equal to zero. Escobar-Berger (2013) showed that this replicate variance estimator is valid for $\alpha_k \geq 0$. In particular, they suggest using $\alpha_k = 1$ for all units in the sample (the default for `VecAlpha.s` if omitted in the function call). Using $\alpha_k > 1$ results in approximating the Demnati-Rao (2004) linearisation variance estimators. There must not be missing values.

Details

For the population total of the variable y:

$$t = \sum_{k \in U} y_k$$

the approximately unbiased Hajek (1971) estimator of t is given by:

$$\hat{t}_{Hajek} = N \frac{\sum_{k \in s} w_k y_k}{\sum_{k \in s} w_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. The variance of $\hat{t}_{Hajek}$ can be estimated by the Escobar-Berger (2013) unequal probability replicate variance estimator (implemented by the current function):

$$\hat{V}(\hat{t}_{Hajek}) = \frac{-1}{2} \sum_{k \in s} \sum_{l \in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} (\breve{\nu}_k - \breve{\nu}_l)^2$$

where

$$\breve{\nu}_k = w_k^{\alpha_k} \left(\hat{t}_{Hajek} - \hat{t}^*_{Hajek,k} \right)$$

for some $\alpha_k \geq 0$ (suggested to be 1, see below comments) and with

$$\hat{t}^*_{Hajek,k} = N \frac{\sum_{l \in s} w_l y_l - w_k^{1-\alpha_k} y_k}{\sum_{l \in s} w_l - w_k^{1-\alpha_k}}$$

Regarding the value of α_k, Escobar-Berger (2013) show that $\hat{V}(\hat{t}_{Hajek})$ is valid for $\alpha_k \geq 0$ but conclude that $\alpha_k > 0$ should be used as $\alpha_k = 0$ corresponds to a naive biased and unstable jackknife. They recommend $\alpha_k = 1$ or $\alpha_k > 1$. If $\alpha_k = 1$, $\hat{V}(\hat{t}_{Hajek})$ reduces to the Escobar-Berger (2011) jackknife. Using $\alpha_k > 1$ results in approximating the empirical influence function, i.e. the Gateaux (1919) derivative, or Demnati-Rao (2004) linearisation variance estimators. The larger the α_k, the closer the approximation. Further, Escobar-Berger (2013) give an intuitive explanation of the replication method from a jackknife and bootstrap perspective.

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Demnati, A. and Rao, J. N. K. (2004) Linearization variance estimators for survey data. *Survey Methodology*, **30**, 17–26.

Escobar, E. L. and Berger, Y. G. (2011) Jackknife variance estimation for functions of Horvitz-Thompson estimators under unequal probability sampling without replacement. In Proceeding of the *58th World Statistics Congress*. Dublin, Ireland: International Statistical Institute.

Escobar, E. L. and Berger, Y. G. (2013) A new replicate variance estimator for unequal probability sampling without replacement. *Canadian Journal of Statistics* **41**, 3, 508–524.

Gateaux, R. (1919) Fonctions d'une infinite de variables indeependantes. *Bulletin de la Societe Mathematique de France*, **47**, 70–96.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Sen, A. R. (1953) On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **5**, 119–127.

Yates, F. and Grundy, P. M. (1953) Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B*, **15**, 253–261.

See Also

VE.Jk.Tukey.Total.Hajek
VE.Jk.CBS.HT.Total.Hajek
VE.Jk.B.Total.Hajek
VE.Jk.EB.SW2.Total.Hajek
VE.EB.SYG.Total.Hajek

Examples

```
data(oaxaca)                                 #Loads the Oaxaca municipalities dataset
pik.U   <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s       <- oaxaca$sHOMES00                   #Defines the sample to be used
N       <- dim(oaxaca)[1]                    #Defines the population size
y1      <- oaxaca$POP10                      #Defines the variable of interest y1
y2      <- oaxaca$POPMAL10                   #Defines the variable of interest y2
Alpha.s <- rep(2, times=373)                 #Defines the vector with Alpha values
#This approximation is only suitable for large-entropy sampling designs
pikl.s  <- Pkl.Hajek.s(pik.U[s==1])          #Approx. 2nd order incl. probs. from s
#Computes the var. est. of the Hajek total point estimator using y1
VE.EB.SYG.Total.Hajek(y1[s==1], pik.U[s==1], pikl.s, N)
#Computes the var. est. of the Hajek total point estimator using y2
VE.EB.SYG.Total.Hajek(y2[s==1], pik.U[s==1], pikl.s, N, Alpha.s)
```

`VE.Hajek.Mean.NHT` *The Hajek variance estimator for the Narain-Horvitz-Thompson point estimator for a mean*

Description

Computes the Hajek (1964) variance estimator for the Narain (1951); Horvitz-Thompson (1952) point estimator for a population mean.

Usage

```
VE.Hajek.Mean.NHT(VecY.s, VecPk.s, N)
```

Arguments

`VecY.s`	vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s`. There must not be missing values.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.
`N`	the population size. It must be an integer or a double-precision scalar with zero-valued fractional part.

Details

For the population mean of the variable y:

$$\bar{y} = \frac{1}{N} \sum_{k \in U} y_k$$

the unbiased Narain (1951); Horvitz-Thompson (1952) estimator of $\bar{y}$ is given by:

$$\hat{\bar{y}}_{NHT} = \frac{1}{N} \sum_{k \in s} \frac{y_k}{\pi_k}$$

where π_k denotes the inclusion probability of the k-th element in the sample s. For large-entropy sampling designs, the variance of $\hat{\bar{y}}_{NHT}$ is approximated by the Hajek (1964) variance:

$$V(\hat{\bar{y}}_{NHT}) = \frac{1}{N(N-1)} \left[\sum_{k \in U} \frac{y_k^2}{\pi_k}(1-\pi_k) - dG^2 \right]$$

with $d = \sum_{k \in U} \pi_k(1-\pi_k)$ and $G = d^{-1} \sum_{k \in U}(1-\pi_k)y_k$.

The variance $V(\hat{t}_{NHT})$ can be estimated by the variance estimator (implemented by the current function):

$$\hat{V}(\hat{\bar{y}}_{NHT}) = \frac{n}{N^2(n-1)} \left[\sum_{k \in s} \left(\frac{y_k}{\pi_k} \right)^2 (1-\pi_k) - \hat{d}\hat{G}^2 \right]$$

where $\hat{d} = \sum_{k \in s}(1-\pi_k)$ and $\hat{G} = \hat{d}^{-1} \sum_{k \in s}(1-\pi)y_k/\pi_k$.

Note that the Hajek (1964) variance approximation is designed for large-entropy sampling designs, large samples and large populations, i.e. care should be taken with highly-stratified samples, e.g. Berger (2005).

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Berger, Y. G. (2005) Variance estimation with highly stratified sampling designs with unequal probabilities. *Australian & New Zealand Journal of Statistics*, **47**, 365–373.

Hajek, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35**, 4, 1491–1523.

Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Narain, R. D. (1951) On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **3**, 169–175.

See Also

VE.HT.Mean.NHT
VE.SYG.Mean.NHT

Examples

```
data(oaxaca)                                 #Loads the Oaxaca municipalities dataset
pik.U <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s     <- oaxaca$sHOMES00                     #Defines the sample to be used
N     <- dim(oaxaca)[1]                      #Defines the population size
y1    <- oaxaca$POP10                        #Defines the variable of interest y1
y2    <- oaxaca$HOMES10                      #Defines the variable of interest y2
#Computes the (approximate) var. est. of the NHT point est. for y1
VE.Hajek.Mean.NHT(y1[s==1], pik.U[s==1], N)
#Computes the (approximate) var. est. of the NHT point est. for y2
VE.Hajek.Mean.NHT(y2[s==1], pik.U[s==1], N)
```

`VE.Hajek.Total.NHT` *The Hajek variance estimator for the Narain-Horvitz-Thompson point estimator for a total*

Description

Computes the Hajek (1964) variance estimator for the Narain (1951); Horvitz-Thompson (1952) point estimator for a population total.

Usage

```
VE.Hajek.Total.NHT(VecY.s, VecPk.s)
```

Arguments

`VecY.s`	vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s`. There must not be missing values.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.

Details

For the population total of the variable y:

$$t = \sum_{k \in U} y_k$$

the unbiased Narain (1951); Horvitz-Thompson (1952) estimator of t is given by:

$$\hat{t}_{NHT} = \sum_{k \in s} \frac{y_k}{\pi_k}$$

where π_k denotes the inclusion probability of the k-th element in the sample s. For large-entropy sampling designs, the variance of $\hat{t}_{NHT}$ is approximated by the Hajek (1964) variance:

$$V(\hat{t}_{NHT}) = \frac{N}{N-1} \left[\sum_{k \in U} \frac{y_k^2}{\pi_k}(1-\pi_k) - dG^2 \right]$$

with $d = \sum_{k \in U} \pi_k(1-\pi_k)$ and $G = d^{-1} \sum_{k \in U}(1-\pi_k)y_k$.

The variance $V(\hat{t}_{NHT})$ can be estimated by the variance estimator (implemented by the current function):

$$\hat{V}(\hat{t}_{NHT}) = \frac{n}{n-1} \left[\sum_{k \in s} \left(\frac{y_k}{\pi_k} \right)^2 (1-\pi_k) - \hat{d}\hat{G}^2 \right]$$

where $\hat{d} = \sum_{k \in s}(1-\pi_k)$ and $\hat{G} = \hat{d}^{-1} \sum_{k \in s}(1-\pi)y_k/\pi_k$.

Note that the Hajek (1964) variance approximation is designed for large-entropy sampling designs, large samples and large populations, i.e. care should be taken with highly-stratified samples, e.g. Berger (2005).

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Berger, Y. G. (2005) Variance estimation with highly stratified sampling designs with unequal probabilities. *Australian & New Zealand Journal of Statistics*, **47**, 365–373.

Hajek, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35**, 4, 1491–1523.

Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Narain, R. D. (1951) On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **3**, 169–175.

See Also

VE.HT.Total.NHT
VE.SYG.Total.NHT

Examples

```
data(oaxaca)                                 #Loads the Oaxaca municipalities dataset
pik.U <- Pk.PropNorm.U(373, oaxaca$SURFAC05) #Reconstructs the 1st order incl. probs.
s     <- oaxaca$sSURFAC                      #Defines the sample to be used
y1    <- oaxaca$POP10                        #Defines the variable of interest y1
y2    <- oaxaca$HOMES10                      #Defines the variable of interest y2
#Computes the (approximate) var. est. of the NHT point est. from y1
VE.Hajek.Total.NHT(y1[s==1], pik.U[s==1])
#Computes the (approximate) var. est. of the NHT point est. from y2
VE.Hajek.Total.NHT(y2[s==1], pik.U[s==1])
```

VE.HT.Mean.NHT — *The Horvitz-Thompson variance estimator for the Narain-Horvitz-Thompson point estimator for a mean*

Description

Computes the Horvitz-Thompson (1952) variance estimator for the Narain (1951); Horvitz-Thompson (1952) point estimator for a population mean.

Usage

```
VE.HT.Mean.NHT(VecY.s, VecPk.s, MatPkl.s, N)
```

Arguments

VecY.s	vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s. There must not be missing values.

VecPk.s	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be missing values.
MatPkl.s	matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in MatPkl.s must be greater than zero and less than or equal to one. There must not be missing values.
N	the population size. It must be an integer or a double-precision scalar with zero-valued fractional part.

Details

For the population mean of the variable y:

$$\bar{y} = \frac{1}{N}\sum_{k\in U} y_k$$

the unbiased Narain (1951); Horvitz-Thompson (1952) estimator of $\bar{y}$ is given by:

$$\hat{\bar{y}}_{NHT} = \frac{1}{N}\sum_{k\in s}\frac{y_k}{\pi_k}$$

where π_k denotes the inclusion probability of the k-th element in the sample s. Let π_{kl} denotes the joint-inclusion probabilities of the k-th and l-th elements in the sample s. The variance of $\hat{\bar{y}}_{NHT}$ is given by:

$$V(\hat{\bar{y}}_{NHT}) = \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}(\pi_{kl}-\pi_k\pi_l)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}$$

which can therefore be estimated by the Horvitz-Thompson variance estimator (implemented by the current function):

$$\hat{V}(\hat{\bar{y}}_{NHT}) = \frac{1}{N^2}\sum_{k\in s}\sum_{l\in s}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}$$

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Narain, R. D. (1951) On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **3**, 169–175.

See Also

VE.SYG.Mean.NHT
VE.Hajek.Mean.NHT

Examples

```
data(oaxaca)                                #Loads the Oaxaca municipalities dataset
pik.U  <- Pk.PropNorm.U(373, oaxaca$SURFAC05) #Reconstructs the 1st order incl. probs.
s      <- oaxaca$sSURFAC                    #Defines the sample to be used
N      <- dim(oaxaca)[1]                    #Defines the population size
y1     <- oaxaca$POP10                      #Defines the variable of interest y1
y2     <- oaxaca$HOMES10                    #Defines the variable of interest y2
#This approximation is only suitable for large-entropy sampling designs
pikl.s <- Pkl.Hajek.s(pik.U[s==1])          #Approx. 2nd order incl. probs. from s
#Computes the variance estimation of the NHT point estimator for y1
VE.HT.Mean.NHT(y1[s==1], pik.U[s==1], pikl.s, N)
#Computes the variance estimation of the NHT point estimator for y2
VE.HT.Mean.NHT(y2[s==1], pik.U[s==1], pikl.s, N)
```

`VE.HT.Total.NHT` *The Horvitz-Thompson variance estimator for the Narain-Horvitz-Thompson point estimator for a total*

Description

Computes the Horvitz-Thompson (1952) variance estimator for the Narain (1951); Horvitz-Thompson (1952) point estimator for a population total.

Usage

```
VE.HT.Total.NHT(VecY.s, VecPk.s, MatPkl.s)
```

Arguments

`VecY.s` vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s`. There must not be missing values.

`VecPk.s` vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.

`MatPkl.s` matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in `MatPkl.s` must be greater than zero and less than or equal to one. There must not be missing values.

Details

For the population total of the variable y:

$$t = \sum_{k \in U} y_k$$

the unbiased Narain (1951); Horvitz-Thompson (1952) estimator of t is given by:

$$\hat{t}_{NHT} = \sum_{k\in s} \frac{y_k}{\pi_k}$$

where π_k denotes the inclusion probability of the k-th element in the sample s. Let π_{kl} denotes the joint-inclusion probabilities of the k-th and l-th elements in the sample s. The variance of $\hat{t}_{NHT}$ is given by:

$$V(\hat{t}_{NHT}) = \sum_{k\in U}\sum_{l\in U}(\pi_{kl} - \pi_k\pi_l)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}$$

which can therefore be estimated by the Horvitz-Thompson variance estimator (implemented by the current function):

$$\hat{V}(\hat{t}_{NHT}) = \sum_{k\in s}\sum_{l\in s}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}$$

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Narain, R. D. (1951) On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **3**, 169–175.

See Also

VE.SYG.Total.NHT
VE.Hajek.Total.NHT

Examples

```
data(oaxaca)                                 #Loads the Oaxaca municipalities dataset
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s      <- oaxaca$sHOMES00                    #Defines the sample to be used
y1     <- oaxaca$POP10                       #Defines the variable of interest y1
y2     <- oaxaca$HOMES10                     #Defines the variable of interest y2
#This approximation is only suitable for large-entropy sampling designs
pikl.s <- Pkl.Hajek.s(pik.U[s==1])           #Approx. 2nd order incl. probs. from s
#Computes the var. est. of the NHT point estimator for y1
VE.HT.Total.NHT(y1[s==1], pik.U[s==1], pikl.s)
#Computes the var. est. of the NHT point estimator for y2
VE.HT.Total.NHT(y2[s==1], pik.U[s==1], pikl.s)
```

`VE.Jk.B.Corr.Hajek` *The Berger (2007) unequal probability jackknife variance estimator for the estimator of a correlation coefficient using the Hajek point estimator*

Description

Computes the Berger (2007) unequal probability jackknife variance estimator for the estimator of a correlation coefficient of two variables using the Hajek (1971) point estimator.

Usage

```
VE.Jk.B.Corr.Hajek(VecY.s, VecX.s, VecPk.s)
```

Arguments

`VecY.s`	vector of the variable of interest Y; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecX.s`. There must not be missing values.
`VecX.s`	vector of the variable of interest X; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecY.s`. There must not be missing values.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.

Details

For the population correlation coefficient of two variables y and x:

$$C = \frac{\sum_{k\in U}(y_k-\bar{y})(x_k-\bar{x})}{\sqrt{\sum_{k\in U}(y_k-\bar{y})^2}\sqrt{\sum_{k\in U}(x_k-\bar{x})^2}}$$

the point estimator of C, assuming that N is unknown (see Sarndal et al., 1992, Sec. 5.9), is:

$$\hat{C}_{Hajek} = \frac{\sum_{k\in s} w_k(y_k-\hat{\bar{y}}_{Hajek})(x_k-\hat{\bar{x}}_{Hajek})}{\sqrt{\sum_{k\in s} w_k(y_k-\hat{\bar{y}}_{Hajek})^2}\sqrt{\sum_{k\in s} w_k(x_k-\hat{\bar{x}}_{Hajek})^2}}$$

where $\hat{\bar{y}}_{Hajek}$ is the Hajek (1971) point estimator of the population mean $\bar{y} = N^{-1}\sum_{k\in U} y_k$,

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

and $w_k = 1/\pi_k$ with π_k denoting the inclusion probability of the k-th element in the sample s. The variance of $\hat{C}_{Hajek}$ can be estimated by the Berger (2007) unequal probability jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{C}_{Hajek}) = \sum_{k\in s}\frac{n}{n-1}(1-\pi_k)\left(\varepsilon_k - \hat{B}\right)^2$$

where

$$\hat{B} = \frac{\sum_{k\in s}(1-\pi_k)\varepsilon_k}{\sum_{k\in s}(1-\pi_k)}$$

and

$$\varepsilon_k = (1-\tilde{w}_k)\left(\hat{C}_{Hajek} - \hat{C}_{Hajek(k)}\right)$$

with

$$\tilde{w}_k = \frac{w_k}{\sum_{l\in s} w_l}$$

and where $\hat{C}_{Hajek(k)}$ has the same functional form as $\hat{C}_{Hajek}$ but omitting the k-th element from the sample s. Note that this variance estimator utilises implicitly the Hajek (1964) approximations that are designed for large-entropy sampling designs, large samples and large populations, i.e. care should be taken with highly-stratified samples, e.g. Berger (2005).

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Berger, Y. G. (2005) Variance estimation with highly stratified sampling designs with unequal probabilities. *Australian & New Zealand Journal of Statistics*, **47**, 365–373.

Berger, Y. G. (2007) A jackknife variance estimator for unistage stratified samples with unequal probabilities. *Biometrika* **94**, 953–964.

Hajek, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35**, 4, 1491–1523.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Sarndal, C.-E. and Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, Inc.

See Also

VE.Jk.Tukey.Corr.Hajek
VE.Jk.CBS.HT.Corr.Hajek
VE.Jk.CBS.SYG.Corr.Hajek
VE.Jk.EB.SW2.Corr.Hajek

Examples

```
data(oaxaca)                                    #Loads the Oaxaca municipalities dataset
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s      <- oaxaca$sHOMES00                       #Defines the sample to be used
y1     <- oaxaca$POP10                          #Defines the variable of interest y1
y2     <- oaxaca$POPMAL10                       #Defines the variable of interest y2
x      <- oaxaca$HOMES10                        #Defines the variable of interest x
#Computes the var. est. of the corr. coeff. point estimator using y1
VE.Jk.B.Corr.Hajek(y1[s==1], x[s==1], pik.U[s==1])
#Computes the var. est. of the corr. coeff. point estimator using y2
VE.Jk.B.Corr.Hajek(y2[s==1], x[s==1], pik.U[s==1])
```

`VE.Jk.B.Mean.Hajek` *The Berger (2007) unequal probability jackknife variance estimator for the Hajek estimator of a mean*

Description

Computes the Berger (2007) unequal probability jackknife variance estimator for the Hajek (1971) estimator of a mean.

Usage

```
VE.Jk.B.Mean.Hajek(VecY.s, VecPk.s)
```

Arguments

`VecY.s`	vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s`. There must not be missing values.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.

Details

For the population mean of the variable y:

$$\bar{y} = \frac{1}{N} \sum_{k \in U} y_k$$

the approximately unbiased Hajek (1971) estimator of $\bar{y}$ is given by:

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k \in s} w_k y_k}{\sum_{k \in s} w_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. The variance of $\hat{\bar{y}}_{Hajek}$ can be estimated by the Berger (2007) unequal probability jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{\bar{y}}_{Hajek}) = \sum_{k\in s} \frac{n}{n-1}(1-\pi_k)\left(\varepsilon_k - \hat{B}\right)^2$$

where

$$\hat{B} = \frac{\sum_{k\in s}(1-\pi_k)\varepsilon_k}{\sum_{k\in s}(1-\pi_k)}$$

and

$$\varepsilon_k = (1-\tilde{w}_k)\left(\hat{\bar{y}}_{Hajek} - \hat{\bar{y}}_{Hajek(k)}\right)$$

with

$$\tilde{w}_k = \frac{w_k}{\sum_{l\in s} w_l}$$

and

$$\hat{\bar{y}}_{Hajek(k)} = \frac{\sum_{l\in s, l\neq k} w_l y_l}{\sum_{l\in s, l\neq k} w_l}$$

Note that this variance estimator utilises implicitly the Hajek (1964) approximations that are designed for large-entropy sampling designs, large samples and large populations, i.e. care should be taken with highly-stratified samples, e.g. Berger (2005).

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Berger, Y. G. (2005) Variance estimation with highly stratified sampling designs with unequal probabilities. *Australian & New Zealand Journal of Statistics*, **47**, 365–373.

Berger, Y. G. (2007) A jackknife variance estimator for unistage stratified samples with unequal probabilities. *Biometrika* **94**, 953–964.

Hajek, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35**, 4, 1491–1523.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

See Also

VE.Jk.Tukey.Mean.Hajek
VE.Jk.CBS.HT.Mean.Hajek
VE.Jk.CBS.SYG.Mean.Hajek
VE.Jk.EB.SW2.Mean.Hajek

Examples

```
data(oaxaca)                                  #Loads the Oaxaca municipalities dataset
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s      <- oaxaca$sHOMES00                     #Defines the sample to be used
y1     <- oaxaca$POP10                        #Defines the variable of interest y1
y2     <- oaxaca$POPMAL10                     #Defines the variable of interest y2
#Computes the var. est. of the Hajek mean point estimator using y1
VE.Jk.B.Mean.Hajek(y1[s==1], pik.U[s==1])
#Computes the var. est. of the Hajek mean point estimator using y2
VE.Jk.B.Mean.Hajek(y2[s==1], pik.U[s==1])
```

VE.Jk.B.Ratio — *The Berger (2007) unequal probability jackknife variance estimator for the estimator of a ratio*

Description

Computes the Berger (2007) unequal probability jackknife variance estimator for the estimator of a ratio of two totals/means.

Usage

```
VE.Jk.B.Ratio(VecY.s, VecX.s, VecPk.s)
```

Arguments

VecY.s	vector of the numerator variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecX.s. There must not be missing values.
VecX.s	vector of the denominator variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecY.s. There must not be missing values. All values of VecX.s should be greater than zero. A warning is displayed if this does not hold and computations continue if mathematical expressions allow this kind of values for the denominator variable.
VecPk.s	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be missing values.

Details

For the population ratio of two totals/means of the variables y and x:

$$R = \frac{\sum_{k\in U} y_k/N}{\sum_{k\in U} x_k/N} = \frac{\sum_{k\in U} y_k}{\sum_{k\in U} x_k}$$

the ratio estimator of R is given by:

$$\hat{R} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k x_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. The variance of $\hat{R}$ can be estimated by the Berger (2007) unequal probability jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{R}) = \sum_{k \in s} \frac{n}{n-1}(1-\pi_k)\left(\varepsilon_k - \hat{B}\right)^2$$

where

$$\hat{B} = \frac{\sum_{k \in s}(1-\pi_k)\varepsilon_k}{\sum_{k \in s}(1-\pi_k)}$$

and

$$\varepsilon_k = (1-\tilde{w}_k)\left(\hat{R} - \hat{R}_{(k)}\right)$$

with

$$\tilde{w}_k = \frac{w_k}{\sum_{l \in s} w_l}$$

and

$$\hat{R}_{(k)} = \frac{\sum_{l \in s, l \neq k} w_l y_l / \sum_{l \in s, l \neq k} w_l}{\sum_{l \in s, l \neq k} w_l x_l / \sum_{l \in s, l \neq k} w_l} = \frac{\sum_{l \in s, l \neq k} w_l y_l}{\sum_{l \in s, l \neq k} w_l x_l}$$

Note that this variance estimator utilises implicitly the Hajek (1964) approximations that are designed for large-entropy sampling designs, large samples and large populations, i.e. care should be taken with highly-stratified samples, e.g. Berger (2005).

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Berger, Y. G. (2005) Variance estimation with highly stratified sampling designs with unequal probabilities. *Australian & New Zealand Journal of Statistics*, **47**, 365–373.

Berger, Y. G. (2007) A jackknife variance estimator for unistage stratified samples with unequal probabilities. *Biometrika* **94**, 953–964.

Hajek, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35**, 4, 1491–1523.

See Also

`VE.Lin.HT.Ratio`
`VE.Lin.SYG.Ratio`
`VE.Jk.Tukey.Ratio`
`VE.Jk.CBS.HT.Ratio`
`VE.Jk.CBS.SYG.Ratio`
`VE.Jk.EB.SW2.Ratio`
`VE.EB.HT.Ratio`
`VE.EB.SYG.Ratio`

Examples

```
data(oaxaca)                               #Loads the Oaxaca municipalities dataset
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s      <- oaxaca$sHOMES00                  #Defines the sample to be used
y1     <- oaxaca$POP10                     #Defines the numerator variable y1
y2     <- oaxaca$POPMAL10                  #Defines the numerator variable y2
x      <- oaxaca$HOMES10                   #Defines the denominator variable x
#Computes the var. est. of the ratio point estimator using y1
VE.Jk.B.Ratio(y1[s==1], x[s==1], pik.U[s==1])
#Computes the var. est. of the ratio point estimator using y2
VE.Jk.B.Ratio(y2[s==1], x[s==1], pik.U[s==1])
```

`VE.Jk.B.RegCo.Hajek` *The Berger (2007) unequal probability jackknife variance estimator for the estimator of the regression coefficient using the Hajek point estimator*

Description

Computes the Berger (2007) unequal probability jackknife variance estimator for the estimator of the regression coefficient using the Hajek (1971) point estimator.

Usage

```
VE.Jk.B.RegCo.Hajek(VecY.s, VecX.s, VecPk.s)
```

Arguments

`VecY.s`	vector of the variable of interest Y; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecX.s`. There must not be missing values.
`VecX.s`	vector of the variable of interest X; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecY.s`. There must not be missing values.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.

Details

From Linear Regression Analysis, for an imposed population model

$$y = \alpha + \beta x$$

the population regression coefficient β, assuming that the population size N is unknown (see Sarndal et al., 1992, Sec. 5.10), can be estimated by:

$$\hat{\beta}_{Hajek} = \frac{\sum_{k\in s} w_k (y_k - \hat{\bar{y}}_{Hajek})(x_k - \hat{\bar{x}}_{Hajek})}{\sum_{k\in s} w_k (x_k - \hat{\bar{x}}_{Hajek})^2}$$

where $\hat{\bar{y}}_{Hajek}$ and $\hat{\bar{x}}_{Hajek}$ are the Hajek (1971) point estimators of the population means $\bar{y} = N^{-1}\sum_{k\in U} y_k$ and $\bar{x} = N^{-1}\sum_{k\in U} x_k$, respectively,

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

$$\hat{\bar{x}}_{Hajek} = \frac{\sum_{k\in s} w_k x_k}{\sum_{k\in s} w_k}$$

and $w_k = 1/\pi_k$ with π_k denoting the inclusion probability of the k-th element in the sample s. The variance of $\hat{\beta}_{Hajek}$ can be estimated by the Berger (2007) unequal probability jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{\beta}_{Hajek}) = \sum_{k\in s} \frac{n}{n-1}(1-\pi_k)\left(\varepsilon_k - \hat{B}\right)^2$$

where

$$\hat{B} = \frac{\sum_{k\in s}(1-\pi_k)\varepsilon_k}{\sum_{k\in s}(1-\pi_k)}$$

and

$$\varepsilon_k = (1-\tilde{w}_k)\left(\hat{\beta}_{Hajek} - \hat{\beta}_{Hajek(k)}\right)$$

with

$$\tilde{w}_k = \frac{w_k}{\sum_{l\in s} w_l}$$

and where $\hat{\beta}_{Hajek(k)}$ has the same functional form as $\hat{\beta}_{Hajek}$ but omitting the k-th element from the sample s. Note that this variance estimator utilises implicitly the Hajek (1964) approximations that are designed for large-entropy sampling designs, large samples and large populations, i.e. care should be taken with highly-stratified samples, e.g. Berger (2005).

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Berger, Y. G. (2005) Variance estimation with highly stratified sampling designs with unequal probabilities. *Australian & New Zealand Journal of Statistics*, **47**, 365–373.

Berger, Y. G. (2007) A jackknife variance estimator for unistage stratified samples with unequal probabilities. *Biometrika* **94**, 953–964.

Hajek, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35**, 4, 1491–1523.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Sarndal, C.-E. and Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, Inc.

See Also

VE.Jk.B.RegCoI.Hajek
VE.Jk.Tukey.RegCo.Hajek
VE.Jk.CBS.HT.RegCo.Hajek
VE.Jk.CBS.SYG.RegCo.Hajek
VE.Jk.EB.SW2.RegCo.Hajek

Examples

```
data(oaxaca)                                #Loads the Oaxaca municipalities dataset
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s      <- oaxaca$sHOMES00                   #Defines the sample to be used
y1     <- oaxaca$POP10                      #Defines the variable of interest y1
y2     <- oaxaca$POPMAL10                   #Defines the variable of interest y2
x      <- oaxaca$HOMES10                    #Defines the variable of interest x
#Computes the var. est. of the regression coeff. point estimator using y1
VE.Jk.B.RegCo.Hajek(y1[s==1], x[s==1], pik.U[s==1])
#Computes the var. est. of the regression coeff. point estimator using y2
VE.Jk.B.RegCo.Hajek(y2[s==1], x[s==1], pik.U[s==1])
```

VE.Jk.B.RegCoI.Hajek — *The Berger (2007) unequal probability jackknife variance estimator for the estimator of the intercept regression coefficient using the Hajek point estimator*

Description

Computes the Berger (2007) unequal probability jackknife variance estimator for the estimator of the intercept regression coefficient using the Hajek (1971) point estimator.

Usage

```
VE.Jk.B.RegCoI.Hajek(VecY.s, VecX.s, VecPk.s)
```

Arguments

VecY.s	vector of the variable of interest Y; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecX.s. There must not be missing values.
VecX.s	vector of the variable of interest X; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecY.s. There must not be missing values.
VecPk.s	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be missing values.

Details

From Linear Regression Analysis, for an imposed population model

$$y = \alpha + \beta x$$

the population intercept regression coefficient α, assuming that the population size N is unknown (see Sarndal et al., 1992, Sec. 5.10), can be estimated by:

$$\hat{\alpha}_{Hajek} = \hat{\bar{y}}_{Hajek} - \frac{\sum_{k\in s} w_k (y_k - \hat{\bar{y}}_{Hajek})(x_k - \hat{\bar{x}}_{Hajek})}{\sum_{k\in s} w_k (x_k - \hat{\bar{x}}_{Hajek})^2} \hat{\bar{x}}_{Hajek}$$

where $\hat{\bar{y}}_{Hajek}$ and $\hat{\bar{x}}_{Hajek}$ are the Hajek (1971) point estimators of the population means $\bar{y} = N^{-1}\sum_{k\in U} y_k$ and $\bar{x} = N^{-1}\sum_{k\in U} x_k$, respectively,

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

$$\hat{\bar{x}}_{Hajek} = \frac{\sum_{k\in s} w_k x_k}{\sum_{k\in s} w_k}$$

and $w_k = 1/\pi_k$ with π_k denoting the inclusion probability of the k-th element in the sample s. The variance of $\hat{\alpha}_{Hajek}$ can be estimated by the Berger (2007) unequal probability jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{\alpha}_{Hajek}) = \sum_{k\in s} \frac{n}{n-1}(1-\pi_k)\left(\varepsilon_k - \hat{B}\right)^2$$

where

$$\hat{B} = \frac{\sum_{k\in s}(1-\pi_k)\varepsilon_k}{\sum_{k\in s}(1-\pi_k)}$$

and

$$\varepsilon_k = (1-\tilde{w}_k)\left(\hat{\alpha}_{Hajek} - \hat{\alpha}_{Hajek(k)}\right)$$

with

$$\tilde{w}_k = \frac{w_k}{\sum_{l\in s} w_l}$$

and where $\hat{\alpha}_{Hajek(k)}$ has the same functional form as $\hat{\alpha}_{Hajek}$ but omitting the k-th element from the sample s. Note that this variance estimator utilises implicitly the Hajek (1964) approximations that are designed for large-entropy sampling designs, large samples and large populations, i.e. care should be taken with highly-stratified samples, e.g. Berger (2005).

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Berger, Y. G. (2005) Variance estimation with highly stratified sampling designs with unequal probabilities. *Australian & New Zealand Journal of Statistics*, **47**, 365–373.

Berger, Y. G. (2007) A jackknife variance estimator for unistage stratified samples with unequal probabilities. *Biometrika* **94**, 953–964.

Hajek, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35**, 4, 1491–1523.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Sarndal, C.-E. and Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, Inc.

See Also

VE.Jk.B.RegCo.Hajek
VE.Jk.Tukey.RegCoI.Hajek
VE.Jk.CBS.HT.RegCoI.Hajek
VE.Jk.CBS.SYG.RegCoI.Hajek
VE.Jk.EB.SW2.RegCoI.Hajek

Examples

```
data(oaxaca)                                  #Loads the Oaxaca municipalities dataset
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s      <- oaxaca$sHOMES00                     #Defines the sample to be used
y1     <- oaxaca$POP10                        #Defines the variable of interest y1
y2     <- oaxaca$POPMAL10                     #Defines the variable of interest y2
x      <- oaxaca$HOMES10                      #Defines the variable of interest x
#Computes the var. est. of the intercept reg. coeff. point estimator using y1
VE.Jk.B.RegCoI.Hajek(y1[s==1], x[s==1], pik.U[s==1])
#Computes the var. est. of the intercept reg. coeff. point estimator using y2
VE.Jk.B.RegCoI.Hajek(y2[s==1], x[s==1], pik.U[s==1])
```

VE.Jk.B.Total.Hajek   *The Berger (2007) unequal probability jackknife variance estimator for the Hajek estimator of a total*

Description

Computes the Berger (2007) unequal probability jackknife variance estimator for the Hajek (1971) estimator of a total.

Usage

```
VE.Jk.B.Total.Hajek(VecY.s, VecPk.s, N)
```

Arguments

`VecY.s`	vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s`. There must not be missing values.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.
`N`	the population size. It must be an integer or a double-precision scalar with zero-valued fractional part.

Details

For the population total of the variable y:

$$t = \sum_{k \in U} y_k$$

the approximately unbiased Hajek (1971) estimator of t is given by:

$$\hat{t}_{Hajek} = N \frac{\sum_{k \in s} w_k y_k}{\sum_{k \in s} w_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. The variance of $\hat{t}_{Hajek}$ can be estimated by the Berger (2007) unequal probability jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{t}_{Hajek}) = \sum_{k \in s} \frac{n}{n-1}(1-\pi_k)\left(\varepsilon_k - \hat{B}\right)^2$$

where

$$\hat{B} = \frac{\sum_{k \in s}(1-\pi_k)\varepsilon_k}{\sum_{k \in s}(1-\pi_k)}$$

and

$$\varepsilon_k = (1-\tilde{w}_k)\left(\hat{t}_{Hajek} - \hat{t}_{Hajek(k)}\right)$$

with

$$\tilde{w}_k = \frac{w_k}{\sum_{l \in s} w_l}$$

and

$$\hat{t}_{Hajek(k)} = N \frac{\sum_{l \in s, l \neq k} w_l y_l}{\sum_{l \in s, l \neq k} w_l}$$

Note that this variance estimator utilises implicitly the Hajek (1964) approximations that are designed for large-entropy sampling designs, large samples and large populations, i.e. care should be taken with highly-stratified samples, e.g. Berger (2005).

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Berger, Y. G. (2005) Variance estimation with highly stratified sampling designs with unequal probabilities. *Australian & New Zealand Journal of Statistics*, **47**, 365–373.

Berger, Y. G. (2007) A jackknife variance estimator for unistage stratified samples with unequal probabilities. *Biometrika* **94**, 953–964.

Hajek, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35**, 4, 1491–1523.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

See Also

VE.Jk.Tukey.Total.Hajek
VE.Jk.CBS.HT.Total.Hajek
VE.Jk.CBS.SYG.Total.Hajek
VE.Jk.EB.SW2.Total.Hajek

Examples

```
data(oaxaca)                                 #Loads the Oaxaca municipalities dataset
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s      <- oaxaca$sHOMES00                    #Defines the sample to be used
N      <- dim(oaxaca)[1]                     #Defines the population size
y1     <- oaxaca$POP10                       #Defines the variable of interest y1
y2     <- oaxaca$POPMAL10                    #Defines the variable of interest y2
#Computes the var. est. of the Hajek total point estimator using y1
VE.Jk.B.Total.Hajek(y1[s==1], pik.U[s==1], N)
#Computes the var. est. of the Hajek total point estimator using y2
VE.Jk.B.Total.Hajek(y2[s==1], pik.U[s==1], N)
```

VE.Jk.CBS.HT.Corr.Hajek

The Campbell-Berger-Skinner unequal probability jackknife variance estimator for the estimator of a correlation coefficient using the Hajek point estimator (Horvitz-Thompson form)

Description

Computes the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator for the estimator of a correlation coefficient of two variables using the Hajek (1971) point estimator. It uses the Horvitz-Thompson (1952) variance form.

Usage

```
VE.Jk.CBS.HT.Corr.Hajek(VecY.s, VecX.s, VecPk.s, MatPkl.s)
```

Arguments

`VecY.s`	vector of the variable of interest Y; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecX.s`. There must not be missing values.
`VecX.s`	vector of the variable of interest X; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecY.s`. There must not be missing values.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.
`MatPkl.s`	matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in `MatPkl.s` must be greater than zero and less than or equal to one. There must not be missing values.

Details

For the population correlation coefficient of two variables y and x:

$$C = \frac{\sum_{k\in U}(y_k - \bar{y})(x_k - \bar{x})}{\sqrt{\sum_{k\in U}(y_k - \bar{y})^2}\sqrt{\sum_{k\in U}(x_k - \bar{x})^2}}$$

the point estimator of C, assuming that N is unknown (see Sarndal et al., 1992, Sec. 5.9), is:

$$\hat{C}_{Hajek} = \frac{\sum_{k\in s} w_k(y_k - \hat{\bar{y}}_{Hajek})(x_k - \hat{\bar{x}}_{Hajek})}{\sqrt{\sum_{k\in s} w_k(y_k - \hat{\bar{y}}_{Hajek})^2}\sqrt{\sum_{k\in s} w_k(x_k - \hat{\bar{x}}_{Hajek})^2}}$$

where $\hat{\bar{y}}_{Hajek}$ is the Hajek (1971) point estimator of the population mean $\bar{y} = N^{-1}\sum_{k\in U} y_k$,

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

and $w_k = 1/\pi_k$ with π_k denoting the inclusion probability of the k-th element in the sample s. The variance of $\hat{C}_{Hajek}$ can be estimated by the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{C}_{Hajek}) = \sum_{k\in s}\sum_{l\in s}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}\varepsilon_k\varepsilon_l$$

where

$$\varepsilon_k = (1 - \tilde{w}_k)\left(\hat{C}_{Hajek} - \hat{C}_{Hajek(k)}\right)$$

with

$$\tilde{w}_k = \frac{w_k}{\sum_{l\in s} w_l}$$

and where $\hat{C}_{Hajek(k)}$ has the same functional form as $\hat{C}_{Hajek}$ but omitting the k-th element from the sample s.

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Campbell, C. (1980) A different view of finite population estimation. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 319–324.

Berger, Y. G. and Skinner, C. J. (2005) A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society B*, **67**, 79–89.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Sarndal, C.-E. and Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, Inc.

See Also

VE.Jk.Tukey.Corr.Hajek
VE.Jk.CBS.SYG.Corr.Hajek
VE.Jk.B.Corr.Hajek
VE.Jk.EB.SW2.Corr.Hajek

Examples

```
data(oaxaca)                                 #Loads the Oaxaca municipalities dataset
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s      <- oaxaca$sHOMES00                    #Defines the sample to be used
y1     <- oaxaca$POP10                       #Defines the variable of interest y1
y2     <- oaxaca$POPMAL10                    #Defines the variable of interest y2
x      <- oaxaca$HOMES10                     #Defines the variable of interest x
#This approximation is only suitable for large-entropy sampling designs
pikl.s <- Pkl.Hajek.s(pik.U[s==1])           #Approx. 2nd order incl. probs. from s
#Computes the var. est. of the corr. coeff. point estimator using y1
VE.Jk.CBS.HT.Corr.Hajek(y1[s==1], x[s==1], pik.U[s==1], pikl.s)
#Computes the var. est. of the corr. coeff. point estimator using y2
VE.Jk.CBS.HT.Corr.Hajek(y2[s==1], x[s==1], pik.U[s==1], pikl.s)
```

`VE.Jk.CBS.HT.Mean.Hajek`

The Campbell-Berger-Skinner unequal probability jackknife variance estimator for the Hajek (1971) estimator of a mean (Horvitz-Thompson form)

Description

Computes the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator for the Hajek estimator of a mean. It uses the Horvitz-Thompson (1952) variance form.

Usage

```
VE.Jk.CBS.HT.Mean.Hajek(VecY.s, VecPk.s, MatPkl.s)
```

Arguments

`VecY.s`	vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s`. There must not be missing values.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.
`MatPkl.s`	matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in `MatPkl.s` must be greater than zero and less than or equal to one. There must not be missing values.

Details

For the population mean of the variable y:

$$\bar{y} = \frac{1}{N}\sum_{k\in U} y_k$$

the approximately unbiased Hajek (1971) estimator of $\bar{y}$ is given by:

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. The variance of $\hat{\bar{y}}_{Hajek}$ can be estimated by the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{\bar{y}}_{Hajek}) = \sum_{k\in s}\sum_{l\in s}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}\varepsilon_k\varepsilon_l$$

where

$$\varepsilon_k = (1 - \tilde{w}_k)\left(\hat{\bar{y}}_{Hajek} - \hat{\bar{y}}_{Hajek(k)}\right)$$

with

$$\tilde{w}_k = \frac{w_k}{\sum_{l \in s} w_l}$$

and

$$\hat{\bar{y}}_{Hajek(k)} = \frac{\sum_{l \in s, l \neq k} w_l y_l}{\sum_{l \in s, l \neq k} w_l}$$

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Campbell, C. (1980) A different view of finite population estimation. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 319–324.

Berger, Y. G. and Skinner, C. J. (2005) A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society B*, **67**, 79–89.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

See Also

VE.Jk.Tukey.Mean.Hajek
VE.Jk.CBS.SYG.Mean.Hajek
VE.Jk.B.Mean.Hajek
VE.Jk.EB.SW2.Mean.Hajek

Examples

```
data(oaxaca)                                 #Loads the Oaxaca municipalities dataset
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s      <- oaxaca$sHOMES00                    #Defines the sample to be used
y1     <- oaxaca$POP10                       #Defines the variable of interest y1
y2     <- oaxaca$POPMAL10                    #Defines the variable of interest y2
#This approximation is only suitable for large-entropy sampling designs
pikl.s <- Pkl.Hajek.s(pik.U[s==1])           #Approx. 2nd order incl. probs. from s
#Computes the var. est. of the Hajek mean point estimator using y1
VE.Jk.CBS.HT.Mean.Hajek(y1[s==1], pik.U[s==1], pikl.s)
#Computes the var. est. of the Hajek mean point estimator using y2
VE.Jk.CBS.HT.Mean.Hajek(y2[s==1], pik.U[s==1], pikl.s)
```

VE.Jk.CBS.HT.Ratio *The Campbell-Berger-Skinner unequal probability jackknife variance estimator for the estimator of a ratio (Horvitz-Thompson form)*

Description

Computes the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator for the estimator of a ratio of two totals/means. It uses the Horvitz-Thompson (1952) variance form.

Usage

```
VE.Jk.CBS.HT.Ratio(VecY.s, VecX.s, VecPk.s, MatPkl.s)
```

Arguments

`VecY.s`	vector of the numerator variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecX.s`. There must not be missing values.
`VecX.s`	vector of the denominator variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecY.s`. There must not be missing values. All values of `VecX.s` should be greater than zero. A warning is displayed if this does not hold and computations continue if mathematical expressions allow this kind of values for the denominator variable.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.
`MatPkl.s`	matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in `MatPkl.s` must be greater than zero and less than or equal to one. There must not be missing values.

Details

For the population ratio of two totals/means of the variables y and x:

$$R = \frac{\sum_{k\in U} y_k/N}{\sum_{k\in U} x_k/N} = \frac{\sum_{k\in U} y_k}{\sum_{k\in U} x_k}$$

the ratio estimator of R is given by:

$$\hat{R} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k x_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. The variance of $\hat{R}$ can be estimated by the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{R}) = \sum_{k\in s}\sum_{l\in s} \frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}} \varepsilon_k \varepsilon_l$$

where
$$\varepsilon_k = (1 - \tilde{w}_k)\left(\hat{R} - \hat{R}_{(k)}\right)$$
with
$$\tilde{w}_k = \frac{w_k}{\sum_{l \in s} w_l}$$
and
$$\hat{R}_{(k)} = \frac{\sum_{l \in s, l \neq k} w_l y_l / \sum_{l \in s, l \neq k} w_l}{\sum_{l \in s, l \neq k} w_l x_l / \sum_{l \in s, l \neq k} w_l} = \frac{\sum_{l \in s, l \neq k} w_l y_l}{\sum_{l \in s, l \neq k} w_l x_l}$$

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Campbell, C. (1980) A different view of finite population estimation. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 319–324.

Berger, Y. G. and Skinner, C. J. (2005) A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society B*, **67**, 79–89.

Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

See Also

VE.Lin.HT.Ratio
VE.Lin.SYG.Ratio
VE.Jk.Tukey.Ratio
VE.Jk.CBS.SYG.Ratio
VE.Jk.B.Ratio
VE.Jk.EB.SW2.Ratio
VE.EB.HT.Ratio
VE.EB.SYG.Ratio

Examples

```
data(oaxaca)                                 #Loads the Oaxaca municipalities dataset
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s      <- oaxaca$sHOMES00                    #Defines the sample to be used
y1     <- oaxaca$POP10                       #Defines the numerator variable y1
y2     <- oaxaca$POPMAL10                    #Defines the numerator variable y2
x      <- oaxaca$HOMES10                     #Defines the denominator variable x
#This approximation is only suitable for large-entropy sampling designs
pikl.s <- Pkl.Hajek.s(pik.U[s==1])           #Approx. 2nd order incl. probs. from s
#Computes the var. est. of the ratio point estimator using y1
VE.Jk.CBS.HT.Ratio(y1[s==1], x[s==1], pik.U[s==1], pikl.s)
```

```
#Computes the var. est. of the ratio point estimator using y2
VE.Jk.CBS.HT.Ratio(y2[s==1], x[s==1], pik.U[s==1], pikl.s)
```

VE.Jk.CBS.HT.RegCo.Hajek
The Campbell-Berger-Skinner unequal probability jackknife variance estimator for the estimator of the regression coefficient using the Hajek point estimator (Horvitz-Thompson form)

Description

Computes the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator for the estimator of the regression coefficient using the Hajek (1971) point estimator. It uses the Horvitz-Thompson (1952) variance form.

Usage

```
VE.Jk.CBS.HT.RegCo.Hajek(VecY.s, VecX.s, VecPk.s, MatPkl.s)
```

Arguments

`VecY.s` vector of the variable of interest Y; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecX.s`. There must not be missing values.

`VecX.s` vector of the variable of interest X; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecY.s`. There must not be missing values.

`VecPk.s` vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.

`MatPkl.s` matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in `MatPkl.s` must be greater than zero and less than or equal to one. There must not be missing values.

Details

From Linear Regression Analysis, for an imposed population model

$$y = \alpha + \beta x$$

the population regression coefficient β, assuming that the population size N is unknown (see Sarndal et al., 1992, Sec. 5.10), can be estimated by:

$$\hat{\beta}_{Hajek} = \frac{\sum_{k\in s} w_k (y_k - \hat{\bar{y}}_{Hajek})(x_k - \hat{\bar{x}}_{Hajek})}{\sum_{k\in s} w_k (x_k - \hat{\bar{x}}_{Hajek})^2}$$

where $\hat{\bar{y}}_{Hajek}$ and $\hat{\bar{x}}_{Hajek}$ are the Hajek (1971) point estimators of the population means $\bar{y} = N^{-1}\sum_{k\in U} y_k$ and $\bar{x} = N^{-1}\sum_{k\in U} x_k$, respectively,

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

$$\hat{\bar{x}}_{Hajek} = \frac{\sum_{k\in s} w_k x_k}{\sum_{k\in s} w_k}$$

and $w_k = 1/\pi_k$ with π_k denoting the inclusion probability of the k-th element in the sample s. The variance of $\hat{\beta}_{Hajek}$ can be estimated by the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{\beta}_{Hajek}) = \sum_{k\in s}\sum_{l\in s} \frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}} \varepsilon_k \varepsilon_l$$

where

$$\varepsilon_k = (1 - \tilde{w}_k)\left(\hat{\beta}_{Hajek} - \hat{\beta}_{Hajek(k)}\right)$$

with

$$\tilde{w}_k = \frac{w_k}{\sum_{l\in s} w_l}$$

and where $\hat{\beta}_{Hajek(k)}$ has the same functional form as $\hat{\beta}_{Hajek}$ but omitting the k-th element from the sample s.

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Campbell, C. (1980) A different view of finite population estimation. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 319–324.

Berger, Y. G. and Skinner, C. J. (2005) A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society B*, **67**, 79–89.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Sarndal, C.-E. and Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, Inc.

See Also

VE.Jk.CBS.HT.RegCoI.Hajek
VE.Jk.Tukey.RegCo.Hajek
VE.Jk.CBS.SYG.RegCo.Hajek
VE.Jk.B.RegCo.Hajek
VE.Jk.EB.SW2.RegCo.Hajek

Examples

```
data(oaxaca)                                 #Loads the Oaxaca municipalities dataset
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s      <- oaxaca$sHOMES00                    #Defines the sample to be used
y1     <- oaxaca$POP10                       #Defines the variable of interest y1
y2     <- oaxaca$POPMAL10                    #Defines the variable of interest y2
x      <- oaxaca$HOMES10                     #Defines the variable of interest x
#This approximation is only suitable for large-entropy sampling designs
pikl.s <- Pkl.Hajek.s(pik.U[s==1])           #Approx. 2nd order incl. probs. from s
#Computes the var. est. of the regression coeff. point estimator using y1
VE.Jk.CBS.HT.RegCo.Hajek(y1[s==1], x[s==1], pik.U[s==1], pikl.s)
#Computes the var. est. of the regression coeff. point estimator using y2
VE.Jk.CBS.HT.RegCo.Hajek(y2[s==1], x[s==1], pik.U[s==1], pikl.s)
```

VE.Jk.CBS.HT.RegCoI.Hajek

The Campbell-Berger-Skinner unequal probability jackknife variance estimator for the estimator of the intercept regression coefficient using the Hajek point estimator (Horvitz-Thompson form)

Description

Computes the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator for the estimator of the intercept regression coefficient using the Hajek (1971) point estimator. It uses the Horvitz-Thompson (1952) variance form.

Usage

```
VE.Jk.CBS.HT.RegCoI.Hajek(VecY.s, VecX.s, VecPk.s, MatPkl.s)
```

Arguments

VecY.s	vector of the variable of interest Y; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecX.s. There must not be missing values.
VecX.s	vector of the variable of interest X; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecY.s. There must not be missing values.

`VecPk.s` vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.

`MatPkl.s` matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in `MatPkl.s` must be greater than zero and less than or equal to one. There must not be missing values.

Details

From Linear Regression Analysis, for an imposed population model

$$y = \alpha + \beta x$$

the population intercept regression coefficient α, assuming that the population size N is unknown (see Sarndal et al., 1992, Sec. 5.10), can be estimated by:

$$\hat{\alpha}_{Hajek} = \hat{\bar{y}}_{Hajek} - \frac{\sum_{k\in s} w_k (y_k - \hat{\bar{y}}_{Hajek})(x_k - \hat{\bar{x}}_{Hajek})}{\sum_{k\in s} w_k (x_k - \hat{\bar{x}}_{Hajek})^2} \hat{\bar{x}}_{Hajek}$$

where $\hat{\bar{y}}_{Hajek}$ and $\hat{\bar{x}}_{Hajek}$ are the Hajek (1971) point estimators of the population means $\bar{y} = N^{-1} \sum_{k\in U} y_k$ and $\bar{x} = N^{-1} \sum_{k\in U} x_k$, respectively,

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

$$\hat{\bar{x}}_{Hajek} = \frac{\sum_{k\in s} w_k x_k}{\sum_{k\in s} w_k}$$

and $w_k = 1/\pi_k$ with π_k denoting the inclusion probability of the k-th element in the sample s. The variance of $\hat{\alpha}_{Hajek}$ can be estimated by the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{\alpha}_{Hajek}) = \sum_{k\in s} \sum_{l\in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \varepsilon_k \varepsilon_l$$

where

$$\varepsilon_k = (1 - \tilde{w}_k) \left(\hat{\alpha}_{Hajek} - \hat{\alpha}_{Hajek(k)} \right)$$

with

$$\tilde{w}_k = \frac{w_k}{\sum_{l\in s} w_l}$$

and where $\hat{\alpha}_{Hajek(k)}$ has the same functional form as $\hat{\alpha}_{Hajek}$ but omitting the k-th element from the sample s.

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Campbell, C. (1980) A different view of finite population estimation. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 319–324.

Berger, Y. G. and Skinner, C. J. (2005) A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society B*, **67**, 79–89.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Sarndal, C.-E. and Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, Inc.

See Also

VE.Jk.CBS.HT.RegCo.Hajek
VE.Jk.Tukey.RegCoI.Hajek
VE.Jk.CBS.SYG.RegCoI.Hajek
VE.Jk.B.RegCoI.Hajek
VE.Jk.EB.SW2.RegCoI.Hajek

Examples

```
data(oaxaca)                                  #Loads the Oaxaca municipalities dataset
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s      <- oaxaca$sHOMES00                     #Defines the sample to be used
y1     <- oaxaca$POP10                        #Defines the variable of interest y1
y2     <- oaxaca$POPMAL10                     #Defines the variable of interest y2
x      <- oaxaca$HOMES10                      #Defines the variable of interest x
#This approximation is only suitable for large-entropy sampling designs
pikl.s <- Pkl.Hajek.s(pik.U[s==1])           #Approx. 2nd order incl. probs. from s
#Computes the var. est. of the intercept reg.  coeff. point estimator using y1
VE.Jk.CBS.HT.RegCoI.Hajek(y1[s==1], x[s==1], pik.U[s==1], pikl.s)
#Computes the var. est. of the intercept reg. coeff. point estimator using y2
VE.Jk.CBS.HT.RegCoI.Hajek(y2[s==1], x[s==1], pik.U[s==1], pikl.s)
```

VE.Jk.CBS.HT.Total.Hajek

The Campbell-Berger-Skinner unequal probability jackknife variance estimator for the Hajek (1971) estimator of a total (Horvitz-Thompson form)

Description

Computes the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator for the Hajek estimator of a total. It uses the Horvitz-Thompson (1952) variance form.

Usage

```
VE.Jk.CBS.HT.Total.Hajek(VecY.s, VecPk.s, MatPkl.s, N)
```

Arguments

`VecY.s` vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s`. There must not be missing values.

`VecPk.s` vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.

`MatPkl.s` matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in `MatPkl.s` must be greater than zero and less than or equal to one. There must not be missing values.

`N` the population size. It must be an integer or a double-precision scalar with zero-valued fractional part.

Details

For the population total of the variable y:

$$t = \sum_{k\in U} y_k$$

the approximately unbiased Hajek (1971) estimator of t is given by:

$$\hat{t}_{Hajek} = N\frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. The variance of $\hat{t}_{Hajek}$ can be estimated by the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{t}_{Hajek}) = \sum_{k\in s}\sum_{l\in s}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}}\varepsilon_k\varepsilon_l$$

where

$$\varepsilon_k = (1-\tilde{w}_k)\left(\hat{t}_{Hajek} - \hat{t}_{Hajek(k)}\right)$$

with

$$\tilde{w}_k = \frac{w_k}{\sum_{l\in s} w_l}$$

and

$$\hat{t}_{Hajek(k)} = N\frac{\sum_{l\in s, l\neq k} w_l y_l}{\sum_{l\in s, l\neq k} w_l}$$

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Campbell, C. (1980) A different view of finite population estimation. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 319–324.

Berger, Y. G. and Skinner, C. J. (2005) A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society B*, **67**, 79–89.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

See Also

`VE.Jk.Tukey.Total.Hajek`
`VE.Jk.CBS.SYG.Total.Hajek`
`VE.Jk.B.Total.Hajek`
`VE.Jk.EB.SW2.Total.Hajek`

Examples

```
data(oaxaca)                                 #Loads the Oaxaca municipalities dataset
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s      <- oaxaca$sHOMES00                    #Defines the sample to be used
N      <- dim(oaxaca)[1]                     #Defines the population size
y1     <- oaxaca$POP10                       #Defines the variable of interest y1
y2     <- oaxaca$POPMAL10                    #Defines the variable of interest y2
#This approximation is only suitable for large-entropy sampling designs
pikl.s <- Pkl.Hajek.s(pik.U[s==1])           #Approx. 2nd order incl. probs. from s
#Computes the var. est. of the Hajek total point estimator using y1
VE.Jk.CBS.HT.Total.Hajek(y1[s==1], pik.U[s==1], pikl.s, N)
#Computes the var. est. of the Hajek total point estimator using y2
VE.Jk.CBS.HT.Total.Hajek(y2[s==1], pik.U[s==1], pikl.s, N)
```

`VE.Jk.CBS.SYG.Corr.Hajek` *The Campbell-Berger-Skinner unequal probability jackknife variance estimator for the estimator of a correlation coefficient using the Hajek point estimator (Sen-Yates-Grundy form)*

Description

Computes the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator for the estimator of a correlation coefficient of two variables using the Hajek (1971) point estimator. It uses the Sen (1953); Yates-Grundy(1953) variance form.

Usage

```
VE.Jk.CBS.SYG.Corr.Hajek(VecY.s, VecX.s, VecPk.s, MatPkl.s)
```

Arguments

`VecY.s`	vector of the variable of interest Y; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecX.s`. There must not be missing values.
`VecX.s`	vector of the variable of interest X; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecY.s`. There must not be missing values.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.
`MatPkl.s`	matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in `MatPkl.s` must be greater than zero and less than or equal to one. There must not be missing values.

Details

For the population correlation coefficient of two variables y and x:

$$C = \frac{\sum_{k\in U}(y_k-\bar{y})(x_k-\bar{x})}{\sqrt{\sum_{k\in U}(y_k-\bar{y})^2}\sqrt{\sum_{k\in U}(x_k-\bar{x})^2}}$$

the point estimator of C, assuming that N is unknown (see Sarndal et al., 1992, Sec. 5.9), is:

$$\hat{C}_{Hajek} = \frac{\sum_{k\in s} w_k(y_k-\hat{\bar{y}}_{Hajek})(x_k-\hat{\bar{x}}_{Hajek})}{\sqrt{\sum_{k\in s} w_k(y_k-\hat{\bar{y}}_{Hajek})^2}\sqrt{\sum_{k\in s} w_k(x_k-\hat{\bar{x}}_{Hajek})^2}}$$

where $\hat{\bar{y}}_{Hajek}$ is the Hajek (1971) point estimator of the population mean $\bar{y} = N^{-1}\sum_{k\in U} y_k$,

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

and $w_k = 1/\pi_k$ with π_k denoting the inclusion probability of the k-th element in the sample s. The variance of $\hat{C}_{Hajek}$ can be estimated by the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{C}_{Hajek}) = \frac{-1}{2}\sum_{k\in s}\sum_{l\in s}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}}(\varepsilon_k-\varepsilon_l)^2$$

where

$$\varepsilon_k = (1-\tilde{w}_k)\left(\hat{C}_{Hajek}-\hat{C}_{Hajek(k)}\right)$$

with

$$\tilde{w}_k = \frac{w_k}{\sum_{l\in s} w_l}$$

and where $\hat{C}_{Hajek(k)}$ has the same functional form as $\hat{C}_{Hajek}$ but omitting the k-th element from the sample s. The Sen-Yates-Grundy form for the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator is proposed in Escobar-Berger (2013) under less-restrictive regularity conditions.

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Campbell, C. (1980) A different view of finite population estimation. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 319–324.

Berger, Y. G. and Skinner, C. J. (2005) A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society B*, **67**, 79–89.

Escobar, E. L. and Berger, Y. G. (2013) A jackknife variance estimator for self-weighted two-stage samples. *Statistica Sinica*, **23**, 595–613.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Sarndal, C.-E. and Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, Inc.

Sen, A. R. (1953) On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **5**, 119–127.

Yates, F. and Grundy, P. M. (1953) Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B*, **15**, 253–261.

See Also

VE.Jk.Tukey.Corr.Hajek
VE.Jk.CBS.HT.Corr.Hajek
VE.Jk.B.Corr.Hajek
VE.Jk.EB.SW2.Corr.Hajek

Examples

```
data(oaxaca)                                 #Loads the Oaxaca municipalities dataset
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s      <- oaxaca$sHOMES00                    #Defines the sample to be used
y1     <- oaxaca$POP10                       #Defines the variable of interest y1
y2     <- oaxaca$POPMAL10                    #Defines the variable of interest y2
x      <- oaxaca$HOMES10                     #Defines the variable of interest x
#This approximation is only suitable for large-entropy sampling designs
pikl.s <- Pkl.Hajek.s(pik.U[s==1])           #Approx. 2nd order incl. probs. from s
#Computes the var. est. of the corr. coeff. point estimator using y1
VE.Jk.CBS.SYG.Corr.Hajek(y1[s==1], x[s==1], pik.U[s==1], pikl.s)
#Computes the var. est. of the corr. coeff. point estimator using y2
VE.Jk.CBS.SYG.Corr.Hajek(y2[s==1], x[s==1], pik.U[s==1], pikl.s)
```

`VE.Jk.CBS.SYG.Mean.Hajek`

The Campbell-Berger-Skinner unequal probability jackknife variance estimator for the Hajek (1971) estimator of a mean (Sen-Yates-Grundy form)

Description

Computes the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator for the Hajek estimator of a mean. It uses the Sen (1953); Yates-Grundy(1953) variance form.

Usage

```
VE.Jk.CBS.SYG.Mean.Hajek(VecY.s, VecPk.s, MatPkl.s)
```

Arguments

`VecY.s`	vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s`. There must not be missing values.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.
`MatPkl.s`	matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in `MatPkl.s` must be greater than zero and less than or equal to one. There must not be missing values.

Details

For the population mean of the variable y:

$$\bar{y} = \frac{1}{N}\sum_{k\in U} y_k$$

the approximately unbiased Hajek (1971) estimator of $\bar{y}$ is given by:

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. The variance of $\hat{\bar{y}}_{Hajek}$ can be estimated by the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{\bar{y}}_{Hajek}) = \frac{-1}{2}\sum_{k\in s}\sum_{l\in s}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}}(\varepsilon_k-\varepsilon_l)^2$$

where

$$\varepsilon_k = (1 - \tilde{w}_k)\left(\hat{\bar{y}}_{Hajek} - \hat{\bar{y}}_{Hajek(k)}\right)$$

with

$$\tilde{w}_k = \frac{w_k}{\sum_{l \in s} w_l}$$

and

$$\hat{\bar{y}}_{Hajek(k)} = \frac{\sum_{l \in s, l \neq k} w_l y_l}{\sum_{l \in s, l \neq k} w_l}$$

The Sen-Yates-Grundy form for the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator is proposed in Escobar-Berger (2013) under less-restrictive regularity conditions.

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Campbell, C. (1980) A different view of finite population estimation. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 319–324.

Berger, Y. G. and Skinner, C. J. (2005) A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society B*, **67**, 79–89.

Escobar, E. L. and Berger, Y. G. (2013) A jackknife variance estimator for self-weighted two-stage samples. *Statistica Sinica*, **23**, 595–613.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Sen, A. R. (1953) On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **5**, 119–127.

Yates, F. and Grundy, P. M. (1953) Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B*, **15**, 253–261.

See Also

`VE.Jk.Tukey.Mean.Hajek`
`VE.Jk.CBS.HT.Mean.Hajek`
`VE.Jk.B.Mean.Hajek`
`VE.Jk.EB.SW2.Mean.Hajek`

Examples

```
data(oaxaca)                                    #Loads the Oaxaca municipalities dataset
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s      <- oaxaca$sHOMES00                       #Defines the sample to be used
y1     <- oaxaca$POP10                          #Defines the variable of interest y1
y2     <- oaxaca$POPMAL10                       #Defines the variable of interest y2
#This approximation is only suitable for large-entropy sampling designs
pikl.s <- Pkl.Hajek.s(pik.U[s==1])          #Approx. 2nd order incl. probs. from s
#Computes the var. est. of the Hajek mean point estimator using y1
VE.Jk.CBS.SYG.Mean.Hajek(y1[s==1], pik.U[s==1], pikl.s)
#Computes the var. est. of the Hajek mean point estimator using y2
VE.Jk.CBS.SYG.Mean.Hajek(y2[s==1], pik.U[s==1], pikl.s)
```

VE.Jk.CBS.SYG.Ratio — *The Campbell-Berger-Skinner unequal probability jackknife variance estimator for the estimator of a ratio (Sen-Yates-Grundy form)*

Description

Computes the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator for the estimator of a ratio of two totals/means. It uses the Sen (1953); Yates-Grundy(1953) variance form.

Usage

```
VE.Jk.CBS.SYG.Ratio(VecY.s, VecX.s, VecPk.s, MatPkl.s)
```

Arguments

`VecY.s`	vector of the numerator variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecX.s`. There must not be missing values.
`VecX.s`	vector of the denominator variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecY.s`. There must not be missing values. All values of `VecX.s` should be greater than zero. A warning is displayed if this does not hold and computations continue if mathematical expressions allow this kind of values for the denominator variable.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.
`MatPkl.s`	matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in `MatPkl.s` must be greater than zero and less than or equal to one. There must not be missing values.

Details

For the population ratio of two totals/means of the variables y and x:

$$R = \frac{\sum_{k\in U} y_k/N}{\sum_{k\in U} x_k/N} = \frac{\sum_{k\in U} y_k}{\sum_{k\in U} x_k}$$

the ratio estimator of R is given by:

$$\hat{R} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k x_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. The variance of $\hat{R}$ can be estimated by the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{R}) = \frac{-1}{2} \sum_{k\in s} \sum_{l\in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} (\varepsilon_k - \varepsilon_l)^2$$

where

$$\varepsilon_k = (1 - \tilde{w}_k) \left(\hat{R} - \hat{R}_{(k)} \right)$$

with

$$\tilde{w}_k = \frac{w_k}{\sum_{l\in s} w_l}$$

and

$$\hat{R}_{(k)} = \frac{\sum_{l\in s, l\neq k} w_l y_l / \sum_{l\in s, l\neq k} w_l}{\sum_{l\in s, l\neq k} w_l x_l / \sum_{l\in s, l\neq k} w_l} = \frac{\sum_{l\in s, l\neq k} w_l y_l}{\sum_{l\in s, l\neq k} w_l x_l}$$

The Sen-Yates-Grundy form for the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator is proposed in Escobar-Berger (2013) under less-restrictive regularity conditions.

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Campbell, C. (1980) A different view of finite population estimation. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 319–324.

Berger, Y. G. and Skinner, C. J. (2005) A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society B*, **67**, 79–89.

Escobar, E. L. and Berger, Y. G. (2013) A jackknife variance estimator for self-weighted two-stage samples. *Statistica Sinica*, **23**, 595–613.

Sen, A. R. (1953) On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **5**, 119–127.

Yates, F. and Grundy, P. M. (1953) Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B*, **15**, 253–261.

See Also

VE.Lin.HT.Ratio
VE.Lin.SYG.Ratio
VE.Jk.Tukey.Ratio
VE.Jk.CBS.HT.Ratio
VE.Jk.B.Ratio
VE.Jk.EB.SW2.Ratio
VE.EB.HT.Ratio
VE.EB.SYG.Ratio

Examples

```
data(oaxaca)                                 #Loads the Oaxaca municipalities dataset
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s      <- oaxaca$sHOMES00                    #Defines the sample to be used for
y1     <- oaxaca$POP10                       #Defines the numerator variable y1
y2     <- oaxaca$POPMAL10                    #Defines the numerator variable y2
x      <- oaxaca$HOMES10                     #Defines the denominator variable x
#This approximation is only suitable for large-entropy sampling designs
pikl.s <- Pkl.Hajek.s(pik.U[s==1])           #Approx. 2nd order incl. probs. from s
#Computes the var. est. of the ratio point estimator using y1
VE.Jk.CBS.SYG.Ratio(y1[s==1], x[s==1], pik.U[s==1], pikl.s)
#Computes the var. est. of the ratio point estimator using y2
VE.Jk.CBS.SYG.Ratio(y2[s==1], x[s==1], pik.U[s==1], pikl.s)
```

VE.Jk.CBS.SYG.RegCo.Hajek

The Campbell-Berger-Skinner unequal probability jackknife variance estimator for the estimator of the regression coefficient using the Hajek point estimator (Sen-Yates-Grundy form)

Description

Computes the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator for the estimator of the regression coefficient using the Hajek (1971) point estimator. It uses the Sen (1953); Yates-Grundy(1953) variance form.

Usage

```
VE.Jk.CBS.SYG.RegCo.Hajek(VecY.s, VecX.s, VecPk.s, MatPkl.s)
```

Arguments

VecY.s	vector of the variable of interest Y; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecX.s. There must not be missing values.

`VecX.s` vector of the variable of interest X; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecY.s`. There must not be missing values.

`VecPk.s` vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.

`MatPkl.s` matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in `MatPkl.s` must be greater than zero and less than or equal to one. There must not be missing values.

Details

From Linear Regression Analysis, for an imposed population model

$$y = \alpha + \beta x$$

the population regression coefficient β, assuming that the population size N is unknown (see Sarndal et al., 1992, Sec. 5.10), can be estimated by:

$$\hat{\beta}_{Hajek} = \frac{\sum_{k\in s} w_k (y_k - \hat{\bar{y}}_{Hajek})(x_k - \hat{\bar{x}}_{Hajek})}{\sum_{k\in s} w_k (x_k - \hat{\bar{x}}_{Hajek})^2}$$

where $\hat{\bar{y}}_{Hajek}$ and $\hat{\bar{x}}_{Hajek}$ are the Hajek (1971) point estimators of the population means $\bar{y} = N^{-1}\sum_{k\in U} y_k$ and $\bar{x} = N^{-1}\sum_{k\in U} x_k$, respectively,

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

$$\hat{\bar{x}}_{Hajek} = \frac{\sum_{k\in s} w_k x_k}{\sum_{k\in s} w_k}$$

and $w_k = 1/\pi_k$ with π_k denoting the inclusion probability of the k-th element in the sample s. The variance of $\hat{\beta}_{Hajek}$ can be estimated by the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{\beta}_{Hajek}) = \frac{-1}{2}\sum_{k\in s}\sum_{l\in s}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}(\varepsilon_k - \varepsilon_l)^2$$

where

$$\varepsilon_k = (1 - \tilde{w}_k)\left(\hat{\beta}_{Hajek} - \hat{\beta}_{Hajek(k)}\right)$$

with

$$\tilde{w}_k = \frac{w_k}{\sum_{l\in s} w_l}$$

and where $\hat{\beta}_{Hajek(k)}$ has the same functional form as $\hat{\beta}_{Hajek}$ but omitting the k-th element from the sample s. The Sen-Yates-Grundy form for the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator is proposed in Escobar-Berger (2013) under less-restrictive regularity conditions.

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Campbell, C. (1980) A different view of finite population estimation. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 319–324.

Berger, Y. G. and Skinner, C. J. (2005) A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society B*, **67**, 79–89.

Escobar, E. L. and Berger, Y. G. (2013) A jackknife variance estimator for self-weighted two-stage samples. *Statistica Sinica*, **23**, 595–613.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Sarndal, C.-E. and Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, Inc.

Sen, A. R. (1953) On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **5**, 119–127.

Yates, F. and Grundy, P. M. (1953) Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B*, **15**, 253–261.

See Also

VE.Jk.CBS.SYG.RegCoI.Hajek
VE.Jk.Tukey.RegCo.Hajek
VE.Jk.CBS.HT.RegCo.Hajek
VE.Jk.B.RegCo.Hajek
VE.Jk.EB.SW2.RegCo.Hajek

Examples

```
data(oaxaca)                                #Loads the Oaxaca municipalities dataset
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s      <- oaxaca$sHOMES00                   #Defines the sample to be used
y1     <- oaxaca$POP10                      #Defines the variable of interest y1
y2     <- oaxaca$POPMAL10                   #Defines the variable of interest y2
x      <- oaxaca$HOMES10                    #Defines the variable of interest x
#This approximation is only suitable for large-entropy sampling designs
pikl.s <- Pkl.Hajek.s(pik.U[s==1])          #Approx. 2nd order incl. probs. from s
#Computes the var. est. of the regression coeff. point estimator using y1
VE.Jk.CBS.SYG.RegCo.Hajek(y1[s==1], x[s==1], pik.U[s==1], pikl.s)
#Computes the var. est. of the regression coeff. point estimator using y2
VE.Jk.CBS.SYG.RegCo.Hajek(y2[s==1], x[s==1], pik.U[s==1], pikl.s)
```

`VE.Jk.CBS.SYG.RegCoI.Hajek` *The Campbell-Berger-Skinner unequal probability jackknife variance estimator for the estimator of the intercept regression coefficient using the Hajek point estimator (Sen-Yates-Grundy form)*

Description

Computes the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator for the estimator of the intercept regression coefficient using the Hajek (1971) point estimator. It uses the Sen (1953); Yates-Grundy(1953) variance form.

Usage

```
VE.Jk.CBS.SYG.RegCoI.Hajek(VecY.s, VecX.s, VecPk.s, MatPkl.s)
```

Arguments

`VecY.s`	vector of the variable of interest Y; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecX.s`. There must not be missing values.
`VecX.s`	vector of the variable of interest X; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecY.s`. There must not be missing values.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.
`MatPkl.s`	matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in `MatPkl.s` must be greater than zero and less than or equal to one. There must not be missing values.

Details

From Linear Regression Analysis, for an imposed population model

$$y = \alpha + \beta x$$

the population intercept regression coefficient α, assuming that the population size N is unknown (see Sarndal et al., 1992, Sec. 5.10), can be estimated by:

$$\hat{\alpha}_{Hajek} = \hat{\bar{y}}_{Hajek} - \frac{\sum_{k\in s} w_k(y_k - \hat{\bar{y}}_{Hajek})(x_k - \hat{\bar{x}}_{Hajek})}{\sum_{k\in s} w_k(x_k - \hat{\bar{x}}_{Hajek})^2}\hat{\bar{x}}_{Hajek}$$

where $\hat{\bar{y}}_{Hajek}$ and $\hat{\bar{x}}_{Hajek}$ are the Hajek (1971) point estimators of the population means $\bar{y} = N^{-1}\sum_{k\in U} y_k$ and $\bar{x} = N^{-1}\sum_{k\in U} x_k$, respectively,

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

$$\hat{\bar{x}}_{Hajek} = \frac{\sum_{k\in s} w_k x_k}{\sum_{k\in s} w_k}$$

and $w_k = 1/\pi_k$ with π_k denoting the inclusion probability of the k-th element in the sample s. The variance of $\hat{\alpha}_{Hajek}$ can be estimated by the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{\alpha}_{Hajek}) = \frac{-1}{2}\sum_{k\in s}\sum_{l\in s}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}(\varepsilon_k - \varepsilon_l)^2$$

where

$$\varepsilon_k = (1 - \tilde{w}_k)\left(\hat{\alpha}_{Hajek} - \hat{\alpha}_{Hajek(k)}\right)$$

with

$$\tilde{w}_k = \frac{w_k}{\sum_{l\in s} w_l}$$

and where $\hat{\alpha}_{Hajek(k)}$ has the same functional form as $\hat{\alpha}_{Hajek}$ but omitting the k-th element from the sample s. The Sen-Yates-Grundy form for the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator is proposed in Escobar-Berger (2013) under less-restrictive regularity conditions.

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Campbell, C. (1980) A different view of finite population estimation. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 319–324.

Berger, Y. G. and Skinner, C. J. (2005) A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society B*, **67**, 79–89.

Escobar, E. L. and Berger, Y. G. (2013) A jackknife variance estimator for self-weighted two-stage samples. *Statistica Sinica*, **23**, 595–613.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Sarndal, C.-E. and Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, Inc.

Sen, A. R. (1953) On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **5**, 119–127.

Yates, F. and Grundy, P. M. (1953) Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B*, **15**, 253–261.

See Also

VE.Jk.CBS.SYG.RegCo.Hajek
VE.Jk.Tukey.RegCoI.Hajek
VE.Jk.CBS.HT.RegCoI.Hajek
VE.Jk.B.RegCoI.Hajek
VE.Jk.EB.SW2.RegCoI.Hajek

Examples

```
data(oaxaca)                                 #Loads the Oaxaca municipalities dataset
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s      <- oaxaca$sHOMES00                    #Defines the sample to be used
y1     <- oaxaca$POP10                       #Defines the variable of interest y1
y2     <- oaxaca$POPMAL10                    #Defines the variable of interest y2
x      <- oaxaca$HOMES10                     #Defines the variable of interest x
#This approximation is only suitable for large-entropy sampling designs
pikl.s <- Pkl.Hajek.s(pik.U[s==1])           #Approx. 2nd order incl. probs. from s
#Computes the var. est. of the intercept reg. coeff. point estimator using y1
VE.Jk.CBS.SYG.RegCoI.Hajek(y1[s==1], x[s==1], pik.U[s==1], pikl.s)
#Computes the var. est. of the intercept reg. coeff. point estimator using y2
VE.Jk.CBS.SYG.RegCoI.Hajek(y2[s==1], x[s==1], pik.U[s==1], pikl.s)
```

VE.Jk.CBS.SYG.Total.Hajek

The Campbell-Berger-Skinner unequal probability jackknife variance estimator for the Hajek (1971) estimator of a total (Sen-Yates-Grundy form)

Description

Computes the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator for the Hajek estimator of a total. It uses the Sen (1953); Yates-Grundy(1953) variance form.

Usage

```
VE.Jk.CBS.SYG.Total.Hajek(VecY.s, VecPk.s, MatPkl.s, N)
```

Arguments

VecY.s	vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s. There must not be missing values.
VecPk.s	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be missing values.

`MatPkl.s` matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in `MatPkl.s` must be greater than zero and less than or equal to one. There must not be missing values.

`N` the population size. It must be an integer or a double-precision scalar with zero-valued fractional part.

Details

For the population total of the variable y:

$$t = \sum_{k \in U} y_k$$

the approximately unbiased Hajek (1971) estimator of t is given by:

$$\hat{t}_{Hajek} = N \frac{\sum_{k \in s} w_k y_k}{\sum_{k \in s} w_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. The variance of $\hat{t}_{Hajek}$ can be estimated by the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{t}_{Hajek}) = \frac{-1}{2} \sum_{k \in s} \sum_{l \in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} (\varepsilon_k - \varepsilon_l)^2$$

where

$$\varepsilon_k = (1 - \tilde{w}_k) \left(\hat{t}_{Hajek} - \hat{t}_{Hajek(k)} \right)$$

with

$$\tilde{w}_k = \frac{w_k}{\sum_{l \in s} w_l}$$

and

$$\hat{t}_{Hajek(k)} = N \frac{\sum_{l \in s, l \neq k} w_l y_l}{\sum_{l \in s, l \neq k} w_l}$$

The Sen-Yates-Grundy form for the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator is proposed in Escobar-Berger (2013) under less-restrictive regularity conditions.

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Campbell, C. (1980) A different view of finite population estimation. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 319–324.

Berger, Y. G. and Skinner, C. J. (2005) A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society B*, **67**, 79–89.

Escobar, E. L. and Berger, Y. G. (2013) A jackknife variance estimator for self-weighted two-stage samples. *Statistica Sinica*, **23**, 595–613.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Sen, A. R. (1953) On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **5**, 119–127.

Yates, F. and Grundy, P. M. (1953) Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B*, **15**, 253–261.

See Also

VE.Jk.Tukey.Total.Hajek
VE.Jk.CBS.HT.Total.Hajek
VE.Jk.B.Total.Hajek
VE.Jk.EB.SW2.Total.Hajek

Examples

```
data(oaxaca)                                 #Loads the Oaxaca municipalities dataset
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s      <- oaxaca$sHOMES00                    #Defines the sample to be used
N      <- dim(oaxaca)[1]                     #Defines the population size
y1     <- oaxaca$POP10                       #Defines the variable of interest y1
y2     <- oaxaca$POPMAL10                    #Defines the variable of interest y2
#This approximation is only suitable for large-entropy sampling designs
pikl.s <- Pkl.Hajek.s(pik.U[s==1])           #Approx. 2nd order incl. probs. from s
#Computes the var. est. of the Hajek total point estimator using y1
VE.Jk.CBS.SYG.Total.Hajek(y1[s==1], pik.U[s==1], pikl.s, N)
#Computes the var. est. of the Hajek total point estimator using y2
VE.Jk.CBS.SYG.Total.Hajek(y2[s==1], pik.U[s==1], pikl.s, N)
```

VE.Jk.EB.SW2.Corr.Hajek   *The self-weighted two-stage sampling Escobar-Berger (2013) jackknife variance estimator for the estimator of a correlation coefficient using the Hajek point estimator*

Description

Computes the self-weighted two-stage sampling Escobar-Berger (2013) jackknife variance estimator for the estimator of a correlation coefficient of two variables using the Hajek (1971) point estimator.

Usage

```
VE.Jk.EB.SW2.Corr.Hajek(VecY.s, VecX.s, VecPk.s, nII, VecPi.s,
                         VecCluLab.s, VecCluSize.s)
```

Arguments

`VecY.s` vector of the variable of interest Y; its length is equal to n, the total sample size. Its length has to be the same as the length of `VecPk.s` and `VecX.s`. There must not be missing values.

`VecX.s` vector of the variable of interest X; its length is equal to n, the total sample size. Its length has to be the same as the length of `VecPk.s` and `VecY.s`. There must not be missing values.

`VecPk.s` vector of the elements' first-order inclusion probabilities; its length is equal to n, the total sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.

`nII` the second stage sample size, i.e. the fixed number of ultimate sampling units that were selected within each cluster. Its size must be less than or equal to the minimum cluster size in the sample.

`VecPi.s` vector of the clusters' first-order inclusion probabilities; its length is equal to n, the total sample size. Hence values are expected to be repeated in the utilised sample dataset. Values in `VecPi.s` must be greater than zero and less than or equal to one. There must not be missing values.

`VecCluLab.s` vector of the clusters' labels for the elements; its length is equal to n, the total sample size. The labels must be integer numbers.

`VecCluSize.s` vector of the clusters' sizes; its length is equal to n, the total sample size. Hence values are expected to be repeated in the utilised sample dataset. None of the sizes must be smaller than `nII`.

Details

For the population correlation coefficient of two variables y and x:

$$C = \frac{\sum_{k\in U}(y_k - \bar{y})(x_k - \bar{x})}{\sqrt{\sum_{k\in U}(y_k - \bar{y})^2}\sqrt{\sum_{k\in U}(x_k - \bar{x})^2}}$$

the point estimator of C, assuming that N is unknown (see Sarndal et al., 1992, Sec. 5.9), is:

$$\hat{C}_{Hajek} = \frac{\sum_{k\in s} w_k(y_k - \hat{\bar{y}}_{Hajek})(x_k - \hat{\bar{x}}_{Hajek})}{\sqrt{\sum_{k\in s} w_k(y_k - \hat{\bar{y}}_{Hajek})^2}\sqrt{\sum_{k\in s} w_k(x_k - \hat{\bar{x}}_{Hajek})^2}}$$

where $\hat{\bar{y}}_{Hajek}$ is the Hajek (1971) point estimator of the population mean $\bar{y} = N^{-1}\sum_{k\in U} y_k$,

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

and $w_k = 1/\pi_k$ with π_k denoting the inclusion probability of the k-th element in the sample s. If s is a self-weighted two-stage sample, the variance of $\hat{C}_{Hajek}$ can be estimated by the Escobar-Berger (2013) jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{C}_{Hajek}) = v_{clu} + v_{obs}$$

$$v_{clu} = \sum_{i\in s}(1-\pi^*_{Ii})\varsigma^2_{(Ii)} - \frac{1}{\hat{d}}\left(\sum_{i\in s}(1-\pi_{Ii})\varsigma_{(Ii)}\right)^2$$

$$v_{obs} = \sum_{k\in s}\phi_k \varepsilon^2_{(k)}$$

where $\hat{d} = \sum_{i\in s}(1-\pi_{Ii})$, $\phi_k = I\{k \in s_i\}\pi^*_{Ii}(M_i - n_{II})/(M_i - 1)$, $\pi^*_{Ii} = \pi_{Ii}n_{II}(M_i - 1)/(n_{II} - 1)M_i$, with s_i denoting the sample elements from the i-th cluster, $I\{k \in s_i\}$ is an indicator that takes the value 1 if the k-th observation is within the i-th cluster and 0 otherwise, π_{Ii} is the inclusion probability of the i-th cluster in the sample s, M_i is the size of the i-th cluster, n_{II} is the sample size within each cluster, n_I is the number of sampled clusters, and where

$$\varsigma_{(Ii)} = \frac{n_I - 1}{n_I}(\hat{C}_{Hajek} - \hat{C}_{Hajek(Ii)})$$

$$\varepsilon_{(k)} = \frac{n-1}{n}(\hat{C}_{Hajek} - \hat{C}_{Hajek(k)})$$

where $\hat{C}_{Hajek(Ii)}$ and $\hat{C}_{Hajek(k)}$ have the same functional form as $\hat{C}_{Hajek}$ but omitting the i-th cluster and the k-th element, respectively, from the sample s. Note that this variance estimator utilises implicitly the Hajek (1964) approximations that are designed for large-entropy sampling designs, large samples and large populations, i.e. care should be taken with highly-stratified samples, e.g. Berger (2005).

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Berger, Y. G. (2005) Variance estimation with highly stratified sampling designs with unequal probabilities. *Australian & New Zealand Journal of Statistics*, **47**, 365–373.

Escobar, E. L. and Berger, Y. G. (2013) A jackknife variance estimator for self-weighted two-stage samples. *Statistica Sinica*, **23**, 595–613.

Hajek, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35**, 4, 1491–1523.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Sarndal, C.-E. and Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, Inc.

See Also

VE.Jk.Tukey.Corr.Hajek
VE.Jk.CBS.HT.Corr.Hajek
VE.Jk.CBS.SYG.Corr.Hajek
VE.Jk.B.Corr.Hajek

Examples

```
data(oaxaca)                         #Loads the Oaxaca municipalities dataset
s          <- oaxaca$sSW_10_3         #Defines the sample to be used
SampData   <- oaxaca[s==1, ]          #Defines the sample dataset
nII        <- 3                       #Defines the 2nd stage fixed sample size
CluLab.s   <- SampData$IDDISTRI       #Defines the clusters' labels
CluSize.s  <- SampData$SIZEDIST       #Defines the clusters' sizes
piIi.s     <- (10 * CluSize.s / 570)  #Reconstructs clusters' 1st order incl. probs.
pik.s      <- piIi.s * (nII/CluSize.s) #Reconstructs elements' 1st order incl. probs.
y1.s       <- SampData$POP10          #Defines the variable y1
y2.s       <- SampData$POPMAL10       #Defines the variable y2
x.s        <- SampData$HOMES10        #Defines the variable x
#Computes the var. est. of the corr. coeff. point estimator using y1
VE.Jk.EB.SW2.Corr.Hajek(y1.s, x.s, pik.s, nII, piIi.s, CluLab.s, CluSize.s)
#Computes the var. est. of the corr. coeff. point estimator using y2
VE.Jk.EB.SW2.Corr.Hajek(y2.s, x.s, pik.s, nII, piIi.s, CluLab.s, CluSize.s)
```

VE.Jk.EB.SW2.Mean.Hajek
The self-weighted two-stage sampling Escobar-Berger (2013) jackknife variance estimator for the Hajek (1971) estimator of a mean

---

Description

Computes the self-weighted two-stage sampling Escobar-Berger (2013) jackknife variance estimator for the Hajek estimator of a mean.

Usage

```
VE.Jk.EB.SW2.Mean.Hajek(VecY.s, VecPk.s, nII, VecPi.s,
                        VecCluLab.s, VecCluSize.s)
```

Arguments

`VecY.s`	vector of the variable of interest; its length is equal to n, the total sample size. Its length has to be the same as the length of `VecPk.s`. There must not be missing values.
`VecPk.s`	vector of the elements' first-order inclusion probabilities; its length is equal to n, the total sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.
`nII`	the second stage sample size, i.e. the fixed number of ultimate sampling units that were selected within each cluster. Its size must be less than or equal to the minimum cluster size in the sample.
`VecPi.s`	vector of the clusters' first-order inclusion probabilities; its length is equal to n, the total sample size. Hence values are expected to be repeated in the utilised sample dataset. Values in `VecPi.s` must be greater than zero and less than or equal to one. There must not be missing values.
`VecCluLab.s`	vector of the clusters' labels for the elements; its length is equal to n, the total sample size. The labels must be integer numbers.
`VecCluSize.s`	vector of the clusters' sizes; its length is equal to n, the total sample size. Hence values are expected to be repeated in the utilised sample dataset. None of the sizes must be smaller than `nII`.

Details

For the population mean of the variable y:

$$\bar{y} = \frac{1}{N}\sum_{k\in U} y_k$$

the approximately unbiased Hajek (1971) estimator of $\bar{y}$ is given by:

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. If s is a self-weighted two-stage sample, the variance of $\hat{\bar{y}}_{Hajek}$ can be estimated by the Escobar-Berger (2013) jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{\bar{y}}_{Hajek}) = v_{clu} + v_{obs}$$

$$v_{clu} = \sum_{i\in s}(1-\pi^*_{Ii})\varsigma^2_{(Ii)} - \frac{1}{\hat{d}}\left(\sum_{i\in s}(1-\pi_{Ii})\varsigma_{(Ii)}\right)^2$$

$$v_{obs} = \sum_{k\in s}\phi_k \varepsilon^2_{(k)}$$

where $\hat{d} = \sum_{i\in s}(1-\pi_{Ii})$, $\phi_k = I\{k \in s_i\}\pi^*_{Ii}(M_i - n_{II})/(M_i-1)$, $\pi^*_{Ii} = \pi_{Ii} n_{II}(M_i - 1)/(n_{II}-1)M_i$, with s_i denoting the sample elements from the i-th cluster, $I\{k \in s_i\}$ is an indicator that takes the value 1 if the k-th observation is within the i-th cluster and 0 otherwise, π_{Ii}

is the inclusion probability of the i-th cluster in the sample s, M_i is the size of the i-th cluster, n_{II} is the sample size within each cluster, n_I is the number of sampled clusters, and where

$$\varsigma_{(Ii)} = \frac{n_I - 1}{n_I}(\hat{\bar{y}}_{Hajek} - \hat{\bar{y}}_{Hajek(Ii)})$$

$$\varepsilon_{(k)} = \frac{n - 1}{n}(\hat{\bar{y}}_{Hajek} - \hat{\bar{y}}_{Hajek(k)})$$

where $\hat{\bar{y}}_{Hajek(Ii)}$ and $\hat{\bar{y}}_{Hajek(k)}$ have the same functional form as $\hat{\bar{y}}_{Hajek}$ but omitting the i-th cluster and the k-th element, respectively, from the sample s. Note that this variance estimator utilises implicitly the Hajek (1964) approximations that are designed for large-entropy sampling designs, large samples and large populations, i.e. care should be taken with highly-stratified samples, e.g. Berger (2005).

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Berger, Y. G. (2005) Variance estimation with highly stratified sampling designs with unequal probabilities. *Australian & New Zealand Journal of Statistics*, **47**, 365–373.

Escobar, E. L. and Berger, Y. G. (2013) A jackknife variance estimator for self-weighted two-stage samples. *Statistica Sinica*, **23**, 595–613.

Hajek, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35**, 4, 1491–1523.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

See Also

VE.Jk.Tukey.Mean.Hajek
VE.Jk.CBS.HT.Mean.Hajek
VE.Jk.CBS.SYG.Mean.Hajek
VE.Jk.B.Mean.Hajek

Examples

```
data(oaxaca)                           #Loads the Oaxaca municipalities dataset
s         <- oaxaca$sSW_10_3           #Defines the sample to be used
SampData  <- oaxaca[s==1, ]            #Defines the sample dataset
nII       <- 3                         #Defines the 2nd stage fixed sample size
CluLab.s  <- SampData$IDDISTRI         #Defines the clusters' labels
CluSize.s <- SampData$SIZEDIST         #Defines the clusters' sizes
piIi.s    <- (10 * CluSize.s / 570)    #Reconstructs clusters' 1st order incl. probs.
```

```
pik.s     <- piIi.s * (nII/CluSize.s) #Reconstructs elements' 1st order incl. probs.
y1.s      <- SampData$POP10           #Defines the variable of interest y1
y2.s      <- SampData$POPMAL10        #Defines the variable of interest y2
#Computes the var. est. of the Hajek mean point estimator using y1
VE.Jk.EB.SW2.Mean.Hajek(y1.s, pik.s, nII, piIi.s, CluLab.s, CluSize.s)
#Computes the var. est. of the Hajek mean point estimator using y2
VE.Jk.EB.SW2.Mean.Hajek(y2.s, pik.s, nII, piIi.s, CluLab.s, CluSize.s)
```

`VE.Jk.EB.SW2.Ratio` *The self-weighted two-stage sampling Escobar-Berger (2013) jackknife variance estimator for the estimator of a ratio*

Description

Computes the self-weighted two-stage sampling Escobar-Berger (2013) jackknife variance estimator for the estimator of a ratio of two totals/means.

Usage

```
VE.Jk.EB.SW2.Ratio(VecY.s, VecX.s, VecPk.s, nII, VecPi.s,
                   VecCluLab.s, VecCluSize.s)
```

Arguments

`VecY.s`	vector of the numerator variable of interest; its length is equal to n, the total sample size. Its length has to be the same as the length of `VecPk.s` and `VecX.s`. There must not be missing values.
`VecX.s`	vector of the denominator variable of interest; its length is equal to n, the total sample size. Its length has to be the same as the length of `VecPk.s` and `VecY.s`. There must not be missing values. All values of `VecX.s` should be greater than zero. A warning is displayed if this does not hold and computations continue if mathematical expressions allow this kind of values for the denominator variable.
`VecPk.s`	vector of the elements' first-order inclusion probabilities; its length is equal to n, the total sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.
`nII`	the second stage sample size, i.e. the fixed number of ultimate sampling units that were selected within each cluster. Its size must be less than or equal to the minimum cluster size in the sample.
`VecPi.s`	vector of the clusters' first-order inclusion probabilities; its length is equal to n, the total sample size. Hence values are expected to be repeated in the utilised sample dataset. Values in `VecPi.s` must be greater than zero and less than or equal to one. There must not be missing values.
`VecCluLab.s`	vector of the clusters' labels for the elements; its length is equal to n, the total sample size. The labels must be integer numbers.
`VecCluSize.s`	vector of the clusters' sizes; its length is equal to n, the total sample size. Hence values are expected to be repeated in the utilised sample dataset. None of the sizes must be smaller than `nII`.

Details

For the population ratio of two totals/means of the variables y and x:

$$R = \frac{\sum_{k\in U} y_k/N}{\sum_{k\in U} x_k/N} = \frac{\sum_{k\in U} y_k}{\sum_{k\in U} x_k}$$

the ratio estimator of R is given by:

$$\hat{R} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k x_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. If s is a self-weighted two-stage sample, the variance of $\hat{R}$ can be estimated by the Escobar-Berger (2013) jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{R}) = v_{clu} + v_{obs}$$

$$v_{clu} = \sum_{i\in s}(1-\pi^*_{Ii})\varsigma^2_{(Ii)} - \frac{1}{\hat{d}}\left(\sum_{i\in s}(1-\pi_{Ii})\varsigma_{(Ii)}\right)^2$$

$$v_{obs} = \sum_{k\in s}\phi_k \varepsilon^2_{(k)}$$

where $\hat{d} = \sum_{i\in s}(1-\pi_{Ii})$, $\phi_k = I\{k \in s_i\}\pi^*_{Ii}(M_i - n_{II})/(M_i - 1)$, $\pi^*_{Ii} = \pi_{Ii} n_{II}(M_i - 1)/(n_{II} - 1)M_i$, with s_i denoting the sample elements from the i-th cluster, $I\{k \in s_i\}$ is an indicator that takes the value 1 if the k-th observation is within the i-th cluster and 0 otherwise, π_{Ii} is the inclusion probability of the i-th cluster in the sample s, M_i is the size of the i-th cluster, n_{II} is the sample size within each cluster, n_I is the number of sampled clusters, and where

$$\varsigma_{(Ii)} = \frac{n_I - 1}{n_I}(\hat{R} - \hat{R}_{(Ii)})$$

$$\varepsilon_{(k)} = \frac{n-1}{n}(\hat{R} - \hat{R}_{(k)})$$

where $\hat{R}_{(Ii)}$ and $\hat{R}_{(k)}$ have the same functional form as $\hat{R}$ but omitting the i-th cluster and the k-th element, respectively, from the sample s. Note that this variance estimator utilises implicitly the Hajek (1964) approximations that are designed for large-entropy sampling designs, large samples and large populations, i.e. care should be taken with highly-stratified samples, e.g. Berger (2005).

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Berger, Y. G. (2005) Variance estimation with highly stratified sampling designs with unequal probabilities. *Australian & New Zealand Journal of Statistics*, **47**, 365–373.

Escobar, E. L. and Berger, Y. G. (2013) A jackknife variance estimator for self-weighted two-stage samples. *Statistica Sinica*, **23**, 595–613.

Hajek, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35**, 4, 1491–1523.

See Also

VE.Jk.Tukey.Ratio
VE.Jk.CBS.HT.Ratio
VE.Jk.CBS.SYG.Ratio
VE.Jk.B.Ratio
VE.EB.HT.Ratio
VE.EB.SYG.Ratio

Examples

```
data(oaxaca)                          #Loads the Oaxaca municipalities dataset
s         <- oaxaca$sSW_10_3          #Defines the sample to be used
SampData  <- oaxaca[s==1, ]           #Defines the sample dataset
nII       <- 3                        #Defines the 2nd stage fixed sample size
CluLab.s  <- SampData$IDDISTRI        #Defines the clusters' labels
CluSize.s <- SampData$SIZEDIST        #Defines the clusters' sizes
piIi.s    <- (10 * CluSize.s / 570)   #Reconstructs clusters' 1st order incl. probs.
pik.s     <- piIi.s * (nII/CluSize.s) #Reconstructs elements' 1st order incl. probs.
y1.s      <- SampData$POP10           #Defines the numerator variable y1
y2.s      <- SampData$POPMAL10        #Defines the numerator variable y2
x.s       <- SampData$HOMES10         #Defines the denominator variable x
#Computes the var. est. of the ratio point estimator using y1
VE.Jk.EB.SW2.Ratio(y1.s, x.s, pik.s, nII, piIi.s, CluLab.s, CluSize.s)
#Computes the var. est. of the ratio point estimator using y2
VE.Jk.EB.SW2.Ratio(y2.s, x.s, pik.s, nII, piIi.s, CluLab.s, CluSize.s)
```

VE.Jk.EB.SW2.RegCo.Hajek

The self-weighted two-stage sampling Escobar-Berger (2013) jackknife variance estimator for the estimator of the regression coefficient using the Hajek point estimator

---

Description

Computes the self-weighted two-stage sampling Escobar-Berger (2013) jackknife variance estimator for the estimator of the regression coefficient using the Hajek (1971) point estimator.

Usage

```
VE.Jk.EB.SW2.RegCo.Hajek(VecY.s, VecX.s, VecPk.s, nII, VecPi.s,
                         VecCluLab.s, VecCluSize.s)
```

Arguments

`VecY.s`	vector of the variable of interest Y; its length is equal to n, the total sample size. Its length has to be the same as the length of `VecPk.s` and `VecX.s`. There must not be missing values.
`VecX.s`	vector of the variable of interest X; its length is equal to n, the total sample size. Its length has to be the same as the length of `VecPk.s` and `VecY.s`. There must not be missing values.
`VecPk.s`	vector of the elements' first-order inclusion probabilities; its length is equal to n, the total sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.
`nII`	the second stage sample size, i.e. the fixed number of ultimate sampling units that were selected within each cluster. Its size must be less than or equal to the minimum cluster size in the sample.
`VecPi.s`	vector of the clusters' first-order inclusion probabilities; its length is equal to n, the total sample size. Hence values are expected to be repeated in the utilised sample dataset. Values in `VecPi.s` must be greater than zero and less than or equal to one. There must not be missing values.
`VecCluLab.s`	vector of the clusters' labels for the elements; its length is equal to n, the total sample size. The labels must be integer numbers.
`VecCluSize.s`	vector of the clusters' sizes; its length is equal to n, the total sample size. Hence values are expected to be repeated in the utilised sample dataset. None of the sizes must be smaller than `nII`.

Details

From Linear Regression Analysis, for an imposed population model

$$y = \alpha + \beta x$$

the population regression coefficient β, assuming that the population size N is unknown (see Sarndal et al., 1992, Sec. 5.10), can be estimated by:

$$\hat{\beta}_{Hajek} = \frac{\sum_{k\in s} w_k (y_k - \hat{\bar{y}}_{Hajek})(x_k - \hat{\bar{x}}_{Hajek})}{\sum_{k\in s} w_k (x_k - \hat{\bar{x}}_{Hajek})^2}$$

where $\hat{\bar{y}}_{Hajek}$ and $\hat{\bar{x}}_{Hajek}$ are the Hajek (1971) point estimators of the population means $\bar{y} = N^{-1}\sum_{k\in U} y_k$ and $\bar{x} = N^{-1}\sum_{k\in U} x_k$, respectively,

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

$$\hat{\bar{x}}_{Hajek} = \frac{\sum_{k\in s} w_k x_k}{\sum_{k\in s} w_k}$$

and $w_k = 1/\pi_k$ with π_k denoting the inclusion probability of the k-th element in the sample s. If s is a self-weighted two-stage sample, the variance of $\hat{\beta}_{Hajek}$ can be estimated by the Escobar-Berger (2013) jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{\beta}_{Hajek}) = v_{clu} + v_{obs}$$

$$v_{clu} = \sum_{i \in s}(1 - \pi^*_{Ii})\varsigma^2_{(Ii)} - \frac{1}{\hat{d}}\left(\sum_{i \in s}(1 - \pi_{Ii})\varsigma_{(Ii)}\right)^2$$

$$v_{obs} = \sum_{k \in s}\phi_k \varepsilon^2_{(k)}$$

where $\hat{d} = \sum_{i \in s}(1 - \pi_{Ii})$, $\phi_k = I\{k \in s_i\}\pi^*_{Ii}(M_i - n_{II})/(M_i - 1)$, $\pi^*_{Ii} = \pi_{Ii}n_{II}(M_i - 1)/(n_{II} - 1)M_i$, with s_i denoting the sample elements from the i-th cluster, $I\{k \in s_i\}$ is an indicator that takes the value 1 if the k-th observation is within the i-th cluster and 0 otherwise, π_{Ii} is the inclusion probability of the i-th cluster in the sample s, M_i is the size of the i-th cluster, n_{II} is the sample size within each cluster, n_I is the number of sampled clusters, and where

$$\varsigma_{(Ii)} = \frac{n_I - 1}{n_I}(\hat{\beta}_{Hajek} - \hat{\beta}_{Hajek(Ii)})$$

$$\varepsilon_{(k)} = \frac{n - 1}{n}(\hat{\beta}_{Hajek} - \hat{\beta}_{Hajek(k)})$$

where $\hat{\beta}_{Hajek(Ii)}$ and $\hat{\beta}_{Hajek(k)}$ have the same functional form as $\hat{\beta}_{Hajek}$ but omitting the i-th cluster and the k-th element, respectively, from the sample s. Note that this variance estimator utilises implicitly the Hajek (1964) approximations that are designed for large-entropy sampling designs, large samples and large populations, i.e. care should be taken with highly-stratified samples, e.g. Berger (2005).

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Berger, Y. G. (2005) Variance estimation with highly stratified sampling designs with unequal probabilities. *Australian & New Zealand Journal of Statistics*, **47**, 365–373.

Escobar, E. L. and Berger, Y. G. (2013) A jackknife variance estimator for self-weighted two-stage samples. *Statistica Sinica*, **23**, 595–613.

Hajek, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35**, 4, 1491–1523.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Sarndal, C.-E. and Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, Inc.

See Also

VE.Jk.EB.SW2.RegCoI.Hajek
VE.Jk.Tukey.RegCo.Hajek
VE.Jk.CBS.HT.RegCo.Hajek
VE.Jk.CBS.SYG.RegCo.Hajek
VE.Jk.B.RegCo.Hajek

Examples

```
data(oaxaca)                          #Loads the Oaxaca municipalities dataset
s         <- oaxaca$sSW_10_3          #Defines the sample to be used
SampData  <- oaxaca[s==1, ]           #Defines the sample dataset
nII       <- 3                        #Defines the 2nd stage fixed sample size
CluLab.s  <- SampData$IDDISTRI        #Defines the clusters' labels
CluSize.s <- SampData$SIZEDIST        #Defines the clusters' sizes
piIi.s    <- (10 * CluSize.s / 570)   #Reconstructs clusters' 1st order incl. probs.
pik.s     <- piIi.s * (nII/CluSize.s) #Reconstructs elements' 1st order incl. probs.
y1.s      <- SampData$POP10           #Defines the variable y1
y2.s      <- SampData$POPMAL10        #Defines the variable y2
x.s       <- SampData$HOMES10         #Defines the variable x
#Computes the var. est. of the regression coeff. point estimator using y1
VE.Jk.EB.SW2.RegCo.Hajek(y1.s, x.s, pik.s, nII, piIi.s, CluLab.s, CluSize.s)
#Computes the var. est. of the regression coeff. point estimator using y2
VE.Jk.EB.SW2.RegCo.Hajek(y2.s, x.s, pik.s, nII, piIi.s, CluLab.s, CluSize.s)
```

VE.Jk.EB.SW2.RegCoI.Hajek

The self-weighted two-stage sampling Escobar-Berger (2013) jackknife variance estimator for the estimator of the intercept regression coefficient using the Hajek point estimator

---

Description

Computes the self-weighted two-stage sampling Escobar-Berger (2013) jackknife variance estimator for the estimator of the intercept regression coefficient using the Hajek (1971) point estimator.

Usage

```
VE.Jk.EB.SW2.RegCoI.Hajek(VecY.s, VecX.s, VecPk.s, nII, VecPi.s,
                          VecCluLab.s, VecCluSize.s)
```

Arguments

VecY.s	vector of the variable of interest Y; its length is equal to n, the total sample size. Its length has to be the same as the length of VecPk.s and VecX.s. There must not be missing values.

`VecX.s` vector of the variable of interest X; its length is equal to n, the total sample size. Its length has to be the same as the length of `VecPk.s` and `VecY.s`. There must not be missing values.

`VecPk.s` vector of the elements' first-order inclusion probabilities; its length is equal to n, the total sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.

`nII` the second stage sample size, i.e. the fixed number of ultimate sampling units that were selected within each cluster. Its size must be less than or equal to the minimum cluster size in the sample.

`VecPi.s` vector of the clusters' first-order inclusion probabilities; its length is equal to n, the total sample size. Hence values are expected to be repeated in the utilised sample dataset. Values in `VecPi.s` must be greater than zero and less than or equal to one. There must not be missing values.

`VecCluLab.s` vector of the clusters' labels for the elements; its length is equal to n, the total sample size. The labels must be integer numbers.

`VecCluSize.s` vector of the clusters' sizes; its length is equal to n, the total sample size. Hence values are expected to be repeated in the utilised sample dataset. None of the sizes must be smaller than `nII`.

Details

From Linear Regression Analysis, for an imposed population model

$$y = \alpha + \beta x$$

the population intercept regression coefficient α, assuming that the population size N is unknown (see Sarndal et al., 1992, Sec. 5.10), can be estimated by:

$$\hat{\alpha}_{Hajek} = \hat{\bar{y}}_{Hajek} - \frac{\sum_{k\in s} w_k (y_k - \hat{\bar{y}}_{Hajek})(x_k - \hat{\bar{x}}_{Hajek})}{\sum_{k\in s} w_k (x_k - \hat{\bar{x}}_{Hajek})^2} \hat{\bar{x}}_{Hajek}$$

where $\hat{\bar{y}}_{Hajek}$ and $\hat{\bar{x}}_{Hajek}$ are the Hajek (1971) point estimators of the population means $\bar{y} = N^{-1}\sum_{k\in U} y_k$ and $\bar{x} = N^{-1}\sum_{k\in U} x_k$, respectively,

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

$$\hat{\bar{x}}_{Hajek} = \frac{\sum_{k\in s} w_k x_k}{\sum_{k\in s} w_k}$$

and $w_k = 1/\pi_k$ with π_k denoting the inclusion probability of the k-th element in the sample s. If s is a self-weighted two-stage sample, the variance of $\hat{\alpha}_{Hajek}$ can be estimated by the Escobar-Berger (2013) jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{\alpha}_{Hajek}) = v_{clu} + v_{obs}$$

$$v_{clu} = \sum_{i\in s}(1-\pi^*_{Ii})\varsigma^2_{(Ii)} - \frac{1}{\hat{d}}\left(\sum_{i\in s}(1-\pi_{Ii})\varsigma_{(Ii)}\right)^2$$

$$v_{obs} = \sum_{k \in s} \phi_k \varepsilon_{(k)}^2$$

where $\hat{d} = \sum_{i \in s}(1 - \pi_{Ii})$, $\phi_k = I\{k \in s_i\}\pi_{Ii}^*(M_i - n_{II})/(M_i - 1)$, $\pi_{Ii}^* = \pi_{Ii} n_{II}(M_i - 1)/(n_{II} - 1)M_i$, with s_i denoting the sample elements from the i-th cluster, $I\{k \in s_i\}$ is an indicator that takes the value 1 if the k-th observation is within the i-th cluster and 0 otherwise, π_{Ii} is the inclusion probability of the i-th cluster in the sample s, M_i is the size of the i-th cluster, n_{II} is the sample size within each cluster, n_I is the number of sampled clusters, and where

$$\varsigma_{(Ii)} = \frac{n_I - 1}{n_I}(\hat{\alpha}_{Hajek} - \hat{\alpha}_{Hajek(Ii)})$$

$$\varepsilon_{(k)} = \frac{n - 1}{n}(\hat{\alpha}_{Hajek} - \hat{\alpha}_{Hajek(k)})$$

where $\hat{\alpha}_{Hajek(Ii)}$ and $\hat{\alpha}_{Hajek(k)}$ have the same functional form as $\hat{\alpha}_{Hajek}$ but omitting the i-th cluster and the k-th element, respectively, from the sample s. Note that this variance estimator utilises implicitly the Hajek (1964) approximations that are designed for large-entropy sampling designs, large samples and large populations, i.e. care should be taken with highly-stratified samples, e.g. Berger (2005).

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Berger, Y. G. (2005) Variance estimation with highly stratified sampling designs with unequal probabilities. *Australian & New Zealand Journal of Statistics*, **47**, 365–373.

Escobar, E. L. and Berger, Y. G. (2013) A jackknife variance estimator for self-weighted two-stage samples. *Statistica Sinica*, **23**, 595–613.

Hajek, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35**, 4, 1491–1523.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Sarndal, C.-E. and Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, Inc.

See Also

`VE.Jk.EB.SW2.RegCo.Hajek`
`VE.Jk.Tukey.RegCoI.Hajek`
`VE.Jk.CBS.HT.RegCoI.Hajek`
`VE.Jk.CBS.SYG.RegCoI.Hajek`
`VE.Jk.B.RegCoI.Hajek`

Examples

```
data(oaxaca)                            #Loads the Oaxaca municipalities dataset
s         <- oaxaca$sSW_10_3            #Defines the sample to be used
SampData  <- oaxaca[s==1, ]             #Defines the sample dataset
nII       <- 3                          #Defines the 2nd stage fixed sample size
CluLab.s  <- SampData$IDDISTRI          #Defines the clusters' labels
CluSize.s <- SampData$SIZEDIST          #Defines the clusters' sizes
piIi.s    <- (10 * CluSize.s / 570)     #Reconstructs clusters' 1st order incl. probs.
pik.s     <- piIi.s * (nII/CluSize.s)   #Reconstructs elements' 1st order incl. probs.
y1.s      <- SampData$POP10             #Defines the variable y1
y2.s      <- SampData$POPMAL10          #Defines the variable y2
x.s       <- SampData$HOMES10           #Defines the variable x
#Computes the var. est. of the intercept reg. coeff. point estimator using y1
VE.Jk.EB.SW2.RegCoI.Hajek(y1.s, x.s, pik.s, nII, piIi.s, CluLab.s, CluSize.s)
#Computes the var. est. of the intercept reg. coeff. point estimator using y2
VE.Jk.EB.SW2.RegCoI.Hajek(y2.s, x.s, pik.s, nII, piIi.s, CluLab.s, CluSize.s)
```

`VE.Jk.EB.SW2.Total.Hajek`

The self-weighted two-stage sampling Escobar-Berger (2013) jackknife variance estimator for the Hajek (1971) estimator of a total

---

Description

Computes the self-weighted two-stage sampling Escobar-Berger (2013) jackknife variance estimator for the Hajek estimator of a total.

Usage

```
VE.Jk.EB.SW2.Total.Hajek(VecY.s, VecPk.s, nII, VecPi.s,
                         VecCluLab.s, VecCluSize.s, N)
```

Arguments

`VecY.s`	vector of the variable of interest; its length is equal to n, the total sample size. Its length has to be the same as the length of `VecPk.s`. There must not be missing values.
`VecPk.s`	vector of the elements' first-order inclusion probabilities; its length is equal to n, the total sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.
`nII`	the second stage sample size, i.e. the fixed number of ultimate sampling units that were selected within each cluster. Its size must be less than or equal to the minimum cluster size in the sample.
`VecPi.s`	vector of the clusters' first-order inclusion probabilities; its length is equal to n, the total sample size. Hence values are expected to be repeated in the utilised sample dataset. Values in `VecPi.s` must be greater than zero and less than or equal to one. There must not be missing values.

`VecCluLab.s`	vector of the clusters' labels for the elements; its length is equal to n, the total sample size. The labels must be integer numbers.
`VecCluSize.s`	vector of the clusters' sizes; its length is equal to n, the total sample size. Hence values are expected to be repeated in the utilised sample dataset. None of the sizes must be smaller than `nII`.
`N`	the population size. It must be an integer or a double-precision scalar with zero-valued fractional part.

Details

For the population total of the variable y:

$$t = \sum_{k \in U} y_k$$

the approximately unbiased Hajek (1971) estimator of t is given by:

$$\hat{t}_{Hajek} = N \frac{\sum_{k \in s} w_k y_k}{\sum_{k \in s} w_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. If s is a self-weighted two-stage sample, the variance of $\hat{t}_{Hajek}$ can be estimated by the Escobar-Berger (2013) jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{t}_{Hajek}) = v_{clu} + v_{obs}$$

$$v_{clu} = \sum_{i \in s} (1 - \pi_{Ii}^*) \varsigma_{(Ii)}^2 - \frac{1}{\hat{d}} \left(\sum_{i \in s} (1 - \pi_{Ii}) \varsigma_{(Ii)} \right)^2$$

$$v_{obs} = \sum_{k \in s} \phi_k \varepsilon_{(k)}^2$$

where $\hat{d} = \sum_{i \in s} (1 - \pi_{Ii})$, $\phi_k = I\{k \in s_i\} \pi_{Ii}^* (M_i - n_{II})/(M_i - 1)$, $\pi_{Ii}^* = \pi_{Ii} n_{II} (M_i - 1)/(n_{II} - 1) M_i$, with s_i denoting the sample elements from the i-th cluster, $I\{k \in s_i\}$ is an indicator that takes the value 1 if the k-th observation is within the i-th cluster and 0 otherwise, π_{Ii} is the inclusion probability of the i-th cluster in the sample s, M_i is the size of the i-th cluster, n_{II} is the sample size within each cluster, n_I is the number of sampled clusters, and where

$$\varsigma_{(Ii)} = \frac{n_I - 1}{n_I} (\hat{t}_{Hajek} - \hat{t}_{Hajek(Ii)})$$

$$\varepsilon_{(k)} = \frac{n - 1}{n} (\hat{t}_{Hajek} - \hat{t}_{Hajek(k)})$$

where $\hat{t}_{Hajek(Ii)}$ and $\hat{t}_{Hajek(k)}$ have the same functional form as $\hat{t}_{Hajek}$ but omitting the i-th cluster and the k-th element, respectively, from the sample s. Note that this variance estimator utilises implicitly the Hajek (1964) approximations that are designed for large-entropy sampling designs, large samples and large populations, i.e. care should be taken with highly-stratified samples, e.g. Berger (2005).

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Berger, Y. G. (2005) Variance estimation with highly stratified sampling designs with unequal probabilities. *Australian & New Zealand Journal of Statistics*, **47**, 365–373.

Escobar, E. L. and Berger, Y. G. (2013) A jackknife variance estimator for self-weighted two-stage samples. *Statistica Sinica*, **23**, 595–613.

Hajek, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35**, 4, 1491–1523.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

See Also

VE.Jk.Tukey.Total.Hajek
VE.Jk.CBS.HT.Total.Hajek
VE.Jk.CBS.SYG.Total.Hajek
VE.Jk.B.Total.Hajek

Examples

```
data(oaxaca)                         #Loads the Oaxaca municipalities dataset
s         <- oaxaca$sSW_10_3         #Defines the sample to be used
N         <- dim(oaxaca)[1]          #Defines the population size
SampData  <- oaxaca[s==1, ]          #Defines the sample dataset
nII       <- 3                       #Defines the 2nd stage fixed sample size
CluLab.s  <- SampData$IDDISTRI       #Defines the clusters' labels
CluSize.s <- SampData$SIZEDIST       #Defines the clusters' sizes
piIi.s    <- (10 * CluSize.s / 570)  #Reconstructs clusters' 1st order incl. probs.
pik.s     <- piIi.s * (nII/CluSize.s) #Reconstructs elements' 1st order incl. probs.
y1.s      <- SampData$POP10          #Defines the variable of interest y1
y2.s      <- SampData$POPMAL10       #Defines the variable of interest y2
#Computes the var. est. of the Hajek total point estimator using y1
VE.Jk.EB.SW2.Total.Hajek(y1.s, pik.s, nII, piIi.s, CluLab.s, CluSize.s, N)
#Computes the var. est. of the Hajek total point estimator using y2
VE.Jk.EB.SW2.Total.Hajek(y2.s, pik.s, nII, piIi.s, CluLab.s, CluSize.s, N)
```

`VE.Jk.Tukey.Corr.Hajek` *The Tukey (1958) jackknife variance estimator for the estimator of a correlation coefficient using the Hajek point estimator*

Description

Computes the Quenouille(1956); Tukey (1958) jackknife variance estimator for the estimator of a correlation coefficient of two variables using the Hajek (1971) point estimator.

Usage

```
VE.Jk.Tukey.Corr.Hajek(VecY.s, VecX.s, VecPk.s, N, FPC= TRUE)
```

Arguments

`VecY.s`	vector of the variable of interest Y; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecX.s`. There must not be missing values.
`VecX.s`	vector of the variable of interest X; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecY.s`. There must not be missing values.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.
`N`	the population size. It must be an integer or a double-precision scalar with zero-valued fractional part. This information is utilised for the finite population correction only, see `FPC` below.
`FPC`	logical value. If an ad hoc finite population correction $FPC = 1 - n/N$ is to be used. The default is TRUE.

Details

For the population correlation coefficient of two variables y and x:

$$C = \frac{\sum_{k\in U}(y_k - \bar{y})(x_k - \bar{x})}{\sqrt{\sum_{k\in U}(y_k - \bar{y})^2}\sqrt{\sum_{k\in U}(x_k - \bar{x})^2}}$$

the point estimator of C, assuming that N is unknown (see Sarndal et al., 1992, Sec. 5.9), is:

$$\hat{C}_{Hajek} = \frac{\sum_{k\in s} w_k(y_k - \hat{\bar{y}}_{Hajek})(x_k - \hat{\bar{x}}_{Hajek})}{\sqrt{\sum_{k\in s} w_k(y_k - \hat{\bar{y}}_{Hajek})^2}\sqrt{\sum_{k\in s} w_k(x_k - \hat{\bar{x}}_{Hajek})^2}}$$

where $\hat{\bar{y}}_{Hajek}$ is the Hajek (1971) point estimator of the population mean $\bar{y} = N^{-1}\sum_{k\in U} y_k$,

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

and $w_k = 1/\pi_k$ with π_k denoting the inclusion probability of the k-th element in the sample s. The variance of $\hat{C}_{Hajek}$ can be estimated by the Quenouille(1956); Tukey (1958) jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{C}_{Hajek}) = \left(1 - \frac{n}{N}\right) \frac{n-1}{n} \sum_{k \in s} \left(\hat{C}_{Hajek(k)} - \hat{C}_{Hajek}\right)^2$$

where $\hat{C}_{Hajek(k)}$ has the same functional form as $\hat{C}_{Hajek}$ but omitting the k-th element from the sample s. Note that we are implementing the Tukey (1958) jackknife variance estimator using the 'ad hoc' finite population correction $1 - n/N$ (see Shao and Tu, 1995; Wolter, 2007). If `FPC=FALSE` then the term $1 - n/N$ is ommited from the above formula.

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Quenouille, M. H. (1956) Notes on bias in estimation. *Biometrika*, **43**, 353–360.

Sarndal, C.-E. and Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, Inc.

Shao, J. and Tu, D. (1995) *The Jackknife and Bootstrap*. Springer-Verlag, Inc.

Tukey, J. W. (1958) Bias and confidence in not-quite large samples (abstract). *The Annals of Mathematical Statistics*, **29**, 2, p. 614.

Wolter, K. M. (2007) *Introduction to Variance Estimation*. 2nd Ed. Springer, Inc.

See Also

VE.Jk.CBS.HT.Corr.Hajek
VE.Jk.CBS.SYG.Corr.Hajek
VE.Jk.B.Corr.Hajek
VE.Jk.EB.SW2.Corr.Hajek

Examples

```
data(oaxaca)                                 #Loads the Oaxaca municipalities dataset
pik.U <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s     <- oaxaca$sHOMES00                     #Defines the sample to be used
N     <- dim(oaxaca)[1]                      #Defines the population size
y1    <- oaxaca$POP10                        #Defines the variable of interest y1
y2    <- oaxaca$POPMAL10                     #Defines the variable of interest y2
x     <- oaxaca$HOMES10                      #Defines the variable of interest x
```

```
#Computes the var. est. of the corr. coeff. point estimator using y1
VE.Jk.Tukey.Corr.Hajek(y1[s==1], x[s==1], pik.U[s==1], N)
#Computes the var. est. of the corr. coeff. point estimator using y2
VE.Jk.Tukey.Corr.Hajek(y2[s==1], x[s==1], pik.U[s==1], N, FPC= FALSE)
```

`VE.Jk.Tukey.Corr.NHT` *The Tukey (1958) jackknife variance estimator for the estimator of a correlation coefficient using the Narain-Horvitz-Thompson point estimator*

Description

Computes the Quenouille(1956); Tukey (1958) jackknife variance estimator for the estimator of a correlation coefficient of two variables using the Narain (1951); Horvitz-Thompson (1952) point estimator.

Usage

```
VE.Jk.Tukey.Corr.NHT(VecY.s, VecX.s, VecPk.s, N, FPC= TRUE)
```

Arguments

`VecY.s`	vector of the variable of interest Y; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecX.s`. There must not be missing values.
`VecX.s`	vector of the variable of interest X; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecY.s`. There must not be missing values.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.
`N`	the population size. It must be an integer or a double-precision scalar with zero-valued fractional part. This information is also utilised for the finite population correction, see `FPC` below.
`FPC`	logical value. If an ad hoc finite population correction $FPC = 1 - n/N$ is to be used. The default is TRUE.

Details

For the population correlation coefficient of two variables y and x:

$$C = \frac{\sum_{k\in U}(y_k - \bar{y})(x_k - \bar{x})}{\sqrt{\sum_{k\in U}(y_k - \bar{y})^2}\sqrt{\sum_{k\in U}(x_k - \bar{x})^2}}$$

the point estimator of C is given by:

$$\hat{C} = \frac{\sum_{k\in s} w_k(y_k - \hat{\bar{y}}_{NHT})(x_k - \hat{\bar{x}}_{NHT})}{\sqrt{\sum_{k\in s} w_k(y_k - \hat{\bar{y}}_{NHT})^2}\sqrt{\sum_{k\in s} w_k(x_k - \hat{\bar{x}}_{NHT})^2}}$$

where $\hat{\bar{y}}_{NHT}$ is the Narain (1951); Horvitz-Thompson (1952) estimator for the population mean $\bar{y} = N^{-1}\sum_{k\in U} y_k$,

$$\hat{\bar{y}}_{NHT} = \frac{1}{N}\sum_{k\in s} w_k y_k$$

and $w_k = 1/\pi_k$ with π_k denoting the inclusion probability of the k-th element in the sample s. The variance of $\hat{C}$ can be estimated by the Quenouille(1956); Tukey (1958) jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{C}) = \left(1 - \frac{n}{N}\right)\frac{n-1}{n}\sum_{k\in s}\left(\hat{C}_{(k)} - \hat{C}\right)^2$$

where $\hat{C}_{(k)}$ has the same functional form as $\hat{C}$ but omitting the k-th element from the sample s. Note that we are implementing the Tukey (1958) jackknife variance estimator using the 'ad hoc' finite population correction $1 - n/N$ (see Shao and Tu, 1995; Wolter, 2007). If `FPC=FALSE` then the term $1 - n/N$ is ommited from the above formula.

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Narain, R. D. (1951) On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **3**, 169–175.

Quenouille, M. H. (1956) Notes on bias in estimation. *Biometrika*, **43**, 353–360.

Shao, J. and Tu, D. (1995) *The Jackknife and Bootstrap*. Springer-Verlag, Inc.

Tukey, J. W. (1958) Bias and confidence in not-quite large samples (abstract). *The Annals of Mathematical Statistics*, **29**, 2, p. 614.

Wolter, K. M. (2007) *Introduction to Variance Estimation*. 2nd Ed. Springer, Inc.

See Also

Est.Corr.Hajek

Examples

```
data(oaxaca)                               #Loads the Oaxaca municipalities dataset
pik.U <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s     <- oaxaca$sHOMES00                   #Defines the sample to be used
N     <- dim(oaxaca)[1]                    #Defines the population size
y1    <- oaxaca$POP10                      #Defines the variable of interest y1
y2    <- oaxaca$POPMAL10                   #Defines the variable of interest y2
```

```
x     <- oaxaca$HOMES10                     #Defines the variable of interest x
#Computes the var. est. of the corr. coeff. point estimator using y1
VE.Jk.Tukey.Corr.NHT(y1[s==1], x[s==1], pik.U[s==1], N)
#Computes the var. est. of the corr. coeff. point estimator using y2
VE.Jk.Tukey.Corr.NHT(y2[s==1], x[s==1], pik.U[s==1], N, FPC= FALSE)
```

`VE.Jk.Tukey.Mean.Hajek` *The Tukey (1958) jackknife variance estimator for the Hajek estimator of a mean*

Description

Computes the Quenouille(1956); Tukey (1958) jackknife variance estimator for the Hajek (1971) estimator of a mean.

Usage

```
VE.Jk.Tukey.Mean.Hajek(VecY.s, VecPk.s, N, FPC= TRUE)
```

Arguments

`VecY.s`	vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s`. There must not be missing values.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.
`N`	the population size. It must be an integer or a double-precision scalar with zero-valued fractional part. This information is also utilised for the finite population correction, see `FPC` below.
`FPC`	logical value. If an ad hoc finite population correction $FPC = 1 - n/N$ is to be used. The default is TRUE.

Details

For the population mean of the variable y:

$$\bar{y} = \frac{1}{N}\sum_{k\in U} y_k$$

the approximately unbiased Hajek (1971) estimator of $\bar{y}$ is given by:

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. The variance of $\hat{\bar{y}}_{Hajek}$ can be estimated by the Quenouille(1956); Tukey (1958) jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{\bar{y}}_{Hajek}) = \left(1 - \frac{n}{N}\right) \frac{n-1}{n} \sum_{k \in s} \left(\hat{\bar{y}}_{Hajek(k)} - \hat{\bar{y}}_{Hajek}\right)^2$$

where

$$\hat{\bar{y}}_{Hajek(k)} = \frac{\sum_{l \in s, l \neq k} w_l y_l}{\sum_{l \in s, l \neq k} w_l}$$

Note that we are implementing the Tukey (1958) jackknife variance estimator using the 'ad hoc' finite population correction $1 - n/N$ (see Shao and Tu, 1995; Wolter, 2007). If `FPC=FALSE` then the term $1 - n/N$ is ommited from the above formula.

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Quenouille, M. H. (1956) Notes on bias in estimation. *Biometrika*, **43**, 353–360.

Shao, J. and Tu, D. (1995) *The Jackknife and Bootstrap*. Springer-Verlag, Inc.

Tukey, J. W. (1958) Bias and confidence in not-quite large samples (abstract). *The Annals of Mathematical Statistics*, **29**, 2, p. 614.

Wolter, K. M. (2007) *Introduction to Variance Estimation*. 2nd Ed. Springer, Inc.

See Also

VE.Jk.CBS.HT.Mean.Hajek
VE.Jk.CBS.SYG.Mean.Hajek
VE.Jk.B.Mean.Hajek
VE.Jk.EB.SW2.Mean.Hajek

Examples

```
data(oaxaca)                                 #Loads the Oaxaca municipalities dataset
pik.U <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s     <- oaxaca$sHOMES00                    #Defines the sample to be used
N     <- dim(oaxaca)[1]                     #Defines the population size
y1    <- oaxaca$POP10                       #Defines the variable of interest y1
y2    <- oaxaca$POPMAL10                    #Defines the variable of interest y2
#Computes the var. est. of the Hajek mean point estimator using y1
```

```
VE.Jk.Tukey.Mean.Hajek(y1[s==1], pik.U[s==1], N)
#Computes the var. est. of the Hajek mean point estimator using y2
VE.Jk.Tukey.Mean.Hajek(y2[s==1], pik.U[s==1], N, FPC= FALSE)
```

`VE.Jk.Tukey.Ratio` *The Tukey (1958) jackknife variance estimator for the estimator of a ratio*

Description

Computes the Quenouille(1956); Tukey (1958) jackknife variance estimator for the estimator of a ratio of two totals/means.

Usage

```
VE.Jk.Tukey.Ratio(VecY.s, VecX.s, VecPk.s, N, FPC= TRUE)
```

Arguments

`VecY.s`	vector of the numerator variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecX.s`. There must not be missing values.
`VecX.s`	vector of the denominator variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecY.s`. There must not be missing values. All values of `VecX.s` should be greater than zero. A warning is displayed if this does not hold and computations continue if mathematical expressions allow this kind of values for the denominator variable.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.
`N`	the population size. It must be an integer or a double-precision scalar with zero-valued fractional part. This information is also utilised for the finite population correction, see `FPC` below.
`FPC`	logical value. If an ad hoc finite population correction $FPC = 1 - n/N$ is to be used. The default is TRUE.

Details

For the population ratio of two totals/means of the variables y and x:

$$R = \frac{\sum_{k\in U} y_k/N}{\sum_{k\in U} x_k/N} = \frac{\sum_{k\in U} y_k}{\sum_{k\in U} x_k}$$

the ratio estimator of R is given by:

$$\hat{R} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k x_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. The variance of $\hat{R}$ can be estimated by the Quenouille(1956); Tukey (1958) jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{R}) = \left(1 - \frac{n}{N}\right) \frac{n-1}{n} \sum_{k \in s} \left(\hat{R}_{(k)} - \hat{R}\right)^2$$

where

$$\hat{R}_{(k)} = \frac{\sum_{l \in s, l \neq k} w_l y_l}{\sum_{l \in s, l \neq k} w_l x_l}$$

Note that we are implementing the Tukey (1958) jackknife variance estimator using the 'ad hoc' finite population correction $1 - n/N$ (see Shao and Tu, 1995; Wolter, 2007). If `FPC=FALSE` then the term $1 - n/N$ is ommited from the above formula.

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Quenouille, M. H. (1956) Notes on bias in estimation. *Biometrika*, **43**, 353–360.

Shao, J. and Tu, D. (1995) *The Jackknife and Bootstrap*. Springer-Verlag, Inc.

Tukey, J. W. (1958) Bias and confidence in not-quite large samples (abstract). *The Annals of Mathematical Statistics*, **29**, 2, p. 614.

Wolter, K. M. (2007) *Introduction to Variance Estimation*. 2nd Ed. Springer, Inc.

See Also

VE.Lin.HT.Ratio
VE.Lin.SYG.Ratio
VE.Jk.CBS.HT.Ratio
VE.Jk.CBS.SYG.Ratio
VE.Jk.B.Ratio
VE.Jk.EB.SW2.Ratio
VE.EB.HT.Ratio
VE.EB.SYG.Ratio

Examples

```
data(oaxaca)                                 #Loads the Oaxaca municipalities dataset
pik.U <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s     <- oaxaca$sHOMES00                     #Defines the sample to be used
N     <- dim(oaxaca)[1]                      #Defines the population size
y1    <- oaxaca$POP10                        #Defines the numerator variable y1
y2    <- oaxaca$POPMAL10                     #Defines the numerator variable y2
```

```
x     <- oaxaca$HOMES10                    #Defines the denominator variable x
#Computes the var. est. of the ratio point estimator using y1
VE.Jk.Tukey.Ratio(y1[s==1], x[s==1], pik.U[s==1], N)
#Computes the var. est. of the ratio point estimator using y2
VE.Jk.Tukey.Ratio(y2[s==1], x[s==1], pik.U[s==1], N, FPC= FALSE)
```

VE.Jk.Tukey.RegCo.Hajek

The Tukey (1958) jackknife variance estimator for the estimator of the regression coefficient using the Hajek point estimator

---

Description

Computes the Quenouille(1956); Tukey (1958) jackknife variance estimator for the estimator of the regression coefficient using the Hajek (1971) point estimator.

Usage

```
VE.Jk.Tukey.RegCo.Hajek(VecY.s, VecX.s, VecPk.s, N, FPC= TRUE)
```

Arguments

VecY.s	vector of the variable of interest Y; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecX.s. There must not be missing values.
VecX.s	vector of the variable of interest X; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecY.s. There must not be missing values.
VecPk.s	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be missing values.
N	the population size. It must be an integer or a double-precision scalar with zero-valued fractional part. This information is utilised for the finite population correction only, see FPC below.
FPC	logical value. If an ad hoc finite population correction $FPC = 1 - n/N$ is to be used. The default is TRUE.

Details

From Linear Regression Analysis, for an imposed population model

$$y = \alpha + \beta x$$

the population regression coefficient β, assuming that the population size N is unknown (see Sarndal et al., 1992, Sec. 5.10), can be estimated by:

$$\hat{\beta}_{Hajek} = \frac{\sum_{k\in s} w_k(y_k - \hat{\bar{y}}_{Hajek})(x_k - \hat{\bar{x}}_{Hajek})}{\sum_{k\in s} w_k(x_k - \hat{\bar{x}}_{Hajek})^2}$$

where $\hat{\bar{y}}_{Hajek}$ and $\hat{\bar{x}}_{Hajek}$ are the Hajek (1971) point estimators of the population means $\bar{y} = N^{-1}\sum_{k\in U} y_k$ and $\bar{x} = N^{-1}\sum_{k\in U} x_k$, respectively,

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

$$\hat{\bar{x}}_{Hajek} = \frac{\sum_{k\in s} w_k x_k}{\sum_{k\in s} w_k}$$

and $w_k = 1/\pi_k$ with π_k denoting the inclusion probability of the k-th element in the sample s. The variance of $\hat{\beta}_{Hajek}$ can be estimated by the Quenouille(1956); Tukey (1958) jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{\beta}_{Hajek}) = \left(1 - \frac{n}{N}\right)\frac{n-1}{n}\sum_{k\in s}\left(\hat{\beta}_{Hajek(k)} - \hat{\beta}_{Hajek}\right)^2$$

where $\hat{\beta}_{Hajek(k)}$ has the same functional form as $\hat{\beta}_{Hajek}$ but omitting the k-th element from the sample s. Note that we are implementing the Tukey (1958) jackknife variance estimator using the 'ad hoc' finite population correction $1 - n/N$ (see Shao and Tu, 1995; Wolter, 2007). If `FPC=FALSE` then the term $1 - n/N$ is ommited from the above formula.

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Quenouille, M. H. (1956) Notes on bias in estimation. *Biometrika*, **43**, 353–360.

Sarndal, C.-E. and Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, Inc.

Shao, J. and Tu, D. (1995) *The Jackknife and Bootstrap*. Springer-Verlag, Inc.

Tukey, J. W. (1958) Bias and confidence in not-quite large samples (abstract). *The Annals of Mathematical Statistics*, **29**, 2, p. 614.

Wolter, K. M. (2007) *Introduction to Variance Estimation*. 2nd Ed. Springer, Inc.

See Also

VE.Jk.Tukey.RegCoI.Hajek
VE.Jk.CBS.HT.RegCo.Hajek
VE.Jk.CBS.SYG.RegCo.Hajek
VE.Jk.B.RegCo.Hajek
VE.Jk.EB.SW2.RegCo.Hajek

Examples

```
data(oaxaca)                                 #Loads the Oaxaca municipalities dataset
pik.U <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s     <- oaxaca$sHOMES00                     #Defines the sample to be used
N     <- dim(oaxaca)[1]                      #Defines the population size
y1    <- oaxaca$POP10                        #Defines the variable of interest y1
y2    <- oaxaca$POPMAL10                     #Defines the variable of interest y2
x     <- oaxaca$HOMES10                      #Defines the variable of interest x
#Computes the var. est. of the regression coeff. point estimator using y1
VE.Jk.Tukey.RegCo.Hajek(y1[s==1], x[s==1], pik.U[s==1], N)
#Computes the var. est. of the regression coeff. point estimator using y2
VE.Jk.Tukey.RegCo.Hajek(y2[s==1], x[s==1], pik.U[s==1], N, FPC= FALSE)
```

VE.Jk.Tukey.RegCoI.Hajek

The Tukey (1958) jackknife variance estimator for the estimator of the intercept regression coefficient using the Hajek point estimator

---

Description

Computes the Quenouille(1956); Tukey (1958) jackknife variance estimator for the estimator of the intercept regression coefficient using the Hajek (1971) point estimator.

Usage

```
VE.Jk.Tukey.RegCoI.Hajek(VecY.s, VecX.s, VecPk.s, N, FPC= TRUE)
```

Arguments

`VecY.s`	vector of the variable of interest Y; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecX.s`. There must not be missing values.
`VecX.s`	vector of the variable of interest X; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s` and `VecY.s`. There must not be missing values.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.
`N`	the population size. It must be an integer or a double-precision scalar with zero-valued fractional part. This information is utilised for the finite population correction only, see `FPC` below.
`FPC`	logical value. If an ad hoc finite population correction $FPC = 1 - n/N$ is to be used. The default is TRUE.

Details

From Linear Regression Analysis, for an imposed population model

$$y = \alpha + \beta x$$

the population intercept regression coefficient α, assuming that the population size N is unknown (see Sarndal et al., 1992, Sec. 5.10), can be estimated by:

$$\hat{\alpha}_{Hajek} = \hat{\bar{y}}_{Hajek} - \frac{\sum_{k\in s} w_k (y_k - \hat{\bar{y}}_{Hajek})(x_k - \hat{\bar{x}}_{Hajek})}{\sum_{k\in s} w_k (x_k - \hat{\bar{x}}_{Hajek})^2} \hat{\bar{x}}_{Hajek}$$

where $\hat{\bar{y}}_{Hajek}$ and $\hat{\bar{x}}_{Hajek}$ are the Hajek (1971) point estimators of the population means $\bar{y} = N^{-1}\sum_{k\in U} y_k$ and $\bar{x} = N^{-1}\sum_{k\in U} x_k$, respectively,

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

$$\hat{\bar{x}}_{Hajek} = \frac{\sum_{k\in s} w_k x_k}{\sum_{k\in s} w_k}$$

and $w_k = 1/\pi_k$ with π_k denoting the inclusion probability of the k-th element in the sample s. The variance of $\hat{\alpha}_{Hajek}$ can be estimated by the Quenouille(1956); Tukey (1958) jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{\alpha}_{Hajek}) = \left(1 - \frac{n}{N}\right)\frac{n-1}{n}\sum_{k\in s}\left(\hat{\alpha}_{Hajek(k)} - \hat{\alpha}_{Hajek}\right)^2$$

where $\hat{\alpha}_{Hajek(k)}$ has the same functional form as $\hat{\alpha}_{Hajek}$ but omitting the k-th element from the sample s. Note that we are implementing the Tukey (1958) jackknife variance estimator using the 'ad hoc' finite population correction $1 - n/N$ (see Shao and Tu, 1995; Wolter, 2007). If `FPC=FALSE` then the term $1 - n/N$ is ommited from the above formula.

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Quenouille, M. H. (1956) Notes on bias in estimation. *Biometrika*, **43**, 353–360.

Sarndal, C.-E. and Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, Inc.

Shao, J. and Tu, D. (1995) *The Jackknife and Bootstrap*. Springer-Verlag, Inc.

Tukey, J. W. (1958) Bias and confidence in not-quite large samples (abstract). *The Annals of Mathematical Statistics*, **29**, 2, p. 614.

Wolter, K. M. (2007) *Introduction to Variance Estimation*. 2nd Ed. Springer, Inc.

See Also

VE.Jk.Tukey.RegCo.Hajek
VE.Jk.CBS.HT.RegCoI.Hajek
VE.Jk.CBS.SYG.RegCoI.Hajek
VE.Jk.B.RegCoI.Hajek
VE.Jk.EB.SW2.RegCoI.Hajek

Examples

```
data(oaxaca)                                #Loads the Oaxaca municipalities dataset
pik.U <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s     <- oaxaca$sHOMES00                    #Defines the sample to be used
N     <- dim(oaxaca)[1]                     #Defines the population size
y1    <- oaxaca$POP10                       #Defines the variable of interest y1
y2    <- oaxaca$POPMAL10                    #Defines the variable of interest y2
x     <- oaxaca$HOMES10                     #Defines the variable of interest x
#Computes the var. est. of the intercept reg. coeff. point estimator using y1
VE.Jk.Tukey.RegCoI.Hajek(y1[s==1], x[s==1], pik.U[s==1], N)
#Computes the var. est. of the intercept reg. coeff. point estimator using y2
VE.Jk.Tukey.RegCoI.Hajek(y2[s==1], x[s==1], pik.U[s==1], N, FPC= FALSE)
```

VE.Jk.Tukey.Total.Hajek

The Tukey (1958) jackknife variance estimator for the Hajek estimator of a total

---

Description

Computes the Quenouille(1956); Tukey (1958) jackknife variance estimator for the Hajek (1971) estimator of a total.

Usage

```
VE.Jk.Tukey.Total.Hajek(VecY.s, VecPk.s, N, FPC= TRUE)
```

Arguments

VecY.s	vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s. There must not be missing values.
VecPk.s	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be missing values.
N	the population size. It must be an integer or a double-precision scalar with zero-valued fractional part. This information is also utilised for the finite population correction, see FPC below.
FPC	logical value. If an ad hoc finite population correction $FPC = 1 - n/N$ is to be used. The default is TRUE.

Details

For the population total of the variable y:

$$t = \sum_{k \in U} y_k$$

the approximately unbiased Hajek (1971) estimator of t is given by:

$$\hat{t}_{Hajek} = N \frac{\sum_{k \in s} w_k y_k}{\sum_{k \in s} w_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. The variance of $\hat{t}_{Hajek}$ can be estimated by the Quenouille(1956); Tukey (1958) jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{t}_{Hajek}) = \left(1 - \frac{n}{N}\right) \frac{n-1}{n} \sum_{k \in s} \left(\hat{t}_{Hajek(k)} - \hat{t}_{Hajek}\right)^2$$

where

$$\hat{t}_{Hajek(k)} = N \frac{\sum_{l \in s, l \neq k} w_l y_l}{\sum_{l \in s, l \neq k} w_l}$$

Note that we are implementing the Tukey (1958) jackknife variance estimator using the 'ad hoc' finite population correction $1 - n/N$ (see Shao and Tu, 1995; Wolter, 2007). If `FPC=FALSE` then the term $1 - n/N$ is ommited from the above formula.

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Quenouille, M. H. (1956) Notes on bias in estimation. *Biometrika*, **43**, 353–360.

Shao, J. and Tu, D. (1995) *The Jackknife and Bootstrap*. Springer-Verlag, Inc.

Tukey, J. W. (1958) Bias and confidence in not-quite large samples (abstract). *The Annals of Mathematical Statistics*, **29**, 2, p. 614.

Wolter, K. M. (2007) *Introduction to Variance Estimation*. 2nd Ed. Springer, Inc.

See Also

`VE.Jk.CBS.HT.Total.Hajek`
`VE.Jk.CBS.SYG.Total.Hajek`
`VE.Jk.B.Total.Hajek`
`VE.Jk.EB.SW2.Total.Hajek`

Examples

```
data(oaxaca)                                 #Loads the Oaxaca municipalities dataset
pik.U <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s     <- oaxaca$sHOMES00                     #Defines the sample to be used
N     <- dim(oaxaca)[1]                      #Defines the population size
y1    <- oaxaca$POP10                        #Defines the variable of interest y1
y2    <- oaxaca$POPMAL10                     #Defines the variable of interest y2
#Computes the var. est. of the Hajek total point estimator using y1
VE.Jk.Tukey.Total.Hajek(y1[s==1], pik.U[s==1], N)
#Computes the var. est. of the Hajek total point estimator using y2
VE.Jk.Tukey.Total.Hajek(y2[s==1], pik.U[s==1], N, FPC= FALSE)
```

`VE.Lin.HT.Ratio` *The unequal probability linearisation variance estimator for the estimator of a ratio (Horvitz-Thompson form)*

Description

Computes the unequal probability Taylor linearisation variance estimator for the estimator of a ratio of two totals/means. It uses the Horvitz-Thompson (1952) variance form.

Usage

```
VE.Lin.HT.Ratio(VecY.s, VecX.s, VecPk.s, MatPkl.s)
```

Arguments

VecY.s	vector of the numerator variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecX.s. There must not be missing values.
VecX.s	vector of the denominator variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecY.s. There must not be missing values. All values of VecX.s should be greater than zero. A warning is displayed if this does not hold and computations continue if mathematical expressions allow this kind of values for the denominator variable.
VecPk.s	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be missing values.
MatPkl.s	matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in MatPkl.s must be greater than zero and less than or equal to one. There must not be missing values.

Details

For the population ratio of two totals/means of the variables y and x:

$$R = \frac{\sum_{k\in U} y_k/N}{\sum_{k\in U} x_k/N} = \frac{\sum_{k\in U} y_k}{\sum_{k\in U} x_k}$$

the ratio estimator of R is given by:

$$\hat{R} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k x_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. The variance of $\hat{R}$ can be estimated by the unequal probability linearisation variance estimator (implemented by the current function). For details see Woodruff (1971); Deville (1999); Demnati-Rao (2004); Sarndal et al., (1992, Secs. 5.5 and 5.6):

$$\hat{V}(\hat{R}) = \sum_{k\in s}\sum_{l\in s} \frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}} w_k u_k w_l u_l$$

where

$$u_k = \frac{y_k - \hat{R}x_k}{\hat{t}_{x,NHT}}$$

with

$$\hat{t}_{x,NHT} = \sum_{k\in s} w_k x_k$$

the unbiased Narain (1951); Horvitz-Thompson (1952) estimator of the population total for the (denominator) variable `VecX.s`.

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Demnati, A. and Rao, J. N. K. (2004) Linearization variance estimators for survey data. *Survey Methodology*, **30**, 17–26.

Deville, J.-C. (1999) Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology*, **25**, 193–203.

Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Narain, R. D. (1951) On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **3**, 169–175.

Sarndal, C.-E. and Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, Inc.

Woodruff, R. S. (1971) A Simple Method for Approximating the Variance of a Complicated Estimate. *Journal of the American Statistical Association*, **66**, 334, 411–414.

See Also

VE.Lin.SYG.Ratio
VE.Jk.Tukey.Ratio
VE.Jk.CBS.SYG.Ratio
VE.Jk.B.Ratio
VE.Jk.EB.SW2.Ratio
VE.EB.HT.Ratio
VE.EB.SYG.Ratio

Examples

```
data(oaxaca)                                 #Loads the Oaxaca municipalities dataset
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s      <- oaxaca$sHOMES00                    #Defines the sample to be used
y1     <- oaxaca$POP10                       #Defines the numerator variable y1
y2     <- oaxaca$POPMAL10                    #Defines the numerator variable y2
x      <- oaxaca$HOMES10                     #Defines the denominator variable x
#This approximation is only suitable for large-entropy sampling designs
pikl.s <- Pkl.Hajek.s(pik.U[s==1])           #Approx. 2nd order incl. probs. from s
#Computes the var. est. of the ratio point estimator using y1
VE.Lin.HT.Ratio(y1[s==1], x[s==1], pik.U[s==1], pikl.s)
#Computes the var. est. of the ratio point estimator using y2
VE.Lin.HT.Ratio(y2[s==1], x[s==1], pik.U[s==1], pikl.s)
```

VE.Lin.SYG.Ratio — *The unequal probability linearisation variance estimator for the estimator of a ratio (Sen-Yates-Grundy form)*

Description

Computes the unequal probability Taylor linearisation variance estimator for the estimator of a ratio of two totals/means. It uses the Sen (1953); Yates-Grundy(1953) variance form.

Usage

```
VE.Lin.SYG.Ratio(VecY.s, VecX.s, VecPk.s, MatPkl.s)
```

Arguments

VecY.s	vector of the numerator variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecX.s. There must not be missing values.
VecX.s	vector of the denominator variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecY.s. There must not be missing values. All values of VecX.s should be greater than zero. A warning is displayed if this does not hold and computations continue if mathematical expressions allow this kind of values for the denominator variable.

`VecPk.s` vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.

`MatPkl.s` matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in `MatPkl.s` must be greater than zero and less than or equal to one. There must not be missing values.

Details

For the population ratio of two totals/means of the variables y and x:

$$R = \frac{\sum_{k\in U} y_k/N}{\sum_{k\in U} x_k/N} = \frac{\sum_{k\in U} y_k}{\sum_{k\in U} x_k}$$

the ratio estimator of R is given by:

$$\hat{R} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k x_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. The variance of $\hat{R}$ can be estimated by the unequal probability linearisation variance estimator (implemented by the current function). For details see Woodruff (1971); Deville (1999); Demnati-Rao (2004); Sarndal et al., (1992, Secs. 5.5 and 5.6):

$$\hat{V}(\hat{R}) = \frac{-1}{2}\sum_{k\in s}\sum_{l\in s}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}}(w_k u_k - w_l u_l)^2$$

where

$$u_k = \frac{y_k - \hat{R}x_k}{\hat{t}_{x,NHT}}$$

with

$$\hat{t}_{x,NHT} = \sum_{k\in s} w_k x_k$$

the unbiased Narain (1951); Horvitz-Thompson (1952) estimator of the population total for the (denominator) variable `VecX.s`.

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Demnati, A. and Rao, J. N. K. (2004) Linearization variance estimators for survey data. *Survey Methodology*, **30**, 17–26.

Deville, J.-C. (1999) Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology*, **25**, 193–203.

Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Narain, R. D. (1951) On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **3**, 169–175.

Sarndal, C.-E. and Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, Inc.

Sen, A. R. (1953) On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **5**, 119–127.

Woodruff, R. S. (1971) A Simple Method for Approximating the Variance of a Complicated Estimate. *Journal of the American Statistical Association*, **66**, 334, 411–414.

Yates, F. and Grundy, P. M. (1953) Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B*, **15**, 253–261.

See Also

VE.Lin.HT.RatioVE.Jk.Tukey.Ratio
VE.Jk.CBS.HT.Ratio
VE.Jk.B.Ratio
VE.Jk.EB.SW2.Ratio
VE.EB.HT.Ratio
VE.EB.SYG.Ratio

Examples

```
data(oaxaca)                                 #Loads the Oaxaca municipalities dataset
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s      <- oaxaca$sHOMES00                    #Defines the sample to be used for
y1     <- oaxaca$POP10                       #Defines the numerator variable y1
y2     <- oaxaca$POPMAL10                    #Defines the numerator variable y2
x      <- oaxaca$HOMES10                     #Defines the denominator variable x
#This approximation is only suitable for large-entropy sampling designs
pikl.s <- Pkl.Hajek.s(pik.U[s==1])           #Approx. 2nd order incl. probs. from s
#Computes the var. est. of the ratio point estimator using y1
VE.Lin.SYG.Ratio(y1[s==1], x[s==1], pik.U[s==1], pikl.s)
#Computes the var. est. of the ratio point estimator using y2
VE.Lin.SYG.Ratio(y2[s==1], x[s==1], pik.U[s==1], pikl.s)
```

VE.SYG.Mean.NHT — *The Sen-Yates-Grundy variance estimator for the Narain-Horvitz-Thompson point estimator for a mean*

Description

Computes the Sen (1953); Yates-Grundy(1953) variance estimator for the Narain (1951); Horvitz-Thompson (1952) point estimator for a population mean.

Usage

```
VE.SYG.Mean.NHT(VecY.s, VecPk.s, MatPkl.s, N)
```

Arguments

`VecY.s`	vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s`. There must not be missing values.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.
`MatPkl.s`	matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in `MatPkl.s` must be greater than zero and less than or equal to one. There must not be missing values.
`N`	the population size. It must be an integer or a double-precision scalar with zero-valued fractional part.

Details

For the population mean of the variable y:

$$\bar{y} = \frac{1}{N}\sum_{k\in U} y_k$$

the unbiased Narain (1951); Horvitz-Thompson (1952) estimator of $\bar{y}$ is given by:

$$\hat{\bar{y}}_{NHT} = \frac{1}{N}\sum_{k\in s}\frac{y_k}{\pi_k}$$

where π_k denotes the inclusion probability of the k-th element in the sample s. Let π_{kl} denotes the joint-inclusion probabilities of the k-th and l-th elements in the sample s. The variance of $\hat{\bar{y}}_{NHT}$ is given by:

$$V(\hat{\bar{y}}_{NHT}) = \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}(\pi_{kl}-\pi_k\pi_l)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}$$

which, if the utilised sampling design is of fixed-size, can therefore be estimated by the Sen-Yates-Grundy variance estimator (implemented by the current function):

$$\hat{V}(\hat{\bar{y}}_{NHT}) = \frac{1}{N^2}\frac{-1}{2}\sum_{k\in s}\sum_{l\in s}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}}\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2$$

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Narain, R. D. (1951) On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **3**, 169–175.

Sen, A. R. (1953) On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **5**, 119–127.

Yates, F. and Grundy, P. M. (1953) Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B*, **15**, 253–261.

See Also

VE.HT.Mean.NHT
VE.Hajek.Mean.NHT

Examples

```
data(oaxaca)                                 #Loads the Oaxaca municipalities dataset
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s      <- oaxaca$sHOMES00                    #Defines the sample to be used
N      <- dim(oaxaca)[1]                     #Defines the population size
y1     <- oaxaca$POP10                       #Defines the variable of interest y1
y2     <- oaxaca$HOMES10                     #Defines the variable of interest y2
#This approx. is only suitable for large-entropy sampling designs
pikl.s <- Pkl.Hajek.s(pik.U[s==1])           #Approx. 2nd order incl. probs. from s
#Computes the var. est. of the NHT point estimator for y1
VE.SYG.Mean.NHT(y1[s==1], pik.U[s==1], pikl.s, N)
#Computes the var. est. of the NHT point estimator for y2
VE.SYG.Mean.NHT(y2[s==1], pik.U[s==1], pikl.s, N)
```

VE.SYG.Total.NHT — *The Sen-Yates-Grundy variance estimator for the Narain-Horvitz-Thompson point estimator for a total*

Description

Computes the Sen (1953); Yates-Grundy(1953) variance estimator for the Narain (1951); Horvitz-Thompson (1952) point estimator for a population total.

Usage

```
VE.SYG.Total.NHT(VecY.s, VecPk.s, MatPkl.s)
```

Arguments

- `VecY.s` vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of `VecPk.s`. There must not be missing values.
- `VecPk.s` vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in `VecPk.s` must be greater than zero and less than or equal to one. There must not be missing values.
- `MatPkl.s` matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in `MatPkl.s` must be greater than zero and less than or equal to one. There must not be missing values.

Details

For the population total of the variable y:

$$t = \sum_{k \in U} y_k$$

the unbiased Narain (1951); Horvitz-Thompson (1952) estimator of t is given by:

$$\hat{t}_{NHT} = \sum_{k \in s} \frac{y_k}{\pi_k}$$

where π_k denotes the inclusion probability of the k-th element in the sample s. Let π_{kl} denotes the joint-inclusion probabilities of the k-th and l-th elements in the sample s. The variance of $\hat{t}_{NHT}$ is given by:

$$V(\hat{t}_{NHT}) = \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \pi_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

which, if the utilised sampling design is of fixed-size, can therefore be estimated by the Sen-Yates-Grundy variance estimator (implemented by the current function):

$$\hat{V}(\hat{t}_{NHT}) = \frac{-1}{2} \sum_{k \in s} \sum_{l \in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2$$

Value

The function returns a value for the estimated variance.

Author(s)

Emilio Lopez Escobar.

References

Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Narain, R. D. (1951) On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **3**, 169–175.

Sen, A. R. (1953) On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **5**, 119–127.

Yates, F. and Grundy, P. M. (1953) Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B*, **15**, 253–261.

See Also

VE.HT.Total.NHT
VE.Hajek.Total.NHT

Examples

```
data(oaxaca)                                 #Loads the Oaxaca municipalities dataset
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00) #Reconstructs the 1st order incl. probs.
s      <- oaxaca$sHOMES00                    #Defines the sample to be used
y1     <- oaxaca$POP10                       #Defines the variable of interest y1
y2     <- oaxaca$HOMES10                     #Defines the variable of interest y2
#This approximation is only suitable for large-entropy sampling designs
pikl.s <- Pkl.Hajek.s(pik.U[s==1])           #Approx. 2nd order incl. probs. from s
#Computes the var. est. of the NHT point estimator for y1
VE.SYG.Total.NHT(y1[s==1], pik.U[s==1], pikl.s)
#Computes the var. est. of the NHT point estimator for y2
VE.SYG.Total.NHT(y2[s==1], pik.U[s==1], pikl.s)
```

Index