Package 'StratifiedSampling'

January 26, 2021

Type Package

Title Different Methods for Stratified Sampling

Version 0.1.0

Description Integrating a stratified structure in the population in a sampling design can considerably reduce the variance of the Horvitz-Thompson estimator. We propose in this package different methods to handle the selection of a balanced sample in stratified population. For more details see Raphaël Jauslin, Esther Eustache and Yves Tillé (2021) <arXiv:2101.05568>.

License GPL (>= 2)

Encoding UTF-8

LazyData true

LinkingTo RcppArmadillo, Rcpp

Imports Rcpp

Depends Matrix, R (>= 3.5.0)

Suggests knitr, rmarkdown, ggplot2, sampling, BalancedSampling, MASS, stats, testthat

RoxygenNote 7.1.1

NeedsCompilation yes

Author Raphael Jauslin [aut, cre] (<https://orcid.org/0000-0003-1088-3356>), Esther Eustache [aut], Yves Tillé [aut] (<https://orcid.org/0000-0003-0904-5523>)

Maintainer Raphael Jauslin <raphael.jauslin@unine.ch>

Repository CRAN

Date/Publication 2021-01-26 12:20:06 UTC

R topics documented:

`balstrat` *Balanced Stratification*

Description

Select a stratified balanced sample. The function is similar to `balancedstratification` of the package sampling.

Usage

```
balstrat(X, strata, pik)
```

Arguments

`X`	A matrix of size (N x p) of auxiliary variables on which the sample must be balanced.
`strata`	A vector of integers that specifies the stratification.
`pik`	A vector of inclusion probabilities.

Details

The function implements the method proposed by Chauvet (2009). Firstly, a flight phase is performed on each strata. Secondly, a flight phase is applied on the whole population by aggregating the strata. Finally, a landing phase is applied by suppression of variables.

Value

A vector with elements equal to 0 or 1. The value 1 indicates that the unit is selected while the value 0 is for rejected units.

Author(s)

Raphaël Jauslin <raphael.jauslin@unine.ch>

References

Chauvet, G. (2009). Stratified balanced sampling. *Survey Methodology*, 35:115-119.

See Also

ffphase,landingRM

Examples

```
N <- 100
n <- 10
p <- 4
X <- matrix(rgamma(N*p,4,25),ncol = p)
strata <- as.matrix(rep(1:n,each = N/n))
pik <- rep(n/N,N)

s <- balstrat(X,strata,pik)

t(X/pik)%*%s
t(X/pik)%*%pik

Xcat <- disj(strata)

t(Xcat)%*%s
t(Xcat)%*%pik
```

---

disj *Disjunctive*

Description

This function transforms a categorical vector into a matrix of indicators.

Usage

```
disj(strata)
```

Arguments

strata	A vector of integers that represents the categories.

Value

A matrix of indicators.

Author(s)

Raphaël Jauslin <raphael.jauslin@unine.ch>

Examples

```
strata <- rep(c(1,2,3),each = 4)
disj(strata)
```

`disjMatrix` *Disjunctive for matrix*

Description

This function transforms a categorical matrix into a matrix of indicators variables.

Usage

```
disjMatrix(strata)
```

Arguments

`strata`	A matrix of integers that contains categorical vector in each column.

Value

A matrix of indicators.

Author(s)

Raphaël Jauslin <raphael.jauslin@unine.ch>

Examples

```
Xcat <-  matrix(c(sample(x = 1:6, size = 100, replace = TRUE),
            sample(x = 1:6, size = 100, replace = TRUE),
            sample(x = 1:6, size = 100, replace = TRUE)),ncol = 3)
disjMatrix(Xcat)
```

fbs *Fast Balanced Sampling*

Description

This function implements the method proposed by Hasler and Tillé (2014). It should be used for selecting a sample from highly stratified population.

Usage

```
fbs(X, strata, pik)
```

Arguments

X	A matrix of size (N x p) of auxiliary variables on which the sample must be balanced.
strata	A vector of integers that specifies the stratification.
pik	A vector of inclusion probabilities.

Details

Firstly a flight phase is performed on each strata. Secondly, several flight phases are applied by adding one by one the stratum. By doing this, some strata are managed on-the-fly. Finally, a landing phase is applied by suppression of the variables. If the number of element selected in each stratum is not equal to an integer, the function can be very time-consuming.

Value

A vector with elements equal to 0 or 1. The value 1 indicates that the unit is selected while the value 0 is for rejected units.

Author(s)

Raphaël Jauslin <raphael.jauslin@unine.ch>

References

Hasler, C. and Tillé Y. (2014). Fast balanced sampling for highly stratified population. *Computational Statistics and Data Analysis*, 74, 81-94

Examples

```
N <- 100
n <- 10
x1 <- rgamma(N,4,25)
x2 <- rgamma(N,4,25)
```

```
strata <- rep(1:n,each = N/n)

pik <- rep(n/N,N)
X <- as.matrix(cbind(matrix(c(x1,x2),ncol = 2)))

s <- fbs(X,strata,pik)

t(X/pik)%*%s
t(X/pik)%*%pik

Xcat <- disj(strata)

t(Xcat)%*%s
t(Xcat)%*%pik
```

`ffphase` *Fast flight phase of the cube method*

Description

This function computes the flight phase of the cube method proposed by Chauvet and Tillé (2006).

Usage

```
ffphase(X, pik)
```

Arguments

`X`	A matrix of size (N x p) of auxiliary variables on which the sample must be balanced.
`pik`	A vector of inclusion probabilities.

Details

This function implements the method proposed by (Chauvet and Tillé 2006). It recursively transforms the vector of inclusion probabilities `pik` into a sample that respects the balancing equations. The algorithm stops when the null space of the sub-matrix B is empty. For more information see (Chauvet and Tillé 2006).

The function uses the function `Null` to find the null space of the sub-matrix B.

Value

Updated vector of `pik` that contains 0 and 1 for unit that are rejected or selected.

Author(s)

Raphaël Jauslin <raphael.jauslin@unine.ch>

References

Chauvet, G. and Tillé, Y. (2006). A fast algorithm of balanced sampling. *Computational Statistics*, 21/1:53-62

See Also

fastflightphase, flightphase.

Examples

```
N <- 100
n <- 10
p <- 4

pik <- rep(n/N,N)
X <- cbind(pik,matrix(rgamma(N*p,4,25),ncol= p))

pikstar <- ffphase(X,pik)
t(X/pik)%*%pikstar
t(X/pik)%*%pik
pikstar
```

---

findB *Find best sub-matrix B in stratifiedcube*

Description

This function is computing a sub-matrix used in stratifiedcube.

Usage

```
findB(X, strata)
```

Arguments

X	A matrix of size (N x p) of auxiliary variables on which the sample must be balanced.
strata	A vector of integers that specifies the stratification.

Details

The function finds the smallest matrix B such that it contains only one more row than the number of columns. It consecutively adds the right number of rows depending on the number of categories that is added.

Value

A list of two components. The sub-matrix of `X` and the corresponding disjunctive matrix. If we use the function `cbind` to combine the two matrices, the resulting matrix has only one more row than the number of columns.

Author(s)

Raphaël Jauslin <raphael.jauslin@unine.ch>

Examples

```
N <- 1000
strata <-  sample(x = 1:6, size = N, replace = TRUE)

p <- 3
X <- matrix(rnorm(N*p),ncol = 3)
findB(X,strata)
```

`landingRM` *Landing by suppression of variables*

Description

This function performs the landing phase of the cube method using suppression of variables proposed by Chauvet and Tillé (2006).

Usage

```
landingRM(X, pikstar)
```

Arguments

`X`	matrix of auxiliary variables on which the sample must be balanced. (The matrix should be divided by the original inclusion probabilities.)
`pikstar`	vector of updated inclusion probabilities by the flight phase. See `ffphase`

Value

A vector with elements equal to 0 or 1. The value 1 indicates that the unit is selected while the value 0 is for rejected units.

Author(s)

Raphaël Jauslin <raphael.jauslin@unine.ch>

References

Chauvet, G. and Tillé, Y. (2006). A fast algorithm of balanced sampling. *Computational Statistics*, 21/1:53-62

See Also

fbs, balstrat.

Examples

```
N <- 1000
n <- 10
p <- 4
pik <- rep(n/N,N)
X <- cbind(pik,matrix(rgamma(N*p,4,25),ncol= p))
pikstar <- ffphase(X,pik)
s <- landingRM(X/pik,pikstar)
sum(s)
t(X/pik)%*%pik
t(X/pik)%*%pikstar
t(X/pik)%*%s
```

---

ncat *Number of categories*

Description

This function returns the number of factor in each column of a categorical matrix.

Usage

```
ncat(Xcat)
```

Arguments

Xcat	A matrix of integers that contains categorical vector in each column.

Value

A row vector that contains the number of categories in each column.

Author(s)

Raphaël Jauslin <raphael.jauslin@unine.ch>

Examples

```
Xcat <-  matrix(c(sample(x = 1:6, size = 100, replace = TRUE),
            sample(x = 1:6, size = 100, replace = TRUE),
            sample(x = 1:6, size = 100, replace = TRUE)),ncol = 3)
ncat(Xcat)
```

`stratifiedcube` *Stratified Sampling*

Description

This function implements a method for selecting a stratified sample. It really improves the performance of the function `fbs` and `balstrat`.

Usage

```
stratifiedcube(X, strata, pik)
```

Arguments

`X`	A matrix of size (N x p) of auxiliary variables on which the sample must be balanced.
`strata`	A vector of integers that specifies the stratification..
`pik`	A vector of inclusion probabilities.

Details

The function is selecting a balanced sample very quickly even if the sum of inclusion probabilities within strata are non-integer. The function should be used in preference. Firstly, a flight phase is performed on each strata. Secondly, the function `findB` is used to find a particular matrix to apply a flight phase by using the cube method proposed by Chauvet, G. and Tillé, Y. (2006). Finally, a landing phase is applied by suppression of variables.

Value

A vector with elements equal to 0 or 1. The value 1 indicates that the unit is selected while the value 0 is for rejected units.

References

Chauvet, G. and Tillé, Y. (2006). A fast algorithm of balanced sampling. *Computational Statistics*, 21/1:53-62

See Also

`fbs`, `balstrat`, `landingRM`, `ffphase`

Examples

```
N <- 100
n <- 10
p <- 4
X <- matrix(rgamma(N*p,4,25),ncol = p)
strata <- as.matrix(rep(1:n,each = N/n))
pik <- rep(n/N,N)

s <- stratifiedcube(X,strata,pik)

t(X/pik)%*%s
t(X/pik)%*%pik

Xcat <- disj(strata)

t(Xcat)%*%s
t(Xcat)%*%pik
```

---

varApp *Approximated variance for balanced sampling*

Description

Approximated variance for balanced sampling

Usage

```
varApp(X, strata, pik, y)
```

Arguments

X	A matrix of size (N x p) of auxiliary variables on which the sample must be balanced.
strata	A vector of integers that represents the categories.
pik	A vector of inclusion probabilities.
y	A variable of interest.

Details

This function gives an approximation of the variance of the Horvitz-Thompson total estimator presented by Hasler and Tillé (2014).

Value

a scalar, the value of the approximated variance.

Author(s)

Raphaël Jauslin <raphael.jauslin@unine.ch>

References

Hasler, C. and Tillé, Y. (2014). Fast balanced sampling for highly stratified population. *Computational Statistics and Data Analysis*, 74:81-94.

See Also

varEst

Examples

```
N <- 1000
n <- 400
x1 <- rgamma(N,4,25)
x2 <- rgamma(N,4,25)

strata <- as.matrix(rep(1:40,each = 25)) # 25 strata
Xcat <- disjMatrix(strata)
pik <- rep(n/N,N)
X <- as.matrix(matrix(c(x1,x2),ncol = 2))

s <- stratifiedcube(X,strata,pik)

y <- 20*strata + rnorm(1000,120) # variable of interest
# y_ht <- sum(y[which(s==1)]/pik[which(s == 1)]) # Horvitz-Thompson estimator
# (sum(y_ht) - sum(y))^2 # true variance
varEst(X,strata,pik,s,y)
varApp(X,strata,pik,y)
```

---

varEst *Estimator of the approximated variance for balanced sampling*

Description

Estimator of the approximated variance for balanced sampling

Usage

```
varEst(X, strata, pik, s, y)
```

Arguments

X	A matrix of size (N x p) of auxiliary variables on which the sample must be balanced.
strata	A vector of integers that represents the categories.
pik	A vector of inclusion probabilities.
s	A sample (vector of 0 and 1, if rejected or selected).
y	A variable of interest.

Details

This function gives an estimator of the approximated variance of the Horvitz-Thompson total estimator presented by Hasler C. and Tillé Y. (2014).

Value

a scalar, the value of the estimated variance.

Author(s)

Raphaël Jauslin <raphael.jauslin@unine.ch>

References

Hasler, C. and Tillé, Y. (2014). Fast balanced sampling for highly stratified population. *Computational Statistics and Data Analysis*, 74:81-94.

See Also

varApp

Examples

```
N <- 1000
n <- 400
x1 <- rgamma(N,4,25)
x2 <- rgamma(N,4,25)

strata <- as.matrix(rep(1:40,each = 25)) # 25 strata
Xcat <- disjMatrix(strata)
pik <- rep(n/N,N)
X <- as.matrix(matrix(c(x1,x2),ncol = 2))

s <- stratifiedcube(X,strata,pik)

y <- 20*strata + rnorm(1000,120) # variable of interest
# y_ht <- sum(y[which(s==1)]/pik[which(s == 1)]) # Horvitz-Thompson estimator
# (sum(y_ht) - sum(y))^2 # true variance
varEst(X,strata,pik,s,y)
varApp(X,strata,pik,y)
```

Index