Package 'optimStrat'

September 10, 2018

Type Package

Title Choosing the Sample Strategy

Version 1.1

Date 2018-09-04

Author Edgar Bueno <edgar.bueno@stat.su.se>

Maintainer Edgar Bueno <edgar.bueno@stat.su.se>

Depends shiny

Description Intended to assist in the choice of the sampling strategy to implement in a survey. It compares five strategies having into account the information available in an auxiliary variable and two superpopulation models, called working and true models.

License GPL-2

NeedsCompilation no

Repository CRAN

Date/Publication 2018-09-10 20:20:10 UTC

R topics documented:

optimStrat-package *optimStrat*

Description

OptimStrat is a package intended to assist in the choice of the sample strategy to implement in a survey. It compares five strategies having into account the information available in an auxiliary variable and two superpopulation models, called working and true models.

Details

The package includes a web-based application where the user can compare five sampling strategies in order to determine which one to implement in a survey.

The package also includes a function to perform the comparison mentioned above, as well as functions for stratifying the auxiliary variable and calculations of the variance of Stratified Simple Random Sampling and Pareto πps.

Author(s)

Edgar Bueno

References

Bueno, E. (2018). *A Comparison of Stratified Simple Random Sampling and Probability Proporional-to-size Sampling.* Research Report, Department of Statistics, Stockholm University 2018:6. http://gauss.stat.su.se/rr/RR2018_6.pdf.

absdif *Absolute Difference*

Description

Compute the absolute differences between `x` and `y`.

Usage

```
absdif(x, y)
```

Arguments

`x`	a numeric vector, matrix or data frame.
`y`	a numeric vector, matrix or data.frame.

Details

Compute the absolute differences between x and y componentwise.

If x and y are vectors of different length, the elements of the shortest one will be recycled as necessary. If matrices or data frames, they should be of the same dimension.

Value

An object with the absolute differences between x and y.

Examples

```
absdif(1:10, 10:1)

x<- matrix(1:12, 4, 3)
y<- matrix(12:1, 4, 3)
absdif(x, y)
```

---

covp *Covariance*

Description

Compute the covariance between x and y.

Usage

```
covp(x, y)
```

Arguments

x	a numeric vector.
y	a numeric vector.

Details

Compute the covariance between x and y using n (instead of $n-1$ as in cov) in the denominator.

If the length of x and y are different, the elements of the shortest one will be recycled as necessary.

Value

An object with the covariance between x and y.

See Also

cov

Examples

```
x<- rnorm(100)
y<- rnorm(100)
covp(x, y)
cov(x, y)
```

`optimApp` *Interactive Web-based Application of optimStrat*

Description

Call Shiny to run a web-based application of `optimStrat`.

Usage

```
optimApp()
```

Author(s)

Edgar Bueno, <edgar.bueno@stat.su.se>

`simulatey` *Simulate the Study Variable*

Description

Simulate values for the study variable based on the auxiliary variable `x` and the parameters of a superpopulation model.

Usage

```
simulatey(x, b0, b1, b2, b4, rho=NULL, b3=NULL)
```

Arguments

`x`	a positive numeric vector giving the values of the auxiliary variable.
`b0`	a number giving the intercept of the trend term in the superpopulation model.
`b1`	a number giving the scale of the trend term in the superpopulation model.
`b2`	a number giving the shape of the trend term in the superpopulation model.
`b4`	a number giving the shape of the spread term in the superpopulation model.
`rho`	a number giving the absolute value of the desired correlation between `x` and the vector to be simulated.
`b3`	a nonnegative number giving the scale of the spread term in the superpopulation model. Ignored if `rho` is given (see 'Details').

Details

The values of the study variable y are simulated using a superpopulation model defined as follows:

$$Y_k = \beta_0 + \beta_1 x_k^{\beta_2} + \epsilon_k$$

with ϵ_k $N(0, \beta_3 x_k^{\beta_4})$.

Note that b3 defines the degree of association between x and y: the larger b3, the smaller the correlation, rho, and vice versa. For this reason only one of them should be defined. If both are defined, b3 will be ignored.

The sign of the correlation should be given through b1 (see 'Examples').

Depending on the value of b2, some correlations cannot be reached, e.g. if b2=2 it is pointless to set rho=1. In those cases, b3 will automatically be set to zero and rho will be ignored (see 'Examples').

Value

A numeric vector giving the simulated value of y associated to each value in x.

Examples

```
#Linear trend and homocedasticity
x<- 1 + sort( rgamma(5000, shape=4/9, scale=108) )
y<- simulatey(x, b0=0, b1=1, b2=1, b4=0, rho=0.90)
plot(x, y)

#Linear trend and heterocedasticity
y<- simulatey(x, b0=0, b1=1, b2=1, b4=1, rho=0.90)
plot(x, y)

#Quadratic trend and homocedasticity
y<- simulatey(x, b0=0, b1=1, b2=2, b4=0, rho=0.80)
plot(x, y)

#Correlation of minus one
y<- simulatey(x, b0=0, b1=-1, b2=1, b4=0, rho=1)
cor(x, y)
plot(x, y)

#Desired correlation cannot be attained
y<- simulatey(x, b0=0, b1=1, b2=3, b4=0, rho=0.99)
cor(x, y)
plot(x, y)
```

---

skewness *Sample Skewness*

Description

Calculate the sample skewness.

Usage

```
skewness(x, na.rm = FALSE)
```

Arguments

x	a numeric vector.
na.rm	a logical value indicating whether NA values should be stripped before the computation proceeds.

Details

Compute the sample skewness of x as

$$\frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^3}{\left[\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2\right]^{3/2}}$$

Value

A vector of length one giving the sample skewness of x.

Examples

```
x<- rnorm(1000)
skewness(x)
```

---

stratify *Stratification of an Auxiliary Variable*

Description

Stratify the auxiliary variable x into H strata using the cum-sqrt-rule.

Usage

```
stratify(x, H, forced = FALSE, J = NULL)
```

Arguments

x	a positive numeric vector giving the values of the auxiliary variable.
H	a positive integer smaller or equal than length(x) giving the desired number of strata.
forced	a logical value indicating if the number of strata *must* be exactly equal to H (see 'Details').
J	a positive integer indicating the number of bins used for the cum-sqrt-rule.

Details

The cum-sqrt-rule is used in order to define `H` strata from the auxiliary vector `x`.

Depending on some characteristics of `x`, e.g. high skewness, few observations or too many ties, the resulting stratification may have a number of strata other than `H`. Using `forced = TRUE` tries its best to obtain exactly `H` strata (see 'Examples').

Note that if `length(x) < H` then `forced` will be set to `FALSE`.

Value

A numeric vector giving the stratum to which each observation in `x` belongs.

References

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

See Also

`varstsi` for computing the variance of Stratified Simple Random Sample.

Examples

```
x<- 1 + sort( rgamma(100, shape=4/9, scale=108) )
stratify(x, H=3)

set.seed(1280)
x<- 1 + sort( rgamma(100, shape=4/100, scale=1200) )
stratify(x, H=4) #Only three strata
stratify(x, H=4, forced=TRUE) #Four strata
```

`stratvar` *Compute the Variance of Five Sampling Strategies*

Description

Simulate the values of a study variable using the auxiliary variable `x` and then compute the design variance of five sampling strategies: πps–reg, STSI–reg, STSI–HT, πps–pos and STSI–pos. The process is iterated `it` times.

Usage

```
stratvar(x, sk = 3, n, H, d2, d4, b2 = d2, b4 = d4, b0 = 0,
         b1 = 1, rho = NULL, b3 = NULL, it = 1)
```

Arguments

x	a positive integer or numeric vector. If an integer, indicates the desired size of the auxiliary variable to simulate. If a vector, gives the values of the auxiliary variable itself.
sk	if x is an integer, indicates the desired skewness of the auxiliary variable to simulate. Ignored otherwise.
n	a positive integer giving the desired sample size.
H	a positive integer giving the desired number of strata/poststrata.
d2	a number giving the *assumed* shape of the trend term in the superpopulation model.
d4	a number giving the *assumed* shape of the spread term in the superpopulation model.
b2	a number giving the shape of the trend term in the superpopulation model.
b4	a number giving the shape of the spread term in the superpopulation model.
b0	a number giving the intercept of the trend term in the superpopulation model.
b1	a number giving the scale of the trend term in the superpopulation model.
rho	a number giving the absolute value of the desired correlation between x and the vector to be simulated.
b3	a nonnegative number giving the scale of the spread term in the superpopulation model. Ignored if rho is given (see 'Details').
it	a positive integer indicating the number of times to iterate the process.

Details

This function allows to study the impact that assuming a misspecified model has on the design variance of five sampling strategies.

If x is a positive integer, the values of an auxiliary variable are simulated as realizations from a gamma distribution with mean 48 and skewness equal to sk, plus one unit. If x is a vector, it is the auxiliary variable itself.

With this auxiliary information, values for the study variable y are simulated using the superpopulation model via simulatey.

The variance of a sample of size n is then computed for five sampling strategies (πps–reg, STSI–reg, STSI–HT, πps–pos and STSI–pos) assuming that the right model has δ_2 instead of β_2 and δ_4 instead of β_4.

The number of strata/poststrata is given by H.

The process is iterated it times.

Value

A matrix of size it x 17. Each row being the results obtained for each iteration. The first eleven columns are the input arguments (with the sample skewness instead of sk) followed by the sample correlation between x and y. The last five columns give the design variance of the five strategies under comparison.

References

Bueno, E. (2018). *A Comparison of Stratified Simple Random Sampling and Probability Proporional-to-size Sampling.* Research Report, Department of Statistics, Stockholm University 2018:6. http://gauss.stat.su.se/rr/RR2018_6.pdf.

See Also

simulatey for the simulation of the y values; stratify for how to define the strata/poststrata boundaries; varstsi for how the sample size is allocated into the strata.

Examples

```
#The assumed model coincides with the true generating model
stratvar(5000, sk = 3, n=100, H=5, d2=1, d4=1, rho=0.8, it=10)

#The assumed model differs with the true generating model
x<- 1 + sort( rgamma(5000, shape=4/9, scale=108) )
stratvar(x, n=100, H=5, d2=1, d4=1, b2=1.5, b4=0.5, rho=0.8, it=10)
```

---

varp *Variance*

Description

Compute the variance of x.

Usage

```
varp(x)
```

Arguments

x	a numeric vector.

Details

Compute the variance of x using n (instead of $n-1$ as in var) in the denominator.

Value

An object with the variance of x.

See Also

var

Examples

```
x<- rnorm(100)
varp(x)
var(x)
```

`varpips` *Variance of Pareto PIps Sampling with the HT Estimator*

Description

Compute the design variance of the Horvitz-Thompson estimator of the total of `y` under Pareto probability proportional-to-size Sampling, where the size variable is indicated by `x` and the sample size is `n`.

Usage

```
varpips(n, x, y)
```

Arguments

`n`	a positive integer indicating the desired sample size.
`x`	a positive numeric vector giving the values of the auxiliary variable that is used in order to define the desired inclusion probabilities.
`y`	a numeric vector giving the values of the study variable.

Details

Target inclusion probabilities are computed as $\pi_k = n \cdot x_k / \sum x_k$.

If $\pi_k > 1$ for at least one element, π_k is set equal to one for those elements and the inclusion probabilities are calculated again for the remaining elements with the remaining sample size.

Once the π_k are obtained, the variance of the Horvitz-Thompson estimator under Pareto probability proportional-to-size Sampling is computed as: $V_{\pi ps}\left[\hat{t}_{HT}\right] = \frac{N}{N-1}(t_1 - \frac{t_2^2}{t_3})$ with

$$t_1 = \sum \frac{y_k^2(1-\pi_k)}{\pi_k}$$

$$t_2 = \sum y_k(1-\pi_k)$$

$$t_3 = \sum \pi_k(1-\pi_k)$$

Value

A list containing the following:

`variance`	a vector of length one giving the variance of the Horvitz-Thompson estimator under Pareto probability proportional-to-size Sampling.
`pinc`	a vector with length `length(x)` giving the target inclusion probabilities of each element..

References

Rosen, B. (1997). *On Sampling with Probability Proportional to Size.* Journal of Statistical Planning and Inference **62**, 159-191.

Examples

```
x<- 1 + sort( rgamma(5000, shape=4/9, scale=108) ) #simulating the auxiliary variable
y<- rgamma(x, shape=1, scale=x) #simulating the study variable
z<- varpips(n=500, x=x, y=y)
z$variance
```

`varstsi` *Variance of STSI Sampling with the HT Estimator*

Description

Compute the design variance of the Horvitz-Thompson estimator of the total of `y` under Stratified Simple Random Sampling, where strata are indicated by `stratum` and the sample of size `n` is allocated using Neyman allocation with respect to `x`.

Usage

```
varstsi(n, x, y = x, stratum)
```

Arguments

`n`	a positive integer indicating the desired sample size.
`x`	a positive numeric vector giving the values of the auxiliary variable that is used in order to allocate the sample size into the strata.
`y`	a numeric vector giving the values of the study variable. By default `y = x`.
`stratum`	a vector indicating the stratum to which each element belongs.

Details

A sample of size n is allocated into the strata using x-optimal allocation, i.e.

$$n_h \propto N_h S_{x,U_h}$$

where N_h is the size of the hth stratum, S_{x,U_h} is the standard deviation of `x` in the hth stratum and $propto$ stands for 'proportional to'.

If $n_h > N_h$ for at least one stratum, n_h is set equal to N_h in those strata and optimal allocation is used again for the remaining strata with the remaining sample size.

Once the n_h are obtained, the variance of the Horvitz-Thompson estimator under Stratified Simple Random Sampling is computed as: $V_{STSI}\left[\hat{t}_{HT}\right] = \sum_h V_h$ with

$$V_h = \frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right) S_{y,U_h}^2$$

where S^2_{y,U_h} is the variance of y in the hth stratum.

Proporional allocation is obtained if x is a constant.

The variance of Simple Random Sampling is computed if `stratum` is a constant.

Value

A list containing the following:

`variance`	a vector of length one giving the variance of the Horvitz-Thompson estimator under Stratified Simple Random Sampling.
`nh`	a vector with length equal to the number of strata giving the size of the sample in each stratum.

References

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

See Also

`stratify` for a method to define the strata.

Examples

```
x<- 1 + sort( rgamma(5000, shape=4/9, scale=108) ) #simulating the auxiliary variable
y<- rgamma(x, shape=1, scale=x) #simulating the study variable

st1<- rep(1:5, each=1000) #defining the strata
z1<- varstsi(n=500, x=x, y=y, stratum=st1)
z1$variance

st2<- stratify(x, H=5) #A better way to stratify
z2<- varstsi(n=500, x=x, y=y, stratum=st2)
z2$variance
```

Index