Package ‘bootsurv’

May 17, 2024

Title Bootstrap Methods for Complete Survey Data

Version 0.0.1

Depends R (>= 3.5.0)

Description Bootstrap resampling methods have been widely studied in the context of survey data. This package implements various bootstrap resampling techniques tailored for survey data, with a focus on stratified simple random sampling and stratified two-stage cluster sampling. It provides tools for precise and consistent bootstrap variance estimation for population totals, means, and quartiles. Additionally, it enables easy generation of bootstrap samples for in-depth analysis.

License GPL-3

Imports MASS, stats

Encoding UTF-8

RoxygenNote 7.3.1

LazyData true

NeedsCompilation no

Author Zeinab Mashreghi [aut, cre] (<`https://orcid.org/0000-0003-0563-0792`>)

Maintainer Zeinab Mashreghi <`z.mashreghi@uwinnipeg.ca`>

Repository CRAN

Date/Publication 2024-05-17 09:40:05 UTC

R topics documented:

`boot.twostage` *Bootstrap methods for two-stage sampling designs*

Description

The function `boot.twostage` applies one of the following bootstrap methods on complete (full response) survey data selected under stratified two-stage cluster sampling SRSWOR/SRSWOR: Rao and Wu (1988), Rao, Wu and Yue (1992), the modified version of Sitter (1992, CJS) (see Chen, Haziza and Mashreghi, 2022), Funaoka, Saigo, Sitter and Toida (2006), Chauvet (2007) or Preston (2009). This function also applies the method of Rao, Wu and Yue (1992) on complete survey data selected under stratified two-stage cluster sampling IPPSWOR/SRSWOR or the method of Chauvet (2007) on complete survey data selected under stratified two-stage cluster sampling CPS/SRSWOR.

Usage

```
boot.twostage(
  data,
  no.cluster,
  cluster.size,
  R,
  parameter = "total",
  bootstrap.method = "Rao.Wu.Yue",
  survey.design = "SRSWOR",
  population.size = NULL,
  boot.sample.size = NULL
)
```

Arguments

`data`	A vector, matrix or data frame. The column of study variable has to be a numeric column named `study.variable` and a column identifying clusters named `cluster` has to be included. If the population is stratified, a column identifying strata named `stratum` has to be included. If an IPPS design is applied on the first stage a column of first stage inclusion probability named `Pi1` has to be included.
`no.cluster`	A vector of the number of clusters within strata.
`cluster.size`	The number of elements within the selected clusters within each stratum. The length of this vector must be the same as the number of all selected clusters in all strata.
`R`	The number of bootstrap replicates. For the Chauvet (2007) method, `R` is a vector with two values: `(R.pop, R.samp)` representing the number of pseudo-populations and the number of bootstrap samples drawn from each pseudo-population.
`parameter`	One of the following population parameters can be applied: `"total"` (population total), `"mean"` (population mean), `"quartile.25"` (population 1st quartile),

"`quartile.50`" or "`median`" (population median) or "`quartile.75`" (population 3rd quartile). If the parameter of interest is the population mean or total, the HT-estimator is applied. If the parameter of interest is a population quartile, the estimator in Sarndal, Swensson, and Wretman (1992, Chapter 5) is applied. The default is the population total.

`bootstrap.method`

One of the following bootstrap methods can be applied in the case of statratified SRS/SRS: "`Rao.Wu`" (Rao and Wu, 1988), "`Rao.Wu.Yue`" (Rao, Wu and Yue, 1992), "`Modified.Sitter`" (the modified version of Sitter 1992 discussed in Chen, Haziza and Mashreghi, 2022), "`Funaoka.etal`" (Funaoka, Saigo, Sitter and Toida, 2006), "`Chauvet`" (Chauvet, 2007) or "`Preston`" (Preston, 2009).

`survey.design` It can be either "`IPPS`" only if the method of Rao, Wu and Yue (1992) is applied or "`CPS`" only if the method of Chauvet (2007) is applied or "`SRSWOR`". The default is "`SRSWOR`".

`population.size`

A vector of stratum population sizes.

`boot.sample.size`

A vector of bootstrap sample sizes within strata. The bootstrap sample size is required only for the method of Rao, Wu and Yue (1988). If it is not specified, the bootstrap sample size will be `nh-1` within each stratum, where `nh` is the original sample size within stratum `h`.

Value

`boot.statistic` A vector of bootstrap statistics of size R.

`boot.var` The bootstrap variance estimator of the estimator of parameter of interest.

`boot.mean` The average of the bootstrap estimator of the parameter of interest.

`boot.sample` A list of results for each iteration. That includes a column of original sample values, a column of cluster identifier and a column of stratum identifier. More information is availble depending on the bootstrap method.

References

Chauvet, G. (2007). Méthodes de bootstrap en population finie. PhD thesis, École Nationale de Statistique et Analyse de l'Information, Bruz, France.

Chen, S., Haziza, D. and Mashreghi, Z., (2022). A Comparison of Existing Bootstrap Algorithms for Multi-Stage Sampling Designs. Stats, 5(2), pp.521-537.

Funaoka, F., Saigo, H., Sitter, R.R., Toida, T. (2006). Bernoulli bootstrap for stratified multistage sampling. Survey Methodology, 32, 151–156.

Rao, J.N.K., Wu, C.F.J. (1998). Resampling inference with complex survey data. Journal of the American Statistical Association, 83, 231–241.

Rao, J.N.K., Wu, C.F.J., Yue, K. (1992). Some recent work on resampling methods for complex surveys. Survey Methodology, 18, 209–217.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). Model-Assisted Survey Sampling. NewYork: Springer.

Sitter, R.R. (1992). Comparing three bootstrap methods for survey data. The Canadian Journal of Statistics, 20, 135–154.

Preston, J. (2009). Rescaled bootstrap for stratified multistage sampling. Survey Methodology, 35, 227–234.

Examples

```
R<- 20

data(data_samp_clust)
data(data_pop_clust)
no_cluster<- 200
cluster_size<- table(data_pop_clust$cluster)[unique(data_samp_clust$cluster)]

# The first stage sampling fraction is about 20% and the overall second stage sampling is about 15%.
# data_samp_clust is a sample taken from data_pop_clust available in the package.

boot.RWY<- boot.twostage(data_samp_clust, no_cluster, cluster_size, R)
boot.RWY$boot.var

boot.Pr<- boot.twostage(data_samp_clust, no_cluster, cluster_size, R, bootstrap.method="Preston")
boot.Pr$boot.var

boot.RWY.med<- boot.twostage(data_samp_clust, no_cluster, cluster_size, R, parameter="median")
boot.RWY.med$boot.var
boot.RWY.med$boot.sample[[5]]

boot.Ch<- boot.twostage(data_samp_clust, no_cluster, cluster_size, R=c(5, 10),
          bootstrap.method="Chauvet")
boot.Ch$boot.mean

data(data_samp_stclust)
data(data_pop_stclust)
# The first stage sampling fraction is about 20% and the overall second stage sampling is about 15%.
# data_samp_stclust is a sample taken from data_pop_stclust available in the package.

no_cluster_stclust<- c(100, 125, 65)
cluster_size_pop_st<- aggregate(data_pop_stclust$cluster,
 by=list(data_pop_stclust$stratum), table)[[2]]
L<- length(unique(data_samp_stclust$stratum))
cluster_size_st<- NULL
for(h in 1:L) cluster_size_st<- c(cluster_size_st,
 cluster_size_pop_st[[h]][unique(data_samp_stclust$cluster[data_samp_stclust$stratum==h])])

boot.RWY.st<- boot.twostage(data_samp_stclust, no_cluster_stclust, cluster_size_st, R)
boot.RWY.st$boot.statistic
```

`boot.weights.stsrs` *Bootstrap Weights Methods for Survey Data*

Description

The function `boot.weights.stsrs` applies one of the following bootstrap weights methods on complete (full response) survey data selected under either SRSWOR or STSRSWOR: Rao, Wu and Yue (1992), Bertail and Combris (1997), Chipperfield and Preston (2007) and Beaumont and Patak (2012)

Usage

```
boot.weights.stsrs(
  data,
  population.size,
  R,
  parameter = "total",
  bootstrap.method = "Rao.Wu.Yue",
  boot.sample.size = NULL,
  distribution.adjust = NULL,
  epsilon = NULL
)
```

Arguments

`data`	A vector, matrix or data frame. If it is a matrix or data frame then the column of study variable has to be named `study.variable`. If the sampling design is STSRSWOR, a column identifying strata named `stratum` has to be included.
`population.size`	A vector of stratum population sizes
`R`	The number of bootstrap replicates
`parameter`	One of the following population parameters can be applied: `"total"` (population total), `"mean"` (population mean), `"quartile.25"` (population 1st quartile), `"quartile.50"` or `"median"` (population median) or `"quartile.75"` (population 3rd quartile). If the parameter of interest is the population mean or total, the HT-estimator is applied. If the parameter of interest is a population quartile, the estimator in Sarndal, Swensson, and Wretman (1992, Chapter 5) is applied. The default is the population total.
`bootstrap.method`	One of the following bootstrap methods can be applied: `"Rao.Wu.Yue"` (Rao, Wu and Yue, 1992),`"Bertail.Combris"` (Bertail and Combris, 1997), `"Chipperfield.Preston"` (Chipperfield and Preston, 2007) or `"Beaumont.Patak"` (Beaumont and Patak, 2012). The default is `"Rao.Wu.Yue"`.
`boot.sample.size`	A vector of bootstrap sample sizes within strata only required for the method of Rao, Wu and Yue (1992). The length of this vector has to be the same as

the number of strata. The default is NULL. If the method of Rao, Wu and Yue (1992) is applied and `boot.sample.size` is not specified, the bootstrap sample size will be `nh-1` within each stratum, where `nh` is the original sample size within stratum `h`.

`distribution.adjust`
The default is NULL. A distribution should be specified for the method of Bertail and Combris (1997) and Beaumont and Patak (2012) to generate the bootstrap weight adjustments if epsilon is NULL. One of the following distribution can be used: `"Normal"`, `"Lognormal"`, `"Exponential"` or `"Uniform"`.

`epsilon` The default is NULL. If either Bertail and Combris (1997) or Beaumont and Patak (2012) is applied and `distribution.adjust` is NULL, a value must be given to epsilon so that Eq(5) in Beaumont and Patak (2012) can be used to generate the bootstrap weight adjustments.

Value

`boot.statistic` A vector of bootstrap statistics

`boot.var` The bootstrap variance estimator of the estimator of parameter of interest.

`boot.mean` The average of the bootstrap estimator of the parameter of interest.

`boot.sample` A list of results for each iteration. That includes a column of original sample values, a column of bootstrap weight adjustments, a column of bootstrap weights and a column of stratum identifier.

References

Beaumont, J.-F. and Patak, Z. (2012). On the generalized bootstrap for sample surveys with special attention to Poisson sampling. International Statistical Review 80 (1), 127–148.

Bertail, P. and Combris, P. (1997). Bootstrap généralisé d'un sondage. Annales d'économie et de statistique 46, 49–83.

Chipperfield, J. and Preston, J. (2007). Efficient bootstrap for business surveys. Survey Methodology 33 (2), 167–172.

Rao, J. N. K., Wu, C. F. J. and Yue, K. (1992). Some recent work on resampling methods for complex surveys. Survey Methodology 18 (2), 209–217.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). Model-Assisted Survey Sampling. NewYork: Springer.

Examples

```
R<- 20

data(data_samp_srs)
population_size<- 6000
# The sampling fraction is about 30%.
# data_samp_srs is a sample taken from data_pop available in the package.

boot.RWY<- boot.weights.stsrs(data_samp_srs, population_size, R)
```

```
boot.RWY$boot.var

boot.CP<- boot.weights.stsrs(data_samp_srs, population_size, R,
           bootstrap.method="Chipperfield.Preston")
boot.CP$boot.var

boot.BP.med<- boot.weights.stsrs(data_samp_srs, population_size, R,
               parameter="median", bootstrap.method="Beaumont.Patak",
               distribution.adjust="Exponential")
boot.BP.med$boot.var
boot.BP.med$boot.sample[[5]]


data(data_samp_stsrs)
population_size_st<- c(4500, 6300, 3500, 2000, 1500)
# The overall sampling fraction is about 30%.
# data_samp_stsrs is a sample taken from data_pop_st available in the package.

boot.RWY.st<- boot.weights.stsrs(data_samp_stsrs, population_size_st, R)
boot.RWY.st$boot.var
boot.RWY.st$boot.statistic
```

`data_pop, data_pop_st, data_pop_clust, data_pop_stclust, data_samp_srs, data_samp_stsrs, data_samp_cl`

Populations and samples gerenated in the `bootsurv` *package*

Description

This package contains multiple datasets described below.

Datasets

`data_pop` This is a population of size 6,000. This data set contains a column of generated study variable, labeled as `study.variable`.

`data_pop_st` This dataset represents a population of size 17,800, divided into 5 strata. It includes a column for the generated study variable, labeled as `study.variable`, and a column identifying the strata, labeled as `stratum`. The subpopulation sizes within each stratum are as follows: 4,500, 6,300, 3,500, 2,000, and 1,500, respectively.

`data_pop_clust` This dataset represents a population consisting of 10,048 units distributed across 200 clusters. The number of units within each cluster was generated using a Poisson distribution with a mean of 50. It includes columns for the generated study variable, labeled as `study.variable`, and cluster identification, denoted as `cluster`.

`data_pop_stclust` This dataset represents a population with 14,511 units distributed across three strata, consisting of 100, 125, and 65 clusters, respectively. The number of units within each cluster was generated using a Poisson distribution with a mean of 50. It includes columns of the generated study variable, labeled as `study.variable`, stratum identification, labeled as `stratum`, and cluster identification within each stratum, labeled as `cluster`.

`data_samp_srs` This dataset comprises a sample of size 1,850, obtained through simple random sampling without replacement from the `data_pop` dataset.

`data_samp_stsrs` This dataset represents a sample of size 5,350 obtained through stratified simple random sampling without replacement from the stratified population `data_pop_st`. The sample consists of subsample sizes of 1,350, 1,900, 1,050, 600, and 450.

`data_samp_clust` This sample was drawn using a two-stage cluster sampling method, with simple random sampling without replacement applied at each stage. The sample is drawn from the `data_pop_clust` dataset. In the first stage, approximately 20% of clusters were selected. Subsequently, within each selected cluster, approximately 15% of units were sampled.

`data_samp_stclust` A stratified two-stage cluster sampling method is applied to draw this sample from the `data_pop_stclust` dataset. In each stratum, simple random sampling without replacement is applied at each stage. The first stage sampling fraction is approximately 20%, and the overall second stage sampling is approximately 15%.

`direct.boot.stsrs` *Direct Bootstrap Methods for Survey Data*

Description

The function `direct.boot.stsrs` applies one of the following bootstrap methods on complete (full response) survey data selected under either SRSWOR or STSRSWOR: Efron (1979), McCarthy and Snowden (1985), Rao and Wu (1988) and Sitter (1992, JASA).

Usage

```
direct.boot.stsrs(
  data,
  population.size,
  R,
  parameter = "total",
  bootstrap.method = "Rao.Wu",
  boot.sample.size = NULL
)
```

Arguments

`data`	A vector, matrix or data frame. If it is a matrix or data frame then the column of study variable has to be named `study.variable`. If the sampling design is STSRSWOR, a column identifying strata named `stratum` has to be included.
`population.size`	A vector of stratum population sizes
`R`	The number of bootstrap replicates

`parameter`	One of the following population parameters can be applied: `"total"` (population total), `"mean"` (population mean), `"quartile.25"` (population 1st quartile), `"quartile.50"` or `"median"` (population median) or `"quartile.75"` (population 3rd quartile). If the parameter of interest is the population mean or total, the HT-estimator is applied. If the parameter of interest is a population quartile, the estimator in Sarndal, Swensson, and Wretman (1992, Chapter 5) is applied. The default is the population total.
`bootstrap.method`	One of the following bootstrap methods can be applied: `"Efron"` (Efron, 1979), `"McCarthy.Snowden"` (McCarthy and Snowden, 1985), `"Rao.Wu"` (Rao and Wu, 1988) or `"Sitter.BMM"` (Sitter, 1992). The default is `"Rao.Wu"`.
`boot.sample.size`	If the method of Rao and Wu (1988) is applied, a vector of bootstrap sample sizes for each stratum may be specified. The length of this vector must match the number of strata. By default, if 'boot.sample.size' is not specified, the bootstrap sample size within each stratum will be 'nh-3', where 'nh' is the original sample size in stratum 'h'.

Value

`boot.statistic` A vector of bootstrap statistics

`boot.var` The bootstrap variance estimator of the estimator of the parameter of interest

`boot.mean` The average of the bootstrap estimator of the parameter of interest

`boot.sample` For each iteration, a list of results is generated, including three columns: bootstrap values (which may be rescaled values if resampling is done on a rescaled version of the original sample), selected indices in each stratum, and a stratum identifier column.

References

Efron, B. (1979). Bootstrap methods: another look at the jackknife. The Annals of Statistics 7 (1), 1–26.

McCarthy, P. J. and C. B. Snowden (1985). The bootstrap and finite population sampling. Vital and Health Statistics, Series 2, No. 95. DHHS Publication No. (PHS) 85–1369. Public Health Service. Washington. U.S. Government Printing Office.

Rao, J. N. K. and C. F. J. Wu (1988). Resampling inference with complex survey data. Journal of the American Statistical Association 83 (401), 231–241.

Särndal, C.-E., Swensson, B. & Wretman, J. (1992). Model-Assisted Survey Sampling. NewYork: Springer.

Sitter, R. R. (1992). A resampling procedure for complex survey data. Journal of the American Statistical Association 87 (419), 755–765.

Examples

```
R<- 20

data(data_samp_srs)
```

```
population_size<- 6000
# The sampling fraction is about 30%.
# data_samp_srs is a sample taken from data_pop available in the package.

boot.RW<- direct.boot.stsrs(data_samp_srs, population_size, R)
boot.RW$boot.var

boot.Efron<- direct.boot.stsrs(data_samp_srs, population_size, R,
              parameter="total", bootstrap.method="Efron")
boot.Efron$boot.var

boot.RW.med<- direct.boot.stsrs(data_samp_srs, population_size, R,
               parameter="median")
boot.RW.med$boot.var

data(data_samp_stsrs)
population_size_st<- c(4500, 6300, 3500, 2000, 1500)
# The overall sampling fraction is about 30%.
# data_samp_stsrs is a sample taken from data_pop_st available in the package.

boot.RW.st<- direct.boot.stsrs(data_samp_stsrs, population_size_st, R,
              parameter="total", bootstrap.method="Rao.Wu")
boot.RW.st$boot.statistic
```

`pseudopop.boot.stsrs` *Pseudo-population Bootstrap Methods for Survey Data*

Description

The function `pseudopop.boot.stsrs` applies one of the following pseudo-population bootstrap methods on complete (full response) survey data selected under either SRSWOR or STSRSWOR: Bickel and Freedman (1984), Chao and Lo (1985), Sitter (1992, CJS), Booth, Butler and Hall (1994) and Chao and Lo (1994).

Usage

```
pseudopop.boot.stsrs(
  data,
  population.size,
  R.pop,
  R.samp,
  parameter = "total",
  bootstrap.method = "Booth.Butler.Hall"
)
```

Arguments

`data`	A vector, matrix or data frame. If it is a matrix or data frame then the column of study variable has to be named `study.variable`. If the sampling design is STSRSWOR, a column identifying strata named `stratum` has to be included.
`population.size`	A vector of stratum population sizes
`R.pop`	The number of bootstrap replicates to create bootstrap pseudo-populations
`R.samp`	The number of bootstrap replicates to draw bootstrap samples from each bootstrap pseudo-population
`parameter`	One of the following population parameters can be applied: `"total"` (population total), `"mean"` (population mean), `"quartile.25"` (population 1st quartile), `"quartile.50"` or `"median"` (population median) or `"quartile.75"` (population 3rd quartile). If the parameter of interest is the population mean or total, the HT-estimator is applied. If the parameter of interest is a population quartile, the estimator in Sarndal, Swensson, and Wretman (1992, Chapter 5) is applied. The default is the population total.
`bootstrap.method`	One of the following bootstrap methods can be applied: `"Bickel.Freedman"` (Bickel and Freedman, 1984),`"Chao.Lo.1985"` (Chao and Lo, 1985), `"Sitter.BWO"` (Sitter, 1992), `"Booth.Butler.Hall"` (Booth, Butler and Hall, 1994) or `"Chao.Lo.1994"` (Chao and Lo, 1994). The default is `"Booth.Butler.Hall"`.

Value

`boot.statistic` A vector of bootstrap statistics

`boot.parameter` A vector of bootstrap parameters computed on bootstrap pseudo-populations

`boot.var` The bootstrap variance estimator of the estimator of parameter of interest

`boot.mean` The average of the bootstrap estimator of the parameter of interest

`boot.sample` A list of size `R.pop`. Each list contains a list of results on each generated bootstrap pseudo-population. This includes three columns: bootstrap values, selected indices in each stratum, and a stratum identifier column.

References

Bickel, P. J. and Freedman, D. A. (1984). Asymptotic normality and the bootstrap in stratified sampling. The Annals of Statistics 12, 470–82.

Booth, J. G., Butler, R. W. and Hall, P. (1994). Bootstrap methods for finite populations. Journal of the American Statistical Association 89 (428), 1282–1289.

Chao, M. T. and Lo, S.-H. (1985). A bootstrap method for finite population. Sankhya: The Indian Journal of Statistics, Series A 47, 399–405.

Chao, M. T. and Lo, S.-H. (1994). Maximum likelihood summary and the bootstrap method in structured finite populations. Statistica Sinica 4 (2), 389–406.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). Model-Assisted Survey Sampling. NewYork: Springer.

Sitter, R. R. (1992). Comparing three bootstrap methods for survey data. The Canadian Journal of Statistics 20 (2), 135–154.

Examples

```
R.pop<- 5
R.samp<- 10

data(data_samp_srs)
population_size<- 6000
# The sampling fraction is about 30%.
# data_samp_srs is a sample taken from data_pop available in the package.

boot.Booth<- pseudopop.boot.stsrs(data_samp_srs, population_size, R.pop, R.samp)
boot.Booth$boot.var

boot.BF<- pseudopop.boot.stsrs(data_samp_srs, population_size, R.pop, R.samp,
           bootstrap.method="Bickel.Freedman")
boot.BF$boot.var

boot.Sitter.med<- pseudopop.boot.stsrs(data_samp_srs, population_size, R.pop,
                   R.samp, parameter="median", bootstrap.method="Sitter.BWO")
boot.Sitter.med$boot.var
boot.Sitter.med$boot.sample[[2]][[5]]

data(data_samp_stsrs)
population_size_st<- c(4500, 6300, 3500, 2000, 1500)
# The overall sampling fraction is about 30%.
# data_samp_stsrs is a sample taken from data_pop_st available in the package.

boot.Booth.st<- pseudopop.boot.stsrs(data_samp_stsrs, population_size_st, R.pop, R.samp)
boot.Booth.st$boot.statistic
```

Index