# Package: winsorize (via r-universe)

September 30, 2024

**Type** Package

**Title** Winsorize Data

**Version** 0.0.2.1

**Date** 2019-05-25

**Author** Andreas Alfons and Dirk Eddelbuettel

**Maintainer** Dirk Eddelbuettel <edd@debian.org>

**Description** Remove outliers by means of winsorization, ie shrinking outlying observations to the border of the main part of the data. This package started from the excellent robustHD package by Andreas Alfons in order to reduce the number dependent package being pulled in. We expect to update the code over time.

**License** GPL (>= 2)

**Imports** Rcpp (>= 0.11.5)

**LinkingTo** Rcpp, RcppArmadillo

**Repository** https://eddelbuettel.r-universe.dev

**RemoteUrl** https://github.com/eddelbuettel/winsorize

**RemoteRef** HEAD

**RemoteSha** 3a1ffddd109fe54244d6d25f71fdb5e1f26618ca

## Contents

winsorize-package . . . . . 2
corHuber . . . . . 2
standardize . . . . . 4
winsorize . . . . . 5

**Index** **8**

`winsorize-package` *Winsorize Data*

## Description

Remove outliers by means of winsorization, ie shrinking outlying observations to the border of the main part of the data.

## Details

This package started from the excellent robustHD package by Andreas Alfons in order to reduce the number dependent package being pulled in. We expect to update the code over time.

## Author(s)

Andreas Alfons wrote `winsorize` as part as his excellent `robustHD` package.

Maintainer: Dirk Eddelbuettel <edd@debian.org>

## See Also

See the `robustHD` package for more.

`corHuber` *Robust correlation based on winsorization.*

## Description

Compute a robust correlation estimate based on winsorization, i.e., by shrinking outlying observations to the border of the main part of the data.

## Usage

```
corHuber(x, y,
  type = c("bivariate", "adjusted", "univariate"),
  standardized = FALSE, centerFun = median,
  scaleFun = mad, const = 2, prob = 0.95,
  tol = .Machine$double.eps^0.5, ...)
```

## Arguments

| | |
|---|---|
| `x` | a numeric vector. |
| `y` | a numeric vector. |
| `type` | a character string specifying the type of winsorization to be used. Possible values are `"univariate"` for univariate winsorization, `"adjusted"` for adjusted univariate winsorization, or `"bivariate"` for bivariate winsorization. |

| | |
|---|---|
| standardized | a logical indicating whether the data are already robustly standardized. |
| centerFun | a function to compute a robust estimate for the center to be used for robust standardization (defaults to median). Ignored if standardized is TRUE. |
| scaleFun | a function to compute a robust estimate for the scale to be used for robust standardization (defaults to mad). Ignored if standardized is TRUE. |
| const | numeric; tuning constant to be used in univariate or adjusted univariate winsorization (defaults to 2). |
| prob | numeric; probability for the quantile of the $\chi^2$ distribution to be used in bivariate winsorization (defaults to 0.95). |
| tol | a small positive numeric value. This is used in bivariate winsorization to determine whether the initial estimate from adjusted univariate winsorization is close to 1 in absolute value. In this case, bivariate winsorization would fail since the points form almost a straight line, and the initial estimate is returned. |
| ... | additional arguments to be passed to robStandardize. |

## Details

The borders of the main part of the data are defined on the scale of the robustly standardized data. In univariate winsorization, the borders for each variable are given by $+/-$const, thus a symmetric distribution is assumed. In adjusted univariate winsorization, the borders for the two diagonally opposing quadrants containing the minority of the data are shrunken by a factor that depends on the ratio between the number of observations in the major and minor quadrants. It is thus possible to better account for the bivariate structure of the data while maintaining fast computation. In bivariate winsorization, a bivariate normal distribution is assumed and the data are shrunken towards the boundary of a tolerance ellipse with coverage probability prob. The boundary of this ellipse is thereby given by all points that have a squared Mahalanobis distance equal to the quantile of the $\chi^2$ distribution given by prob. Furthermore, the initial correlation matrix required for the Mahalanobis distances is computed based on adjusted univariate winsorization.

## Value

The robust correlation estimate.

## Author(s)

Andreas Alfons, based on code by Jafar A. Khan, Stefan Van Aelst and Ruben H. Zamar

## References

Khan, J.A., Van Aelst, S. and Zamar, R.H. (2007) Robust linear model selection based on least angle regression. *Journal of the American Statistical Association*, **102**(480), 1289–1299.

## See Also

winsorize

## Examples

```
  ## Not run:
## generate data
library("mvtnorm")
set.seed(1234)  # for reproducibility
Sigma <- matrix(c(1, 0.6, 0.6, 1), 2, 2)
xy <- rmvnorm(100, sigma=Sigma)
x <- xy[, 1]
y <- xy[, 2]

## introduce outlier
x[1] <- x[1] * 10
y[1] <- y[1] * (-5)

## compute correlation
cor(x, y)
corHuber(x, y)

## End(Not run)
```

---

`standardize` *Data standardization*

---

## Description

Standardize data with given functions for computing center and scale.

## Usage

```
standardize(x, centerFun = mean, scaleFun = sd)

robStandardize(x, centerFun = median, scaleFun = mad, fallback = FALSE,
  eps = .Machine$double.eps, ...)
```

## Arguments

| | |
|---|---|
| `x` | a numeric vector, matrix or data frame to be standardized. |
| `centerFun` | a function to compute an estimate of the center of a variable (defaults to `mean`). |
| `scaleFun` | a function to compute an estimate of the scale of a variable (defaults to `sd`). |
| `fallback` | a logical indicating whether standardization with `mean` and `sd` should be performed as a fallback mode for variables whose robust scale estimate is too small. This is useful, e.g., for data containing dummy variables. |
| `eps` | a small positive numeric value used to determine whether the robust scale estimate of a variable is too small (an effective zero). |
| `...` | currently ignored. |

## Details

`robStandardize` is a wrapper function for robust standardization, hence the default is to use `median` and `mad`.

## Value

An object of the same type as the original data `x` containing the centered and scaled data. The center and scale estimates of the original data are returned as attributes `"center"` and `"scale"`, respectively.

## Note

The implementation contains special cases for the typically used combinations `mean`/`sd` and `median`/`mad` in order to reduce computation time.

## Author(s)

Andreas Alfons

## See Also

`scale`, `sweep`

## Examples

```
## generate data
set.seed(1234)     # for reproducibility
x <- rnorm(10)     # standard normal
x[1] <- x[1] * 10  # introduce outlier

## standardize data
x
standardize(x)     # mean and sd
robStandardize(x)  # median and MAD
```

---

# winsorize *Data cleaning by winsorization*

---

## Description

Clean data by means of winsorization, i.e., by shrinking outlying observations to the border of the main part of the data.

## Usage

```
winsorize(x, ...)

## Default S3 method:
winsorize(x, standardized = FALSE, centerFun = median,
  scaleFun = mad, const = 2, return = c("data", "weights"), ...)

## S3 method for class 'matrix'
winsorize(x, standardized = FALSE, centerFun = median,
  scaleFun = mad, const = 2, prob = 0.95, tol = .Machine$double.eps^0.5,
  return = c("data", "weights"), ...)

## S3 method for class 'data.frame'
winsorize(x, ...)
```

## Arguments

| | |
|---|---|
| x | a numeric vector, matrix or data frame to be cleaned. |
| standardized | a logical indicating whether the data are already robustly standardized. |
| centerFun | a function to compute a robust estimate for the center to be used for robust standardization (defaults to median). Ignored if standardized is TRUE. |
| scaleFun | a function to compute a robust estimate for the scale to be used for robust standardization (defaults to mad). Ignored if standardized is TRUE. |
| const | numeric; tuning constant to be used in univariate winsorization (defaults to 2). |
| return | character string; if standardized is TRUE, this specifies the type of return value. Possible values are "data" for returning the cleaned data, or "weights" for returning data cleaning weights. |
| prob | numeric; probability for the quantile of the $\chi^2$ distribution to be used in multivariate winsorization (defaults to 0.95). |
| tol | a small positive numeric value used to determine singularity issues in the computation of correlation estimates based on bivariate winsorization (see corHuber). |
| ... | for the generic function, additional arguments to be passed down to methods. For the "data.frame" method, additional arguments to be passed down to the "matrix" method. For the other methods, additional arguments to be passed down to robStandardize. |

## Details

The borders of the main part of the data are defined on the scale of the robustly standardized data. In the univariate case, the borders are given by $+/-$const, thus a symmetric distribution is assumed. In the multivariate case, a normal distribution is assumed and the data are shrunken towards the boundary of a tolerance ellipse with coverage probability prob. The boundary of this ellipse is thereby given by all points that have a squared Mahalanobis distance equal to the quantile of the $\chi^2$ distribution given by prob.

## Value

If `standardize` is `TRUE` and `return` is `"weights"`, a set of data cleaning weights. Multiplying each observation of the standardized data by the corresponding weight yields the cleaned standardized data.

Otherwise an object of the same type as the original data `x` containing the cleaned data is returned.

## Note

Data cleaning weights are only meaningful for standardized data. In the general case, the data need to be standardized first, then the data cleaning weights can be computed and applied to the standardized data, after which the cleaned standardized data need to be backtransformed to the original scale.

## Author(s)

Andreas Alfons, based on code by Jafar A. Khan, Stefan Van Aelst and Ruben H. Zamar

## References

Khan, J.A., Van Aelst, S. and Zamar, R.H. (2007) Robust linear model selection based on least angle regression. *Journal of the American Statistical Association*, **102**(480), 1289–1299.

## See Also

`corHuber`

## Examples

```
## generate data
set.seed(1234)     # for reproducibility
x <- rnorm(10)     # standard normal
x[1] <- x[1] * 10  # introduce outlier

## winsorize data
x
winsorize(x)
```

# Index

∗ **array**
standardize, 4
∗ **multivariate**
corHuber, 2
∗ **package**
winsorize-package, 2
∗ **robust**
corHuber, 2
winsorize, 5

corHuber, 2, *6, 7*

mad, *3*, *5, 6*
mean, *4, 5*
median, *3*, *5, 6*

robStandardize, *3*, *6*
robStandardize *(*standardize*)*, 4
robustHD, *2*

scale, *5*
sd, *4, 5*
standardize, 4
sweep, *5*

winsorize, *3*, 5
winsorize-package, 2