Package 'reldist'

May 15, 2022

Version 1.7-1

Date 2022-05-14

Title Relative Distribution Methods

Author Mark S. Handcock <handcock@stat.ucla.edu>

Maintainer Mark S. Handcock <handcock@stat.ucla.edu>

Description Tools for the comparison of distributions. This includes nonparametric estimation of the relative distribution PDF and CDF and numerical summaries as described in ``Relative Distribution Methods in the Social Sciences'' by Mark S. Handcock and Martina Morris, Springer-Verlag, 1999, Springer-Verlag, ISBN 0387987789.

Imports mgcv, densEstBayes

Suggests locfit

License GPL-3 + file LICENSE

URL http://www.stat.ucla.edu/~handcock/RelDist/

NeedsCompilation no

Repository CRAN

Date/Publication 2022-05-14 23:20:09 UTC

R topics documented:

gini *Compute the Gini Coefficient*

Description

Computes the Gini coefficient based on (possibly weighted) sample data

Usage

```
gini(x,  weights=rep(1,length=length(x)))
```

Arguments

`x`	a vector containing at least non-negative elements
`weights`	an optional vector of sample weights for `x`

Details

`Gini` is the Gini coefficient, a common measure of inequality within a distribution. It is commonly used to measure income inequality. It is defined as twice the area between the 45 degree line and a Lorenz curve, where the Lorenz curve is a graph describing the share of total income T accruing to the poorest fraction p of the population.

In typical use the values of `x` are the incomes of individuals from a survey and the `weights` are the corresponding survey weights. If the values of `x` are the mean incomes within income classes and the weights `weights` are the corresponding population proportions within those classes, the function computes an estimate of the Gini coefficient of the underlying income distribution.

Value

the Gini coefficient (between 0 and 1).

Author(s)

Mark S. Handcock <`handcock@stat.ucla.edu`>

Source

Relative Distribution Methods in the Social Sciences, by Mark S. Handcock and Martina Morris, Springer-Verlag, Inc., New York, 1999. ISBN 0387987789.

References

Relative Distribution Methods in the Social Sciences, by Mark S. Handcock and Martina Morris, Springer-Verlag, Inc., New York, 1999. ISBN 0387987789.

Divergent Paths: Economic Mobility in the New American Labor Market, Russell Sage Foundation, New York, June 2001 Annette D. Bernhardt, Martina Morris, Mark S. Handcock and Marc Scott.

Measurement of Inequality, by F. A. Cowell, in A. B. Atkinson / F. Bourguignon (Eds): Handbook of Income Distribution, Amsterdam, 2000.

Measuring Inequality, by F. A. Cowell, Prentice Hall/Harvester Wheatshef, 1995.

See Also

`reldist`, `nls`

Examples

```
# generate vector (of incomes)
x <- c(541, 1463, 2445, 3438, 4437, 5401, 6392, 8304, 11904, 22261)
# compute Gini coefficient
gini(x)
# generate a vector of weights.
w <- runif(n=length(x))
gini(x,w)
#
# Compute the inequality in income growth for the recent cohort of the
# National Longitudinal Survey (NLS) initiated in 1979.
#
library(reldist)
data(nls)
help(nls)
# Compute the wage growth
y <- exp(recent$chpermwage)
# Compute the unweighted estimate
gini(y)
# Compute the weighted estimate
gini(y,w=recent$wgt)
```

`nls` *Permanent wage growth in two cohorts of the NLS*

Description

These data are from two cohorts of the National Longitudinal Survey (NLS) initiated in 1966 and 1979. The cohorts are referred to as the ‘original’ and the ‘recent’ cohort, respectively. The data represents the permanent wage growth of each individual in the cohort from age 16 through 36. This was used in Handcock and Morris (1999) and Bernhardt, Morris, Handcock and Scott (2001) to study the question of wage mobility. A development of the estimation of these permanent wages and their relevance to the study of wage mobility is given in Handcock and Morris (1999). For the purposes of this `reldist` package, we can regard the permanent wages as measurements on two groups that we wish to compare.

The data set is comprised of two `data.frames` called ‘original’ and ‘recent’. Each has three columns: `chpermwage`: the change in permanent wages (in log-dollars), `endeduc`: the final achieved educational level (in years), and `wgt`: the sample weight.

Usage

```
data(nls)
```

Source

Relative Distribution Methods in the Social Sciences, by Mark S. Handcock and Martina Morris, Springer-Verlag, Inc., New York, 1999. ISBN 0387987789.

References

Relative Distribution Methods in the Social Sciences, by Mark S. Handcock and Martina Morris, Springer-Verlag, Inc., New York, 1999. ISBN 0387987789.

Divergent Paths: Economic Mobility in the New American Labor Market, Russell Sage Foundation, New York, June 2001 Annette D. Bernhardt, Martina Morris, Mark S. Handcock and Marc Scott.

See Also

reldist

`precipitation` *Annual Precipitation in US Cities*

Description

The average amount of precipitation (rainfall) in inches for each of 70 United States (and Puerto Rico) cities.

Usage

```
data(precipitation)
```

Format

A named vector of length 70.

Details

This is a clone of the `precip` R dataset to avoid a bug in R.

Source

Statistical Abstracts of the United States, 1975.

References

McNeil, D. R. (1977) *Interactive Data Analysis*. New York: Wiley.

Examples

```
require(graphics)
data(precipitation)
dotchart(precipitation[order(precipitation)], main="precipitation data")
title(sub = "Average annual precipitation (in.)")
```

reldist *Inference for Relative Distributions*

Description

Estimate and graph relative distribution and density functions for continuous or discrete data.

Usage

```
reldist(y, yo=FALSE, ywgt=FALSE,yowgt=FALSE,
  show="none", decomp="locadd",
  location="median", scale="IQR",
  rpmult=FALSE,
  z=FALSE, zo=FALSE,
  smooth = 0.35,
  quiet = TRUE,
  cdfplot=FALSE,
  ci=FALSE,
  bar="no",
  add=FALSE,
  graph=TRUE, type="l",
  xlab="Reference proportion",ylab="Relative Density",yaxs="r",
  yolabs=pretty(yo), yolabslabs=NULL,
  ylabs=pretty(y), ylabslabs=NULL,
  yolabsloc=0.6, ylabsloc=1,
  ylim=NULL, cex=0.8, lty=1,
  binn=5000,
  aicc=seq(0.0001, 5, length=30),
  deciles=(0:10)/10,
  discrete=FALSE,
  method="Bayes",
  y0=NULL,
  control = list(samples = 4000, burnin = 1000),
  ...)
```

Arguments

`y` Sample from comparison distribution.

`yo` Sample from reference distribution.

`discrete` Do `y` and `yo` refer to a discrete distribution? If `TRUE` a discrete estimator is used instead of the default continuous one.

smooth Degree of smoothness required in the fit. Higher values lead to smoother curves, lower positive values lead to closer fits to the observed data. If it is not specified the value that minimizes GCV is used. If a value less than zero is specified then the value is chosen to minimize a corrected AIC. If discrete=TRUE it is the minimum number of values to pool in the reference distribution in the probability mass function estimate.

method Method used to estimate the relative density. The default ("Bayes") uses a density estimator based on Poisson Nonparametric Regression and Bayesian inference developed by Wand and Wu (2022). The option ("bgk") uses a Gaussian kernel density estimator for bounded domain one-dimensional data developed by Botev, Grotowski and Kroese (2010). The option ("gam") uses a local likelihood approach based on smoothed Poisson regression. The option "loclik" uses log-splines. The option "quick" uses the Anscombe transformation to stabilize variances. In versions prior to 1.3 the "quick" approach was used.

graph Graph the results on the current device.

bar Graph the deciles on the current device. Possible values of bar are "no" (no deciles plotted), "yes" (deciles plotted with the non-parametric fit, "only"(deciles plotted without non-parametric fit).

add Add the density to the current plot?

ylim plotting limit for the vertical axis.

lty Line type to be used for the density.

xlab Horizontal label.

ylab Vertical label.

ylabs Locations for label to be added to the right axis.

ylabslabs Labels indicating the original scale for the comparison distribution.

ylabsloc Distance of labels to right of axis (in lines).

yolabs locations for labels to be added to the tip axis.

yolabslabs Labels indicating the original scale for the reference distribution.

yolabsloc Distance of labels above axis (in lines).

yaxs Style of vertical axis.

cdfplot calculate and plot the CDF rather than the density.

quiet Should the output be returned invisibly?

ci Plot (pointwise) 95% confidence intervals?

ywgt Weights on the comparison sample.

yowgt Weights on the reference sample.

z Covariate on the comparison sample to be used to adjust it to the reference distribution. Only used if the form of matching specified in decomp="covariate".

zo Covariate on the reference sample to be used in the adjustment. to the reference distribution. Only used if the form of matching specified in decomp="covariate".

`show` Type of relative distribution to produce. Possible values are `"none"` (comparisons to reference), `residual` (location-matched reference to reference), `effect` (comparison to location-matched reference).

`decomp` Form of matching to the comparison sample. Possible values are `locmult` (multiplicatively scale the reference), `locadd` (additively shift the reference), `lsadd` (location/scale additive shift), `covariate` (covariate adjust the reference (requires z and zo to be specified)).

`location` How to measure location. Possible values are `"mean"` and `"median"`.

`scale` How to measure the scale. Possible values are `"standev"` (standard deviation) and `IQR` (interquartile range).

`rpmult` Only in calculation of polarization indices: multiplicatively scale the reference sample to the comparison sample before comparing the two distributions?

`binn` Number of bins used in the smoother.

`deciles` The percentiles used for the histogram bins. Typically deciles (i.e., 0.0, 0.1, 0.2,. . . ,0.9, 1.0), but any set can be used (e.g., quintiles, terciles).

`aicc` Values of the smoothing parameter to search over in minimizing the corrected AIC. Only used if `method="gam"` and `smooth` is less than 0.

`type` Type of plot to use. See `par()`.

`cex` Character expansion to use in plots. See `par()`.

`y0` A test to see if `yo` was passed. If `y0` is passed it prints a warning. If `y0` is passed and `yo` is not passed then `y0` is used as the sample from reference distribution.

`control` list; A simple list of control options for the STAN HMC computation used when `method="Bayes"`, the default. The two components are `samples=4000`, the number of samples drawn and `burnin=1000`, the burn-in used.

`...` Additional arguments to the plot functions. See `par()`.

Value

`x` Horizontal coordinates for the density (typically percentages).

`y` Density at x.

`rp` 95% confidence interval for the median relative polarization as lower bound, estimate, upper bound.

`rpl` 95% confidence interval for the lower relative polarization as lower bound, estimate, upper bound.

`rpu` 95% confidence interval for the upper relative polarization as lower bound, estimate, upper bound.

`cdf` x coordinates for the CDF (typically percentages) and y CDF at x.

Note

Most of the code is for the plotting and tinkering. The guts of the method are forming the relative data at the top. The rest is a standard fixed interval density estimation with a few bells and whistles.

References

For more examples see the tech report

Mark S. Handcock and Eric Mark Aldrich *Applying Relative Distribution Methods in R* University of Washington CSSS Working Paper No. 27, Available at SSRN: doi: 10.2139/ssrn.1515775.

Z. I. Botev, J. F. Grotowski and D. P. Kroese *Kernel Density Estimation Via Diffusion* Annals of Statistics, 2010, Volume 38, Number 5, Pages 2916-2957.

M. Wand and J. C. F. Yu *Density Estimation via Bayesian Inference Engines* AStA Advances in Statistical Analysis, 2021, doi: 10.1007/s10182021004228.

Examples

```
#
# First load the data.
#

data(nls, package="reldist")

#
# A simple example comparing permanent wages of the original to the
# recent cohort in the NLS.  See H&M (1999) for details.

reldist(y=recent$chpermwage,yo=original$chpermwage,method="bgk")

#
# A more sophisticated version of the same.
#

reldist(y=recent$chpermwage, yo=original$chpermwage,
        yowgt=original$wgt, ywgt=recent$wgt,
        bar=TRUE,
        smooth=0.1, method="bgk",
        yolabs=seq(-1, 3, by=0.5),
        ylim=c(0, 3.0),cex=0.8,
        ylab="Relative Density",
        xlab="Proportion of the Original Cohort")

#
# A CDF version.
#

reldist(y=recent$chpermwage, yo=original$chpermwage,
    yowgt=original$wgt, ywgt=recent$wgt,
    cdfplot=TRUE,
    smooth=0.4,
    yolabs=seq(-1,3,by=0.5),
    ylabs=seq(-1,3,by=0.5),
    cex=0.8,
    method="bgk",
    ylab="proportion of the recent cohort",
    xlab="proportion of the original cohort")
```

resplot *Relative distribution plot to a Standard Normal*

Description

`resplot` produces a relative distribution of the values to a standard normal.

Graphical parameters may be given as arguments to `resplot`.

Usage

```
resplot(x,
        standardize=TRUE,
        xlab="Gaussian Cumulative Proportion",
        method="Bayes",
        ...)
```

Arguments

`x`	The first sample for `resplot`.
`standardize`	Should the sample be converted to standard units first?
`xlab`	plot labels.
`method`	Method used to estimate the relative density. The default (`"Bayes"`) uses a density estimator based on Poisson Nonparametric Regression and Bayesian inference developed by Wand and Wu (2022). The option (`"bgk"`) uses a Gaussian kernel density estimator for bounded domain one-dimensional data developed by Botev, Grotowski and Kroese (2010). The option (`"gam"`) uses a local likelihood approach based on smoothed Poisson regression. The option `"loclik"` uses log-splines. The option `"quick"` uses the Anscombe transformation to stabilize variances. In versions prior to 1.3 the `"quick"` approach was used.
`...`	graphical parameters.

Value

A list with components summarizing the relative distribution. See `reldist` for the details.

See Also

`reldist`.

Examples

```
y <- rnorm(2000)
resplot(y, method="bgk")
data(precipitation)
resplot(precipitation, ylab = "Precipitation [in/yr] for 70 US cities", method="bgk")
```

`wtd.iqr` *Weighted Interquartile range*

Description

Compute weighted Interquartile range (iqr)

Usage

```
wtd.iqr (x, na.rm = FALSE, weight=NULL)
```

Arguments

`x` Vector of data, same length as `weight`

`na.rm` Logical: Should NAs be stripped before computation proceeds?

`weight` Vector of weights. If NULL, the default, weights are not used.

Details

Uses a simple algorithm based on sorting.

Value

Returns an empirical interquartile range from a weighted sample.

`wtd.quantile` *Weighted Quantiles*

Description

Compute weighted quantile

Usage

```
wtd.quantile (x, q=0.5, na.rm = FALSE, weight=NULL)
```

Arguments

`x` Vector of data, same length as `weight`

`q` Quantile to compute

`na.rm` Logical: Should NAs be stripped before computation proceeds?

`weight` Vector of weights. If NULL, the default, weights are not used.

Details

Uses a simple algorithm based on sorting.

Value

Returns an empirical q quantile from a weighted sample.

Index