Package ‘snpar’

February 20, 2015

Type Package

Title Supplementary Non-parametric Statistics Methods

Version 1.0

Date 2014-08-11

Author Debin Qiu

Maintainer Debin Qiu <debinqiu@uga.edu>

Description contains several supplementary non-parametric statistics methods including quantile test, Cox-Stuart trend test, runs test, normal score test, kernel PDF and CDF estimation, kernel regression estimation and kernel Kolmogorov-Smirnov test.

License GPL (>= 2)

NeedsCompilation no

Repository CRAN

Date/Publication 2014-08-16 10:45:40

R topics documented:

snpar-package *Supplementary Non-parametric Statistics Methods*

Description

Provide supplementary non-parametric statistics methods to perform several non-parametric tests based on rank score, estimate kernel probability density and cumulative distribution function, and fit kernel regression.

Details

Package:	snpar
Type:	Package
Version:	1.0
Date:	2014-08-12
License:	GPL (>= 2)

This package contains several supplementary non-parametric statistics methods, including one- or two-sample quantile test and normal score test as well as multiple-sample normal score test, Cox-Stuart trend test, runs test for randomness, kernel PDF and CDF estimation, kernel regression estimation and kernel Kolmogorov-Smirov test.

For a complete list of functions, use `library(help = snpar)`.

Author(s)

Debin Qiu

Maintainer: Debin Qiu <`<debinqiu@uga.edu>`>

References

Abdi, H. (2007). Bonferroni and Sidak corrections for multiple comparisons. In Salkind, N. J. *Encyclopedia of Measurement and Statistics*. Thousand Oaks, CA: Sage.

Conover, W. J. (1999). *Practical Nonparameteric Statistics* (Third Edition ed.). Wiley. pp. 396-406.

D.R. Cox and A. Stuart (1955). Some quick sign tests for trend in location and dispersion. *Biometrika*, Vol. 42, pp. 80-95.

Fan, I. Gijbels (1996). *Local Polynomial Modeling and its Applications*. Chapman & Hall, London.

Li, Qi; Racine, Jeffrey S. (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press. ISBN 0-691-12161-3.

Nadaraya, E. A. (1964). On Estimating Regression. *Theory of Probability and its Applications* 9(1): 141-2.

Wald, A. and Wolfowitz, J. (1940). On a test whether two samples are from the same population. *Ann. Math Statist.* 11, 147-162.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman and Hall, London.

Wang, J., Cheng, F. and Yang, L. (2013). Smooth simultaneous confidence bands for cumulative distribution functions. *Journal of Nonparametric Statistics*. 25, 395-407.

Wu, X. and Zhao, B. (2013). *Nonparametric Statistics* (Fourth Edition ed). China Statistics Press.

`cs.test` *Cox-Stuart Trend Test*

Description

Perform one-sample Cox-Stuart trend test.

Usage

```
cs.test(x, alternative = c("two.sided", "increasing", "decreasing"),
        exact = TRUE, correct = TRUE)
```

Arguments

`x` a numeric vector of data values.

`alternative` indicates the alternative hypothesis and must be one of `"two.sided"` (default), `"increasing"`, or `"decreasing"`.

`exact` `TRUE` (default) or `FALSE` indicating whether an exact p-value should be computed. See 'Details' for the meaning of `TRUE`.

`correct` a logical indicating whether to apply continuity correction in the normal approximation for the p-value. The default is `TRUE`.

Details

Cox-Stuart trend analysis is a robust method to detect the presence of the trend regardless of the distribution of the data. Given the independent data, i.e., $X[1], ..., X[n]$, one can divide the data into two sequences with equal number of observations cutted in the midpoint and then take the paired difference, i.e., $D = X[i] - X[i + c], i = 1, ..., floor(n/2)$, where c is the index of midpoint. The totals of the positive or negative sign in D is defined as $S+$ or $S-$. Under null hypothesis, $S+$ or $S-$ has a binomial distribution with the number of experiment being the number of elements in D after removing element(s) 0 and probability $p = 0.5$. The exact method (`exact = TRUE`) is based on binomial distribution of statistic $S+$ (`"increasing"`) or $S-$ (`"decreasing"`) or $S = min(S+, S-)$ (`"two.sided"`) and one can thus compute the exact p-value. When the sample size is large, one can also use the normal approximation (argument `exact = TRUE`) to the binomial distribution with or without continuity correction. Missing values have been removed.

Value

A list with class "htest" containing the following components:

`data.name`	a character string giving the names of the data.
`method`	the type of test applied.
`alternative`	a character string describing the alternative hypothesis.
`p.value`	the p-value for the test.
`statistic`	the value of the test statistic with a name describing it.

Author(s)

Debin Qiu <<debinqiu@uga.edu>>

References

D.R. Cox and A. Stuart (1955). Some quick sign tests for trend in location and dispersion. *Biometrika*, Vol. 42, pp. 80-95.

Examples

```
x <- 0.5*c(1:100) + rnorm(100,2,20)
# exact method
cs.test(x)
# approximate method
cs.test(x, exact = FALSE)
```

kde *Kernel Density and Distribution Estimation*

Description

To compute the non-parametric kernel estimation of the probability density function (PDF) and cumulative distribution function (CDF).

Usage

```
kde(x, h, xgrid, ngrid, kernel = c("epan", "unif", "tria", "quar",
    "triw", "tric", "gaus", "cos"), plot = FALSE)
```

Arguments

`x`	a numeric vector of data values.
`h`	the smoothing bandwidth. See 'Details' of the default bandwidth.
`xgrid`	the user-defined data points at which the PDF and CDF are to be evaluated. The default is the data values `x`.

`ngrid` the number of equally spaced points at which the PDF and CDF are to be evaluated. The default is `NULL`.

`kernel` a character string which determines the smoothing kernel function. This must be one of `"unif"` (uniform), `"tria"` (triangular), `"epan"` (epanechnikov), `"quar"` (quartic), `"triw"` (triweight), `"tric"` (tricube), `"gaus"` (gaussian) and `"cos"` (cosine). The default is `"epan"`.

`plot` a logical indicating whether to plot the estimated PDF and CDF graphs.

Details

Kernel density and distribution estimation is a non-parametric method to estimate the probability density function (PDF) and cumulative distribution function (CDF) by using kernel function for a continuous random variable. The default smoothing bandwidth is the plug-in optimal one in Fan and Gijbels (1996), i.e., $h = c*n^{(}-1/5)$, where the constant is replaced by (8*pi/3)^(1/5)*2.0362*(((quantile(x, 0.75) - quantile(x, 0.25))/1.349)^(2/3)) in this function. Missing values have been removed.

Value

`x` the original data values.

`xgrid` the points where the PDF and CDF are to be evaluated.

`fhat` the estimated PDF values at the specified points.

`Fhat` the estimated CDF values at the specified points.

`bw` the smoothing bandwidth used.

Warning

The smoothing bandwidth is always a critical issue in non-parametric statistics. The default smoothing bandwidth suggested by Fan and Gijbels (1996) may not perform the best in some cases. You are recommended to provide one obtained by other methods.

Author(s)

Debin Qiu <`<debinqiu@uga.edu>`>

References

Fan, I. Gijbels (1996). *Local Polynomial Modeling and its Applications*. Chapman & Hall, London. pp. 47.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman and Hall, London.

Examples

```
x <- rnorm(200,2,3)
# with default bandwidth
kde(x, kernel = "quar", plot = TRUE)

# with specified bandwidth
kde(x, h = 4, kernel = "quar", plot = TRUE)
```

`kre` *Kernel Regression Estimation*

Description

To fit a non-parametric relation between a pair of random variables by using kernel method.

Usage

```
kre(x, y, h, kernel = c("epan", "unif", "tria", "quar",
     "triw", "tric", "gaus", "cos"), plot = FALSE)
```

Arguments

`x` input of explanatory values.

`y` input of response values.

`h` the smoothing bandwidth. See 'Details' of the default bandwidth.

`kernel` a character string which determines the smoothing kernel function. This must be one of `"unif"` (uniform), `"tria"` (triangular), `"epan"` (epanechnikov), `"quar"` (quartic), `"triw"` (triweight), `"tric"` (tricube), `"gaus"` (gaussian) and `"cos"` (cosine). The default is `"epan"`.

`plot` a logical indicating whether to plot the graph with fitted line.

Details

Kernel regression is a non-parametric technique to find a non-linear relation between a pair of random variables X and Y. It is also called Nadaraya-Watson kernel regression which estimates the conditional expectation of a random variable, i.e., $E(Y|X)$. The default smoothing bandwidth is the optimal plug-in bandwidth in Fan and Gijbels (1996), i.e., $h = c * n^{(} - 1/5)$, where c is a constant and replaced by IQR in this function. Note that it provides the initial bandwidth and may not be the best one. Missing values have been removed.

Value

`results` a matrix including the original data of explanatory variable `x` (first column), the original data of response variable `y` (second column), and the fitted values of response `yhat` (third column).

`bw` the smoothing bandwidth used.

Warning

The smoothing bandwidth is always a critical issue in non-parametric statistics. The default smoothing bandwidth suggested by Fan and Gijbels (1996) may not perform well in some cases. You are recommended to provide one obtained by other methods.

Author(s)

Debin Qiu <<debinqiu@uga.edu>>

References

Fan, I. Gijbels (1996). *Local Polynomial Modeling and its Applications*. Chapman & Hall, London.

Li, Q., Racine, J. (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press. ISBN 0-691-12161-3.

Nadaraya, E. A. (1964). On Estimating Regression. *Theory of Probability and its Applications* 9(1): 141-2.

Examples

```
x <- rnorm(100)
y <- 1 + 4*x^2 + rnorm(100)
kr <- kre(x,y, kernel = "epan", plot = TRUE)
```

`KS.test` *Kolmogorov-Smirnov Test*

Description

Perform a Kolmogorov-Smirnov test for one sample or two samples using kernel method.

Usage

```
KS.test(x, y, ..., kernel = c("epan", "unif", "tria",
          "quar", "triw", "tric", "gaus", "cos"), hx, hy,
          alternative = c("two.sided", "less", "greater"))
```

Arguments

`x`	a numeric vector of data values.
`y`	either a numeric vector of data values, or a character string naming a cumulative distribution function or an actual cumulative distribution function such as `"pnorm"`. Only continuous CDFs are valid.
`...`	parameters of the distribution specified (as a character string) by `y`.
`kernel`	a character string which determines the smoothing kernel function. TThis must be one of `"unif"` (uniform), `"tria"` (triangular), `"epan"` (epanechnikov), `"quar"` (quartic), `"triw"` (triweight), `"tric"` (tricube), `"gaus"` (gaussian) and `"cos"` (cosine). The default is `"epan"`.
`hx`	the smoothing bandwidth for `x`. See 'Details' of the default bandwidth.
`hy`	the smoothing bandwidth for `y`. See 'Details' of the default bandwidth.
`alternative`	indicates the alternative hypothesis and must be one of `"two.sided"` (default), `"less"`, or `"greater"`.

Details

The traditional Kolmogorov-Smirnov test is based on the empirical cumulative distribution function (CDF) which is not continuous and may not provide good estimations to the true CDF. However, the CDF estimated by kernel method overcomes this shortcoming and generally performs much better than the empirical CDF. Namely, the kernel CDF is closer to the true CDF than the empirical CDF. Therefore, applying the kernel CDF is more reasonable than using the empirical CDF in Kolmogorov-Smirnov test. The test statistic is defined as the maximum difference in value and depends on the form of the alternative hypothesis. When the sample size is large, the test statistic has the following Kolmogorov-Smirnov distribution function:

$$K(x) = \sum(-1)^(j) * exp{-2 * j^2 * x^2}, j = -inf, ..., inf, x \geq 0,$$

and $K(x) = 0, x < 0$. See Conover, W. J. (1999) for more details. The default smoothing bandwidth is the plug-in optimal bandwidth used in Wang, Cheng and Yang (2013). Missing values have been removed.

Value

A list with class "htest" containing the following components:

`data.name` a character string giving the name(s) of the data.

`statistic` the value of the test statistic.

`p.value` the p-value of the test.

`method` a character string indicating what type of test was performed.

`alternative` a character string describing the alternative hypothesis.

Warning

The smoothing bandwidth is always a critical issue in non-parametric statistics. The default smoothing bandwidth suggested by Wang, Cheng and Yang (2013) may not perform well. This only gives the initial bandwidth in some cases. You are recommended to provide one obtained by other methods.

Note

This function only computes the p-value for large sample size. For small sample size, you can use `ks.test` to compute the exact p-value. Missing values have been removed.

Author(s)

Debin Qiu <`<debinqiu@uga.edu>`>

References

Conover, W. J. (1999). *Practical Nonparameteric Statistics* (Third Edition ed.). Wiley. pp. 396-406.

Wang, J., Cheng, F. and Yang, L. (2013). Smooth simultaneous confidence bands for cumulative distribution functions. *Journal of Nonparametric Statistics*. 25, 395-407.

See Also

`ks.test`

Examples

```
# one-sample Kolmogorov-Smirnov test
x <- rnorm(100,2,3)
KS.test(x, "pnorm", 2, 3)

# two-sample Kolmogorov-Smirnov test
y <- rgamma(100,1,6)
KS.test(x,y)
```

`ns.test` *Normal Score (Van der Waerden) Test*

Description

Perform a normal score (Van der Waerden) test of location(s) for one sample or two or multiple samples.

Usage

```
ns.test(x, g = NULL, q, alternative = c("two.sided", "less", "greater"),
        paired = FALSE, compared = FALSE, alpha = 0.05)
```

Arguments

`x`	a numeric vector of data values.
`g`	a factor of group values. The default is `NULL` for one-sample test.
`q`	a number used to form the null hypothesis for one-sample test only.
`paired`	a logical indicating whether you want a paired test for two samples.
`compared`	a logical indicating whether you want to compare the location of each group for multiple-sample test.
`alpha`	the Type I error for the pairwise comparision in multiple-sample test.
`alternative`	a character string specifying the alternative hypothesis, must be one of `"two.sided"` (default), `"greater"` or `"less"`.

Details

Normal score (Van der Waerden) test examines the location of one sample or equality of locations for two or multiple samples regardless of the distributions of the numeric data based on Van der Waerden rank scores. The Van der Waerden rank scores are defined as the ranks of data, i.e., $R[i], i = 1, 2, ..., n$, divided by $1 + n$ transformed to a normal score by applying the inverse of the normal distribution function, i.e., $\Phi^{(} - 1)(R[i]/(1 + n))$. For two-sample or multiple-sample test, the ranks of data are obtained by ordering the observations from all groups. Note that the statistic for one-sample and two-sample test is normally distributed, but has a chi-squared distribution with $k-1$ degrees of freedom for multiple-sample test, where k is the number of groups. For multiple-sample test, the pairwise comparsion is applied by controlling the familywise error rate with Bonferroni correction method. See Abdi, H. (2007) for more details.

Value

For one- or two-sample test, a list with class "htest" containing the following components:

`data.name` a character string giving the names of the data.

`method` the type of test applied.

`statistic` the value of the test statistic with a name describing it.

`p.value` the p-value for the test.

`alternative` a character string describing the alternative hypothesis.

For multiple-sample test, a list containing the above components plus the following components:

`df` degrees of freedom of chi-squared distribution for multiple-sample test.

`compare` a matrix indicting whether the locations of two different groups are equal.

Note

The normal score test provides high efficiency compared to other nonparametric test methods based on ranks of data when the normality assumptions are nealy or in fact satisfied. It is also roubust to the violation of normality assumtions.

Author(s)

Debin Qiu <`<debinqiu@uga.edu>`>

References

Abdi, H. (2007). Bonferroni and Sidak corrections for multiple comparisons". In Salkind, N. J. *Encyclopedia of Measurement and Statistics*. Thousand Oaks, CA: Sage.

Conover, W. J. (1999). *Practical Nonparameteric Statistics* (Third Edition ed.). Wiley. pp. 396-406.

Wu, X. and Zhao, B. (2013). *Nonparametric Statistics* (Fourth Edition ed). China Statistics Press.

See Also

`wilcox.test`, `kruskal.test`

Examples

```
# one-sample test
x <- c(14.22, 15.83, 17.74, 19.88, 20.42, 21.96, 22.33, 22.79, 23.56, 24.45)
ns.test(x, q = 19)

# two-sample test
y <- c(5.54, 5.52, 5.00, 4.89, 4.95, 4.85, 4.80, 4.78, 4.82, 4.85, 4.72, 4.48,
       4.39, 4.36, 4.30, 4.26, 4.25, 4.22)
group <- gl(2,9)
## independent two-sample test
ns.test(y, group)
## paired two-sample test
ns.test(y,group, paired = TRUE)

# multiple-sample test
z <- c(10.7, 10.8, 10.5, 10.9, 9.7, 14.5, 12.2, 12.4, 12.8, 12.7, 15.2, 12.3,
       13.5, 14.7, 15.6)
gr <- gl(3,5)
ns.test(z, gr, compared = TRUE)
```

`quant.test` *Quantile Test*

Description

Perform a one-sample quantile test and two-sample equality of quantiles.

Usage

```
quant.test(x, y = NULL, q, paired = FALSE, p = 0.5,
           alternative = c("two.sided", "less", "greater"),
           exact = FALSE, correct = TRUE)
```

Arguments

`x` a numeric vector of data values.

`y` a numeric vector of data or group values.

`q` a quantile used to form the null hypothesis for one-sample test only.

`paired` a logical indicating whether you want a paired test for two samples.

`p` probability of the quantile, must be between 0 and 1. The default is 0.50 which is the median.

`alternative` a character string specifying the alternative hypothesis, must be one of `"two.sided"` (default), `"greater"` or `"less"`.

`exact` a logical indicating whether an exact p-value should be computed. See 'Details' for the meaning of `TRUE`.

`correct` a logical indicating whether to apply continuity correction in the normal approximation for the p-value. The default is `TRUE`.

Details

Quantile test examines the location of one sample or the equality of locations for two samples. Compared to the t-test, quantile test does not require the normality assumptions of the data. So it is more general than the t-test, although it lacks statistical power when the normality assumptions do hold. For exact one-sample test, it is also called sign test. For this test, a new sequence is obtained by subtracting the location q from the original numeric sample and removing element(s) 0. The totals of positive or negative signs of the new sequence is defined as $S+$ or $S-$. Thus, $S+$ or $S-$ has a binomial distribution with probability p. The paired-sample test has the similar procedures for the sequence of paired difference except setting $q = 0$. For exact two-sample quantile test, it is also called Brwon-Mood test which is based on hypergeometric distribution for the total numbers of the first sample greater than the pooled median. When the sample size is large, one can also use the normal approximation, i.e., argument `exact = FALSE`, to the binomal distribution (one-sample case) and hypergeometric distribution (two-sample case) with or without continuity correction. See Conover, W. J. (1999) for more details. Missing values have been removed.

Value

A list with class "htest" containing the following components:

`data.name`	a character string giving the names of the data.
`method`	the type of test applied.
`statistic`	the value of the test statistic with a name describing it.
`p.value`	the p-value for the test.
`alternative`	a character string describing the alternative hypothesis.

Author(s)

Debin Qiu <`<debinqiu@uga.edu>`>

References

Conover, W. J. (1999). *Practical Nonparameteric Statistics* (Third Edition ed.). Wiley. pp. 396-406.

See Also

`wilcox.test`, `ns.test`

Examples

```
# one-sample test
x <- c(14.22, 15.83, 17.74, 19.88, 20.42, 21.96, 22.33, 22.79, 23.56, 24.45)
## normal approximation test
quant.test(x, q = 19)
## exact quantile test
quant.test(x, q = 19, exact = TRUE)

# two-sample test
y <- c(5.54, 5.52, 5.00, 4.89, 4.95, 4.85, 4.80, 4.78, 4.82, 4.85, 4.72, 4.48,
       4.39, 4.36, 4.30, 4.26, 4.25, 4.22)
group <- as.numeric(gl(2,9))
## independent two-sample test
quant.test(y, group, exact = TRUE)
## paired two-sample test
quant.test(y,group, paired = TRUE)
```

runs.test *Runs Test for Randomness*

Description

Perform the runs test for randomness of a numeric sequence.

Usage

```
runs.test(x, exact = FALSE, alternative = c("two.sided", "less", "greater"))
```

Arguments

x	a numeric vector of data values.
exact	TRUE or FALSE (default) indicating whether an exact p-value should be computed. See 'Details' for the meaning of TRUE.
alternative	indicates the alternative hypothesis and must be one of "two.sided" (default), "less", or "greater". See 'Details' for the meanings of the possible values.

Details

Runs test examines the randomness of a numeric sequence x by studying the frequency of runs R. Generally, every numeric sequence can be transformed into dichotomous (binary) data defined as 0 and 1 by comparing each element of the sequence to its median (default threshold). Given m 0 and n 1, the runs R is defined as a series of similar responses and has a statistical distribution. See Wald, A. and Wolfowitz, J. (1940) for more details of this distribution. Based on the known distribution, the exact p-value can be computed for the data with small sample size. When the sample size is large, one can use the normal approximation (argument `exact = TRUE`) with mean $2mn/(m+n)+1$ and variance $2mn(2mn-m-n)/((m+n)^2*(m+n-1))$. The null of randomness is tested against the "under-mixing" trend and "over-mixing" trend by using alternative "less" and "greater". Missing values have been removed.

Value

A list with class "htest" containing the following components:

data.name	a character string giving the names of the data.
method	the type of test applied.
alternative	a character string describing the alternative hypothesis.
statistic	the value of the test statistic with a name describing it.
p.value	the p-value for the test.

Warning

When the runs R is large, the exact p-value cannot be computed as the combination in the distribution function of R will be infinity. Please use argument "`exact = F`" or "`exact = FALSE`" in this case.

Author(s)

Debin Qiu <<`debinqiu@uga.edu`>>

References

Wald, A. and Wolfowitz, J. (1940). On a test whether two samples are from the same population. *Ann. Math Statist.* 11, 147-162.

Wu, X. and Zhao, B. (2013). *Nonparametric Statistics* (Fourth Edition ed). China Statistics Press. pp. 40-42.

Examples

```
x <- rnorm(100)
runs.test(x)

y <- c(12.85, 13.29, 12.41, 15.21, 14.23, 13.56)
runs.test(y, exact = TRUE)
```

Index