Package 'Rita'

October 12, 2022

Type Package

Title Automated Transformations, Normality Testing, and Reporting

Version 1.2.0

Description Automated performance of common transformations used to fulfill parametric assumptions of normality and identification of the best performing method for the user. Output for various normality tests (Thode, 2002) corresponding to the best performing method and a descriptive statistical report of the input data in its original units (5-number summary and mathematical moments) are also presented. Lastly, the Rankit, an empirical normal quantile transformation (ENQT) (Soloman & Sawilowsky, 2009), is provided to accommodate non-standard use cases and facilitate adoption.
<DOI:10.1201/9780203910894>.
<DOI:10.22237/jmasm/1257034080>.

License MIT + file LICENSE

Encoding UTF-8

Imports base, stats, lattice

RoxygenNote 7.1.2

NeedsCompilation no

Author Daniel Mattei [aut, cre],
John Ruscio [aut]

Maintainer Daniel Mattei <DMattei@live.com>

Repository CRAN

Date/Publication 2022-03-15 14:50:07 UTC

R topics documented:

`ADTest` *Anderson-Darling Test*

Description

This function computes the one-sample Anderson-Darling test statistic and p-value for fit to a normal distribution.

Usage

```
ADTest(data, alpha = 0.05, j = 1)
```

Arguments

`data`	Data of a univariate distribution for which the test statistic is computed (vector)
`alpha`	The two-sided decision threshold used for hypothesis-testing
`j`	The # hypotheses tested; used to compute a Bonferonni correction, if applicable; should remain at its default if multiple testing is not an issue (scalar)

Details

An adjusted statistic provided by D'agostino & Stephens (1986) is used, where the mean and variance of the population are treated as unknown. D'agostino & Stephen's (1986) text provides the equations used to obtain the function's p-values.

Value

An object including the test statistic, p-value, and a significance flag (list)

References

D'agostino, R. B., & Stephens, M. A. (1986). Goodness-of-fit-techniques (Vol. 68). CRC press.

Examples

```
values <- rnorm(100)
x <- ADTest(data = values)
```

`arcsineXform` *Arcsine Transformation*

Description

This function transforms the scale, if needed, to values of unity. Then, the data is transformed by taking the arcsine of each value. Per the recommendations of Osborne(2002), data points are left-anchored at 0 to maximize the efficacy of the square-root transformation used enroute to the arcsine.

Usage

```
arcsineXform(sample)
```

Arguments

`sample`	The input data (vector)

Value

The arcsine-transformed data (vector)

References

Osborne, J. W. (2002). Notes on the use of data transformations. Practical Assessment, Research and Evaluation, 9(1), 42-50.

Osborne, J. W. (2002). The Effects of Minimum Values on Data Transformations. Retrieved from https://files.eric.ed.gov/fulltext/ED463313.pdf

Examples

```
values <- rnorm(100)
x <- arcsineXform(values)
```

`chisqTest` *Chi-Square Test*

Description

This function computes the chi-square test for normality.

Usage

```
chisqTest(data, alpha = 0.05, j = 1, df = 3)
```

Arguments

`data`	Data of a univariate distribution for which the test statistic is computed (vector)
`alpha`	The two-sided decision threshold used for hypothesis-testing
`j`	The # hypotheses tested; used to compute a Bonferonni correction, if applicable; should remain at its default if multiple testing is not an issue (scalar)
`df`	The degrees of freedom used to test for significance against the sampling distribution (scalar)

Details

Bins are created by cutting the data to ensure that values within these intervals would be equally probable if data are normal (Moore, 1986). By default, this function assumes that all relevant parameters (mu, sigma) are estimators, fixing the degrees of freedom at df = 3.

Value

An object including the test statistic, p-value, and a significance flag (list)

References

Moore, D.S., (1986) Tests of the chi-squared type. In: D'agostino, R.B. and Stephens, M.A., eds.: Goodness-of-Fit Techniques. Marcel Dekker, New York.

Examples

```
values <- rnorm(100)
x <- chisqTest(data = values)
```

DPTest *D'agostino Pearson Omnibus Test*

Description

This function computes the D'agostino Pearson omnibus test using adjusted Fisher- Pearson skewness and kurtosis estimators.

Usage

```
DPTest(data, alpha = 0.05, j = 1, warn = T)
```

Arguments

data	Data of a univariate distribution for which the test statistic is computed (vector)
alpha	The two-sided decision threshold used for hypothesis-testing
j	The # hypotheses tested; used to compute a Bonferonni correction, if applicable; should remain at its default if multiple testing is not an issue (scalar)
warn	Used for printing a warning message when testing is terminated for N < 8 (boolean)

Value

An object including the test statistic, p-value, and a significance flag (list)

References

D'agostino, R. B., & Stephens, M. A. (1986). Goodness-of-fit-techniques (Vol. 68). CRC press.

D'agostino, R. B., & Belanger, A. (1990). A Suggestion for Using Powerful and Informative Tests of Normality. The American Statistician, 44(4), 316–321. https://doi.org/10.2307/2684359

Shreve, Joni N. and Donna Dea Holland . 2018. SAS® Certification Prep Guide: Statistical Business Analysis Using SAS®9. Cary, NC: SAS Institute Inc.

Examples

```
values <- rnorm(100)
x <- DPTest(data = values)
```

inverseXform *Inverse/Reciprocal Transformation*

Description

This function imputes minimum values per the recommendations of Osborne (2002) and subsequently transforms the data using the reciprocal.

Usage

```
inverseXform(sample)
```

Arguments

sample	The input data (vector)

Value

The reciprocal-transformed data (vector)

References

Osborne, J. W. (2002). Notes on the use of data transformations. Practical Assessment, Research and Evaluation, 9(1), 42-50.

Osborne, J. W. (2002). The Effects of Minimum Values on Data Transformations. Retrieved from https://files.eric.ed.gov/fulltext/ED463313.pdf

Examples

```
values <- rnorm(100)
x <- inverseXform(values)
```

JBTest *Jarque-Bera Test*

Description

This function performs the Jarque-Bera test for normality using adjusted Fisher- Pearson skewness and kurtosis coefficients.

Usage

```
JBTest(data, alpha = 0.05, j = 1, N_Sample = 10000, warn = T)
```

Arguments

data	Data of a univariate distribution for which the test statistic is computed (vector)
alpha	The two-sided decision threshold used for hypothesis-testing
j	The # hypotheses tested; used to compute a Bonferonni correction, if applicable; should remain at its default if multiple testing is not an issue (scalar)
N_Sample	The # samples used to generate the bootstrapped sampling distribution, in cases when N < 2000 (scalar)
warn	Used for printing a warning message when boostrapping is performed for sample-sizes < 2000 or when testing is terminated for N < 4 (boolean)

Details

Large samples (N >= 2000) use p-values obtained with reference to the chi-square distribution, whereas smaller samples output p-values obtained via bootstrapping. When N < 4, testing is terminated.

Value

An object including the test statistic, p-value, and a significance flag (list)

References

Jarque, C. M. and Bera, A. K. (1980). Efficient test for normality, homoscedasticity and serial independence of residuals. Economic Letters, 6(3), pp. 255-259.

Shreve, Joni N. and Donna Dea Holland . 2018. SAS® Certification Prep Guide: Statistical Business Analysis Using SAS®9. Cary, NC: SAS Institute Inc.

Examples

```
values <- rnorm(100)
x <- JBTest(data = values)
```

KSLTest *Kolmogorov-Smirnov-Lilliefors Test*

Description

This function computes the Lilliefors variant of the one-sample Kolmogorov-Smirnov test.

Usage

```
KSLTest(data, alpha = 0.05, j = 1, warn = T)
```

Arguments

data	The data of a univariate distribution for which the test statistic is computed (vector)
alpha	The two-sided decision threshold used for hypothesis-testing (scalar)
j	The # hypotheses tested; used to compute a Bonferonni correction, if applicable; should remain at its default if multiple testing is not an issue (scalar)
warn	Used for printing a warning message when negative values are imputed to 0.0 (boolean)

Details

Molin & Abdi's (1998) algorithmic approximation of p-values is used for hypothesis-testing. Note that this algorithm requires the imputation of 0.0 for negative output when p-values would otherwise be low in value (< 0.001) using other methods. A similar issue with extremely large values requires the imputation of 1.0 for values larger than 1.0 when p > .99.

Value

An object including the test statistic, p-value, and a significance flag (list)

References

Lilliefors, H.W. (1967). On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. Journal of the American Statistical Association, 62, 399-402.

Molin, P., & Abdi, H. (1998). New Tables and numerical approximation for the KolmogorovSmirnov/Lillierfors/Van Soest test of normality.

Examples

```
values <- rnorm(100)
x <- KSLTest(data = values)
```

---

kurtCoeff *Adjusted Fisher-Pearson Excess Sample Kurtosis*

Description

Adjusted Fisher-Pearson Excess Sample Kurtosis

Usage

```
kurtCoeff(data, sd)
```

Arguments

data	The data for which kurtosis is computed (vector)
sd	The population standard deviation, used to compute kurtosis (scalar)

Value

The kurtosis value (scalar)

References

Shreve, Joni N. and Donna Dea Holland . 2018. SAS® Certification Prep Guide: Statistical Business Analysis Using SAS®9. Cary, NC: SAS Institute Inc.

Examples

```
values <- rnorm(100)
x <- kurtCoeff(data = values, sd = sd(values))
```

---

logitXform *Logit/Log-Odds Transformation*

Description

This function transforms data via the logit/log-odds transformation.

Usage

```
logitXform(sample, divisor = 2)
```

Arguments

sample	The input data (vector, matrix, or dataframe)
divisor	Number used to modify epsilon enroute to the empirical logit, in cases of output consisting of a single distinct value (scalar)

Details

Initially, features of the input data are extracted and used to determine an initial transformation to perform.

All forms of data representing an underlying discrete scale are converted to proportions of the total sample size, if needed. In these cases, values should be stored such that elements are in absolute frequency, relative frequency, or percentage form.

For non-count data, variables are shifted and bounded at [0,1] in a manner analogous to the potential transformations of the scale performed by arcsineXform() prior to the arcine, although transformed values are not expected to outperform more suitable transformations.

Then, the empirical logit transformation is applied to avoid zeroes or ones, and the data are transformed by taking the log-odds/logit of each value.

Value

The logit-transformed data (vector)

References

Stevens, S., Valderas, J. M., Doran, T., Perera, R., & Kontopantelis, E. (2016). Analysing indicators of performance, satisfaction, or safety using empirical logit transformation. bmj, 352.

Osborne, J. W. (2002). Notes on the use of data transformations. Practical Assessment, Research and Evaluation, 9(1), 42-50.

Osborne, J. W. (2002). The Effects of Minimum Values on Data Transformations. Retrieved from https://files.eric.ed.gov/fulltext/ED463313.pdf

Warton, D. I., & Hui, F. K. (2011). The arcsine is asinine: the analysis of proportions in ecology. Ecology, 92(1), 3-10.

Examples

```
values <- rnorm(100)
x <- logitXform(values)
```

`logXform` *Logarithmic Transformation*

Description

This function imputes minimum values per the recommendations of Osborne (2002) and subsequently transforms the data to a base-10 logarithmic scale.

Usage

```
logXform(sample)
```

Arguments

`sample`	The input data (vector)

Value

The log-transformed data (vector)

References

Osborne, J. W. (2002). Notes on the use of data transformations. Practical Assessment, Research and Evaluation, 9(1), 42-50.

Osborne, J. W. (2002). The Effects of Minimum Values on Data Transformations. Retrieved from https://files.eric.ed.gov/fulltext/ED463313.pdf

Examples

```
values <- rnorm(100)
x <- logXform(values)
```

MasterTest *Master Normality Testing Function*

Description

This is a master function to call the appropriate test(s) to be used in the 'Rita' function.

Usage

```
MasterTest(c, data, alpha = 0.05, j = 1)
```

Arguments

c	Input specifying the test to run (scalar)
data	The data of a univariate distribution for which the test statistic is computed (vector)
alpha	The two-sided decision threshold used for hypothesis-testing (scalar)
j	The # hypotheses tested; used to compute a Bonferonni correction, if applicable; should remain at its default if multiple testing is not an issue (scalar)

Value

An results object specific to the test designated with the 'c' argument (list)

Examples

```
values <- rnorm(100)
x <- MasterTest(c = 1, data = values)
```

MasterXform *Master Transformation Function*

Description

This is a master function used to perform the appropriate transformation(s) within the 'Rita' function.

Usage

```
MasterXform(c, data)
```

Arguments

c	Input specifying the test to run (scalar)
data	The data of a univariate distribution for which the test statistic is computed (vector)

Value

Output from the appropriate subfunction (list)

Examples

```
values <- rnorm(100)
x <- MasterXform(c = 2, data = values)
```

---

popSD *Converts Sample Standard Deviations into Population Equivalents*

Description

This function converts a sample standard deviation (SD) input into the population equivalent. This code is vectorized to convert several sample standard deviations for univariate distributions of identical sample-sizes, if desired.

Usage

```
popSD(s, n)
```

Arguments

s	The sample SD(s) (vector)
n	The sample-size for each SD to be converted (vector)

Value

The population SD(s) (vector)

References

Ruscio, J. (2021). Fundamentals of research design and statistical analysis. Ewing, NJ: The College of New Jersey, Psychology Department.

Examples

```
values <- rnorm(100)
x <- popSD(s = sd(values),n = 100)
```

rankitXform *Rankit Transformation*

Description

This function transforms data via the Rankit, a member of the families of 'rank-based normalization methods' and 'empirical normal quantile transformations' employed in both the social sciences and quantitative genetics.

Usage

```
rankitXform(sample)
```

Arguments

sample	The input data (vector)

Value

The Rankit-transformed data (vector)

References

Soloman, S. R., & Sawilowsky, S. S. (2009). Impact of rank-based normalizing transformations on the accuracy of test scores. Journal of Modern Applied Statistical Methods, 8(2), 9.

Peng, B., Robert, K. Y., DeHoff, K. L., & Amos, C. I. (2007, December). Normalizing a large number of quantitative traits using empirical normal quantile transformation. In BMC proceedings (Vol. 1, No. 1, p. S156). BioMed Central. doi: 10.1186/1753-6561-1-s1-s156

Bliss, C. I., Greenwood, M. L., & White, E. S. (1956). A rankit analysis of paired comparisons for measuring the effect of sprays on flavor. Biometrics, 12(4), 381-403.

Examples

```
values <- rnorm(100)
x <- rankitXform(values)
```

Rita *Rita*

Description

R Exploratory Data Analysis (REDA; pronounced "rita") summarizes an input dataset by the M, SD + 5-number summary + third and fourth moments and visualizes the data according to an algorithm or as specified by the user. In addition, Rita will provide the results of one or several normality tests. Lastly, Rita normalizes the dataset with several methods and provides visualizations of the best performing method to the user.

Usage

```
Rita(
  data,
  test = 1,
  xform = 1,
  alpha = 0.05,
  j = 1,
  autoPlot = T,
  histPlot = F,
  densPlot = F,
  stripPlot = F,
  violinPlot = F,
  xformPlot = F,
  return = T,
  seed = 10
)
```

Arguments

data	Input dataset (matrix, dataframe, or vector). For a univariate distribution, submit a vector or a subsetted matrix or dataframe. If results for many univariate distributions are desired, submit a matrix or dataframe with each column representing a given variable if all distributions are of the same sample-size. If not, it is recommended to call Rita repeatedly for each variable.
test	Desired normality test (scalar). By default (test = 1), Rita will present the results of the Shapiro-wilk test to the user. test = 1: Shapiro-Wilk (SW) test = 2: Kolmogorov-Smirnov/Lilliefors (KSL) test = 3: Anderson-Darling (AD) test = 4: Jarque-Bera (JB) test = 5: D'Agostino Pearson Omnibus (DP) test = 6: Chi-square test (chiSq) test = 7: Results of all tests for the best performing transformation The order of the tests printed corresponds to the order of the variables stored within the input dataset.
xform	Desired normalization method (scalar). By default (xform = 1), Rita will assess which method performs best and (a.) return the transformed data to the user, and (b.) visualize the data according to the settings of the plot argument. Please note that, per the recommendations of Osborne (2002), a constant is added prior to logarithmic and inverse transformations to ensure that the minimum value is anchored at 1, and prior to the square-root transformation to ensure a left anchor of 0. Similarly, the arc-sine and logit transformations are applied after converting the units, if needed, to ensure that variables are bounded between 0 and 1. The "best performing" method is identified by comparing goodness-of-fit to the straight line of the QQ plot for the quantiles of the data normalized by a given

	method and the standard normal distribution. If a tie is present between transformations for a variable, one of the best performing transformations is arbitrarily selected. xform = 1: Best performing method is presented (excluding the Rankit) xform = 2: Logarithmic transform xform = 3: Inverse/reciprocal transform xform = 4: Square-root transform xform = 5: Arc-sine transform xform = 6: Logit transform xform = 7: Rankit transform
`alpha`	The two-sided decision threshold used for normality hypothesis-testing (scalar)
`j`	The # hypotheses tested; used to compute a Bonferonni correction, if applicable; should remain at its default if multiple testing is not an issue (scalar)
`autoPlot`	Desired plotting method (boolean). By default (plot = 1), the visualization will be implicitly chosen based on extracted features of the dataset. When autoPlot = F, values of additional plotting arguments are used to determine the visualizations provided to the user. When autoPlot = T: Histograms are always generated for discrete data. Density plots are always generated for continuous data. Strip plots are generated when the # distinct values are <= 20 AND the # datapoints are 15 <= x <= 150. Violin plots are instead generated in lieu of the strip plots created when the above conditions are not met. Lastly, density plots for each (transformed*) variable are generated. *Transformed variables correspond to the choice made by the user for the xform argument or to the best-performing transformation for each variable when xform = 1. All plots are drawn in the R console and saved as plotting objects.
`histPlot`	Whether to generate histograms for each variable (boolean).
`densPlot`	Whether to generate density plots for each variable (boolean).
`stripPlot`	Whether to draw strip plots for each variable (boolean).
`violinPlot`	Whether to draw violin plots for each variable (boolean).
`xformPlot`	Whether to draw density plots for each transformed variable (boolean).
`return`	Whether to return the transformed variables of the best performing method (return = T; default), or the cleaned, untransformed variables eligible for transformation (return = F) (boolean).
`seed`	Number used for reproduction of random number generator results (scalar).

Details

Any rows with missing values (NAs) are removed for calculation purposes; if desired, incomplete records should be imputed or removed with subsetting prior to calling Rita. In addition, note that any columns not numeric type or coercible to numeric are excluded from analysis, as are any numeric columns with 2 distinct values or less.

Value

An object containing the dataset of the best performing transformation for each variable and the specified plots (list)

Examples

```
values <- rnorm(100)
x <- Rita(data = values)
```

`skewCoeff` *Adjusted Fisher-Pearson Skewness Coefficient with Sample-size Correction Factor*

Description

Adjusted Fisher-Pearson Skewness Coefficient with Sample-size Correction Factor

Usage

```
skewCoeff(data, sd)
```

Arguments

`data`	The data for which skewness is computed (vector)
`sd`	The population standard deviation, used to compute skewness (scalar)

Value

The skewness value (scalar)

References

Shreve, Joni N. and Donna Dea Holland . 2018. SAS® Certification Prep Guide: Statistical Business Analysis Using SAS®9. Cary, NC: SAS Institute Inc.

Examples

```
values <- rnorm(100)
x <- skewCoeff(data = values,sd = sd(values))
```

squareXform *Square-root Transformation*

Description

This function left anchors the minimum value to 0 per the recommendations of Osborne (2002) and subsequently transforms the data by taking the square-root of each value.

Usage

```
squareXform(sample)
```

Arguments

sample	The input data (vector)

Value

The square-transformed data (vector)

References

Osborne, J. W. (2002). Notes on the use of data transformations. Practical Assessment, Research and Evaluation, 9(1), 42-50.

Osborne, J. W. (2002). The Effects of Minimum Values on Data Transformations. Retrieved from https://files.eric.ed.gov/fulltext/ED463313.pdf

Examples

```
values <- rnorm(100)
x <- squareXform(values)
```

SWTest *Shapiro-Wilk Test*

Description

This function is a wrapper for shapiro.test() from the stats package. Options added include an ability to toggle a Bonferonni correction for significance, a corresponding significance flag, and reorganized output to facilitate integration with the Rita package.

Usage

```
SWTest(data, alpha = 0.05, j = 1, warn = T)
```

Arguments

data	Data of a univariate distribution for which the test statistic is computed (vector)
alpha	The two-sided decision threshold used for hypothesis-testing
j	The # hypotheses tested; used to compute a Bonferonni correction, if applicable; should remain at its default if multiple testing is not an issue (scalar)
warn	Used for printing a warning message when resampling is performed on sample-sizes > 5000 or when testing is terminated for N < 3 (boolean)

Details

Note that when the sample-size of the input vector is > 5000, resampling with replacement is used to proceed with hypothesis-testing with a vector of 5000 elements. When N < 3, testing is terminated.

Value

An object including the test statistic, p-value, and a significance flag (list)

References

Patrick Royston (1982). An extension of Shapiro and Wilk's W test for normality to large samples. Applied Statistics, 31, 115–124. 10.2307/2347973

Patrick Royston (1982). Algorithm AS 181: The W test for Normality. Applied Statistics, 31, 176–180. 10.2307/2347986

Patrick Royston (1995). Remark AS R94: A remark on Algorithm AS 181: The W test for normality. Applied Statistics, 44, 547–551. 10.2307/2986146

Examples

```
values <- rnorm(100)
x <- SWTest(data = values)
```

Index