Package 'Dpit'

March 3, 2017

Date 2017-02-10

Version 1.0

Title Distribution Pitting

Author Harry Joo [aut, cre], Herman Aguinis [aut, dtc], Kyle J. Bradley [aut], Takuya Noguchi [ctb]

Maintainer Harry Joo <harryjoo19@gmail.com>

Depends R (>= 3.1.1)

Description
Compares distributions with one another in terms of their fit to each sample in a dataset that contains multiple samples, as described in Joo, Aguinis, and Bradley (in press). Users can examine the fit of seven distributions per sample: pure power law, lognormal, exponential, power law with an exponential cutoff, normal, Poisson, and Weibull. Automation features allow the user to compare all distributions for all samples with a single command line, which creates a separate row containing results for each sample until the entire dataset has been analyzed.

License GPL (>= 2)

Imports VGAM, gsl, moments, utils, fitdistrplus

Repository CRAN

NeedsCompilation no

Date/Publication 2017-03-03 21:25:42

R topics documented:

`Dpit-package` *Distribution Pitting*

Description

Compares distributions with one another in terms of their fit to each sample in a dataset that contains multiple samples, as described in Joo, Aguinis, and Bradley (2017). Users can examine the fit of seven distributions per sample: pure power law, lognormal, exponential, power law with an exponential cutoff, normal, Poisson, and Weibull.

Details

Based on the R code found at: `http://tuvalu.santafe.edu/~aaronc/powerlaws/`. In particular, we borrowed heavily from Cosma R. Shalizi's code. Finally, we owe much gratitude to Cosma R. Shalizi, Aaron Clauset, and Yogesh Virkar for their kind responses to our questions regarding the code we borrowed from.

Author(s)

Harry Joo, Herman Aguinis, Kyle J. Bradley

Maintainer: Harry Joo <harryjoo19@gmail.com>

References

Joo, H., Aguinis, H., & Bradley, K. J. 2017. Not all nonnormal distributions are created equal: Improved theoretical and measurement precision. Journal of Applied Psychology. Advance online publication. doi: 10.1037/apl0000214

`descriptives` *descriptives*

Description

Returns a data frame containing descriptive statistics per sample.

Usage

```
descriptives(x)
```

Arguments

`x`	A data set

Value

`Negative.values?` Whether or not there are any negative values in a sample. If there is at least one negative value, then the program will print out "Negative values detected" in the relevant cell. If there are no negative values, then the program will leave the cell blank.

`zeros?` Whether or not there are any zero values in a sample. If there is at least one zero value, then the program will print out "zero values detected" in the relevant cell. If there are no negative values, then the program will leave the cell blank.

`N` Number of observations in a sample–before the package removes any non-positive values that lead to incalculable expressions (e.g., the log of zero is undefined).

`median` Median value of a sample–before removing any non-positive values.

`mean` Mean value of a sample–before removing any non-positive values.

`SD` Standard deviation in a sample–before removing any non-positive values.

`skewness` Skew, or the amount of non-symmetry, in a sample–before removing any non-positive values.

`kurtosis` Kurtosis in a sample–before removing any non-positive values.

`minimum` Minimum value of a sample–before removing any non-positive values.

`maximum` Maximum value of a sample–before removing any non-positive values.

`No.of.SDs` Number of standard deviations contained by a sample. That is: (maximum value - minimum value) / standard deviation.

Examples

```
## Not run:
#The following example uses FourSamples.rda, which is a data set included in the package.

data(file = "FourSamples.rda")
out<-descriptives(FourSamples)

## End(Not run)
```

`Dpit` *Distribution Pitting*

Description

Compares distributions with one another in terms of their fit to each sample in a dataset that contains multiple samples, as described in Joo, Aguinis, and Bradley (2017). Users can examine the fit of seven distributions per sample: pure power law, lognormal, exponential, power law with an exponential cutoff, normal, Poisson, and Weibull. Automation features allow the user to compare all distributions for all samples with a single command line, which creates a separate row containing results for each sample until the entire dataset has been analyzed. Automation features also skip over any unsuccessful calculations and continues analyzing the remainder of the samples. When calculations fail (e.g., sample size was too small), "NA" entries will be printed in the relevant cells of the results matrix before continuing with subsequent calculations.

Usage

```
Dpit(x)
```

Arguments

x	A data set

Details

For a given sample, the Dpit() function does not truncate (i.e., discard) data points that fall below a certain threshold, or xmin. More precisely, Dpit() sets xmin at the lowest positive number in the sample. This is because, theoretically, the function focuses on assessing the fit of distributions in their entirety rather than their tail ends. In other words, the goal of Dpit() is to conclude whether a sample itself follows a certain type of distribution, not whether the tail end of the sample follows a certain type of distribution.

Value

This function returns a data frame containing the complete detailed results of distribution pitting. In the data frame, each row corresponds to a sample, or data vector. That is, the first row in the data frame is sample #1, the second row is sample #2, etc. Each column in the data frame shows a distribution pitting statistic per sample. The data frame contains a large number of columns–as described in detail below:

`PLvCut.rawLR` A loglikelihood ratio (LR) quantifying the degree to which a pure power law (PL) fits the focal sample better than a power law with an exponential cutoff (Cut). So, a positive value means that the power law with an exponential cutoff fits worse, whereas a negative value means that the pure power law fits worse. This loglikelihood ratio is not normalized–hence, the word "raw" in the label.

`PLvCut.p` The p value associated with the previously mentioned loglikelihood ratio, or PLvCut.rawLR. The p value indicates the extent to which random fluctuations alone likely explain the presence of a non-zero loglikelihood ratio, such that loglikelihood ratio = 0 constitutes the null hypothesis. The lower the p value, the less likely that the loglikelihood ratio is simply due to chance.

`PLvWeib.rawLR` A loglikelihood ratio (LR) quantifying the degree to which a pure power law (PL) fits the focal sample better than a Weibull distribution (Weib). So, a positive value means that the Weibull distribution fits worse, whereas a negative value means that the pure power law fits worse.

`PLvWeib.normLR` Normalized value of the previously mentioned loglikelihood ratio, or PLvWeib.rawLR.

`PLvWeib.p` The p value associated with the previously mentioned loglikelihood ratio, or PLvWeib.normLR.

`PLvLgN.rawLR` A loglikelihood ratio (LR) quantifying the degree to which a pure power law (PL) fits the focal sample better than a lognormal distribution (LgN). So, a positive value means that the lognormal distribution fits worse, whereas a negative value means that the pure power law fits worse.

`PLvLgN.normLR` Normalized value of the previously mentioned loglikelihood ratio, or PLvLgN.rawLR.

`PLvLgN.p` The p value associated with the previously mentioned loglikelihood ratio, or PLvLgN.normLR.

`PLvExp.rawLR` A loglikelihood ratio (LR) quantifying the degree to which a pure power law (PL) fits the focal sample better than an exponential distribution (Exp). So, a positive value means that the exponential distribution fits worse, whereas a negative value means that the pure power law fits worse.

`PLvExp.normLR` Normalized value of the previously mentioned loglikelihood ratio, or PLvExp.rawLR.

`PLvExp.p` The p value associated with the previously mentioned loglikelihood ratio, or PLvExp.normLR.

`PLvPoi.rawLR` A loglikelihood ratio (LR) quantifying the degree to which a pure power law (PL) fits the focal sample better than a Poisson distribution (Poi). So, a positive value means that the Poisson distribution fits worse, whereas a negative value means that the pure power law fits worse.

`PLvPoi.normLR` Normalized value of the previously mentioned loglikelihood ratio, or PLvPoi.rawLR.

`PLvPoi.p` The p value associated with the previously mentioned loglikelihood ratio, or PLvPoi.normLR.

`CutvWeib.rawLR` A loglikelihood ratio (LR) quantifying the degree to which a power law with an exponential cutoff (Cut) fits the focal sample better than a Weibull distribution (Weib). So, a positive value means that the Weibull distribution fits worse, whereas a negative value means that the power law with an exponential cutoff fits worse.

`CutvWeib.normLR` Normalized value of the previously mentioned loglikelihood ratio, or CutvWeib.rawLR.

`CutvWeib.p` The p value associated with the previously mentioned loglikelihood ratio, or CutvWeib.normLR.

`CutvLgN.rawLR` A loglikelihood ratio (LR) quantifying the degree to which a power law with an exponential cutoff (Cut) fits the focal sample better than a lognormal distribution (LgN). So, a positive value means that the lognormal distribution fits worse, whereas a negative value means that the power law with an exponential cutoff fits worse.

`CutvLgN.normLR` Normalized value of the previously mentioned loglikelihood ratio, or CutvLgN.rawLR.

`CutvLgN.p` The p value associated with the previously mentioned loglikelihood ratio, or CutvLgN.normLR.

`CutvExp.rawLR` A loglikelihood ratio (LR) quantifying the degree to which a power law with an exponential cutoff (Cut) fits the focal sample better than an exponential distribution (Exp). So, a positive value means that the exponential distribution fits worse, whereas a negative value means that the power law with an exponential cutoff fits worse.

`CutvExp.normLR` Normalized value of the previously mentioned loglikelihood ratio, or CutvExp.rawLR.

`CutvExp.p` The p value associated with the previously mentioned loglikelihood ratio, or CutvExp.normLR.

`CutvPoi.rawLR` A loglikelihood ratio (LR) quantifying the degree to which a power law with an exponential cutoff (Cut) fits the focal sample better than a Poisson distribution (Poi). So, a positive value means that the Poisson distribution fits worse, whereas a negative value means that the power law with an exponential cutoff fits worse.

`CutvPoi.normLR` Normalized value of the previously mentioned loglikelihood ratio, or CutvPoi.rawLR.

`CutvPoi.p` The p value associated with the previously mentioned loglikelihood ratio, or CutvPoi.normLR.

`WeibvLgN.rawLR` A loglikelihood ratio (LR) quantifying the degree to which a Weibull distribution (Weib) fits the focal sample better than a lognormal distribution (LgN). So, a positive value means that the lognormal distribution fits worse, whereas a negative value means that the Weibull distribution fits worse.

`WeibvLgN.normLR` Normalized value of the previously mentioned loglikelihood ratio, or WeibvLgN.rawLR.

`WeibvLgN.p` The p value associated with the previously mentioned loglikelihood ratio, or WeibvLgN.normLR.

`WeibvExp.rawLR` A loglikelihood ratio (LR) quantifying the degree to which a Weibull distribution (Weib) fits the focal sample better than an exponential distribution (Exp). So, a positive value means that the exponential distribution fits worse, whereas a negative value means that the Weibull distribution fits worse.

`WeibvExp.normLR` Normalized value of the previously mentioned loglikelihood ratio, or WeibvExp.rawLR.

`WeibvExp.p` The p value associated with the previously mentioned loglikelihood ratio, or WeibvExp.normLR.

`WeibvPoi.rawLR` A loglikelihood ratio (LR) quantifying the degree to which a Weibull distribution (Weib) fits the focal sample better than a Poisson distribution (Poi). So, a positive value means that the Poisson distribution fits worse, whereas a negative value means that the Weibull distribution fits worse.

`WeibvPoi.normLR` Normalized value of the previously mentioned loglikelihood ratio, or WeibvPoi.rawLR.

`WeibvPoi.p` The p value associated with the previously mentioned loglikelihood ratio, or WeibvPoi.normLR.

`LgNvExp.rawLR` A loglikelihood ratio (LR) quantifying the degree to which a lognormal distribution (LgN) fits the focal sample better than an exponential distribution (Exp). So, a positive value means that the exponential distribution fits worse, whereas a negative value means that the lognormal distribution fits worse.

`LgNvExp.normLR` Normalized value of the previously mentioned loglikelihood ratio, or LgNvExp.rawLR.

`LgNvExp.p` The p value associated with the previously mentioned loglikelihood ratio, or LgNvExp.normLR.

`LgNvPoi.rawLR` A loglikelihood ratio (LR) quantifying the degree to which a lognormal distribution (LgN) fits the focal sample better than a Poisson distribution (Poi). So, a positive value means that the Poisson distribution fits worse, whereas a negative value means that the lognormal distribution fits worse.

`LgNvPoi.normLR` Normalized value of the previously mentioned loglikelihood ratio, or LgNvPoi.rawLR.

`LgNvPoi.p` The p value associated with the previously mentioned loglikelihood ratio, or LgNvPoi.normLR.

`ExpvPoi.rawLR` A loglikelihood ratio (LR) quantifying the degree to which an exponential distribution (Exp) fits the focal sample better than a Poisson distribution (Poi). So, a positive value means that the Poisson distribution fits worse, whereas a negative value means that the exponential distribution fits worse.

`ExpvPoi.normLR` Normalized value of the previously mentioned loglikelihood ratio, or ExpvPoi.rawLR.

`ExpvPoi.p` The p value associated with the previously mentioned loglikelihood ratio, or ExpvPoi.normLR.

`NvPL.rawLR` A loglikelihood ratio (LR) quantifying the degree to which a normal distribution (N) fits the focal sample better than a pure power law (PL). So, a positive value means that the pure power law fits worse, whereas a negative value means that the normal distribution fits worse.

`NvPL.normLR` Normalized value of the previously mentioned loglikelihood ratio, or NvPL.rawLR.

NvPL.p The p value associated with the previously mentioned loglikelihood ratio, or NvPL.normLR.

NvCut.rawLR A loglikelihood ratio (LR) quantifying the degree to which a normal distribution (N) fits the focal sample better than a power law with an exponential cutoff (Cut). So, a positive value means that the power law with an exponential cutoff fits worse, whereas a negative value means that the normal distribution fits worse.

NvCut.normLR Normalized value of the previously mentioned loglikelihood ratio, or NvCut.rawLR.

NvCut.p The p value associated with the previously mentioned loglikelihood ratio, or NvCut.normLR.

NvWeib.rawLR A loglikelihood ratio (LR) quantifying the degree to which a normal distribution (N) fits the focal sample better than a Weibull disribution (Weib). So, a positive value means that the Weibull distribution fits worse, whereas a negative value means that the normal distribution fits worse.

NvWeib.normLR Normalized value of the previously mentioned loglikelihood ratio, or NvWeib.rawLR.

NvWeib.p The p value associated with the previously mentioned loglikelihood ratio, or NvWeib.normLR.

NvLgN.rawLR A loglikelihood ratio (LR) quantifying the degree to which a normal distribution (N) fits the focal sample better than a lognormal disribution (LgN). So, a positive value means that the lognormal distribution fits worse, whereas a negative value means that the normal distribution fits worse.

NvLgN.normLR Normalized value of the previously mentioned loglikelihood ratio, or NvLgN.rawLR.

NvLgN.p The p value associated with the previously mentioned loglikelihood ratio, or NvLgN.normLR.

NvExp.rawLR A loglikelihood ratio (LR) quantifying the degree to which a normal distribution (N) fits the focal sample better than an exponential disribution (Exp). So, a positive value means that the exponential distribution fits worse, whereas a negative value means that the normal distribution fits worse.

NvExp.normLR Normalized value of the previously mentioned loglikelihood ratio, or NvExp.rawLR.

NvExp.p The p value associated with the previously mentioned loglikelihood ratio, or NvExp.normLR.

NvPoi.rawLR A loglikelihood ratio (LR) quantifying the degree to which a normal distribution (N) fits the focal sample better than a Poisson disribution (Poi). So, a positive value means that the Poisson distribution fits worse, whereas a negative value means that the normal distribution fits worse.

NvPoi.normLR Normalized value of the previously mentioned loglikelihood ratio, or NvPoi.rawLR.

NvPoi.p The p value associated with the previously mentioned loglikelihood ratio, or NvPoi.normLR.

Note

The Dpit() function is based on the R code found at: http://tuvalu.santafe.edu/~aaronc/powerlaws/. In particular, we borrowed heavily from Cosma R. Shalizi's code. But our code differs from the aforementioned code in mainly three ways. First, the Dpit() function sets xmin at the lowest positive number in a sample to assess the fit of distributions in their entirety rather than their tail ends. Second, our code allows for the comparison of non-pure power law distributions with one another. Third, our code includes automation features that allow the user to compare all distributions for all samples with a single command line, which creates a separate row containing results for each sample until the entire dataset has been analyzed. Automation features clean each sample by removing missing cases and non-positive values that lead to incalculable expressions (e.g., the log of zero is undefined). Automation features also skip over any unsuccessful calculations and continues analyzing the remainder of the samples. When calculations fail (e.g., sample size was too small), "NA" entries will be printed in the relevant cells of the results matrix before continuing with subsequent calculations.

References

Clauset, A., Shalizi, C. R., & Newman, M. E. J. 2009. Power-law distributions in empirical data. SIAM Review, 51, 661-703. Available at: http://arxiv.org/abs/0706.1062

Joo, H., Aguinis, H., & Bradley, K. J. 2017. Not all nonnormal distributions are created equal: Improved theoretical and measurement precision. Journal of Applied Psychology. Advance online publication. doi: 10.1037/apl0000214

Examples

```
## Not run:
#The following example uses FourSamples.rda, which is a data set included in the package.

data(file = "FourSamples.rda")
out<-Dpit(FourSamples)

#Full results are shown in Table 4 in Joo, Aguinis, and Bradley (2017)

#Next, to draw conclusions regarding the fit of a certain type of distribution per sample,
#we suggest that users implement three decision rules, which are described in
#the Analysis section in Joo, Aguinis, and Bradley (2017).

#Conclusions regarding the fit of distributions to the four samples in the focal data set
#--after applying the three decision rules--can be found in
#the first two and last two rows in Table 3, in Joo, Aguinis, and Bradley (2017).

## End(Not run)
```

FourSamples *A part of the data set used in Joo, Aguinis, & Bradley (2017).*

Description

A part of the data set used in Joo, Aguinis, & Bradley (2017).

Usage

```
data(FourSamples)
```

Format

A data frame consisting of four variables with varying number of observations.

v1 number of publications in top five journals in the field of Agriculture.

v2 number of publications in top five journals in the field of Agronomy.

v228 number electrical fixtures assembled by assemblers.

v229 wirer's ratio of production time per unit assembled to standard.

Details

Running the Dpit() function on the FourSamples dataset will produce the results reported in Table 4 in Joo, Aguinis, & Bradley (2017). Though the precise results the user obtains may very slightly differ from those in Joo et al.'s Table 4 due to statistical fluctuations, substantive conclusions remain the same regardless of such fluctuations.

Author(s)

Harry Joo, Herman Aguinis, Kyle J. Bradley

Maintainer: Harry Joo <harryjoo19@gmail.com>

Source

Joo, H., Aguinis, H., & Bradley, K. J. 2017. Not all nonnormal distributions are created equal: Improved theoretical and measurement precision. Journal of Applied Psychology. Advance online publication. doi: 10.1037/apl0000214

Index