Package 'BeyondBenford'

July 24, 2020

Type Package

Title Compare the Goodness of Fit of Benford's and Blondeau Da Silva's Digit Distributions to a Given Dataset

Version 1.4

Date 2020-07-24

Author Blondeau Da Silva Stephane

Maintainer Blondeau Da Silva Stephane <Stephane.Blondeau-Da-Silva@ac-limoges.fr>

Imports ggplot2 (>= 2.1.0)

Description Allows to compare the goodness of fit of Benford's and Blondeau Da Silva's digit distributions in a dataset. It is used to check whether the data distribution is consistent with theoretical distributions highlighted by Blondeau Da Silva or not (through the dat.distr() function): this ideal theoretical distribution must be at least approximately followed by the data for the use of Blondeau Da Silva's model to be well-founded. It also enables to plot histograms of digit distributions, both observed in the dataset and given by the two theoretical approaches (with the digit.ditr() function). Finally, it proposes to quantify the goodness of fit via Pearson's chi-squared test (with the chi2() function).

License GPL-2

NeedsCompilation no

Repository CRAN

Date/Publication 2020-07-24 17:22:10 UTC

R topics documented:

`BeyondBenford-package` *Compare the goodness of fit of Benford's and Blondeau Da Silva's digit distributions to a given dataset*

Description

The purpose of this package is to compare the goodness of fit of Benford's and Blondeau Da Silva's digit distributions in a dataset. The package is used to check whether the data distribution is consistent with theoretical distributions highlighted by Blondeau Da Silva or not (through the function 'dat.distr'): this ideal theoretical distribution must be at least approximately followed by the data for the use of Blondeau Da Silva's model to be well-founded. It also enables to plot histograms of digit distributions, both observed in the dataset and given by the two theoretical approaches (with the function 'digit.ditr'). Finally, it proposes to quantify the goodness of fit via Pearson's chi-squared test (with the function 'chi2').

Author(s)

Blondeau Da Silva

Maintainer: Blondeau Da Silva

References

F. Benford (1938). The law of anomalous numbers. Proceedings of the American Philosophical Society, 78:127-131.

A. Berger and T. Hill (2015). An introduction to Benford's Law. Princeton University Press, Princeton, NJ. ISSN/ISBN: 978-0-691-16306-2.

S. Blondeau Da Silva (2020). Benford or not Benford: a systematic but not always well-founded use of an elegant law in experimental fields. Communications in Mathematics and Statistics, 8:167-201. doi: 10.1007/s40304-018-00172-1.

S. Blondeau Da Silva (2018). Benford or not Benford: new results on digits beyond the first. https://arxiv.org/abs/1805.01291.

S. Blondeau Da Silva (2019). BeyondBenford: An R Package to Determine Which of Benford's or BDS's Distributions is the Most Relevant. https://arxiv.org/abs/1910.06104. https://hal.archives-ouvertes.fr/hal-02310013.

T. Hill (1995). The significant-digit phenomenon. The American Mathematical Monthly, 102(4):322-327.

S. J. Miller, editor (2015). Benford's Law: Theory and Applications. Princeton University Press, Princeton, NJ. ISSN/ISBN: 978-0-691-14761-1.

R. Newcomb (1881). Note on the frequency of use of the different digits in natural numbers. American Journal of Mathematics, 4:39-40.

K. Pearson (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine, 50(302):157-175.

Examples

```
data(address_PierreBuffiere)
data(census)
data(address_AixesurVienne)

dat.distr(address_PierreBuffiere,nchi=6)
dat.distr(census,theor=0,nclass=100,dig=3)
dat.distr(address_AixesurVienne,upbound=75)

digit.distr(address_AixesurVienne,mod="ben&blo",lwbound=5,No.sd=1,Sd.pr=1)
digit.distr(address_PierreBuffiere,mod="blo",dig=2)

chi2(address_PierreBuffiere,dig=2,pval=1)
chi2(address_PierreBuffiere,dig=2,pval=1,mod="blo")
```

---

address_AixesurVienne *Street addresses of Aixe-sur-Vienne*

Description

Street addresses of Aixe-sur-Vienne, a town of approximately 5800 inhabitants in Haute-Vienne (France).

Usage

```
data(address_AixesurVienne)
```

Format

A factor containing all 1911 existing street address numbers.

Source

From an open platform for French public data:

https://www.data.gouv.fr/fr/datasets/base-d-adresses-nationale-ouverte-bano/ (http://bano.openstreetmap.fr/data/).

address_Limoges *Street addresses of Limoges*

Description

Street addresses of Limoges, a city of approximately 133600 inhabitants in Haute-Vienne (France).

Usage

```
data(address_Limoges)
```

Format

A factor containing all 35975 existing street address numbers.

Source

From an open platform for French public data:

https://www.data.gouv.fr/fr/datasets/base-d-adresses-nationale-ouverte-bano/ (http://bano.openstreetmap.fr/data/).

address_PierreBuffiere *Street addresses of Pierre-Buffiere*

Description

Street addresses of Pierre-Buffiere, a small town of approximately 1200 inhabitants in Haute-Vienne (France).

Usage

```
data(address_PierreBuffiere)
```

Format

A factor containing all 346 existing street address numbers.

Source

From an open platform for French public data:

https://www.data.gouv.fr/fr/datasets/base-d-adresses-nationale-ouverte-bano/ (http://bano.openstreetmap.fr/data/).

`Benf.val` *Benford's values*

Description

The function returns Benford's probability that a figure is at a given position.

Usage

```
Benf.val(fig, dig = 1)
```

Arguments

`fig`	The considered figure.
`dig`	The chosen position of the digit (from the left).

Value

The function returns Benford's probability.

Author(s)

Blondeau Da Silva St\'ephane

References

F. Benford (1938). The law of anomalous numbers. Proceedings of the American Philosophical Society, 78:127-131.

T. Hill (1995). The significant-digit phenomenon. The American Mathematical Monthly, 102(4):322-327.

R. Newcomb (1881). Note on the frequency of use of the different digits in natural numbers. American Journal of Mathematics, 4:39-40.

Examples

```
Benf.val(7, dig = 2)
```

`Blon.val` *Blondeau Da Silva's values*

Description

The function returns Blondeau Da Silva's probability that a figure is at a given position (once the associated lower and upper bounds have been specified) and, if requested, the associated standard deviation.

Usage

```
Blon.val(lwbound = 10^(dig - 1), upbound, fig, dig = 1, sd = 0)
```

Arguments

`lwbound`	A positive integer, which characterizes the data. All (or most) of the data are greater than this "lower bound".
`upbound`	A positive integer, which characterizes the data. All (or most) of the data are lower than this "upper bound".
`fig`	The considered figure.
`dig`	The chosen position of the digit (from the left).
`sd`	If sd=0, only the probability is returned. Else, the function returns a dataframe containing the probability and the standard deviation of the expected digit frequency.

Value

The function returns Blondeau Da Silva's probability and, if requested, its standard deviation.

Author(s)

Blondeau Da Silva St\'ephane

References

S. Blondeau Da Silva (2020). Benford or not Benford: a systematic but not always well-founded use of an elegant law in experimental fields. Communications in Mathematics and Statistics, 8:167-201. doi: 10.1007/s40304-018-00172-1.

S. Blondeau Da Silva (2018). Benford or not Benford: new results on digits beyond the first. https://arxiv.org/abs/1805.01291.

Examples

```
Blon.val(171,825, 5, dig = 3)
```

Blon.val.sd *Blondeau Da Silva's standard deviations*

Description

The function returns the Blondeau Da Silva's standard deviation of the frequency of a digit at a given position (once the associated lower and upper bounds have been specified).

Usage

```
Blon.val.sd(lwbound = 10^(dig - 1), upbound, fig, dig = 1)
```

Arguments

lwbound	A positive integer, which characterizes the data. All (or most) of the data are greater than this "lower bound".
upbound	A positive integer, which characterizes the data. All (or most) of the data are lower than this "upper bound".
fig	The considered figure.
dig	The chosen position of the digit (from the left).

Value

The function returns Blondeau Da Silva's standard deviations of digit frequencies.

Author(s)

Blondeau Da Silva St\'ephane

Examples

```
Blon.val.sd(171,825, 5, dig = 3)
```

census *Alabama census*

Description

Populations in Alabama cities and towns.

Usage

```
data(census)
```

Format

A data frame containing the populations of all 460 Alabama cities or towns (dimension: one row and 460 columns).

Source

From the United States Census Bureau:

https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk.

---

chi2 *Pearson's chi-squared test*

Description

It is a test of goodness of fit to find out whether the distribution of first (second, third or fourth) digit in the studied data differs from two theoretical distributions (that of Benford and that of Blondeau Da Silva) or not. The null hypothesis states that the studied distribution is consistent with the considered theoretical distribution.

Usage

```
chi2(dat, mod = "ben", lwbound = max(floor(min(abs(dat))) + 1, (10^(dig - 1))),
upbound = ceiling(max(dat)), dig = 1, pval = 0)
```

Arguments

dat	The considered dataset, a data frame containing non-zero real numbers.
mod	A character string. If mod="ben", the theorical distribution considered is that of Benford, else it is Blondeau Da Silva's ones which is chosen.
lwbound	A positive integer, which characterizes the data. All (or most) of the data are greater than this "lower bound".
upbound	A positive integer, which characterizes the data. All (or most) of the data are lower than this "upper bound".
dig	The chosen position of the digit (from the left).
pval	If pval=0, the p-value is not returned, else it is available.

Value

A data frame containing the Pearson chi-squared statistic (and the associated p-value if requested).

Note

This warning message can appear: NAs introduced during the automatic conversion. This is due to the fact that some data are not numerical in the entered dataset. Non numerical values and zeros are not counted.

Author(s)

Blondeau Da Silva St\'ephane

References

K. Pearson (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine, 50(302):157-175.

Examples

```
data(address_PierreBuffiere)
chi2(address_PierreBuffiere,dig=2,pval=1)
chi2(address_PierreBuffiere,dig=2,pval=1,mod="blo")
```

---

dat.distr *Data distribution*

Description

The function returns the histogram of the data. It can also plot one of the Blondeau Da Silva's theoretical distributions (thanks to a lower and an upper bound): this ideal theoretical distribution must be at least approximately followed by the data for the use of Blondeau Da Silva's model to be well-founded. A specific chi-squared statistic can also be computed to find out whether the data distribution is consistent with the theoretical distribution or not.

Usage

```
dat.distr(dat, xlab = "Data", ylab = "Frequency", main = "Distribution of data",
theor = TRUE, nclass = 50, col = "lightblue", conv = 0,
lwbound = max(floor(min(abs(dat))) + 1, (10^(dig - 1))),
upbound = ceiling(max(dat)), dig = 1, colt = "red", ylim = NULL, border = "blue",
nchi = 0, legend = TRUE, bg.leg = "gray85")
```

Arguments

dat	The considered dataset, a data frame containing non-zero real numbers.
xlab	The x-axis label.
ylab	The y-axis label.
main	The title of the graph.
theor	If theor=TRUE Blondeau Da Silva's theoretical distribution is plotted, otherwise only the histogram is represented.
nclass	A strictly positive integer: the number of classes in the histogram.
col	The color used to fill the bars of the histogram. NULL yields unfilled bars.

conv	If conv=1, all values of the dataset are multiplied by 10^k where k is the smallest positive integer such that all non-zero numerical values in the newly multiplied data frame have an absolute value greater than or equal to 1.
lwbound	A positive integer, which characterizes the data. All (or most) of the data are greater than this "lower bound".
upbound	A positive integer, which characterizes the data. All (or most) of the data are lower than this "upper bound".
dig	The chosen position of the digit (from the left).
colt	The color used to plot Blondeau Da Silva's theoretical distribution.
ylim	A two-components vector: the range of y values.
border	The color of the border around the bars.
nchi	A positive integer: the number of classes for values from 10^(p-1) to max(max(data),upbound). If nchi>0, the function returns the chi-squared statistic (with nchi-1 degrees of freedom) of goodness of fit determined by the different classes. The null hypothesis states that the studied distribution is consistent with the considered theoretical distribution.
legend	If legend=TRUE, the legend is displayed.
bg.leg	The background color for the legend box.

Value

The histogram of the data along with optional Blondeau Da Silva's theoretical distributions and a data frame containing the chi-squared statistic and its associated p-value if requested.

Note

This warning message can appear: NAs introduced during the automatic conversion. This is due to the fact that some data are not numerical in the entered dataset. Non numerical values and zeros are not counted.

Author(s)

Blondeau Da Silva St\'ephane

References

S. Blondeau Da Silva (2020). Benford or not Benford: a systematic but not always well-founded use of an elegant law in experimental fields. Communications in Mathematics and Statistics, 8:167-201. doi: 10.1007/s40304-018-00172-1.

S. Blondeau Da Silva (2018). Benford or not Benford: new results on digits beyond the first. https://arxiv.org/abs/1805.01291.

K. Pearson (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine, 50(302):157-175.

Examples

```
data(address_PierreBuffiere)
dat.distr(address_PierreBuffiere,nchi=6)

data(census)
dat.distr(census,theor=0,nclass=100,dig=3)

data(address_AixesurVienne)
dat.distr(address_AixesurVienne,lwbound=3,upbound=75)
```

`digit.distr` *Distribution of figures in a given position*

Description

The function returns histograms of distribution of figures in a given position: (i) in the dataset, (ii) due to Benford, (iii) due to Blondeau Da Silva. Error bars can be added to the plotted histogram.

Usage

```
digit.distr(dat, mod = "ben", lwbound = max(floor(min(abs(dat))) + 1, (10^(dig - 1))),
upbound = ceiling(max(dat)), dig = 1, col = c("#E69F00", "#999999"),
colbl = c("#AAFFAA", "#999999"), colbebl = c("#E69F00", "#AAFFAA", "#999999"),
main = "Distribution of digits", No.sd = 0, Sd.pr = 0)
```

Arguments

`dat`	The considered dataset, a data frame containing non-zero real numbers.
`mod`	A character string. If mod="ben", the data histogram and that of Benford are displayed, if mod="ben&blo", the data histogram, that of Benford and that of Blondeau Da Silva are plotted, and otherwise the data histogram and that of Blondeau Da Silva are given.
`lwbound`	A positive integer, which characterizes the data. All (or most) of the data are greater than this "lower bound".
`upbound`	A positive integer, which characterizes the data. All (or most) of the data are lower than this "upper bound".
`dig`	The chosen position of the digit (from the left).
`col`	A vector containing two colors used to fill the bars of the histogram, if mod="ben".
`colbl`	A vector containing two colors used to fill the bars of the histogram, if both the data histogram and Blondeau Da Silva's histogram are plotted.
`colbebl`	A vector containing three colors used to fill the bars of the histogram, if mod="ben&blo".
`main`	The title of the graph.
`No.sd`	The positive decimal number of standard deviation that defines the confidence intervals i.e. the error bars. If No.sd=0, no error bars are drawn.
`Sd.pr`	If Sd.pr=1, error bars for proportions are plotted (with No.sd standard deviation confidence intervals). If Sd.pr=0, they are not plotted.

Value

Histograms of distribution of figures in a given position: (i) in the dataset, (ii) due to Benford, (iii) due to Blondeau Da Silva.

Note

This warning message can appear: NAs introduced during the automatic conversion. This is due to the fact that some data are not numerical in the entered dataset. Non numerical values and zeros are not counted.

Author(s)

Blondeau Da Silva St\'ephane

References

F. Benford (1938). The law of anomalous numbers. Proceedings of the American Philosophical Society, 78:127-131.

S. Blondeau Da Silva (2020). Benford or not Benford: a systematic but not always well-founded use of an elegant law in experimental fields. Communications in Mathematics and Statistics, 8:167-201. doi: 10.1007/s40304-018-00172-1.

S. Blondeau Da Silva (2018). Benford or not Benford: new results on digits beyond the first. https://arxiv.org/abs/1805.01291.

T. Hill (1995). The significant-digit phenomenon. The American Mathematical Monthly, 102(4):322-327.

R. Newcomb (1881). Note on the frequency of use of the different digits in natural numbers. American Journal of Mathematics, 4:39-40.

Examples

```
data(address_AixesurVienne)
digit.distr(address_AixesurVienne,mod="ben&blo",lwbound=2,No.sd=1, Sd.pr=1)

data(address_PierreBuffiere)
digit.distr(address_PierreBuffiere,mod="blo",dig=2)
```

---

obs.numb.dig *Frequency of each figure at a given position*

Description

The function returns the frequencies of each figure at a given position in the considered dataset.

Usage

```
obs.numb.dig(dat, dig = 1)
```

Arguments

dat	The considered dataset, a data frame containing non-zero real numbers.
dig	The chosen position of the digit (from the left).

Value

A vector containing the frequencies of each figure in ascending order. Its length is 9 if dig=1 (the figures ranging from 1 to 9) and 10 if dig>1 (the figures ranging from 0 to 9).

Note

This warning message can appear: NAs introduced during the automatic conversion. This is due to the fact that some data are not numerical in the entered dataset. Non numerical values and zeros are not counted.

Author(s)

Blondeau Da Silva St\'ephane

Examples

```
data(census)
obs.numb.dig(census, dig=2)
```

---

prep *Data set preparation*

Description

The function returns a prepared data frame that can be used by the other functions of the package.

Usage

```
prep(dat)
```

Arguments

dat	The considered dataset, a data frame.

Value

The prepared dataset, a data frame containing only numerical values: character strings and NA values are all replaced by 0.

Author(s)

Blondeau Da Silva St\'ephane

`theor.distr.val` *Theoretical distribution*

Description

The function returns the theorical probability distribution described by Blondeau Da Silva for data. If the dataset follows this particular distribution well enough, it enables not to use Benford's values of first (second, third or fourth) digit distribution but rather Blondeau Da Silva's ones. The distribution depends on a lower and an upper bound, which characterize the data.

Usage

```
theor.distr.val(lwbound, upbound, dig = 1)
```

Arguments

`lwbound`	A positive integer, which characterizes the data. All (or most) of the data are greater than this "lower bound".
`upbound`	A positive integer, which characterizes the data. All (or most) of the data are lower than this "upper bound".
`dig`	The chosen position of the digit (from the left).

Value

The function returns a vector contening the probability distribution of the model determined by the upper bound value.

Author(s)

Blondeau Da Silva St\'ephane

References

S. Blondeau Da Silva (2020). Benford or not Benford: a systematic but not always well-founded use of an elegant law in experimental fields. Communications in Mathematics and Statistics, 8:167-201. doi: 10.1007/s40304-018-00172-1.

S. Blondeau Da Silva (2018). Benford or not Benford: new results on digits beyond the first. https://arxiv.org/abs/1805.01291.

Examples

```
theor.distr.val(10,27)
```

Index