

Package ‘corpora’

January 2, 2012

Type Package

Title Statistics for corpus linguists

Version 0.3-2.1

Depends R (>= 2.0.0)

Date 2009-02-25

Author Stefan Evert <stefan.evert@uos.de>

Maintainer Stefan Evert <stefan.evert@uos.de>

Description Utility functions for the statistical analysis of corpus frequency data

Encoding latin1

License GPL

URL <http://purl.org/stefan.evert/SIGIL/>

Repository CRAN

Date/Publication 2009-02-25 08:29:17

R topics documented:

binom.pval	2
BNCcomparison	3
BNCdomains	4
BNCInChargeOf	4
chisq	5
chisq.pval	7
cont.table	8
fisher.pval	9
prop.cint	10
rel.risk.cint	11
VSS	13
z.score	14
z.score.pval	15

binom.pval	<i>P-values of the binomial test for frequency counts (corpora)</i>
------------	---

Description

This function computes the p-value of a binomial test for frequency counts. In the two-sided case, a fast approximation is used that may be inaccurate for small samples.

Usage

```
binom.pval(k, n, p = 0.5,  
           alternative = c("two.sided", "less", "greater"))
```

Arguments

k	frequency of a type in the corpus (or an integer vector of frequencies)
n	number of tokens in the corpus, i.e. sample size (or an integer vector specifying the sizes of different samples)
p	null hypothesis, giving the assumed proportion of this type in the population (or a vector of proportions for different types and/or different populations)
alternative	a character string specifying the alternative hypothesis; must be one of <code>two.sided</code> (default), <code>less</code> or <code>greater</code>

Details

When `alternative` is `two.sided`, a fast approximation of the two-sided p-value is used (multiplying the appropriate single-sided tail probability by two), which may be inaccurate for small samples. Unlike the exact algorithm of [binom.test](#), this implementation can be applied to large frequencies and samples without a serious impact on performance.

Value

The p-value of a binomial test applied to the given data (or a vector of p-values).

Author(s)

Stefan Evert

See Also

[z.score.pval](#), [prop.cint](#)

BNCcomparison

Comparison of written and spoken frequencies (BNC)

Description

This data set compares the frequencies of 60 selected nouns in the written and spoken parts of the British National Corpus, World Edition (BNC). Nouns were chosen from three frequency bands, namely the 20 most frequent nouns in the corpus, 20 nouns with approximately 1000 occurrences, and 20 nouns with approximately 100 occurrences.

See Aston & Burnard (1998) for more information about the BNC, or go to <http://www.natcorp.ox.ac.uk/>.

Usage

data(BNCcomparison)

Format

A data set with 61 rows and the following columns:

noun: lemmatised noun (aka stem form)

written: frequency in the written part of the BNC

spoken: frequency in the spoken part of the BNC

Details

In addition to the 60 nouns, the data set contains a column labelled OTHER, which represents the total frequency of all other nouns in the BNC. This value is needed in order to calculate the sample sizes of the written and spoken part for frequency comparison tests.

Author(s)

Stefan Evert (<http://purl.org/stefan.evert>)

References

Aston, Guy and Burnard, Lou (1998). *The BNC Handbook*. Edinburgh University Press, Edinburgh. See also the BNC homepage at <http://www.natcorp.ox.ac.uk/>.

BNCdomains

Distribution of domains in the British National Corpus (BNC)

Description

This data set gives the number of documents and tokens in each of the 18 domains represented in the British National Corpus, World Edition (BNC). See Aston & Burnard (1998) for more information about the BNC and the domain classification, or go to <http://www.natcorp.ox.ac.uk/>.

Usage

data(BNCdomains)

Format

A data set with 19 rows and the following columns:

domain: name of the respective domain in the BNC

documents: number of documents from this domain

tokens: total number of tokens in all documents from this domain

Details

For one document in the BNC, the domain classification is missing. This document is represented by the code Unlabeled in the data set.

Author(s)

Marco Baroni (<baroni@sslmit.unibo.it>)

References

Aston, Guy and Burnard, Lou (1998). *The BNC Handbook*. Edinburgh University Press, Edinburgh. See also the BNC homepage at <http://www.natcorp.ox.ac.uk/>.

BNCInChargeOf

Collocations of the phrase "in charge of" (BNC)

Description

This data set lists collocations (in the sense of Sinclair 1991) of the phrase *in charge of* found in the British National Corpus, World Edition (BNC). A span size of 3 and a frequency threshold of 5 were used, i.e. all words that occur at least five times within a distance of three tokens from the key phrase *in charge of* are listed as collocates. Note that collocations were not allowed to cross sentence boundaries.

See Aston & Burnard (1998) for more information about the BNC, or go to <http://www.natcorp.ox.ac.uk/>.

Usage

```
data(BNCInChargeOf)
```

Format

A data set with 250 rows and the following columns:

collocate: a collocate of the key phrase *in charge of* (word form)

f.in: occurrences of the collocate within a distance of 3 tokens from the key phrase, i.e. *inside* the span

N.in: total number of tokens inside the span

f.out: occurrences of the collocate *outside* the span

N.out: total number of tokens outside the span

Details

Punctuation, numbers and any words containing non-alphabetic characters (except for -) were not considered as potential collocates. Likewise, the number of tokens inside / outside the span given in the columns N.in and N.out only includes simple alphabetic word forms.

Author(s)

Stefan Evert (<http://purl.org/stefan.evert>)

References

Aston, Guy and Burnard, Lou (1998). *The BNC Handbook*. Edinburgh University Press, Edinburgh. See also the BNC homepage at <http://www.natcorp.ox.ac.uk/>.

Sinclair, John (1991). *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.

chisq	<i>Pearson's chi-squared statistic for frequency comparisons (corpora)</i>
-------	--

Description

This function computes Pearson's chi-squared statistic (often written as X^2) for frequency comparison data, with or without Yates' continuity correction. The implementation is based on the formula given by Evert (2004, 82).

Usage

```
chisq(k1, n1, k2, n2, correct = TRUE, one.sided=FALSE)
```

Arguments

k1	frequency of a type in the first corpus (or an integer vector of type frequencies)
n1	the sample size of the first corpus (or an integer vector specifying the sizes of different samples)
k2	frequency of the type in the second corpus (or an integer vector of type frequencies, in parallel to k1)
n2	the sample size of the second corpus (or an integer vector specifying the sizes of different samples, in parallel to n1)
correct	if TRUE, apply Yates' continuity correction (default)
one.sided	if TRUE, compute the <i>signed square root</i> of X^2 as a statistic for a one-sided test (see details below; the default value is FALSE)

Details

The X^2 values returned by this function are identical to those computed by `chisq.test`. Unlike the latter, `chisq` accepts vector arguments so that a large number of frequency comparisons can be carried out with a single function call.

The one-sided test statistic (for `one.sided=TRUE`) is the signed square root of X^2 . It is positive for $k_1/n_1 > k_2/n_2$ and negative for $k_1/n_1 < k_2/n_2$. Note that this statistic has a *standard normal distribution* rather than a chi-squared distribution under the null hypothesis of equal proportions.

Value

The chi-squared statistic X^2 corresponding to the specified data (or a vector of X^2 values). This statistic has a *chi-squared distribution* with $df = 1$ under the null hypothesis of equal proportions.

Author(s)

Stefan Evert

References

Evert, Stefan (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart. Published in 2005, URN urn:nbn:de:bsz:93-opus-23714. Available from <http://www.collocations.de/phd.html>.

See Also

[chisq.pval](#), [chisq.test](#), [cont.table](#)

chisq.pval	<i>P-values of Pearson's chi-squared test for frequency comparisons (corpora)</i>
------------	---

Description

This function computes the p-value of Pearson's chi-squared test for the comparison of corpus frequency counts (under the null hypothesis of equal population proportions). It is based on the chi-squared statistic X^2 implemented by the [chisq](#) function.

Usage

```
chisq.pval(k1, n1, k2, n2, correct = TRUE,  
           alternative = c("two.sided", "less", "greater"))
```

Arguments

k1	frequency of a type in the first corpus (or an integer vector of type frequencies)
n1	the sample size of the first corpus (or an integer vector specifying the sizes of different samples)
k2	frequency of the type in the second corpus (or an integer vector of type frequencies, in parallel to k1)
n2	the sample size of the second corpus (or an integer vector specifying the sizes of different samples, in parallel to n1)
correct	if TRUE, apply Yates' continuity correction (default)
alternative	a character string specifying the alternative hypothesis; must be one of <code>two.sided</code> (default), <code>less</code> or <code>greater</code>

Details

The p-values returned by this functions are identical to those computed by [chisq.test](#) (two-sided only) and [prop.test](#) (one-sided and two-sided) for two-by-two contingency tables.

Value

The p-value of Pearson's chi-squared test applied to the given data (or a vector of p-values).

Author(s)

Stefan Evert

See Also

[chisq](#), [fisher.pval](#), [chisq.test](#), [prop.test](#), [rel.risk.cint](#)

`cont.table`*Build contingency table for frequency comparison (corpora)*

Description

This is a convenience function which constructs contingency tables needed for frequency comparisons with `chisq.test` and `fisher.test`.

Usage

```
cont.table(k1, n1, k2, n2)
```

Arguments

<code>k1</code>	frequency of a type in the first corpus
<code>n1</code>	the size of the first corpus (sample size)
<code>k2</code>	frequency of the type in the second corpus
<code>n2</code>	the size of the second corpus (sample size)

Details

Because matrices cannot easily be vectorized, this function does not accept vector arguments and will abort with an error message in this case.

Value

A numeric matrix containing the two-by-two contingency table for the frequency comparison.

Author(s)

Stefan Evert

See Also

`chisq.test`, `fisher.test`

`fisher.pval`*P-values of Fisher's exact test for frequency comparisons (corpora)*

Description

This function computes the p-value of Fisher's exact test (Fisher 1934) for the comparison of corpus frequency counts (under the null hypothesis of equal population proportions). In the two-sided case, a fast approximation is used that may be inaccurate for small samples.

Usage

```
fisher.pval(k1, n1, k2, n2,  
            alternative = c("two.sided", "less", "greater"))
```

Arguments

<code>k1</code>	frequency of a type in the first corpus (or an integer vector of type frequencies)
<code>n1</code>	the sample size of the first corpus (or an integer vector specifying the sizes of different samples)
<code>k2</code>	frequency of the type in the second corpus (or an integer vector of type frequencies, in parallel to <code>k1</code>)
<code>n2</code>	the sample size of the second corpus (or an integer vector specifying the sizes of different samples, in parallel to <code>n1</code>)
<code>alternative</code>	a character string specifying the alternative hypothesis; must be one of <code>two.sided</code> (default), <code>less</code> or <code>greater</code>

Details

When `alternative` is `two.sided`, a fast approximation of the two-sided p-value is used (multiplying the appropriate single-sided tail probability by two), which may be inaccurate for small samples. Unlike the exact algorithm of [fisher.test](#), this implementation is memory-efficient and can be applied to large samples and/or large frequency counts.

For one-sided tests, the p-values returned by this functions are identical to those computed by [fisher.test](#) on two-by-two contingency tables.

Value

The p-value of Fisher's exact test applied to the given data (or a vector of p-values).

Author(s)

Stefan Evert

References

Fisher, R. A. (1934). *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh, 2nd edition (1st edition 1925, 14th edition 1970).

See Also

[fisher.test](#), [chisq.pval](#), [rel.risk.cint](#)

prop.cint	<i>Confidence interval for proportion based on frequency counts (corpora)</i>
-----------	---

Description

This function computes a confidence interval for a population proportion from the corresponding frequency count in a corpus. The confidence interval can be based on a binomial test or on a z-score test (with or without continuity correction).

Usage

```
prop.cint(k, n, method = c("binomial", "z.score"), correct = TRUE,
          conf.level = 0.95, alternative = c("two.sided", "less", "greater"))
```

Arguments

k	frequency of a type in the corpus (or an integer vector of frequencies)
n	number of tokens in the corpus, i.e. sample size (or an integer vector specifying the sizes of different samples)
method	a character string specifying whether the confidence interval is based on the binomial test (binomial) or the z-score test (z.score)
correct	if TRUE, apply Yates' continuity correction for the z-score test (default)
conf.level	the desired confidence level (defaults to 95%)
alternative	a character string specifying the alternative hypothesis, yielding a two-sided (two.sided, default), lower one-sided (less) or upper one-sided (greater) confidence interval

Details

The confidence intervals computed by this function correspond to those returned by [binom.test](#) and [prop.test](#), respectively. However, `prop.cint` accepts vector arguments, allowing many confidence intervals to be computed with a single function call. In addition, it uses a fast approximation of the two-sided binomial test that can safely be applied to large samples.

The confidence interval for a z-score test is computed by solving the z-score equation

$$\frac{k - np}{\sqrt{np(1-p)}} = \alpha$$

for p , where α is the z -value corresponding to the chosen confidence level (e.g. ± 1.96 for a two-sided test with 95% confidence). This leads to the quadratic equation

$$p^2(n + \alpha^2) + p(-2k - \alpha^2) + \frac{k^2}{n} = 0$$

whose two solutions correspond to the lower and upper boundary of the confidence interval.

When Yates' continuity correction is applied, the value k in the numerator of the z -score equation has to be replaced by k^* , with $k^* = k - 1/2$ for the *lower* boundary of the confidence interval (where $k > np$) and $k^* = k + 1/2$ for the *upper* boundary of the confidence interval (where $k < np$). In each case, the corresponding solution of the quadratic equation has to be chosen (i.e., the solution with $k > np$ for the lower boundary and vice versa).

Value

A data frame with two columns, labelled `lower` for the lower boundary and `upper` for the upper boundary of the confidence interval. The number of rows is determined by the length of the longest input vector (`k`, `n` and `conf.level`).

Author(s)

Stefan Evert

See Also

[z.score.pval](#), [prop.test](#), [binom.pval](#), [binom.test](#)

rel.risk.cint

Conservative confidence interval for the relative risk ratio (corpora)

Description

This function approximates a conservative confidence interval for the relative risk coefficient, i.e. the ratio $r = p_1/p_2$ between two population proportions, based on frequency counts from two corpora. The approximation is computed from individual confidence intervals for the two proportions, with confidence levels adjusted accordingly.

Usage

```
rel.risk.cint(k1, n1, k2, n2,
             conf.level = 0.95, alternative = c("two.sided", "less", "greater"),
             method = c("binomial", "z.score"), correct = TRUE)
```

Arguments

k1	frequency of a type in the first corpus (or an integer vector of type frequencies)
n1	the sample size of the first corpus (or an integer vector specifying the sizes of different samples)
k2	frequency of the type in the second corpus (or an integer vector of type frequencies, in parallel to k1)
n2	the sample size of the second corpus (or an integer vector specifying the sizes of different samples, in parallel to n1)
conf.level	the desired confidence level (defaults to 95%)
alternative	a character string specifying the alternative hypothesis, yielding a two-sided (two.sided, default), lower one-sided (less) or upper one-sided (greater) confidence interval
method	a character string specifying whether the individual confidence intervals for the two proportions are based on the binomial test (binomial) or the z-score test (z.score)
correct	if TRUE, apply Yates' continuity correction for the z-score test (default)

Details

This function computes individual confidence intervals for the two population proportions p_1 (from k_1 and n_1) and p_2 (from k_2 and n_2). Then, a confidence interval for the relative risk ratio $r = p_1/p_2$ is determined in such a way, that r lies within the interval whenever p_1 and p_2 lie in their respective confidence intervals.

Thus, when these intervals are computed with a confidence level of e.g. .975, r is certain to fall within its confidence interval in $.975^2 = .95$ of all cases. This adjustment of confidence levels is made automatically. Note that r *might* fall within its confidence interval even when either p_1 or p_2 is outside the respective interval, hence `rel.risk.cint` computes a *conservative* confidence interval that will be larger than necessary.

Exact confidence intervals for the *odds ratio* coefficient $\theta = (p_1/(1 - p_1))/(p_2/(1 - p_2))$ can be computed with the `fisher.test` function. However, these exact intervals are computationally *very* expensive and may cause R to run out of memory for large frequency counts. In addition, `fisher.test` only computes a single confidence interval for each function call (i.e., it cannot be applied to vectorised data).

Value

A data frame with two columns, labelled lower for the lower boundary and upper for the upper boundary of the confidence interval. The number of rows is determined by the length of the longest input vector (k1, n1, k2, n2 and conf.level).

Author(s)

Stefan Evert

See Also

[prop.cint](#), [chisq.pval](#), [fisher.pval](#), [fisher.test](#)

Description

This data set contains a small corpus (8043 tokens) of short stories from the collection *Very Short Stories* (VSS, see <http://www.schtepf.de/pages/stories.html>). The text was automatically segmented (tokenised) and annotated with part-of-speech tags (from the Penn tagset) and lemmas (base forms), using the IMS TreeTagger (Schmid 1994).

Usage

data(VSS)

Format

A data set with 8043 rows corresponding to tokens and the following columns:

word: the word form (or surface form) of the token

pos: the part-of-speech tag of the token (using the Penn tagset)

word: the lemma (or base form) of the token

Details

The Penn tagset defines the following part-of-speech tags:

CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential <i>there</i>
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NP	Proper noun, singular
NPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PP	Personal pronoun
PP\$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative

RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	<i>to</i>
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP\$	Possessive wh-pronoun
WRB	Wh-adverb

Author(s)

Stefan Evert (<http://purl.org/stefan.evert>)

References

Schmid, Helmut (1994). Probabilistic part-of-speech tagging using decision trees. In: *Proceedings of the International Conference on New Methods in Language Processing (NeMLaP)*, pages 44-49.

z.score

The z-score statistic for frequency counts (corpora)

Description

This function computes a z-score statistic for frequency counts, based on a normal approximation to the correct binomial distribution under the random sampling model.

Usage

```
z.score(k, n, p = 0.5, correct = TRUE)
```

Arguments

k	frequency of a type in the corpus (or an integer vector of frequencies)
n	number of tokens in the corpus, i.e. sample size (or an integer vector specifying the sizes of different samples)
p	null hypothesis, giving the assumed proportion of this type in the population (or a vector of proportions for different types and/or different populations)
correct	if TRUE, apply Yates' continuity correction (default)

Details

The z statistic is given by

$$z := \frac{k - np}{\sqrt{np(1-p)}}$$

When Yates' continuity correction is enabled, the *absolute value* of the numerator $d := k - np$ is reduced by $1/2$, but clamped to a non-negative value.

Value

The z -score corresponding to the specified data (or a vector of z -scores).

Author(s)

Stefan Evert

See Also

[z.score.pval](#)

z.score.pval

P-values of the z-score test for frequency counts (corpora)

Description

This function computes the p-value of a z -score test for frequency counts, based on the z -score statistic implemented by [z.score](#).

Usage

```
z.score.pval(k, n, p = 0.5, correct = TRUE,
             alternative = c("two.sided", "less", "greater"))
```

Arguments

k	frequency of a type in the corpus (or an integer vector of frequencies)
n	number of tokens in the corpus, i.e. sample size (or an integer vector specifying the sizes of different samples)
p	null hypothesis, giving the assumed proportion of this type in the population (or a vector of proportions for different types and/or different populations)
correct	if TRUE, apply Yates' continuity correction (default)
alternative	a character string specifying the alternative hypothesis; must be one of two.sided (default), less or greater

Value

The p-value of a *z*-score test applied to the given data (or a vector of p-values).

Author(s)

Stefan Evert

See Also

[z.score](#), [binom.pval](#), [prop.cint](#)

Index

*Topic **array**

cont.table, 8

*Topic **datasets**

BNCcomparison, 3

BNCdomains, 4

BNCInChargeOf, 4

VSS, 13

*Topic **htest**

binom.pval, 2

chisq, 5

chisq.pval, 7

cont.table, 8

fisher.pval, 9

prop.cint, 10

rel.risk.cint, 11

z.score, 14

z.score.pval, 15

binom.pval, 2, 11, 16

binom.test, 2, 10, 11

BNCcomparison, 3

BNCdomains, 4

BNCInChargeOf, 4

chisq, 5, 7

chisq.pval, 6, 7, 10, 12

chisq.test, 6–8

cont.table, 6, 8

fisher.pval, 7, 9, 12

fisher.test, 8–10, 12

prop.cint, 2, 10, 12, 16

prop.test, 7, 10, 11

rel.risk.cint, 7, 10, 11

VSS, 13

z.score, 14, 15, 16

z.score.pval, 2, 11, 15, 15