Package 'chisquare'

April 4, 2022

Title Chi-Square and G-Square Test of Independence, Residual Analysis, and Measures of Categorical Association

Version 0.3

Description Provides the facility to perform the chi-square and G-square test of independence, calculates permutation-based p value, and provides measures of association such as Phi, odds ratio with 95 percent CI and p value, adjusted contingency coefficient, Cramer's V and 95 percent CI, bias-corrected Cramer's V, Cohen's w, Goodman-Kruskal's lambda, gamma and its p value, and tau, Cohen's k and its 95 percent CI. It also calculates standardized, moment-corrected standardized, and adjusted standardized residuals, and their significance. Different outputs are returned in nicely formatted tables.

Depends R (>= 4.0.0)

Imports gt (>= 0.3.1)

License GPL (>= 2)

Encoding UTF-8

LazyData true

RoxygenNote 7.1.1

NeedsCompilation no

Author Gianmarco Alberti [aut, cre]

Maintainer Gianmarco Alberti <gianmarcoalberti@gmail.com>

Repository CRAN

Date/Publication 2022-04-04 17:30:02 UTC

R topics documented:

chisquare *R function for chi-square and G-square test of independence, measures of association, and standardized/moment-corrected standardized/adjusted standardized residuals*

Description

The function performs the chi-square (and the G-square) test of independence on the input contingency table, calculates various measures of categorical association, returns standardized, moment-corrected standardized, and adjusted standardized residuals (with indication of their significance), and calculates relative and absolute contributions to the chi-square. The p value associated to the chi-square statistic is also calculated on the basis of a permutation-based procedure. Nicely-formatted output tables are rendered.

Usage

```
chisquare(
  data,
  B = 999,
  adj.alpha = FALSE,
  format = "short",
  graph = FALSE,
  tfs = 14
)
```

Arguments

data	Dataframe containing the input contingency table.
B	Number of simulated tables to be used to calculate the Monte Carlo-based p value (999 by default).
adj.alpha	Takes TRUE or FALSE (defult) if the user wants or does not want the significance level of the residuals (both standarized and adjusted standardized) to be corrected using the Sidak's adjustment method (see Details).
format	Takes *short* (default) if the dataset is a dataframe storing a contingency table; if the input dataset is a dataframe storing two columns that list the levels of the two categorical variables, *long* will preliminarily cross-tabulate the levels of the categorical variable in the 1st column against the levels of the variable stored in the 2nd column.
graph	Takes TRUE or FALSE (default) if the user wants or does not want to chart the distribution of the permuted chi-square statistic accross the number of randomized tables set by the B parameter.
tfs	Numerical value to set the size of the font used in the main body of the various output tables (14 by default).

Details

The following **measures of categorical associations** are produced by the function:

- Phi (only for 2x2 tables)
- Phi signed (only for 2x2 tables)
- Yule's Q (and p value; only for 2x2 tables)
- Odds ratio (with 95perc confidence interval and p value; only for 2x2 tables)
- Adjusted contingency coefficient C
- Cramer's V (with 95perc confidence interval; indication of the magnitude of the effect size according to Cohen is provided for tables with up to 5 degrees of freedom)
- Bias-corrected Cramer's V (indication of the magnitude of the effect size according to Cohen is provided for tables with up to 5 degrees of freedom)
- Cohen's w (with indication of the magnitude of the effect size)
- Goodman-Kruskal's lambda (asymmetric)
- Goodman-Kruskal's lambda (symmetric)
- Goodman-Kruskal's tau (asymmetric)
- Goodman-Kruskal's gamma (and p value)
- Cohen's k (and 95perc confidence interval)

The **p value** of the observed chi-square statistic is also calculated on the basis of a **permutation-based approach**, using *B* random tables created under the Null Hypothesis of independence. For the rationale of this approach, see for instance the description in Beh E.J., Lombardo R. 2014, Correspondence Analysis: Theory, Practice and New Strategies, Chichester, Wiley: 62-64.

The **permutation-based p value** is calculated as follows:
$(1 + sum(chistat.perm > chisq.stat))/(1 + B)$, where *chistat.perm* is a vector storing the B chi-square statistics generated under the Null Hypothesis, and *chisq.stat* is the observed chi-square statistic. For the logic of the calculation, see for example Baddeley et al., "Spatial Point Patterns. Methodology and Applications with R", CRC Press 2016: 387.

The **moment-corrected standardized residuals** are calculated as follows:
$stand.res/(sqrt((nr - 1) * (nc - 1)/(nr * nc)))$, where *stand.res* is each cell's standardized residual, *nr* and *nc* are the number of rows and columns respectively; see Garcia-Perez, MA, and Nunez-Anton, V (2003). Cellwise Residual Analysis in Two-Way Contingency Tables. Educational and Psychological Measurement, 63(5): 827.

The **adjusted standardized residuals** are calculated as follows:
$stand.res[i, j]/sqrt((1 - sr[i]/n) * (1 - sc[j]/n))$, where *stand.res* is the standardized residual for cell *ij*, *sr* is the row sum for row *i*, *sc* is the column sum for column *j*, and *n* is the table grand total. The *adjusted standardized residuals* may prove useful since it has been demonstrated that the standardized residuals tend to underestimate the significance of differences in small samples. The adjusted standardized residuals correct that deficiency.

The **significance of the residuals** (standardized, moment-corrected standardized, and adjusted standardized) is assessed using alpha 0.05 or, optionally (by setting the parameter 'adj.alpha' to TRUE), using an adjusted alpha calculated using the Sidak's method:
$alpha.adj = 1 - (1 - 0.05)^(1/(nr * nc))$, where *nr* and *nc* are the number of rows and columns in the table respectively. The adjusted alpha is then converted into a critical two-tailed z value. See: Beasley TM and Schumacker RE (1995), Multiple Regression Approach to Analyzing Contingency Tables: Post Hoc and Planned Comparison Procedures, The Journal of Experimental Education, 64(1): 86, 89.

The **cells' relative contribution (in percent) to the chi-square statistic** is calculated as:
$chisq.values/chisq.stat * 100$, where *chisq.values* and *chisq.stat* are the chi-square value in each individual cell of the table and the value of the chi-square statistic, respectively. The *average contribution* is calculated as $100/(nr * nc)$, where *nr* and *nc* are the number of rows and columns in the table respectively.

The **cells' absolute contribution (in percent) to the chi-square statistic** is calculated as:
$chisq.values/n * 100$, where *chisq.values* and *n* are the chi-square value in each individual cell of the table and the table's grant total, respectively. The *average contribution* is calculated as sum of all the absolute contributions divided by the number of cells in the table. For both the relative and absolute contributions to the chi-square, see Beasley and Schumacker (1995): 90.

The calculation of the **95perc confidence interval around Cramer's V** is based on Smithson M.J. (2003). Confidence Intervals, Quantitative Applications in the Social Sciences Series, No. 140. Thousand Oaks, CA: Sage, 39-41, and builds on the R code made available by the author on the web (http://www.michaelsmithson.online/stats/CIstuff/CI.html).

The **bias-corrected Cramer's V** is based on Bergsma, W. (2013). "A bias correction for Cramér's V and Tschuprow's T". Journal of the Korean Statistical Society. 42 (3): 323–328, https://doi.org/10.1016

For the **other measures of categorical association** provided by the function, see for example Sheskin, D. J. (2011). Handbook of Parametric and Nonparametric Statistical Procedures, Fifth Edition (5th ed.). Chapman and Hall/CRC: 675-679, 1415-1427.

Note that:
-the **Phi** coefficient is based on the chi-square statistic as per Sheskin's equation 16.21, whereas the **Phi signed** is after Sheskin's equation 16.20;

-the **2-sided p value of Yule's Q** is calculated following Sheskin's equation 16.24;

-**Cohen's w** is calculate as $V * sqrt(min(nr, nc) - 1)$, where *V* is Cramer's V, and *nr* and *nc* are the number of rows and columns respectively; see Sheskin 2011: 679;

-in the output table reporting the result of the chi-square test and the various measures of association, the magnitude of the **effect size** according to *Cohen*'s guidelines is reported for Cramer's V, Cramer's V biase-corrected, and for Cohen's w; the effect size for the former two coefficients is only reported for tables featuring up to 5 degrees of freedom (see Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed). Hillsdale, N.J: L. Erlbaum Associates);

-the **2-tailed p value** of **Goodman-Kruskal's gamma** is based on the associated z-score calculated as per Sheskin's equation 32.2;

-the **symmetric** version of **Goodman-Kruskal's lambda** is calculated as per Reynolds, H. T. (1984). Analysis of Nominal Data (Quantitative Applications in the Social Sciences) (1st ed.). SAGE Publications: 55-57;

-**Goodman-Kruskal's tau** is calculated as per Reynolds 1984: 57-60;

-**Cohen's k** is calculated as per Sheskin 2011: 688-689 (equation 16.30).

Value

The function produces an **optional chart** (distribution of the permuted chi-square statistic), and a number of **output tables** that are nicely formatted with the help of the *gt* package. The output tables are listed below:

- Input contingency table (with some essential analytical results annotated at the bottom)
- Expected frequencies
- Cells' chi-square value
- Cells' relative contribution (in percent) to the chi-square statistic (cells in RED feature a larger-than-average contribution)
- Cells' absolute contribution (in percent) to the chi-square statistic (colour same as above)
- Standardized residuals (RED for large significant residuals, BLUE for small significant residuals)
- Moment-corrected standardized residuals (colour same as above)
- Adjusted standardized residuals (colour same as above)
- Table of output statistics, p values, and association measures

Also, the function returns a **list containing the following elements**:

- *crosstab*: input contingency table
- *exp.freq.*: table of expected frequencies
- *chisq.values*: cells' chi-square value
- *chisq.relat.contrib.*: cells' relative contribution (in percent) to the chi-square statistic
- *chisq.abs.contrib.*: cells' absolute contribution (in percent) to the chi-square statistic
- *chisq.statistic*: observed chi-square value
- *chisq.p.value*: p value of the chi-square statistic
- *chisq.p.value.perm.*: p value based on B permuted tables
- *Gsq.statistic*: observed G-square value
- *Gsq.p.value*: p value of the G-square statistic
- *stand.resid.*: table of standardized residuals
- *mom.corr.stand.resid.*: table of moment-corrected standardized residuals

- *adj.stand.resid.*: table of adjusted standardized residuals
- *Phi*: Phi coefficient (only for 2x2 tables)
- *Phi signed*: signed Phi coefficient (only for 2x2 tables)
- *Yule's Q*: Q coefficient (only for 2x2 tables)
- *Yule's Q p.value*: 2-tailed p value of Yule's Q
- *Odds ratio*: odds ratio (only for 2x2 tables)
- *Odds ratio CI lower boundary*: lower boundary of the 95perc CI
- *Odds ratio CI upper boundary*: upper boundary of the 95perc CI
- *Odds ratio p.value*: p value of the odds ratio
- *Cadj*: adjusted contigency coefficient C
- *Cramer's V*: Cramer's V coefficient
- *Cramer's V CI lower boundary*: lower boundary of the 95perc CI
- *Cramer's V CI upper boundary*: upper boundary of the 95perc CI
- *Cramer's Vbc*: bias-corrected Cramer's V coefficient
- *w*: Cohen's w
- *lambda (rows dep.)*: Goodman-Kruskal's lambda coefficient (considering the rows being the dependent variable)
- *lambda (cols dep.)*: Goodman-Kruskal's lambda coefficient (considering the columns being the dependent variable)
- *lambda.symmetric*: Goodman-Kruskal's symmetric lambda coefficient
- *tau (rows dep.)*: Goodman-Kruskal's tau coefficient (considering the rows being the dependent variable)
- *tau (cols dep.)*: Goodman-Kruskal's tau coefficient (considering the columns being the dependent variable)
- *gamma*: Goodman-Kruskal's gamma coefficient
- *gamma.p.value*: 2-sided p value for the Goodman-Kruskal's gamma coefficient
- *k*: Cohen'k
- *k CI lower boundary*: lower boundary of the 95perc CI
- *k CI upper boundary*: upper boundary of the 95perc CI

Note that the *p values* returned in the above list are expressed in scientific notation, whereas the ones reported in the output table featuring the tests' result and measures of association are reported as broken down into classes (e.g., <0.05, or <0.01, etc).

The **following examples**, which use in-built datasets, can be run to fimiliarise with the function:

-perform the test on the in-built 'social_class' dataset
result <- chisquare(social_class)

-perform the test on a 2x2 subset of the 'diseases' dataset
mytable <- diseases[3:4,1:2]

result <- chisquare(mytable)

-perform the test on a 2x2 subset of the 'safety' dataset
mytable <- safety[c(4,1),c(1,6)]
result <- chisquare(mytable)

-build a toy dataset in 'long' format (gender vs. opinion about death sentence)
mytable <- data.frame(GENDER=c(rep("F", 360), rep("M", 340)),
OPINION=c(rep("oppose", 235),
rep("favour", 125),
rep("oppose", 160),
rep("favour", 180)))

-perform the test specifying that the input table is in 'long' format
result <- chisquare(mytable, format="long")

Examples

```
#perform the test on the in-built 'diseases' dataset
result <- chisquare(diseases, B=99)
```

`diseases` *Dataset: Cross-tabulation of quantity of tobacco smoked daily vs. cause of death*

Description

Cross-tabulation (15x4) of the amount of tobacco smoked on a daily basis (in gramms) against cause of death.
After: Velleman P F, Hoaglin D C, Applications, Basics, and Computing of Exploratory Data Analysis, Wadsworth Pub Co 1984 (Exhibit 8-1)

Usage

```
data(diseases)
```

Format

dataframe

safety — *Dataset: Cross-tabulation of people's feeling of safety vs. town size*

Description

Cross-tabulation (4x6).

Usage

```
data(safety)
```

Format

dataframe

social_class — *Dataset: Cross-tabulation of social class vs. diagnostic category for a sample of psychiatric patients*

Description

Cross-tabulation (3x4) after: Everitt B.S (1992), The Analysis of Contingency Tables, Chapman&Hall/CRC, second edition, table 3.13.

Usage

```
data(social_class)
```

Format

dataframe

Index