Package 'raters'

July 23, 2025

Type Package

Title A Modification of Fleiss' Kappa in Case of Nominal and Ordinal Variables

Version 2.1.1

Author Daniele Giardiello [cre],
Piero Quatto [aut],
Enrico Ripamonti [aut],
Stefano Vigliani [ctb]

Maintainer Daniele Giardiello <daniele.giardiello1@gmail.com>

License GPL (>= 2)

Description The kappa statistic implemented by Fleiss is a very popular index for assessing the reliability of agreement among multiple observers. It is used both in the psychological and in the psychiatric field. Other fields of application are typically medicine, biology and engineering. Unfortunately,the kappa statistic may behave inconsistently in case of strong agreement between raters, since this index assumes lower values than it would have been expected. We propose a modification kappa implemented by Fleiss in case of nominal and ordinal variables. Monte Carlo simulations are used both to testing statistical hypotheses and to calculating percentile bootstrap confidence intervals based on proposed statistic in case of nominal and ordinal data.

Encoding UTF-8

LazyData true

NeedsCompilation no

Repository CRAN

Date/Publication 2024-09-02 13:00:02 UTC

Contents

`raters-package` *A Modification of Fleiss' Kappa in case of Nominal and Ordinal Variables*

Description

Computes a statistic as an index of inter-rater agreement among a set of raters in case of nominal or ordinal data.This procedure is based on a statistic not affected by Kappa paradoxes. In case of ordinal data, the weighted versions of the statistic has been developed using a matrix of linear or quadratic weights. The percentile Boostrap confidence interval is computed and the test argument allows to perform if the agreement is nil. The p value can be approximated using the Normal, Chi-squared distribution or using Monte Carlo algorithm in case of nominal data. Otherwise, the approximation and the Monte Carlo algorithm is computed. Fleiss' Kappa index is also shown in case of nominal data. In a nutshell, the function `concordance` can be used in case of nominal scale while the functions `wlin.conc` and `wquad.conc` can be used in case of ordinal data using linear or quadratic weights, respectively.

Details

Package:	raters
Type:	Package
Version:	2.1.1
License:	GPL(>=2)

Author(s)

Daniele Giardiello, Piero Quatto, Enrico Ripamonti and Stefano Vigliani

Maintainer: Daniele Giardiello <daniele.giardiello1@gmail.com>

References

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. Psychological Bulletin **76, 378-382**.

Falotico, R. Quatto, P. (2010). On avoiding paradoxes in assessing inter-rater agreement. Italian Journal of Applied Statistics **22, 151-160.**

Falotico, R., Quatto, P. (2014). Fleiss' kappa statistic without paradoxes. Quality & Quantity, **1-8**.

Marasini, D. Quatto, P. Ripamonti, E. (2014). Assessing the inter-rater agreement for ordinal data through weighted indexes. Statistical methods in medical research.

Examples

```
# Nominal data
data(diagnostic)
concordance(diagnostic,test="Normal")

# Ordinal data with linear weights
data(winetable)
set.seed(12345)
wlin.conc(winetable,test="MC")

# Ordinal data with quadratic weights
data(winetable)
set.seed(12345)
wquad.conc(winetable,test="MC")
```

`concordance` *Inter-rater agreement among a set of raters for nominal data*

Description

Computes a statistic as an index of inter-rater agreement among a set of raters in case of nominal data. This procedure is based on a statistic not affected by paradoxes of Kappa. It is also possible to get the confidence interval at level alpha using the percentile Bootstrap and to evaluate if the agreement is nil using the test argument. The p value can be approximated using the Normal, Chi squared distribution or using Monte Carlo algorithm. Normal approximation and Monte Carlo procedure can be calculated even though the number of observers is not the same for each evaluated subject. Fleiss Kappa is also shown and its confidence interval, standard error and pvalue using Normal approximation are available when the number of observes is the same for each classified subject and the test argument is specified. The functions `wlin.conc` and `wquad.conc` can be used in case of ordinal data using linear or quadratic weight matrix, respectively.

Usage

```
concordance(db, test = "Default", B = 1000, alpha = 0.05)
```

Arguments

`db`	n*c matrix or data frame, n subjects c categories. The numbers inside the matrix or data frame indicate how many raters chose a specific category for a given subject. A sum of row indicates the total number of raters who evaluated a given subject.
`test`	Statistical test to evaluate if the raters make random assignment regardless of the characteristic of each subject. Under null hypothesis, it corresponds to a high percentage of assignment errors. Thus, the expected agreement is weak. Normal approximation is advisable when the number of subject is pretty large while a Chi square approximation when the number of raters is large. Monte Carlo test is useful for small samples even though the higher number of simulations, the more time the procedure takes. If test is not mentioned, no test will be computed.

B	Number of iterations for the percentile Bootstrap and for Monte Carlo test.
alpha	Level of significance for Bootstrap confidence interval

Value

A list containing the following components:

$Fleiss	A list with Kappa of Fleiss index. When the number of raters is the same for every evaluated subject, the standard deviation, the Z Wald test and the p value are also shown.
$Statistic	A list with the index of inter-rater agreement not affected by Kappa paradoxes and the percentile Bootstrap confidence interval. If the test argument is specified the p value is also shown.

Author(s)

Piero Quatto <piero.quatto@unimib.it>, Daniele Giardiello <daniele.giardiello1@gmail.com>, Stefano Vigliani

References

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. Psychological Bulletin **76, 378-382**

Falotico, R. Quatto, P. (2010). On avoiding paradoxes in assessing inter-rater agreement. Italian Journal of Applied Statistics **22, 151-160**

Examples

```
data(diagnostic)
concordance(diagnostic, test = "Chisq")
concordance(diagnostic, test = "Normal")
concordance(diagnostic, test = "MC", B = 100)
```

---

diagnostic *Frequency of assignment of patients to diagnostic categories*

Description

Six psychiatrists classified 30 patients into 5 diagnostic categories.

Usage

```
data(diagnostic)
```

Format

A matrix with 30 rows and 5 columns.

`Depression` number of psychiatrists who judged a given patient affected by “Depression”

`Personality disorders` number of psychiatrists who judged a given patient affected by “Personality disorders”

`Schizophrenia` number of psychiatrists who judged a given patient affected by “Schizophrenia”

`Neurosis` number of psychiatrists who judged a given patient affected by “Neurosis”

`Other` number of psychiatrists who judged a given patient affected by “Other” diseases

References

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. Psychological Bulletin **76, 378-382**

---

uterine *Variability in classification of carcinoma in situ of the uterine cervix*

Description

Seven oncologists classified 118 patients into five stages of carcinoma classification

Usage

```
data(uterine)
```

Format

A data frame with 118 observations on the following 5 variables.

`Negative` number of doctors who judged a given patient as "Negative"

`Atypical Squamous Hyperplasia` number of doctors who judged a given patient affected by "Atypical Squamous Hyperplasia"

`Carcinoma in Situ` number of doctors who judged a given patient affected by "Carcinoma in Situ"

`Squamous Carcinoma with Early Stromal Invasion` number of doctors who judged a given patient affected by "Squamous Carcinoma with Early Stromal"

`Invasive Carcinoma` number of doctors who judged a given patient affected by "Invasive Carcinoma"

References

Holmquist, N.D. et al. (1967). Variability in classification of carcinoma in situ of the uterine cervix. Archives of Pathology **84,334-345**

`winetable` *Sensory wine evaluation*

Description

The data wine in ordinal package represent a factorial experiment on factors determining the bitterness of wine with 1 as "Least bitter"and 5 as "Most bitter". In this case, we supposed that eight different bottles of wine were evaluated by nine judges according to the bitterness scale described above. It is possible to get this data using the dataframe wine included in ordinal package using `table(wine$bottle,wine$rating)`

Usage

```
data(winetable)
```

Format

A data frame with 8 observations on 5 variables.

References

Randall J.H.(1989) The Analysis of Sensory Data by Generalized Linear Model. Biometrical Journal **vol 31,issue 7, pp 781-793**

`wlin.conc` *Inter-rater agreement among a set of raters for ordinal data using linear weights*

Description

Computes a statistic as an index of inter-rater agreement among a set of raters in case of ordinal data using linear weights. The matrix of linear weights is defined inside the function. This procedure is based on a statistic not affected by Kappa paradoxes. It is also possible to get the confidence interval at level alpha using the percentile Bootstrap and to evaluate if the agreement is nil using the Monte Carlo algorithm. Fleiss' Kappa cannot be used in case of ordinal data. It is advisable to use `set.seed` to get the same replications for Bootstrap confidence limits and Montecarlo test.

Usage

```
wlin.conc(db, test = "Default", B = 1000, alpha = 0.05)
```

Arguments

db	n*c matrix or data frame, n subjects c categories. The numbers inside the matrix or data frame indicate how many raters chose a specific category for a given subject. A sum of row indicates the total number of raters who evaluated a given subject. In case of ordinal data, the c categories can be sorted according to a specific scale.
test	Statistical test to evaluate if the raters make random assignment regardless of the characteristic of each subject. Under null hypothesis, it corresponds to a high percentage of assignment errors. Thus, the expected agreement is weak. If this argument is not specified the p value are not being computed.
B	Number of iterations for the percentile Bootstrap and for Monte Carlo test.
alpha	Level of significance for Bootstrap confidence interval and for Monte Carlo algorithm if it is specified

Value

A list containing the following components:

$Statistic	A list with the index of inter-rater agreement not affected by Kappa paradoxes for ordinal data and the percentile Bootstrap confidence interval. If the test argument is specified the p value is also shown.

Author(s)

Piero Quatto <piero.quatto@unimib.it>, Daniele Giardiello <daniele.giardiello1@gmail.com>, Stefano Vigliani

References

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. Psychological Bulletin **76, 378-382**

Falotico, R. Quatto, P. (2010). On avoiding paradoxes in assessing inter-rater agreement. Italian Journal of Applied Statistics **22, 151-160**

Marasini, D. Quatto, P. Ripamonti, E. (2014). Assessing the inter-rater agreement for ordinal data through weighted indexes. Statistical methods in medical research.

Examples

```
data(uterine)
set.seed(12345)
wlin.conc(uterine, test = "MC", B = 25)
```

wquad.conc *Inter-rater agreement among a set of raters for ordinal data using quadratic weights*

Description

Computes a statistic as an index of inter-rater agreement among a set of raters in case of ordinal data using quadratic weights. The matrix of quadratic weights is defined inside the function. This procedure is based on a statistic not affected by Kappa paradoxes. It is also possible to get the confidence interval at level alpha using the percentile Bootstrap and to evaluate if the agreement is nil using the Monte Carlo algorithm. Fleiss' Kappa cannot be used in case of ordinal data. It is advisable to use `set.seed` to get the same replications for Bootstrap confidence limits and Montecarlo test.

Usage

```
wquad.conc(db, test = "Default", B = 1000, alpha = 0.05)
```

Arguments

db	n*c matrix or data frame, n subjects c categories. The numbers inside the matrix or data frame indicate how many raters chose a specific category for a given subject. A sum of row indicates the total number of raters who evaluated a given subject. In case of ordinal data, the c categories can be sorted according to a specific scale.
test	Statistical test to evaluate if the raters make random assignment regardless of the characteristic of each subject. Under null hypothesis, it corresponds to a high percentage of assignment errors. Thus, the expected agreement is weak. If this argument is not specified the p value are not being computed.
B	Number of iterations for the percentile Bootstrap and for Monte Carlo test.
alpha	Level of significance for Bootstrap confidence interval and for Monte Carlo algorithm if it is specified

Value

A list containing the following components:

$Statistic	A list with the index of inter-rater agreement not affected by Kappa paradoxes for ordinal data and the percentile Bootstrap confidence interval. If the test argument is specified the p value is also shown.

Author(s)

Piero Quatto <piero.quatto@unimib.it>, Daniele Giardiello <daniele.giardiello1@gmail.com>, Stefano Vigliani

References

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. Psychological Bulletin **76, 378-382**

Falotico, R. Quatto, P. (2010). On avoiding paradoxes in assessing inter-rater agreement. Italian Journal of Applied Statistics **22, 151-160**

Marasini, D. Quatto, P. Ripamonti, E. (2014). Assessing the inter-rater agreement for ordinal data through weighted indexes. Statistical methods in medical research.

Examples

```
data(uterine)
set.seed(12345)
wquad.conc(uterine, test = "MC", B = 25)
```

Index