Package 'irrNA': Coefficients of Interrater Reliability – Generalized for Randomly Incomplete Datasets (Version 0.2.2)

Technical Report · April 2021

CITATIONS
0

READS
2

1 author:

Markus Brückl
Technische Universität Berlin
31 PUBLICATIONS **103** CITATIONS

SEE PROFILE

Package 'irrNA'

April 5, 2021

Type Package

Title Coefficients of Interrater Reliability – Generalized for Randomly Incomplete Datasets

Version 0.2.2

Date 2021-04-05

Author Markus Brueckl [aut, cre], Florian Heuer [aut, trl]

Maintainer Markus Brueckl <markus.brueckl@tu-berlin.de>

Description Provides coefficients of interrater reliability that are generalized to cope with randomly incomplete (i.e. unbalanced) datasets without any imputation of missing values or any (row-wise or column-wise) omissions of actually available data. Applied to complete (balanced) datasets, these generalizations yield the same results as the common procedures, namely the Intraclass Correlation according to McGraw & Wong (1996) <doi:10.1037/1082-989X.1.1.30> and the Coefficient of Concordance according to Kendall & Babington Smith (1939) <doi:10.1214/aoms/1177732186>.

Depends R (>= 2.10.0)

License GPL-3

LazyData true

Imports irr, stats

RoxygenNote 7.1.1

Encoding UTF-8

URL https://CRAN.R-project.org/package=irrNA

NeedsCompilation no

Repository CRAN

Date/Publication 2021-04-05 12:00:02 UTC

R topics documented:

`Consist` *irrNA example data, showing perfect consistency between raters*

Description

This data set shows perfect consistency and moderate agreement between raters.

Usage

```
data(Consist)
```

Format

A 2-dimensional data frame including column and row headers.

`ConsistNA` *irrNA example data, showing perfect consistency between raters and NAs*

Description

This data set shows missing values (NAs) and perfect consistency and moderate agreement between raters.

Usage

```
data(ConsistNA)
```

Format

A 2-dimensional data frame including column and row headers and NAs.

Ebel51 *Example data, given by Ebel (1951, Table 2)*

Description

This data set is used by Ebel (1951) to demonstate the computation of an intraclass correlation on incomplete data sets.

Usage

```
data(Ebel51)
```

Format

A 2-dimensional data frame including column and row headers and NAs.

Source

Psychometrika

References

Ebel, R.L. (1951). Estimation of the reliability of ratings. Psychometrika, 16(4), 407–424.

EbelFILL *Example data, based on Ebel (1951, Table 2)*

Description

This data set is the same as Ebel51, but with the missing data filled up with arbitrary values.

Usage

```
data(EbelFILL)
```

Format

A 2-dimensional data frame including column and row headers.

References

Ebel, R.L. (1951). Estimation of the reliability of ratings. Psychometrika, 16(4), 407–424.

iccNA *Intraclass correlation coefficients (ICCs) – generalized for randomly incomplete datasets*

Description

This function computes intraclass correlation coefficients (ICCs) as indices of interrater reliability or agreement based on cardinally scaled data. This function also works on (unbalanced) incomplete datasets without any imputation of missing values (*NA*s) or (row- or cloumn-wise) omissions of data! p-values and confidence intervals are provided. In case of extreme input data (e.g. zero variances) output *NaN*s are avoided by approximation.

Usage

```
iccNA(ratings, rho0 = 0, conf = 0.95, detail = FALSE, oneG = TRUE, Cs = 10000)
```

Arguments

ratings	n*m matrix or data frame; n objects (rows), m raters (columns)
rho0	numeric value; correlation in population (ρ) according to the null hypothesis (0 is default)
conf	numeric value; confidence level (95% is default)
detail	logical; if TRUE, returns additional information (sums of squares, degrees of freedom, the means per object, data corrected for the raters' biases)
oneG	logical; if TRUE, the ipsation (correction for the raters' effects) is done the simple way, using the difference of each raters mean to the one grand mean (G) of all values to estimate the raters' biases. If FALSE the weighted sub-means (G_j of those objects that an individual rater j rated are used instead (cp. Brueckl, 2011, Equation 4.30).
Cs	numeric value; denominator (10000 is default) of the effect-size-criterion to stop iteration of the correction for the raters' biases; the enumerator denotes a small effect (η-squared = 1%)

Details

This function is able to compute ICCs on randomly incomplete (i.e. unbalanced) data sets. Thus, both an imputation of missing values (*NA*s) and row-wise or column-wise omissions of data are obsolete. Working on complete datasets, it yields the same results as the common functions, e.g. icc_corr.

The method of Ebel (1951) is used to calculate the oneway ICCs. The solution for the twoway ICCs is derived from the oneway solution (cp. Brueckl, 2011, p. 96 ff.): The raters' individual effects (biases) are estimated, reducing this problem again to the oneway problem (cp. Greer & Dunlap, 1997).

This estimation can be done using the difference of a certain (j) rater's mean to the grand mean (G) or to the sub-mean (G_j) representing only those objects that were rated by this rater. The first method is fail-safe. The second method is thought to provide the more precise estimates (of the

raters' biases), the more the mean of the true values of the objects that each rater rated differ from the grand mean, e.g. if there are raters that only rate objects with low true values (and therefore also other raters that only rate objects with high true values).
If the second method is chosen and if the ratings are unbalanced, which happens most of the time if not all raters rated all objects, the raters' biases cannot be determined exactly – but as approximately as desired. This approximation needs an iteration, thus a stop criterion (`Cs`): The iteration is stopped, when the difference in the raters' effect size (η-squared) between subsequent iterations would be equal to or smaller than the `Cs`th part of a small effect (i.e. η-squared = 1%).

Just as in icc_corr and icc, the designation established by McGraw & Wong (1996) – *A* for *absolute agreement* and *C* for *consistency* – is used to differ between the (twoway) ICCs that rely on different cases and thus must be interpreted differently.

The generalization of the procedure entails a generalization of the three cases that differentiate the ICCs (cp. Shrout & Fleiss, 1979):
- Case 1 (oneway case, treated by ICC(1) and ICC(k)):
Each object – of a sample that was randomly drawn from the population of objects; also holds true for case 2 and case 3 – is rated by (a different number of) different raters that were randomly drawn from the population of raters.
- Case 2 (twoway case, treated by ICC(A,1) and ICC(A,k)):
Each object is rated by a random subset of the group of raters that is drawn randomly from the population of raters.
- Case 3 (twoway case, treated by ICC(C,1) and ICC(C,k)):
Each object is rated by a random subset of the group of all relevant (i.e. fixed) raters.

Output NaNs, that usually occur (see e.g. icc or icc_corr) in case of extreme input data (e.g. in case of zero variance(s), within or between objects) are avoided by approximation from little less extreme input data. Warning messages are given in these cases.

Value

`ICCs`	data frame containing the intraclass correlation coefficients, the corresponding p-values, and confidence intervals
`n`	number of rated objects
`k`	maximum number of raters per object
`amk`	mean number of ratings per object
`k_0`	approximate harmonic mean (cp. Ebel, 1951) of the number of ratings per object
`n_iter`	number of iterations for correcting for the raters' biases
`corr_ratings`	ratings, corrected for the individual raters' biases
`am0`	means of ratings for each object, based on (1) the original data and on (2) the data that are corrected for the raters' biases
`oneway`	statistics for the oneway ICCs
`twoway`	statistics for the twoway ICCs

Author(s)

Markus Brueckl

References

Brueckl, M. (2011). Statistische Verfahren zur Ermittlung der Urteileruebereinstimmung. in: Altersbedingte Veraenderungen der Stimme und Sprechweise von Frauen, Berlin: Logos, 88–103.

Ebel, R.L. (1951). Estimation of the reliability of ratings. Psychometrika, 16(4), 407–424.

Greer, T., & Dunlap, W.P. (1997). Analysis of variance with ipsative measures. Psychological Methods, 2, 200–207.

McGraw, K.O., & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. Psychological Methods, 1, 30–46.

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin, 86(2), 420–428.

See Also

kendallNA,icc_corr,icc

Examples

```
# Example 1:
data(ConsistNA)
# ConsistNA exhibits missing values, a perfect consistency, and
# a moderate agreement between raters:
ConsistNA
# Common ICC-algorithms fail, since each row as well as each
# column of ConsistNA exhibits unfilled cells and these missing
# data are omitted column-wise or row-wise:
library(irr)
icc(ConsistNA, r0=0.3)
# Ebel's (1951) method for computing ICC(1) and ICC(k) that is
# implemented in iccNA can cope with such data without an
# omission or an imputation of missing values, but still can
# not depict the raters' interdependency...
iccNA(ConsistNA, rho0=0.3)
# ...but generalizations of Ebel's method for the twoway ICCs
# are able to assess moderate agreement (ICC(A,1) and ICC(A,k))
# and perfect consistency (ICC(C,1) and ICC(C,k)), assuming that
# the data were acquired under case 2 or case 3, see Details in
# the Help file.
#
# Example 2:
data(IndepNA)
# IndepNA exhibits missing values and zero variance between
# the raters just as well as between the objects:
IndepNA
# Again, common ICC-algorithms fail:
icc(IndepNA)
# But iccNA is able to include all available data in its
```

```
# calculation and thereby to show the perfect independence of
# the ratings:
iccNA(IndepNA)
#
# Example 3:
# The example provided by Ebel (1951, Tables 2 and 3):
# data(Ebel51)
Ebel51
# iCCNA achieves to include all available ratings and to assess
# twoway ICCs, assuming that the data were acquired under
# case 2 or case 3:
iccNA(Ebel51, detail=TRUE)
```

`icc_corr` *Intraclass correlation coefficients (ICCs) for oneway and twoway models – corrected version of icc{irr}*

Description

Computes single score or average score ICCs as an index of interrater reliability of quantitative data. Additionally, F-test and confidence interval are computed. icc_corr{irrNA} corrects 3 errors of Matthias Gamer's function icc (version 0.84.1).

Usage

```
icc_corr(
  ratings,
  model = c("oneway", "twoway"),
  type = c("consistency", "agreement"),
  unit = c("single", "average"),
  r0 = 0,
  conf.level = 0.95
)
```

Arguments

`ratings`	n*m matrix or dataframe, n subjects m raters.
`model`	a character string specifying if a "oneway" model (default) with row effects random, or a "twoway" model with column and row effects random should be applied. You can specify just the initial letter.
`type`	a character string specifying if "consistency" (default) or "agreement" between raters should be estimated. If a '"oneway"' model is used, only "consistency" could be computed. You can specify just the initial letter.
`unit`	a character string specifying the unit of analysis: Must be one of "single" (default) or "average". You can specify just the initial letter.
`r0`	specification of the null hypothesis r $\leq$ r0. Note that a one sided test (H1: r > r0) is performed.
`conf.level`	confidence level of the interval.

Details

By this ICC-function three bugs are corrected that were found in the function icc of the irr package (version 0.84.1):
Due to the first bug the p-values of ICC(A,1) and ICC(A,k) are computed wrongly: McGraw & Wong (1996) use the variable "v" both for the computation of the CIs and for the computation of the p-values. But "v" takes different values in these calculations. In the implementation of icc{irr} (version 0.84.1) this fact is missed.
The second correction only affects the rare case of the residual mean square (of the twoway model) being zero, i.e. the case that the variance in the data may be explained completely by the two factors (Raters and Objects). In this case the F-value for determining all four twoway p-values is not correctly computet by icc.
The third correction addresses the problems arising in the rare cases of (a) no part or (b) nearly no part of variance may be explained by both factors.

Value

A list with class '"icclist"' containing the following components:

`$subjects`	the number of subjects examined.
`$raters`	the number of raters.
`$model`	a character string describing the selected model for the analysis.
`$type`	a character string describing the selected type of interrater reliability.
`$unit`	a character string describing the unit of analysis.
`$icc.name`	a character string specifying the name of ICC according to McGraw & Wong (1996).
`$value`	the intraclass correlation coefficient.
`$r0`	the specified null hypothesis.
`$Fvalue`	the value of the F-statistic.
`$df1`	the numerator degrees of freedom.
`$df2`	the denominator degrees of freedom.
`$p.value`	the p-value for a two-sided test.
`$conf.level`	the confidence level for the interval.
`$lbound`	the lower bound of the confidence interval.
`$ubound`	the upper bound of the confidence interval.

Author(s)

Matthias Gamer, Markus Brueckl

References

McGraw, K.O., & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. Psychological Methods, 1, 30–46.

Shrout, P.E., & Fleiss, J.L. (1979), Intraclass correlation: uses in assessing rater reliability. Psychological Bulletin, 86, 420–428.

See Also

icc, iccNA

Examples

```
# Example 1:
data(EbelFILL)
# EbelFILL is a rather arbitrary data set:
EbelFILL
# If twoway agreement ICCs are computed (e.g. the single
# measure) with icc{irr}, the 2nd df of F and thus the
# p-value is erroneous:
library(irr)
icc(EbelFILL, model="twoway", type="agreement")
# icc_corr calculates correctly:
icc_corr(EbelFILL, model="twoway", type="agreement")
#
# Example 2:
data(Consist)
# Consist exhibits a perfect consistency and
# a moderate absolute agreement between raters:
Consist
# If twoway ICCs are computed with icc{irr}, the F-value is smaller
# than zero (!) and thus the p-value is enourmously erroneous:
library(irr)
icc(Consist, model="twoway", type="consistency", unit="average")
# icc_corr calculates correctly:
icc_corr(Consist, model="twoway", type="consistency", unit="average")
#
# Example 3:
data(Indep)
# Indep exhibits zero variance between the raters just as
# well as between the objects:
Indep
# Errors occur, if twoway agreement ICCs are computed with icc{irr}:
# ICC(A,k) just as well as its CI-bounds are (falsely) positive
# and greater than 1...
icc(Indep, model="twoway", type="agreement", unit="average")
# ...but must be -Inf, just as icc_corr shows:
icc_corr(Indep, model="twoway", type="agreement", unit="average")
# ICC(A,1): 2nd df of F and thus the p-value are NaN
icc(Indep, model="twoway", type="agreement")
# icc_corr calculates correlctly:
icc_corr(Indep, model="twoway", type="agreement")
```

Indep *irrNA example data, showing perfect independence among raters and objects*

Description

This data set shows perfect independance among raters and objects.

Usage

```
data(Indep)
```

Format

A 2-dimensional data frame including column and row headers.

`IndepNA` *irrNA example data, showing NAs and perfect independence among raters and objects*

Description

This data set shows missing values (NAs) and perfect independance among raters and objects.

Usage

```
data(IndepNA)
```

Format

A 2-dimensional data frame including column and row headers and NAs.

`IndepW` *irrNA example data, showing perfect independence among raters and NAs*

Description

This data set shows missing values (NAs) and perfect independance among raters.

Usage

```
data(IndepW)
```

Format

A 2-dimensional data frame including column and row headers and NAs.

kendallNA *Kendall's coefficient of concordance W – generalized for randomly incomplete datasets*

Description

This function computes Kendall's coefficient of concordance W that is an index of interrater reliability for ordinal ratings. This function also works on incomplete datasets without any imputation of missing values or (row- or cloumn-wise) omissions of data.

Usage

```
kendallNA(X)
```

Arguments

X	n*m matrix or dataframe; n objects (rows), k raters (columns)

Details

This function is able to calculate W, also on randomly incomplete (i.e. unbalanced) data sets. Therefor it uses the mean Spearman's ρ of all pairwise comparisons, see Kendall (1962):

$$W = [1 + mean\rho_S * (k - 1)]/k$$

where k is the mean number of (pairwise) ratings per object and $mean\rho_S$ is calculated weighted, according to Taylor (1987), since the pairwise ρ_S are possibly based on a different number of ratings, what must be reflected in weights.
Thus, an imputation of missing values or (row- or cloumn-wise) omissions of data are obsolete. In case of complete datasets, it yields the same results as usual implementations of Kendall's W, except for tied ranks. In case of tied ranks, the (pairwise) correction of ρ_S is used, which (already with complete datasets) results in slightly different values than the tie correction explicitly specified for W.
More details are given in Brueckl (2011).

Value

amrho	mean Spearman's ρ
amk	mean number of (pairwise) ratings per object
W	Kendall's coefficient of concordance among raters
chisqu	value of the χ-squared test statistic
df	degrees of freedom
p	one-tailed type I error probability (statistical significance)

Author(s)

Markus Brueckl

References

Brueckl, M. (2011). Statistische Verfahren zur Ermittlung der Urteileruebereinstimmung. in: Altersbedingte Veraenderungen der Stimme und Sprechweise von Frauen, Berlin: Logos, 88–103.

Kendall, M.G. (1962). Rank correlation methods (3rd ed.). London: Griffin.

Taylor, J.M.G. (1987). Kendall's and Spearman's correlation coefficients in the presence of a blocking variable. Biometrics, 43, 409–416.

See Also

iccNA,kendall

Examples

```
# Example 1:
data(ConsistNA)
# ConsistNA exhibits missing values and a perfect concordance
# between raters:
ConsistNA
# Common W-algorithms fail, since each row as well as each
# column of ConsistNA exhibits unfilled cells and these missing
# data are omitted column-wise or row-wise:
library(irr)
# try here: kendall(ConsistNA)
# But the generalization of Kendall's W implemeted in irrNA
# is able to assess the perfect concordance, assuming that
# the data were at least ordinally scaled and not tied, e.g.
# that each rater really ranked the objects that he rated
# without giving equal ranks to two or more objects.
kendallNA(ConsistNA)
#
# Example 2:
data(IndepNA)
# IndepNA exhibits missing values and zero variance between
# the raters (just as well as between the objects):
IndepNA
# Common W-algorithms fail:
# try here: kendall(IndepNA)
# kendallNA includes all (rater-pairwise) available data in
# its calculation (e.g. only Objects 1--4 when Rater1 and
# Rater2 are correlated):
kendallNA(IndepNA)
#
# Example 3:
data(IndepW)
# IndepW exhibits missing values and a mean Spearman's rho,
# that equals zero:
IndepW
# Again, common W-algorithms fail:
# try here: kendall(IndepW)
# kendallNA includes all (rater-pairwise) available
# data:
```

```
kendallNA(IndepW)
```

Index