# Package: kappaGold (via r-universe)

November 19, 2024

**Title** Agreement of Nominal Scale Raters with a Gold Standard

**Version** 0.3.2

**Date** 2024-09-27

**Description** Estimate agreement of a group of raters with a gold standard rating on a nominal scale. For a single gold standard rater the average pairwise agreement of raters with this gold standard is provided. For a group of gold standard raters the approach of S. Vanbelle, A. Albert (2009) <doi:10.1007/s11336-009-9116-1> is implemented. Bias and standard error are estimated via delete-1 jackknife.

**License** MIT + file LICENSE

**Encoding** UTF-8

**LazyData** true

**Imports** future.apply (>= 1.6), purrr (>= 0.3), stats, tibble, tidyr

**Suggests** dplyr, irr, knitr, testthat (>= 3.0.0)

**RoxygenNote** 7.3.2

**Config/testthat/edition** 3

**Depends** R (>= 4.0)

**NeedsCompilation** no

**Author** Matthias Kuhn [aut, cre] (<https://orcid.org/0000-0003-2868-5155>), Jonas Breidenstein [aut]

**Maintainer** Matthias Kuhn <matthias.kuhn@tu-dresden.de>

**Repository** CRAN

**Date/Publication** 2024-09-30 16:50:02 UTC

**Config/pak/sysreqs** libicu-dev

# Contents



---

`diagnoses` *Psychiatric diagnoses*

---

## Description

N = 30 patients were given one of k = 5 diagnoses by some n = 6 psychiatrists out of 43 psychiatrists in total. The diagnoses are

1. Depression
2. PD (=Personality Disorder)
3. Schizophrenia
4. Neurosis
5. Other

## Usage

```
diagnoses
```

## Format

`diagnoses`**:**

A matrix with 30 rows and 6 columns:

**rater1** 1st rating of some six raters

**rater2** 2nd rating of some six raters

**rater3** 3rd rating of some six raters

**rater4** 4th rating of some six raters

**rater5** 5th rating of some six raters

**rater6** 6th rating of some six raters

## Details

A total of 43 psychiatrists provided diagnoses. In the actual study (Sandifer, Hordern, Timbury, & Green, 1968), between 6 and 10 psychiatrists from the pool of 43 were unsystematically selected to diagnose a subject. Fleiss randomly selected six diagnoses per subject to bring the number of assignments per patient down to a constant of six.

As there is not a fixed set of six raters the ratings from the same column are not related to each other. Therefore, compared to the dataset with the same name in package `irr`, we applied a permutation of the six ratings.

## References

Sandifer, M. G., Hordern, A., Timbury, G. C., & Green, L. M. Psychiatric diagnosis: A comparative study in North Carolina, London and Glasgow. British Journal of Psychiatry, 1968, 114, 1-9.

Fleiss, J. L. Measuring nominal scale agreement among many raters. Psychological Bulletin, 1971, 76(5), 378–382. doi:10.1037/h0031619

## See Also

This dataset is also available as `diagnoses` in the irr-package on CRAN.

---

`kappa2` *Cohen's kappa for nominal data*

---

## Description

The data of ratings must be stored in a two column object, each rater is a columns and the subjects are in the rows.

## Usage

```
kappa2(ratings, robust = FALSE, ratingScale = NULL)
```

## Arguments

| | |
|---|---|
| `ratings` | matrix (dimension nx2), containing the ratings as subjects by raters |
| `robust` | flag. Use robust estimate for random chance of agreement by Brennan-Prediger? |
| `ratingScale` | Possible levels for the rating. Or `NULL`. |

## Details

Every rating category is used and the levels are sorted. Weighting is currently not implemented.

## Value

list containing Cohen's kappa agreement measure (value) or `NULL` if no valid subjects

## See Also

irr::kappa2()

## Examples

```
# 2 raters have assessed 4 subjects into categories "A", "B" or "C"
# organize ratings as two column matrix, one row per subject rated
m <- rbind(sj1 = c("A", "A"),
           sj2 = c("C", "B"),
           sj3 = c("B", "C"),
           sj4 = c("C", "C"))

# Cohen's kappa -----
kappa2(ratings = m)

# robust variant ---------
kappa2(ratings = m, robust = TRUE)
```

---

# kappaGold *kappaGold package*

---

## Description

Estimate agreement with a gold-standard rating for nominal categories.

## Author(s)

**Maintainer**: Matthias Kuhn <matthias.kuhn@tu-dresden.de> (ORCID)

Authors:

- Jonas Breidenstein <jonas.breidenstein@tu-dresden.de>

---

# kappam_fleiss *Fleiss' kappa for multiple nominal-scale raters*

---

## Description

When multiple raters judge subjects on a nominal scale we can assess their agreement with Fleiss' kappa. It is a generalization of Cohen's Kappa for two raters and there are different variants how to assess chance agreement.

## Usage

```
kappam_fleiss(
  ratings,
  variant = c("fleiss", "conger", "robust", "uniform"),
  detail = FALSE,
  ratingScale = NULL
)
```

## Arguments

| | |
|---|---|
| ratings | matrix (subjects by raters), containing the ratings |
| variant | Which variant of kappa? Default is Fleiss (1971). Other options are Conger (1980) or robust variant. |
| detail | Should category-wise Kappas be computed? Only available for the Fleiss (1971) variant. |
| ratingScale | Specify possible levels for the rating. Default NULL means to use all unique levels from the sample. |

## Details

Different **variants** of Fleiss' kappa are implemented. By default (variant="fleiss"), the original Fleiss Kappa (1971) is calculated, together with an asymptotic standard error and test for kappa=0. It assumes that the raters involved are not assumed to be the same (one-way ANOVA setting). The marginal category proportions determine the chance agreement. Setting variant="conger" gives the variant of Conger (1980) that reduces to Cohen's kappa when m=2 raters. It assumes identical raters for the different subjects (two-way ANOVA setting). The chance agreement is based on the category proportions of each rater separately. Typically, the Conger variant yields slightly higher values than Fleiss kappa. variant="robust" assumes a chance agreement of two raters to be simply 1/q, where q is the number of categories (uniform model).

## Value

list containing Fleiss's kappa agreement measure (value) or NULL if no subjects

## See Also

irr::kappam.fleiss()

## Examples

```
# 4 subjects were rated by 3 raters in categories "1", "2" or "3"
# organize ratings as matrix with subjects in rows and raters in columns
m <- matrix(c("3", "2", "3",
              "2", "2", "1",
              "1", "3", "1",
              "2", "2", "3"), ncol = 3, byrow = TRUE)
kappam_fleiss(m)

# show category-wise kappas -----
```

```
kappam_fleiss(m, detail = TRUE)
```

---

kappam_gold *Agreement of a group of nominal-scale raters with a gold standard*

---

## Description

First, Cohen's kappa is calculated between each rater against the gold standard which is taken from the 1st column. The average of these kappas is returned as 'kappam_gold0'. The variant setting (`robust=`) is forwarded to Cohen's kappa. A bias-corrected version 'kappam_gold' and a corresponding confidence interval are provided as well via the jackknife method.

## Usage

```
kappam_gold(ratings, robust = FALSE, ratingScale = NULL, conf.level = 0.95)
```

## Arguments

| | |
|---|---|
| `ratings` | matrix subjects by raters |
| `robust` | flag. Use robust estimate for random chance of agreement by Brennan-Prediger? |
| `ratingScale` | Possible levels for the rating. Or `NULL`. |
| `conf.level` | confidence level for confidence interval |

## Value

list. agreement measures (raw and bias-corrected) kappa with confidence interval. Entry `raters` refers to the number of tested raters, not counting the reference rater

## Examples

```
# matrix with subjects in rows and raters in columns.
# 1st column is taken as goldstandard
m <- matrix(c("O", "G", "O",
              "G", "G", "R",
              "R", "R", "R",
              "G", "G", "O"), ncol = 3, byrow = TRUE)
kappam_gold(m)
```

kappam_vanbelle *Agreement between two groups of raters*

## Description

This function expands upon Cohen's and Fleiss' Kappa as measures for interrater agreement while taking into account the heterogeneity within each group.

## Usage

```
kappam_vanbelle(
  ratingsGr1,
  ratingsGr2,
  ratingScale = NULL,
  weights = c("unweighted", "linear", "quadratic"),
  conf.level = 0.95
)
```

## Arguments

| | |
|---|---|
| ratingsGr1 | matrix of subjects x raters for 1st group of raters |
| ratingsGr2 | matrix of subjects x raters for 2nd group of raters |
| ratingScale | character vector of the levels for the rating. Or NULL. |
| weights | optional weighting schemes: "unweighted", "linear","quadratic" |
| conf.level | confidence level for interval estimation |

## Details

Data need to be stored with raters in columns.

## Value

list. kappa agreement between two groups of raters

## References

Vanbelle, S., Albert, A. Agreement between Two Independent Groups of Raters. Psychometrika 74, 477–491 (2009). doi:10.1007/s11336-009-9116-1

## Examples

```
# compare rater1-rater2 vs rater3-rater6 from diagnoses-data
# (there is no systematic difference between both groups
#+as the raters are randomly selected per subject)
kappam_vanbelle(diagnoses[,1:2], diagnoses[,3:6])
```

kappa_test *Significance test for homogeneity of kappa coefficients*

## Description

When groups of different subjects are rated on a nominal scale. Assuming independence of subjects and their ratings between groups a chi-squared test for equality of kappa between these groups is performed. The test requires estimates of kappa and its standard error per group.

## Usage

```
kappa_test(kappas, val = "value0", se = "se0", conf.level = 0.95)
```

## Arguments

| | |
|---|---|
| kappas | list of kappas from different groups. It uses the kappa estimate and its standard error. |
| val | character. Name of field to extract kappa coefficient estimate. |
| se | character. Name of field to extract standard error of kappa. |
| conf.level | numeric. confidence level of confidence interval for overall kappa |

## Details

A common overall kappa coefficient across groups is estimated. The test statistic assesses the weighted squared deviance of the individual kappas from the overall kappa estimate. The weights depend on the provided standard errors.

## Value

list containing the test results, including the entries `statistic` and `p.value` (class `htest`)

## References

Joseph L. Fleiss, Statistical Methods for Rates and Proportions, 3rd ed., 2003, section 18.1

## Examples

```
# script concordance test on 34 clinical situations,
# rated by 39 students and 11 experts
kappa_stud <- kappam_fleiss(SC_test[, 1:39])
kappa_expert <- kappam_fleiss(SC_test[, 40:50])

# compare student and expert agreement
kappa_test(kappas = list(kappa_stud, kappa_expert))
```

SC_test *Script concordance test (SCT).*

## Description

In medical education, the script concordance test (SCT) (Charlin, Gagnon, Sibert, & Van der Vleuten, 2002) is used to score physicians or medical students in their ability to solve clinical situations as compared to answers given by experts. The test consists of a number of items to be evaluated on a 5-point Likert scale.

## Usage

```
SC_test
```

## Format

A matrix with 34 rows and 50 columns. Columns 1 to 39 are student raters, columns 40 to 50 are experts. Each rater applies to each clinical situation one of five levels ranging from -2 to 2 with the following meaning:

**-2** The assumption is practically eliminated;

**-1** The assumption becomes less likely;

**0** The information has no effect on the assumption;

**+1** The assumption becomes more likely;

**+2** The assumption is virtually the only possible one.

## Details

Each item represents a clinical situation (called an 'assumption') likely to be encountered in the physician's practice. The situation has to be unclear, even for an expert. The task of the subjects being evaluated is to consider the effect of new information on the assumption to solve the situation. The data incorporates 50 raters, 39 students and 11 experts.

Each rater judges the same 34 assumptions.

## Source

Sophie Vanbelle (personal communication, 2021)

## References

Vanbelle, S., Albert, A. Agreement between Two Independent Groups of Raters. Psychometrika 74, 477–491 (2009). doi:10.1007/s1133600991161

simulKappa *Simulate rating data and calculate agreement with gold standard*

## Description

The function generates simulation data according to given categories and probabilities. and can repeatedly apply function `kappam_gold()`. Currently, there is no variation in probabilities from rater to rater, only sampling variability from multinomial distribution is at work.

## Usage

```
simulKappa(nRater, cats, nSubj, probs, mcSim = 10, simOnly = FALSE)
```

## Arguments

| | |
|---|---|
| `nRater` | numeric. number of raters. |
| `cats` | categories specified either as character vector or just the numbers of categories. |
| `nSubj` | numeric. number of subjects per gold standard category. Either a single number or as vector of numbers per category, e.g. for non-balanced situation. |
| `probs` | numeric square matrix (nCat x nCat) with classification probabilities. Row `i` has probabilities of rater categorization for subjects of category `i` (gold standard). |
| `mcSim` | numeric. Number of Monte-Carlo simulations. |
| `simOnly` | logical. Need only simulation data? Default is `FALSE`. |

## Details

This function is future-aware for the repeated evaluation of `kappam_gold()` that is triggered by this function.

## Value

dataframe of kappa-gold on the simulated datasets or (when `simOnly=TRUE`) list of length `mcSim` with each element a simulated data set with goldrating in first column and then the raters.

## Examples

```
# repeatedly estimate agreement with goldstandard for simulated data
simulKappa(nRater = 8, cats = 3, nSubj = 11,
           # assumed prob for classification by raters
           probs = matrix(c(.6, .2, .1, # subjects of cat 1
                            .3, .4, .3, # subjects of cat 2
                            .1, .4, .5  # subjects of cat 3
           ), nrow = 3, byrow = TRUE))
```

`stagingData` *Staging of colorectal carcinoma*

## Description

Staging of carcinoma is done by different medical professions. Gold standard is the (histo-)pathological rating of a tissue sample but this information typically only becomes available late, after surgery. However prior to surgery the carcinoma is also staged by radiologists in the clinical setting on the basis of MRI scans.

## Usage

```
stagingData
```

## Format

A data frame with 21 observations and 6 variables:

**patho** the (histo-)pathological staging (gold standard) with categories I, II or III

**rad1** the clinical staging with categories I, II or III by radiologist 1

**rad2** the clinical staging with categories I, II or III by radiologist 2

**rad3** the clinical staging with categories I, II or III by radiologist 3

**rad4** the clinical staging with categories I, II or III by radiologist 4

**rad5** the clinical staging with categories I, II or III by radiologist 5

## Details

These fictitious data were inspired by the OCUM trial. The simulation uses the following two assumptions: over-staging occurs more frequently than under-staging and an error by two categories is less likely than an error by only one category.

Stages conform to the UICC classification according to the TNM classification. Note that cases in stage IV do not appear in this data set and that the following description of stages is simplified.

1. **I** Until T2, N0, M0
2. **II** From T3, N0, M0
3. **III** Any T, N1/N2, M0

## Source

simulated data

## References

Kreis, M. E. et al., MRI-Based Use of Neoadjuvant Chemoradiotherapy in Rectal Carcinoma: Surgical Quality and Histopathological Outcome of the OCUM Trial doi:10.1245/s1043401907696y

`victorinox` *delete-1 jackknife estimator*

## Description

Quick simple jackknife routine to estimate bias and standard error of an estimator.

## Usage

```
victorinox(est, idx)
```

## Arguments

| | |
|---|---|
| `est` | estimator function |
| `idx` | maximal index vector for data of estimator |

## Value

list with jackknife information, bias and SE

## References

https://de.wikipedia.org/wiki/Jackknife-Methode

# Index