PART IV

APPENDICES

List of Appendices

APPENDIX A

Data Tables

This appendix contains 4 tables of ratings that were used in the text to illustrate the implementation of some procedures. Some of these tables were deemed too large to be included in the chapters they were used and are therefore presented here in their entirety. Here are the 4 tables:

- Table A.1 contains rating data on the classification of 29 Stickleback fishes' color displays into 5 intensity Levels, by 4 Experienced Raters.
- Table A.2 contains the distribution of raters by fish (used as subject) and the fish degree of nuptial coloration.
- Table A.3 contains 11 students' linguistics test scores presented in a wide format type.
- Table A.4 contains 11 students' linguistics test scores presented in a long format type.

Table A.1

Classification of 29 Stickleback Fishes' Color Displays into 5 Intensity Levels, by 4 Experienced Raters[a].

Fish	Rater 1	Rater 2	Rater 3	Rater 4
1	5	5	5	5
2	3	1	3	1
3	5	5	5	5
4	3	1	3	1
5	5	5	4	5
6	1	2	3	3
7	3	1	1	1
8	1	1	1	3
9	3	3	4	4
10	1	3	1	1
11	5	5	5	5
12	1	1	1	1
13	1	1	1	1
14	1	1	1	1
15	3	3	4	3
16	1	3	3	4
17	4	4	5	5
18	5	5	5	5
19	5	3	3	3
20	3	2	3	3
21	5	3	5	5
22	3	3	3	4
23	1	1	1	1
24	1	1	1	1
25	1	1	3	3
26	3	3	3	1
27	3	3	1	1
28	3	3	1	1
29	3	3	2	5

[a]To download this table, follow the link https://agreestat.com/books/cac5/

Table A.2: Distribution of Raters by Fish and their Degree of Nuptial Coloration

Fish	1	2	3	4	5	Total
1	0	0	0	0	4	4
2	2	0	2	0	0	4
3	0	0	0	0	4	4
4	2	0	2	0	0	4
5	0	0	0	1	3	4
6	1	1	2	0	0	4
7	3	0	1	0	0	4
8	3	0	1	0	0	4
9	0	0	2	2	0	4
10	3	0	1	0	0	4
11	0	0	0	0	4	4
12	4	0	0	0	0	4
13	4	0	0	0	0	4
14	4	0	0	0	0	4
15	0	0	3	1	0	4
16	1	0	2	1	0	4
17	0	0	0	2	2	4
18	0	0	0	0	4	4
19	0	0	3	0	1	4
20	0	1	3	0	0	4
21	0	0	1	0	3	4
22	0	0	3	1	0	4
23	4	0	0	0	0	4
24	4	0	0	0	0	4
25	2	0	2	0	0	4
26	1	0	3	0	0	4
27	2	0	2	0	0	4
28	2	0	2	0	0	4
29	0	1	2	0	1	4

Table A.3: Eleven Students' Linguistics Test Scores in a Wide Format[a]

NAME	VERSION	COMPONENT	RATER1	RATER2	RATER3	RATER4
Suzan	M	Grammar	3		3.5	3
Suzan	M	Vocabulary	3		3.5	3.5
Suzan	M	Fluency	3		3.5	3
Suzan	M	Pronunciation	3		3.5	3
Suzan	M	Overall	3		3.5	3
Mary	M	Grammar	4	4	4	
Mary	M	Vocabulary	3.5	4	3.5	
Mary	M	Fluency	3	3.5	3.5	
Mary	M	Pronunciation	3	4	3.5	
Mary	M	Overall	3.5	3.5	3.5	
Lee	M	Grammar		3.5	3.5	3.5
Lee	M	Vocabulary		3.5	3.5	3.5
Lee	M	Fluency		3.5	3.5	3
Lee	M	Pronunciation		3	3.5	3.5
Lee	M	Overall		3.5	3.5	3.5
Amber	VCP	Listening	5	5		5
Amber	VCP	Grammar	5	5		5
Amber	VCP	Vocabulary	5	5		5
Amber	VCP	Fluency	5	5		4.5
Amber	VCP	Pronunciation	5	5		5
Amber	VCP	Overall	5	5		5
Ricardo	VCP	Listening	4.5		5	4.5
Ricardo	VCP	Grammar	4.5		4.5	4
Ricardo	VCP	Vocabulary	4		4.5	4
Ricardo	VCP	Fluency	4		5	4
Ricardo	VCP	Pronunciation	4.5		4.5	4
Ricardo	VCP	Overall	4.5		4.5	4
Lee	VCP	Listening	4.5	5	4.5	
Lee	VCP	Grammar	4.5	4.5	4	
Lee	VCP	Vocabulary	4.5	4.5	4.5	
Lee	VCP	Fluency	4	4.5	4	
Lee	VCP	Pronunciation	5	4.5	4.5	
Lee	VCP	Overall	4.5	4.5	4.5	
Mary	VCP	Listening	4	4.5		4.5
Mary	VCP	Grammar	4	4		4

Continued on next page

[a]To download this table, follow the link https://agreestat.com/books/cac5/

Table A.3 – continued from previous page

NAME	VERSION	COMPONENT	RATER1	RATER2	RATER3	RATER4
Mary	VCP	Vocabulary	4	4.5		4.5
Mary	VCP	Fluency	4	4.5		4
Mary	VCP	Pronunciation	4	4.5		4
Mary	VCP	Overall	4	4.5		4.5
Suzan	VCP	Listening	4.5	4.5		4.5
Suzan	VCP	Grammar	4	3.5		4
Suzan	VCP	Vocabulary	3.5	3.5		4
Suzan	VCP	Fluency	3.5	4		4
Suzan	VCP	Pronunciation	3.5	3.5		4
Suzan	VCP	Overall	4	3.5		4
Yin	VCP	Listening		3.5	3.5	4
Yin	VCP	Grammar		4	3.5	3
Yin	VCP	Vocabulary		3.5	3.5	3.5
Yin	VCP	Fluency		3.5	3.5	3.5
Yin	VCP	Pronunciation		3.5	3	3.5
Yin	VCP	Overall		3.5	3.5	3.5
Isaac	VCP	Listening	4.5	4.5	4.5	
Isaac	VCP	Grammar	4.5	4.5	4	
Isaac	VCP	Vocabulary	4.5	4.5	4	
Isaac	VCP	Fluency	4.5	4.5	4.5	
Isaac	VCP	Pronunciation	4.5	4.5	4.5	
Isaac	VCP	Overall	4.5	4.5	4.5	
Viktor	VCP	Listening	4.5	4.5		4.5
Viktor	VCP	Grammar	4.5	4		4
Viktor	VCP	Vocabulary	3.5	4		4.5
Viktor	VCP	Fluency	4	4		4.5
Viktor	VCP	Pronunciation	4	4.5		4.5
Viktor	VCP	Overall	4	4		4.5
David	VCP	Listening	4.5	4.5	4	
David	VCP	Grammar	3.5	4.5	4	
David	VCP	Vocabulary	4	4	4	
David	VCP	Fluency	4	4	4	
David	VCP	Pronunciation	3.5	3.5	3	
David	VCP	Overall	4	4	4	
Yanick	VCP	Listening	4.5	4.5	4	

Continued on next page

[a]To download this table, follow the link https://agreestat.com/books/cac5/

Table A.3 – continued from previous page

NAME	VERSION	COMPONENT	RATER1	RATER2	RATER3	RATER4
Yanick	VCP	Grammar	3.5	4	4	
Yanick	VCP	Vocabulary	4	4	3.5	
Yanick	VCP	Fluency	3.5	4.5	4	
Yanick	VCP	Pronunciation	4	3.5	4	
Yanick	VCP	Overall	4	4	4	
Jasmine	VCP	Listening		4.5	4.5	4.5
Jasmine	VCP	Grammar		4.5	4.5	4.5
Jasmine	VCP	Vocabulary		5	4.5	5
Jasmine	VCP	Fluency		4	4.5	4.5
Jasmine	VCP	Pronunciation		4	4	4
Jasmine	VCP	Overall		4.5	4.5	4.5
David	VCE	Listening	4.5	4.5		4.5
David	VCE	Grammar	3.5	4.5		4
David	VCE	Vocabulary	4	4		3.5
David	VCE	Fluency	3.5	3.5		3.5
David	VCE	Pronunciation	3.5	3.5		3.5
David	VCE	Overall	3.5	4		3.5
Amber	VCE	Listening	5	4.5		5
Amber	VCE	Grammar	5	5		5
Amber	VCE	Vocabulary	5	5		5
Amber	VCE	Fluency	5	4		5
Amber	VCE	Pronunciation	5	5		5
Amber	VCE	Overall	5	5		5
Isaac	VCE	Listening	4	4		3.5
Isaac	VCE	Grammar	4	3.5		3.5
Isaac	VCE	Vocabulary	4	3.5		3.5
Isaac	VCE	Fluency	4	4		4
Isaac	VCE	Pronunciation	4	4		4
Isaac	VCE	Overall	4	4		3.5
Mary	VCE	Listening	4	4.5		4.5
Mary	VCE	Grammar	4	4		4
Mary	VCE	Vocabulary	4	4		4
Mary	VCE	Fluency	4	4		4
Mary	VCE	Pronunciation	4	4		4.5
Mary	VCE	Overall	4	4		4

Continued on next page

[a]To download this table, follow the link https://agreestat.com/books/cac5/

Table A.3 – continued from previous page

NAME	VERSION	COMPONENT	RATER1	RATER2	RATER3	RATER4
Suzan	VCE	Listening	4		4.5	4
Suzan	VCE	Grammar	3.5		4	3.5
Suzan	VCE	Vocabulary	3.5		4	3.5
Suzan	VCE	Fluency	3.5		4	3.5
Suzan	VCE	Pronunciation	3.5		4	3.5
Suzan	VCE	Overall	3.5		4	3.5
Yin	VCE	Listening		2.5	3	3
Yin	VCE	Grammar		2.5	3	3
Yin	VCE	Vocabulary		2.5	3	3
Yin	VCE	Fluency		2.5	3	2.5
Yin	VCE	Pronunciation		2.5	3	2.5
Yin	VCE	Overall		2.5	3	3
Viktor	VCE	Listening		4.5	4	3.5
Viktor	VCE	Grammar		3.5	3.5	3
Viktor	VCE	Vocabulary		3.5	3	3
Viktor	VCE	Fluency		3	3.5	3
Viktor	VCE	Pronunciation		3	3.5	3
Viktor	VCE	Overall		3.5	3.5	3.5

[a]To download this table, follow the link https://agreestat.com/books/cac5/

Table A.4: Eleven Students' Linguistics Test Scores in a Long Format[a]

NAME	VERSION	RATER	FLUENCY	GRAMMAR	LISTEN	OVERALL	PRONU	VOCAB
Suzan	M	Rater 1	3	3		3	3	3
Suzan	M	Rater 3	3.5	3.5		3.5	3.5	3.5
Suzan	M	Rater 4	3	3		3	3	3.5
Mary	M	Rater 1	3	4		3.5	3	3.5
Mary	M	Rater 2	3.5	4		3.5	4	4
Mary	M	Rater 3	3.5	4		3.5	3.5	3.5
Lee	M	Rater 2	3.5	3.5		3.5	3	3.5
Lee	M	Rater 3	3.5	3.5		3.5	3.5	3.5
Lee	M	Rater 4	3	3.5		3.5	3.5	3.5
Amber	VCP	Rater 1	5	5	5	5	5	5
Amber	VCP	Rater 2	5	5	5	5	5	5
Amber	VCP	Rater 4	4.5	5	5	5	5	5
Ricardo	VCP	Rater 1	4	4.5	4.5	4.5	4.5	4
Ricardo	VCP	Rater 3	5	4.5	5	4.5	4.5	4.5
Ricardo	VCP	Rater 4	4	4	4.5	4	4	4
Lee	VCP	Rater 1	4	4.5	4.5	4.5	5	4.5
Lee	VCP	Rater 2	4.5	4.5	5	4.5	4.5	4.5
Lee	VCP	Rater 3	4	4	4.5	4.5	4.5	4.5
Mary	VCP	Rater 1	4	4	4	4	4	4
Mary	VCP	Rater 2	4.5	4	4.5	4.5	4.5	4.5
Mary	VCP	Rater 4	4	4	4.5	4.5	4	4.5
Suzan	VCP	Rater 1	3.5	4	4.5	4	3.5	3.5
Suzan	VCP	Rater 2	4	3.5	4.5	3.5	3.5	3.5
Suzan	VCP	Rater 4	4	4	4.5	4	4	4
Yin	VCP	Rater 2	3.5	4	3.5	3.5	3.5	3.5
Yin	VCP	Rater 3	3.5	3.5	3.5	3.5	3	3.5
Yin	VCP	Rater 4	3.5	3	4	3.5	3.5	3.5
Isaac	VCP	Rater 1	4.5	4.5	4.5	4.5	4.5	4.5
Isaac	VCP	Rater 2	4.5	4.5	4.5	4.5	4.5	4.5
Isaac	VCP	Rater 3	4.5	4	4.5	4.5	4.5	4
Viktor	VCP	Rater 1	4	4.5	4.5	4	4	3.5
Viktor	VCP	Rater 2	4	4	4.5	4	4.5	4
Viktor	VCP	Rater 4	4.5	4	4.5	4.5	4.5	4.5
David	VCP	Rater 1	4	3.5	4.5	4	3.5	4
David	VCP	Rater 2	4	4.5	4.5	4	3.5	4

Continued on next page

[a]This table can be downloaded at the following URL: https://agreestat.com/books/cac5/

Table A.4 – continued from previous page

NAME	VERSION	RATER	FLUENCY	GRAMMAR	LISTEN	OVERALL	PRONU	VOCAB
Yanick	VCP	Rater 1	3.5	3.5	4.5	4	4	4
Yanick	VCP	Rater 2	4.5	4	4.5	4	3.5	4
Yanick	VCP	Rater 3	4	4	4	4	4	3.5
Jasmine	VCP	Rater 2	4	4.5	4.5	4.5	4	5
Jasmine	VCP	Rater 3	4.5	4.5	4.5	4.5	4	4.5
Jasmine	VCP	Rater 4	4.5	4.5	4.5	4.5	4	5
David	VCE	Rater 1	3.5	3.5	4.5	3.5	3.5	4
David	VCE	Rater 2	3.5	4.5	4.5	4	3.5	4
David	VCE	Rater 4	3.5	4	4.5	3.5	3.5	3.5
Amber	VCE	Rater 1	5	5	5	5	5	5
Amber	VCE	Rater 2	4	5	4.5	5	5	5
Amber	VCE	Rater 4	5	5	5	5	5	5
Isaac	VCE	Rater 1	4	4	4	4	4	4
Isaac	VCE	Rater 2	4	3.5	4	4	4	3.5
Isaac	VCE	Rater 4	4	3.5	3.5	3.5	4	3.5
Mary	VCE	Rater 1	4	4	4	4	4	4
Mary	VCE	Rater 2	4	4	4.5	4	4	4
Mary	VCE	Rater 4	4	4	4.5	4	4.5	4
Suzan	VCE	Rater 1	3.5	3.5	4	3.5	3.5	3.5
Suzan	VCE	Rater 3	4	4	4.5	4	4	4
Suzan	VCE	Rater 4	3.5	3.5	4	3.5	3.5	3.5
Yin	VCE	Rater 2	2.5	2.5	2.5	2.5	2.5	2.5
Yin	VCE	Rater 3	3	3	3	3	3	3
Yin	VCE	Rater 4	2.5	3	3	3	2.5	3
Viktor	VCE	Rater 2	3	3.5	4.5	3.5	3	3.5
Viktor	VCE	Rater 3	3.5	3.5	4	3.5	3.5	3
Viktor	VCE	Rater 4	3	3	3.5	3.5	3	3

[a]To download this table, follow the link https://agreestat.com/books/cac5/

APPENDIX B

Software Solutions

This appendix provides a brief discussion of the software solutions available to researchers for computing inter-rater reliability coefficients. The list of software packages presented here is far from being exhaustive. It merely represents my short list of products that I recommend the readers of this book to consider. Many packages offer several options for computing agreement coefficients, although the number of built-in procedures is quite limited. Specialized add-in packages, functions, or macros written by independent researcher-programmers compensate this deficiency to some extent. Among the statistical packages considered here are R, SAS, SPSS, and STATA, with a particular emphasis on R and SAS. I will also mention some freely-available online calculators, which generally have limited capability. For MS Excel, AgreeStat developed by the author of this book, is a user-friendly Excel-based software that is commercially available in Windows and Mac versions. The Mac version of AgreeStat requires Mac Office 2011 or a more recent version.

B.1 The R Software

The R package has become an immensely popular statistical package across the world. If you are going to do statistical analysis on a regular basis for many years, and you do not know which statistical software to learn, this is one you will want to give a very serious consideration to. No doubt. You will enjoy the assistance of an extended online support group where you will be able to ask questions. Moreover, the product is entirely free, and can be downloaded at `http://www.r-project.com`. Numerous quality books have been published to help practitioners and scientists learn how to use it.

R is an interactive computing environment that makes a large collection of statistical functions available to you. Using R is about finding the right function and learning how to use it. R gives you the opportunity to develop your own functions for performing routine tasks as well as develop completely new packages for advanced

users. Those who are new to R might be interested in the PDF file entitled "Using R for Introductory Statistics" prepared by John Verzani. It provides a short and friendly introduction to the R package as well as a good overview of its capabilities. It can be downloaded at,

http://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf

A key feature of R is the opportunity to write your own functions to perform special analyses. Moreover, several functions aimed at performing similar analyses are often organized in a package. You must install the R package in an R session so that you can access its many functions. I will first review so R packages of interest developed by various authors, before describing a few specific R function that are freely available online.

B.1.1 *R Packages for Computing Inter-Rater Reliability Coefficients*

Several R packages are available to practitioners for computing agreement coefficients. Here is a non-exhaustive list of them:

- `irrCAC`: *Computing Chance-Corrected Agreement Coefficients (CAC)*, by Gwet, K.L.
  This R package was developed by the author of this book and is available on the Comprehensive R Archive Network (CRAN). This package implements most agreement coefficients discussed in this book with the exception of those special conditional coefficients discussed in chapter 8. A detailed description of this package can be obtained from the following URL:

  https://cran.r-project.org/web/packages/irrCAC/index.html

- `irr`: *Various Coefficients of Interrater Reliability and Agreement,* by Matthias Gamer, Jim Lemon, Ian Fellows, and Puspendra Singh.
  This R package offers several functions that implement various agreement coefficients. A few coefficients implemented in `irr` are not implemented in `irrCAC`, and vice-versa, there many coefficients in `irrCAC` are not implemented in the `irr` package. A detailed description of this package can be obtained from the following URL:

  https://cran.r-project.org/web/packages/irr/index.html.

- `rel`: *Reliability Coefficients*, by Riccardo LoMartire.
  This package provides point estimates with confidence intervals for agreement coefficients proposed by Bennett et al. (1954), Cohen (1960), Conger (1980),

Fleiss (1971), Gwet (2008a), Krippendorff (1970) and a few more. Note that this package implemented Fleiss' generalized kappa using the standard error expression proposed by Fleiss et al. (1979). However, the correct expression of this standard error was proposed by Gwet (2020). A detailed description of this package can be obtained from the following URL:

https://cran.r-project.org/web/packages/rel/index.html.

- `icr`: *Compute Krippendorff's Alpha,* by Alexander Staudt, and Pierre L'Ecuyer. This small package focuses on Krippendorff's alpha coefficient and is described in details on the following webpage:

https://cran.r-project.org/web/packages/icr/index.html.

B.1.2 *Some R Functions for Computing Inter-Rater Reliability Coefficients*

I wrote several stand-alone R functions to compute most chance-corrected agreement coefficients presented in this book. Readers familiar with the R environment, could download these functions and modify them to fit their specific needs or just use them unmodified. Standard errors, confidence intervals, as well as p-values associated with these coefficients are calculated by these functions. These R functions can all be downloaded from the following URL:

`https://agreestat.com/software/default.html#rpackage`

They are organized in three R script files, corresponding to the three ways your ratings must be organized:

- **agree.coeff2.r**
  The functions contained in this script file compute various agreement coefficients and their standard errors when dealing with two raters, and ratings that are organized in a contingency table (or a square matrix) showing counts of subjects by rater and category. You may use this format if each rater rated all subjects. Otherwise subjects rated by one rater and not by the other may not be properly classified. In this case, you should have two columns of raw scores, and use one of the functions in the script file agree.coeff3.raw.r.

- **agree.coeff3.dist.r**
  The functions contained in this script file compute various agreement coefficients and their standard errors when dealing with multiple raters, and ratings that are organized in the form of an $n \times q$ table showing counts of raters by subject and category. Here n represents the number of subjects and q the number of categories.

- **agree.coeff3.raw.r**
  The functions contained in this script file compute various agreement coefficients and their standard errors when dealing with multiple raters, and rating data organized in the form of an $n \times r$ table showing the (alphanumeric) raw ratings that the raters assigned to the subjects. Here n represents the number of subjects and r the number of raters. The data is presented in the form of n rows containing r ratings each.

- **weights.gen.r**
  The functions in this script file generate various weights to be used when computing weighted agreement coefficients.

In order to use any of the functions contained in these script files, you need to read the appropriate script into R. If you want to use the functions contained in "agree.coeff2.r" for example, then you will read this file into R as follows:

```
>source("C:\\AdvancedAnalytics\\R Scripts\\agree.coeff2.r")
```

R FUNCTIONS IN SCRIPT FILE **agree.coeff2.r**

If your analysis is limited to two raters, then you may organize your data in a contingency table that shows the count of subjects by rater and by category. Table B.1.1 is an example of such data where two neurologists classified 65 patients who suffer from Multiple Sclerosis into 4 diagnostic categories.

Table B.1.1: Diagnostic Classification of Multiple Sclerosis Patients by Two Neurologists[a]

New Orleans Neurologist	Winnipeg Neurologist 1	2	3	4
1	5	3	0	0
2	3	11	4	0
3	2	13	3	4
4	1	2	4	14

[a]From Landis and G. (1977)

Here is the list of functions in the script file `agree.coeff2.r`:

(1) `kappa2.table` (for Cohen's unweighted and weighted kappa coefficients)

(2) `scott2.table` (for Scott's unweighted and weighted Pi coefficients)

(3) `gwet.ac1.table` (for Gwet's unweighted and weighted AC_1 coefficients)

(4) `bp2.table` (for Brennan-Prediger unweighted and weighted coefficients)

(5) `krippen2.table` (for Krippendorff alpha coefficients)

All these functions operate the same way. Therefore, only the first of these functions named `kappa2.table` is discussed here in details. The same discussion applies to the other functions as well.

`kappa2.table`: *Cohen's kappa coefficient for 2 raters*

Description

This function calculates the unweighted as well as the weighted Cohen's kappa coefficients for 2 raters whose ratings are summarized in a square contingency table such as Table B.1.1. Some cells may have 0 values. However, the number of rows and columns must be equal.

Usage

```
kappa2.table(ratings,weights=diag(ncol(ratings)),conflev=0.95,
          N=Inf,print=TRUE)
```

Arguments

Of all arguments that this function takes, only the first one is required. The remaining arguments are all optional.

- `ratings`: A $q \times q$ matrix, where q is the number of categories. This is the only argument you must specify if you want the unweighted analysis,
- `weights`: A $q \times q$ matrix of weights. The default argument is the diagonal matrix where all diagonal numbers equal to 1, and all off-diagonal numbers equal to 0. This special weight matrix leads to the unweighted analysis. You may specify your own $q \times q$ weight matrix here as `weights=own.weights`. If you want to use quadratic weights with Table B.1.1 data for example, then the weights parameter would be `weights=quadratic.weights(1:4)`. You may want to look at the `weights.gen.r` script for a complete reference of all weight functions.
- `conflev`: The confidence level associated with the agreement coefficient's confidence interval.

- `N`: An optional parameter representing the total number of subjects in the target subject population. Its default value is infinity, which for all practical purposes assumes the target subject population to be very large and will not require any finite-population correction when computing the standard error.

- `print`: An optional logical parameter which takes the default value of TRUE if you also want the function to output the results on the screen. Set this parameter to FALSE if you do not want the results to be displayed on the screen. Setting this parameter to FALSE is recommended if this function is used as part of another routine.

Details

`kappa2.table` can accept data in the form of a matrix or in the form of a data frame as long as the input data supplied can be interpreted as a square matrix. To do the weighted analysis, you may create your own weight matrix, or use one of the many existing weight-generating functions in the `weights.ge.r` script file. Each weight function takes a single mandatory parameter, which is a vector containing all categories used in the study. *The weight functions always sort all numeric-type category vectors in ascending order. Consequently, the weighted coefficients are computed properly only if the positions of the columns and rows in the input dataset are in the same order as the corresponding categories in the sorted category vector. For alphanumeric-type category vectors, they are assumed to be already ranked following an order that is meaningful to the researcher. That is adjacent columns and adjacent rows are associated with categories that can be considered as partial agreement.*

Value

Calling the function `kappa2.table` returns the following 5 values:

- `pa`: the percent agreement.

- `pe`: the percent chance agreement.

- `kappa`: Cohen's kappa coefficient.

- `stderr`: the standard error of Cohen's kappa.

- `p.value`: the p-value of the kappa coefficient.

Examples

```
>ratings<-matrix(c(5, 3, 0, 0, # creates a matrix with Table B.1.1 data
+                  3, 11, 4, 0,
+                  2, 13, 3, 4,
+                  1, 2, 4, 14),ncol=4,byrow=T)

# to compute unweighted kappa, its standard error and more
>kappa2.table(ratings)
```

The results displayed on the screen will look like this:

```
Cohen's Kappa Coefficient
=========================
Percent agreement: 0.4782609 Percent chance agreement: 0.2583491
Kappa coefficient: 0.2965166 Standard error: 0.07850387
95% Confidence Interval: ( 0.1398645, 0.4531686)
P-value: 0.0003361083
```

```
# to compute weighted kappa with quadratic weights
>kappa2.table(ratings,quadratic.weights(1:4))
# the above call assumes the script file weights.gen.r was read into R, and
```

the results obtained are the following:

```
Cohen's Kappa Coefficient
=========================
Percent agreement: 0.9098229 Percent chance agreement: 0.7591542
Kappa coefficient: 0.6255814 Standard error: 0.07873187
95% Confidence Interval: ( 0.4684744, 0.7826884)
P-value: 2.749756e-11
```

R FUNCTIONS IN SCRIPT FILE **agree.coeff3.dist.r**

If your experiment involves three raters or more you can no longer summarize the ratings in a contingency table as previously done for the case of two raters. One option is to present that data in the form of a table where each row represents one subject, each column represents one category, and each table cell represents the number of raters who classified the specified subject into the specified category. Such a table shows the distribution of raters by subject and by category. Table B.1.2 is an example of such data where six raters classified 4 patients into 5 diagnostic categories.

Table B.1.2: Distribution of 6 Raters by Subject and Category[a]

Subject	Category				
	Depression	Personality Disorder	Schizophrenia	Neurosis	Other
A	0	0	0	6	0
B	0	1	4	0	1
C	2	0	4	0	0
D	0	3	3	0	0

[a]An extract of Table 1 of Fleiss (1971)

The following functions contained in the script file `agree.coeff3.dist.r` are what you will need to analyze rating data such as described in Table B.1.2:

(1) `fleiss.kappa.dist` (for Fleiss's unweighted and weighted kappa coefficients)

(3) `gwet.ac1.dist` (for Gwet's unweighted and weighted AC_1 coefficients)

(4) `bp.coeff.dist` (for Brennan-Prediger unweighted and weighted coefficients)

(5) `krippen.alpha.dist` (for Krippendorff unweighted and weighted alpha coefficients)

All these functions operate the same way. Therefore, only the first of these functions named `fleiss.kappa.dist` is discussed here in details. The same discussion applies to the other functions as well.

`fleiss.kappa.dist`: *Fleiss' kappa coefficient for multiple raters*

Description

This function calculates the unweighted as well as the weighted Fleiss' generalized kappa coefficients for multiple raters whose ratings are presented in the form of a distribution of raters by subject and category such as in Table B.1.2. A table cell may have a 0 value if none of the raters classified the subject into the category associated with that cell. The number of raters may vary by subject leading to a table with different row totals. That will be the case when the experiment generated missing ratings, with subjects being rated by a different number of raters.

Usage

```
fleiss.kappa.dist(ratings,weights="unweighted",conflev=0.95,
          N=Inf,print=TRUE)
```

Arguments

Of all arguments used by this function, only the first one is required, the remaining arguments being all optional. If your goal is limited to unweighted statistics, then the simple function call `fleiss.kappa.dist(ratings)` *is sufficient to produce Fleiss' generalized kappa along with it standard error, confidence interval, and p-value.*

- `ratings`: This is an $n \times q$ matrix or data frame (or matrix), where n is the number of subjects, and q the number of categories. This is the only argument that must be specified if you want an unweighted analysis,

- `weights`: This is a $q \times q$ matrix of weights. The default argument is "unweighted". With this option, the function will create a diagonal weight matrix with all diagonal numbers equal to 1, and all off-diagonal numbers equal to 0. This special weight matrix leads to the unweighted analysis. You may create your own $q \times q$ weight matrix (e.g. `own.weights`) and assign it to the weights parameter as `weights=own.weights`. If you want to use quadratic weights with Table B.1.2 data for example, then the weights parameter would be `weights=quadratic.weights(1:5)`. You may want to look at the `weights.gen.r` script for a complete reference of all weight functions available.

- `conflev`: The confidence level associated with the agreement coefficient's confidence interval.

- `N`: An optional parameter representing the total number of subjects in the target subject population. Its default value is infinity, which for all practical purposes assumes the target subject population to be very large and will not require any finite-population correction when computing the standard error.

- `print`: An optional logical parameter which takes the default value of TRUE if you also want the function to output the results on the screen. Set this parameter to FALSE if you do not want the results to be displayed on the screen.

Details

`fleiss.kappa.dist` can accept input data in the form of a matrix or in the form of a data frame as long as the input data supplied can be interpreted as a

matrix. To do the weighted analysis, you may create your own weight matrix, or use one of the many existing weight-generating functions in the `weights.ge.r` script. Each weight function takes a single mandatory parameter, which is a vector containing all categories used in the study. *The weight functions always sort all numeric-type category vectors in ascending order. Consequently, the weighted coefficients are computed properly only if column positions in the input dataset match those of the corresponding categories in the sorted category vector. For alphanumeric-type category vectors, they are assumed to already be ranked following an order that is meaningful to the researcher.*

Value

Calling function `fleiss.kappa.dist` returns the following 5 values:

- `pa`: the percent agreement.
- `pe`: the percent chance agreement.
- `fleiss.kappa`: Fleiss' generalized kappa coefficient.
- `stderr`: the standard error of Fleiss' kappa.
- `p.value`: the p-value of Fleiss' kappa coefficient.

Examples

```
# creates a matrix with Table B.1.2 data
>ratings<-matrix(c(0, 0, 0, 6, 0,
+                  0, 1, 4, 0, 1,
+                  2, 0, 4, 0, 0,
+                  0, 3, 3, 0, 0),ncol=5,byrow=T)

# to compute unweighted Fleiss' kappa, its standard error and more
>fleiss.kappa.dist(ratings)
```

The results displayed on the screen will look like this:

```
Fleiss' Kappa Coefficient
=========================
Percent agreement: 0.5666667 Percent chance agreement: 0.3090278
Fleiss kappa coefficient: 0.3728643 Standard error: 0.2457742
95% Confidence Interval: ( -0.409299 , 1 )
P-value: 0.2265189

# to compute weighted kappa with quadratic weights
>fleiss.kappa.dist(ratings,quadratic.weights(1:5))
# the call above assumes the script file weights.gen.r was read into R, and
```

generates the following results:

```
Fleiss' Kappa Coefficient
=========================
Percent agreement: 0.9270833 Percent chance agreement: 0.8854167
Fleiss kappa coefficient: 0.3636364 Standard error: 0.2525845
Weights:
1 0.9375 0.75 0.4375 0
0.9375 1 0.9375 0.75 0.4375
0.75 0.9375 1 0.9375 0.75
0.4375 0.75 0.9375 1 0.9375
0 0.4375 0.75 0.9375 1

95% Confidence Interval: ( -0.4402002 , 1)
P-value: 0.2455769
```

R FUNCTIONS IN SCRIPT FILE **agree.coeff3.raw.r**

If your analysis is based on three raters or more we previously saw that one option is to organize your data as a distribution of raters by subject and by category. Alternatively, you may report the raw ratings in a table where each row represents a subject, each column a rater, and each table cell the actual rating assigned by the rater to the subject. Table B.1.3 is an example of such data where 5 raters classified 4 subjects into 3 categories labeled as {1, 2, 3}.

Table B.1.3: Rating of Four Subjects by Five Raters[a]

	Raters				
Subject	I	II	III	IV	V
A	2	2	3	2	2
B	2	2	2	2	2
C	2	2	2	2	1
D	1	2	2	2	2

[a]This is Table 2 of Finn (1970)

The following functions contained in the script file `agree.coeff3.raw.r` are what you will need to analyze rating data such as described in Table B.1.3:

(1) `fleiss.kappa.raw` (for Fleiss's unweighted and weighted kappa coefficients)

(3) `gwet.ac1.raw` (for Gwet's unweighted and weighted AC_1 coefficients)

(4) `bp.coeff.raw` (for Brennan-Prediger unweighted and weighted coefficients)

(5) `krippen.alpha.raw` (for Krippendorff unweighted and weighted alpha coefficients)

(5) `conger.kappa.raw` (for Conger's unweighted and weighted kappa coefficients)

All these functions operate the same way. Therefore, only the first of these functions named `fleiss.kappa.raw` is discussed here in details. The same discussion applies to the other functions as well.

`fleiss.kappa.raw`: *Fleiss' kappa coefficient for multiple raters & raw ratings*

Description

This function calculates the unweighted as well as the weighted Fleiss' generalized kappa coefficients for multiple raters whose raw ratings are listed horizontally for each subject such as in Table B.1.3. A table cell may be missing if a rater did not rate a particular subject. When the ratings are alphanumeric then the blank character is treated as a missing value.

Usage

```
fleiss.kappa.raw(ratings,weights="unweighted",conflev=0.95,
          N=Inf,print=TRUE)
```

Arguments

Of all arguments used by this function, only the first one is required. The remaining arguments being all optional. If your goal is limited to unweighted statistics, then the simple function call `fleiss.kappa.raw(ratings)` *is sufficient to produce Fleiss' generalized kappa along with its standard error, confidence interval, and p-value.*

- `ratings`: This is an $n \times r$ matrix or data frame (or matrix), where n is the number of subjects, and r the number of raters. This is the only argument that is required if you want an unweighted analysis.

- `weights`: This is a $q \times q$ matrix of weights. The default argument is "unweighted", and there is no need to specify it explicitly when the unweighted analysis is what you want. The `weights` parameter can take any of the following values "quadratic", "linear", "ordinal", "radical", "ratio", "circular", or "bipolar". You may refer to the previous chapters for an explicit definition

of these different weights. You will need to read the `weights.gen.r` script into R before this function can perform a weighted analysis.

When the input data is in the form of raw ratings, you may not have a direct way of obtaining a list of all categories involved in the experiment, especially if the dataset is large. This makes it more difficult although not impossible to define your own weight matrix.

- `conflev`: The confidence level associated with the agreement coefficient's confidence interval.
- `N`: An optional parameter representing the total number of subjects in the target subject population. Its default value is infinity, which for all practical purposes assumes the target subject population to be very large and will not require any finite-population correction when computing the standard error.
- `print`: An optional logical parameter which takes the default value of TRUE if you also want the function to output the results on the screen. Set this parameter to FALSE if you do not want the results to be displayed on the screen.

Details

`fleiss.kappa.raw` can accept data in the form of a matrix or in the form of a data frame as long as the input data supplied can be interpreted as a matrix. The ratings may be of numeric or alphanumeric types. To perform the weighted analysis, you need to assign one the values mentioned above to the weights parameter. If you have the list of categories in your dataset, you may even create your own weight matrix, or use one of the many existing weight-generating functions in the `weights.ge.r` script. Each weight function takes a single mandatory parameter, which is a vector containing all categories used in the study. *The weight functions always sort all numeric-type category vectors in ascending order. I assume here that adjacent categories on the sorted list represent a higher degree of agreement than two categories that are farther apart.*

Value

Calling function `fleiss.kappa.raw` returns the following 5 values:

- `pa`: the percent agreement.
- `pe`: the percent chance agreement.
- `fleiss.kappa`: Fleiss' generalized kappa coefficient.

- `stderr`: the standard error of Fleiss' kappa.
- `p.value`: the p-value of Fleiss' kappa coefficient.

Examples

```
# creates a matrix with Table B.1.3 data
>table.b.3<-matrix(c(
+                   2, 2, 3, 2, 2,
+                   2, 2, 2, 2, 2,
+                   2, 2, 2, 2, 1,
+                   1, 2, 2, 2, 2),ncol=5,byrow=TRUE)

# to compute unweighted Fleiss' kappa, its standard error and more
>fleiss.kappa.raw(table.b.3)
```

The results displayed on the screen will look like this:

```
Fleiss' Kappa Coefficient
=========================
Percent agreement: 0.7 Percent chance agreement: 0.735
Fleiss kappa coefficient: -0.1320755 Standard error: 0.05375461
95% Confidence Interval: ( -0.3031466 , 0.03899568 )
P-value: 0.09110958
```

```
# to compute weighted kappa with quadratic weights
>fleiss.kappa.raw(table.b.3,weights="quadratic")
# the above call assumes that the script file weights.gen.r was previously
read into R, and generates the following results:
Fleiss' Kappa Coefficient
=========================
Percent agreement: 0.925 Percent chance agreement: 0.92625
Fleiss kappa coefficient: -0.01694915 Standard error: 0.06525606
Weights: quadratic
95% Confidence Interval: ( -0.5307952 , 0.1907247 )
P-value: 0.8118745
```

B.2 AgreeStat for Excel

At the time this book was published, there were 2 known Excel solutions that I would recommend practitioners to consider. One of these solutions is non-commercial while the other is commercial. The non-commercial solution is the *Real Statistics Data Analysis Tools* and the commercial solution is *AgreeStat360*.

- *Real Statistics Data Analysis Tools*, by Charles Zaiontz.

  The "Real Statistics Data Analysis Tools" is an Excel add-in, which implements a large number of statistical techniques, including the calculation of several inter-rater reliability coefficients. Interested Excel users can obtain more information regarding the installation of this software from the following URL:

  https://www.real-statistics.com/free-download/real-statistics-resource-pack/

  For a detailed description of the capability of this Excel add-in with respect to the computation of various inter-rater reliability coefficients, you may visit the following page:

  https://www.real-statistics.com/reliability/interrater-reliability/

- *AgreeStat360*, by K. Gwet.

  AgreeStat360 is a commercial Excel-based software for Windows developed by the author of this book. It is menu-driven and very user-friendly. It can compute various chance-corrected agreement coefficients, and many versions of intraclass correlation coefficients. A detailed description of its features can be obtained from the following URL:

  https://agreestat.com/software/default.html#excel

  AgreeStat360 implements all chance-corrected agreement coefficients discussed in this book, including the special conditional agreement coefficients of chapter 8. Moreover, various forms of the intraclass correlation can be calculated as well, including the optimal sample sizes. Intraclass correlation coefficients are discussed in more details by Gwet (2021).

B.3 Online Calculators

There are very few non-commercial online inter-rater reliability calculators. Although most of them are limited to the original two-rater version of Cohen (1960), a few have implemented Fleiss' extension to multiple raters as well. These online calculators are rarely maintained and do not always implement the latest techniques. Nevertheless, readers interested in the non-commercial options may want to consider the following 2 solutions:

- *StatsToDo:* http://www.statstodo.com/ResourceIndex_Subjects.php
- *ReCal:* http://dfreelon.org/utils/recalfront/

The only genuine cloud-based software for inter-rater reliability assessment is the commercial **AgreeStat360.com**. It is regularly maintained, and can be accessed from the URL agreestat360.com. To compute most agreement coefficients discussed in this book, all you need is your browser. No installation is required.

If your goal is to compare the difference between 2 agreement coefficients for statistical significance, you may consider using the free cloud-based **AgreeTest** developed by K. Gwet, and which can be accessed from the following URL:

https://agreestat.net/agreetest/

B.4 SAS Software

SAS is one of the major statistical software packages on the market today. It is a massive software system that has been around for many decades, and which is very expensive. It would be unwise to consider acquiring this product for the sole purpose of computing inter-rater reliability coefficients. Those who already have access to it, will certainly want to know about its capability as far as computing inter-rater reliability coefficients is concerned.

One of the many SAS modules is known as SAS/STAT. As of its version 14.2, SAS/STAT offers with the FREQ Procedure, options for computing the AC_1 coefficient (see Gwet, 2008a) as well as the PABAK coefficient[1] (see Byrt et al., 1993), in addition to Cohen's Kappa by Cohen (1960) , which was already implemented in the previous versions. Therefore, SAS users do not need to use another software to obtain theses statistics.

Note the FREQ Procedure will only compute the extent of agreement between 2 raters and presents some limitations regarding its treatment of missing values. By default, the FREQ procedure systematically deletes all observations with a missing rating. Consequently, the results obtained with SAS may differ from those obtained with functions available in several R packages, if your dataset contains missing ratings. An option is available for instructing the FREQ procedure to treat missing values as true categories. However, this option is useless for the analysis of agreement among raters. What would be of interest is for Proc FREQ developers to allow for the marginals associated with both raters to be calculated independently. That is, if a rating is available from one rater, then it should be used for calculating that rater's marginal probability whether or not the other rater rated the same subject or not.

[1] The coefficient often referred to by researchers as PABAK is also known (perhaps more rightfully so) as the Brennan-Prediger coefficient. It was formally studied by Brennan and Prediger (1981) , 13 years earlier.

If you want to compute the extent of agreement among 3 raters or more, then the FREQ Procedure can no longer be used. Fortunately, SAS Institute has provided a very useful Macro program called *MAGREE*, which is well documented and implemented several agreement coefficients used in the literature. For more information about this SAS macro, read *Sample 25006: Compute estimates and tests of agreement among multiple raters* from the following URL:

https://support.sas.com/kb/25/006.html

Here is a summary of the purpose of this macro as provided by SAS Institute:

> Compute estimates and tests of agreement among multiple raters when responses (ratings) are on a nominal or ordinal scale. For a nominal response, kappa and Gwet's AC1 agreement coefficient are available. For an ordinal response, Gwet's weighted agreement coefficient (AC2) and statistics based on a cumulative probit model with random effects for raters and items are available. If the response is numerically-coded (and possibly continuous), Kendall's coefficient of concordance is also available.

B.5 SPSS & STATA

STATA is another major statistical software packages, which is more recent than SAS and SPSS, and which specializes in the medical field. I strongly advise STATA users with interest in interrater reliability assessment to start by reading the very interesting article written by Klein (2018). Moreover, the document http://www.stata.com/manuals13/rkappa.pdf summarizes the built-in STATA commands related to inter-rater reliability assessment. Unlike SAS and SPSS, STATA has a built-in procedure for computing the multiple-rater version of kappa proposed by Fleiss (1971). Unless you are already a STATA user, it would be unwise to acquire this major software for the sole purpose of computing inter-rater reliability coefficients.

SPSS is also one of the major statistical software packages on the market today. Just like SAS, SPSS has been around for a while, and specializes in the social sciences. It offers some limited built-in procedures for computing inter-rater reliability coefficients and will be useful to researchers who are primarily interested in Cohen's kappa (see Cohen, 1960) or its generalized version by Fleiss (see Fleiss, 1971).

For computing Cohen's Kappa, one can follow the detailed instructions provided on the following page:

https://statistics.laerd.com/spss-tutorials/cohens-kappa-in-spss-statistics.php

You can compute Fleiss' generalized Kappa with SPSS. However, the procedure for doing it depends on the version of SPSS you are using. Researchers interested in this procedure can obtain more information from the following URL:

https://statistics.laerd.com/spss-tutorials/fleiss-kappa-in-spss-statistics.php

B.6 Concluding Remarks

In this appendix, I reviewed some of what I consider to be among the most interesting software options for researchers involved in inter-rater reliability assessment. The number of software solutions available to researchers for computing inter-rater reliability coefficients has increased dramatically since the first edition of this book. R and Excel users have more options to choose from than others. However, SAS and STATA users can also rely on very useful macro programs and functions developed by independent researchers. SPSS on the other hand, offers very few options, which are all limited to Cohen's Kappa and its generalized Fleiss' version. More details regarding all these software solutions can be obtained from the following URL:

https://agreestat.com/software/default.html

Before using a particular software package for calculating inter-rater reliability coefficients, researchers need to find out how that package handles missing ratings. Many programs made available to the general public do not have a well-defined strategy for dealing with missing ratings, which are known to be an important problem in many inter-rater reliability experiments. Even some existing R functions proposed in some publicly-available R packages such as `'irr'` or `'concord'` tend to exclude from analysis any subject that was not rated by all raters. This crude strategy may eliminate a substantial amount of data collected during the experiment. This may not a problem if it is the way you want missing data to be handled. But a better strategy would be to use every single data point that was gathered as recommended throughout this book.

APPENDIX C

Sample Size Calculations

This appendix contains some of the datasets used in section 6.5 of chapter 6, to model the maximum variance of agreement coefficients. Modeling the maximum variance is an essential task for computing the optimal number of subjects n needed to ensure an adequate precision level for the agreement coefficient of interest. It is through modeling that it became possible to estimate the constant C used in equations 6.5.2 and 6.5.3.

The datasets were created separately for each of the 4 agreement coefficients that were considered in section 6.5. These 4 coefficients are the AC_1, the Brennan-Prediger statistic, Fleiss' generalized kappa and the percent agreement. This appendix contains the following 8 tables:

1) Table C.1:
2) Table C.2:
3) Table C.3:
4) Table C.4:
5) Table C.5:
6) Table C.6:
7) Table C.7:
8) Table C.8:

Table C.1

Maximum standard error of Gwet's AC_1 and Brennan-Prediger coefficients by number of raters (r) and number of subjects (n) when the number of categories is $\boldsymbol{q = 2}$[a]

n	Number of raters (r)																		
	Gwet's AC_1									Brennan-Prediger									
	2	3	4	5	6	7	8	9	10	2	3	4	5	6	7	8	9	10
10	0.373	0.280	0.249	0.244	0.224	0.227	0.213	0.218	0.207	0.333	0.222	0.222	0.200	0.200	0.190	0.190	0.185	0.185
15	0.308	0.226	0.206	0.199	0.185	0.187	0.176	0.180	0.171	0.267	0.178	0.178	0.160	0.160	0.152	0.152	0.148	0.148
20	0.257	0.193	0.171	0.168	0.154	0.157	0.147	0.150	0.143	0.229	0.153	0.153	0.138	0.138	0.131	0.131	0.127	0.127
25	0.233	0.173	0.155	0.151	0.140	0.141	0.133	0.136	0.129	0.204	0.136	0.136	0.122	0.122	0.117	0.117	0.113	0.113
30	0.208	0.156	0.139	0.136	0.125	0.127	0.119	0.122	0.116	0.186	0.124	0.124	0.111	0.111	0.106	0.106	0.103	0.103
35	0.195	0.144	0.130	0.126	0.117	0.118	0.111	0.114	0.108	0.171	0.114	0.114	0.103	0.103	0.098	0.098	0.095	0.095
40	0.179	0.134	0.120	0.117	0.108	0.109	0.102	0.105	0.100	0.160	0.107	0.107	0.096	0.096	0.091	0.091	0.089	0.089
45	0.171	0.125	0.114	0.111	0.102	0.104	0.098	0.100	0.095	0.151	0.100	0.100	0.090	0.090	0.086	0.086	0.084	0.084
50	0.160	0.117	0.107	0.103	0.096	0.097	0.091	0.094	0.089	0.143	0.095	0.095	0.086	0.086	0.082	0.082	0.079	0.079
55	0.154	0.115	0.103	0.098	0.092	0.094	0.088	0.090	0.085	0.136	0.091	0.091	0.082	0.082	0.078	0.078	0.076	0.076
60	0.146	0.109	0.097	0.095	0.087	0.088	0.083	0.085	0.081	0.130	0.087	0.087	0.078	0.078	0.074	0.074	0.072	0.072
65	0.141	0.103	0.094	0.091	0.085	0.084	0.081	0.082	0.078	0.125	0.083	0.083	0.075	0.075	0.071	0.071	0.069	0.069
70	0.135	0.098	0.090	0.088	0.081	0.082	0.077	0.078	0.075	0.120	0.080	0.080	0.072	0.072	0.069	0.069	0.067	0.067
75	0.131	0.096	0.087	0.083	0.079	0.080	0.075	0.075	0.073	0.116	0.077	0.077	0.070	0.070	0.066	0.066	0.065	0.065
80	0.126	0.094	0.084	0.082	0.076	0.077	0.072	0.074	0.070	0.112	0.075	0.075	0.067	0.067	0.064	0.064	0.062	0.062
85	0.123	0.092	0.082	0.080	0.074	0.075	0.070	0.070	0.068	0.109	0.073	0.073	0.065	0.065	0.062	0.062	0.061	0.061
90	0.119	0.087	0.079	0.076	0.071	0.072	0.068	0.068	0.066	0.106	0.071	0.071	0.064	0.064	0.061	0.061	0.059	0.059
95	0.116	0.087	0.077	0.076	0.070	0.069	0.066	0.067	0.064	0.103	0.069	0.069	0.062	0.062	0.059	0.059	0.057	0.057
100	0.113	0.082	0.075	0.073	0.068	0.067	0.064	0.064	0.063	0.100	0.067	0.067	0.060	0.060	0.057	0.057	0.056	0.056

[a]To download this table or any other table in appendix C, follow the link https://agreestat.com/books/cac5/

Table C.2

Maximum standard error of the Percent Agreement p_a and Fleiss' Kappa coefficients by number of raters (r) and number of subjects (n) when the number of categories is $\boldsymbol{q = 2}$[a]

n	Number of raters (r)								
	Percent Agreement p_a								
	2	3	4	5	6	7	8	9	10
10	0.167	0.111	0.111	0.100	0.100	0.095	0.095	0.093	0.093
15	0.133	0.089	0.089	0.080	0.080	0.076	0.076	0.074	0.074
20	0.115	0.076	0.076	0.069	0.069	0.066	0.066	0.064	0.064
25	0.102	0.068	0.068	0.061	0.061	0.058	0.058	0.057	0.057
30	0.093	0.062	0.062	0.056	0.056	0.053	0.053	0.052	0.052
35	0.086	0.057	0.057	0.051	0.051	0.049	0.049	0.048	0.048
40	0.080	0.053	0.053	0.048	0.048	0.046	0.046	0.044	0.044
45	0.075	0.050	0.050	0.045	0.045	0.043	0.043	0.042	0.042
50	0.071	0.048	0.048	0.043	0.043	0.041	0.041	0.040	0.040
55	0.068	0.045	0.045	0.041	0.041	0.039	0.039	0.038	0.038
60	0.065	0.043	0.043	0.039	0.039	0.037	0.037	0.036	0.036
65	0.062	0.042	0.042	0.037	0.037	0.036	0.036	0.035	0.035
70	0.060	0.040	0.040	0.036	0.036	0.034	0.034	0.033	0.033
75	0.058	0.039	0.039	0.035	0.035	0.033	0.033	0.032	0.032
80	0.056	0.037	0.037	0.034	0.034	0.032	0.032	0.031	0.031
85	0.055	0.036	0.036	0.033	0.033	0.031	0.031	0.030	0.030
90	0.053	0.035	0.035	0.032	0.032	0.030	0.030	0.029	0.029
95	0.052	0.034	0.034	0.031	0.031	0.029	0.029	0.029	0.029
100	0.050	0.033	0.033	0.030	0.030	0.029	0.029	0.028	0.028

n	Fleiss' Kappa								
	2	3	4	5	6	7	8	9	10
10	0.378	0.342	0.321	0.308	0.296	0.283	0.277	0.275	0.273
15	0.352	0.325	0.308	0.297	0.290	0.285	0.281	0.277	0.272
20	0.342	0.316	0.301	0.292	0.286	0.281	0.278	0.275	0.273
25	0.336	0.312	0.298	0.289	0.283	0.279	0.276	0.273	0.271
30	0.332	0.309	0.295	0.287	0.281	0.277	0.274	0.272	0.270
35	0.330	0.306	0.293	0.285	0.280	0.276	0.273	0.271	0.269
40	0.328	0.305	0.292	0.284	0.279	0.275	0.272	0.270	0.268
45	0.326	0.303	0.291	0.283	0.278	0.274	0.271	0.269	0.268
50	0.325	0.302	0.290	0.282	0.277	0.274	0.271	0.269	0.267
55	0.324	0.302	0.289	0.282	0.277	0.273	0.270	0.268	0.267
60	0.323	0.301	0.289	0.281	0.276	0.273	0.270	0.268	0.266
65	0.322	0.300	0.288	0.281	0.276	0.272	0.270	0.268	0.266
70	0.322	0.300	0.288	0.281	0.276	0.272	0.270	0.267	0.266
75	0.321	0.300	0.288	0.280	0.275	0.272	0.269	0.267	0.266
80	0.321	0.299	0.287	0.280	0.275	0.272	0.269	0.267	0.265
85	0.320	0.299	0.287	0.280	0.275	0.271	0.269	0.267	0.265
90	0.320	0.299	0.287	0.279	0.275	0.271	0.269	0.267	0.265
95	0.320	0.298	0.286	0.279	0.275	0.271	0.269	0.267	0.265
100	0.319	0.298	0.286	0.279	0.274	0.271	0.268	0.266	0.265

[a]To download this table or any other table in appendix C, follow the link https://agreestat.com/books/cac5/

Table C.3

Maximum standard error of Gwet's AC_1 and Brennan-Prediger coefficients by number of raters (r) and number of subjects (n) when the number of categories is $\boldsymbol{q = 3}$[a]

n	Number of raters (r)																	
	Gwet's AC_1									Brennan-Prediger								
	2	3	4	5	6	7	8	9	10	2	3	4	5	6	7	8	9	10
10	0.287	0.272	0.233	0.226	0.217	0.210	0.210	0.204	0.202	0.250	0.250	0.208	0.200	0.200	0.190	0.187	0.187	0.183
15	0.230	0.221	0.189	0.183	0.177	0.171	0.170	0.166	0.164	0.200	0.200	0.167	0.160	0.160	0.152	0.150	0.150	0.147
20	0.198	0.187	0.160	0.156	0.150	0.145	0.144	0.140	0.139	0.172	0.172	0.143	0.138	0.138	0.131	0.129	0.129	0.126
25	0.176	0.168	0.144	0.139	0.134	0.130	0.129	0.126	0.125	0.153	0.153	0.127	0.122	0.122	0.117	0.115	0.115	0.112
30	0.160	0.151	0.130	0.126	0.121	0.117	0.117	0.113	0.112	0.139	0.139	0.116	0.111	0.111	0.106	0.104	0.104	0.102
35	0.147	0.141	0.120	0.116	0.113	0.109	0.108	0.106	0.104	0.129	0.129	0.107	0.103	0.103	0.098	0.096	0.096	0.094
40	0.138	0.130	0.112	0.107	0.104	0.101	0.100	0.098	0.097	0.120	0.120	0.100	0.096	0.096	0.091	0.090	0.090	0.088
45	0.130	0.123	0.106	0.101	0.099	0.096	0.094	0.093	0.092	0.113	0.113	0.094	0.090	0.090	0.086	0.085	0.085	0.083
50	0.123	0.116	0.100	0.097	0.093	0.090	0.089	0.087	0.086	0.107	0.107	0.089	0.086	0.086	0.082	0.080	0.080	0.079
55	0.117	0.111	0.095	0.093	0.089	0.086	0.085	0.084	0.083	0.102	0.102	0.085	0.082	0.082	0.078	0.077	0.077	0.075
60	0.112	0.106	0.091	0.087	0.085	0.082	0.081	0.080	0.079	0.098	0.098	0.081	0.078	0.078	0.074	0.073	0.073	0.072
65	0.108	0.102	0.087	0.084	0.082	0.079	0.078	0.077	0.076	0.094	0.094	0.078	0.075	0.075	0.071	0.070	0.070	0.069
70	0.103	0.098	0.084	0.082	0.078	0.076	0.076	0.074	0.073	0.090	0.090	0.075	0.072	0.072	0.069	0.068	0.068	0.066
75	0.100	0.095	0.081	0.078	0.076	0.074	0.072	0.071	0.070	0.087	0.087	0.073	0.070	0.070	0.066	0.065	0.065	0.064
80	0.096	0.092	0.078	0.075	0.073	0.071	0.071	0.069	0.068	0.084	0.084	0.070	0.067	0.067	0.064	0.063	0.063	0.062
85	0.094	0.089	0.076	0.074	0.071	0.069	0.069	0.067	0.066	0.082	0.082	0.068	0.065	0.065	0.062	0.061	0.061	0.060
90	0.091	0.086	0.073	0.070	0.069	0.067	0.066	0.065	0.064	0.079	0.079	0.066	0.064	0.064	0.061	0.060	0.060	0.058
95	0.089	0.084	0.072	0.069	0.067	0.065	0.065	0.063	0.062	0.077	0.077	0.064	0.062	0.062	0.059	0.058	0.058	0.057
100	0.087	0.082	0.070	0.067	0.065	0.063	0.062	0.061	0.061	0.075	0.075	0.063	0.060	0.060	0.057	0.057	0.057	0.055

[a]To download this table or any other table in appendix C, follow the link https://agreestat.com/books/cac5/

Table C.4

Maximum standard error of the Percent Agreement p_a and Fleiss' Kappa coefficients by number of raters (r) and number of subjects (n) when the number of categories is $\boldsymbol{q = 3}$[a]

n	Number of raters (r)								
	Percent Agreement p_a								
	2	3	4	5	6	7	8	9	10
10	0.167	0.167	0.139	0.133	0.133	0.127	0.125	0.125	0.122
15	0.133	0.133	0.111	0.107	0.107	0.102	0.100	0.100	0.098
20	0.115	0.115	0.096	0.092	0.092	0.087	0.086	0.086	0.084
25	0.102	0.102	0.085	0.082	0.082	0.078	0.076	0.076	0.075
30	0.093	0.093	0.077	0.074	0.074	0.071	0.070	0.070	0.068
35	0.086	0.086	0.071	0.069	0.069	0.065	0.064	0.064	0.063
40	0.080	0.080	0.067	0.064	0.064	0.061	0.060	0.060	0.059
45	0.075	0.075	0.063	0.060	0.060	0.057	0.057	0.057	0.055
50	0.071	0.071	0.060	0.057	0.057	0.054	0.054	0.054	0.052
55	0.068	0.068	0.057	0.054	0.054	0.052	0.051	0.051	0.050
60	0.065	0.065	0.054	0.052	0.052	0.050	0.049	0.049	0.048
65	0.062	0.062	0.052	0.050	0.050	0.048	0.047	0.047	0.046
70	0.060	0.060	0.050	0.048	0.048	0.046	0.045	0.045	0.044
75	0.058	0.058	0.048	0.046	0.046	0.044	0.044	0.044	0.043
80	0.056	0.056	0.047	0.045	0.045	0.043	0.042	0.042	0.041
85	0.055	0.055	0.045	0.044	0.044	0.042	0.041	0.041	0.040
90	0.053	0.053	0.044	0.042	0.042	0.040	0.040	0.040	0.039
95	0.052	0.052	0.043	0.041	0.041	0.039	0.039	0.039	0.038
100	0.050	0.050	0.042	0.040	0.040	0.038	0.038	0.038	0.037

n	Number of raters (r)								
	Fleiss' Kappa								
	2	3	4	5	6	7	8	9	10
10	0.378	0.342	0.321	0.308	0.296	0.283	0.277	0.275	0.273
15	0.352	0.325	0.308	0.297	0.290	0.285	0.281	0.277	0.272
20	0.342	0.316	0.301	0.292	0.286	0.281	0.278	0.275	0.273
25	0.336	0.312	0.298	0.289	0.283	0.279	0.276	0.273	0.271
30	0.332	0.309	0.295	0.287	0.281	0.277	0.274	0.272	0.270
35	0.330	0.306	0.293	0.285	0.280	0.276	0.273	0.271	0.269
40	0.328	0.305	0.292	0.284	0.279	0.275	0.272	0.270	0.268
45	0.326	0.303	0.291	0.283	0.278	0.274	0.271	0.269	0.268
50	0.325	0.302	0.290	0.282	0.277	0.274	0.271	0.269	0.267
55	0.324	0.302	0.289	0.282	0.277	0.273	0.270	0.268	0.267
60	0.323	0.301	0.289	0.281	0.276	0.273	0.270	0.268	0.266
65	0.322	0.300	0.288	0.281	0.276	0.272	0.270	0.268	0.266
70	0.322	0.300	0.288	0.281	0.276	0.272	0.270	0.267	0.266
75	0.321	0.300	0.288	0.280	0.275	0.272	0.269	0.267	0.266
80	0.321	0.299	0.287	0.280	0.275	0.272	0.269	0.267	0.265
85	0.320	0.299	0.287	0.280	0.275	0.271	0.269	0.267	0.265
90	0.320	0.299	0.287	0.279	0.275	0.271	0.269	0.267	0.265
95	0.320	0.298	0.286	0.279	0.275	0.271	0.269	0.267	0.265
100	0.319	0.298	0.286	0.279	0.274	0.271	0.268	0.266	0.265

[a]To download this table or any other table in appendix C, follow the link https://agreestat.com/books/cac5/

Table C.5

Maximum standard error of Gwet's AC_1 and Brennan-Prediger coefficients by number of raters (r) and number of subjects (n) when the number of categories is $\boldsymbol{q = 4}$[a]

n	Number of raters (r)																	
	Gwet's AC_1									Brennan-Prediger								
	2	3	4	5	6	7	8	9	10	2	3	4	5	6	7	8	9	10
10	0.244	0.245	0.237	0.215	0.209	0.207	0.203	0.198	0.197	0.222	0.222	0.222	0.200	0.193	0.190	0.190	0.185	0.183
15	0.196	0.196	0.191	0.174	0.168	0.167	0.164	0.160	0.159	0.178	0.178	0.178	0.160	0.154	0.152	0.152	0.148	0.146
20	0.168	0.168	0.163	0.148	0.144	0.142	0.140	0.137	0.135	0.153	0.153	0.153	0.138	0.133	0.131	0.131	0.127	0.126
25	0.150	0.150	0.146	0.132	0.128	0.127	0.125	0.122	0.121	0.136	0.136	0.136	0.122	0.118	0.117	0.117	0.113	0.112
30	0.136	0.136	0.132	0.120	0.116	0.115	0.113	0.111	0.110	0.124	0.124	0.124	0.111	0.107	0.106	0.106	0.103	0.102
35	0.126	0.126	0.122	0.111	0.108	0.106	0.105	0.103	0.101	0.114	0.114	0.114	0.103	0.099	0.098	0.098	0.095	0.094
40	0.117	0.116	0.114	0.103	0.100	0.099	0.097	0.095	0.094	0.107	0.107	0.107	0.096	0.092	0.091	0.091	0.089	0.088
45	0.111	0.111	0.107	0.098	0.094	0.094	0.092	0.090	0.089	0.100	0.100	0.100	0.090	0.087	0.086	0.086	0.084	0.083
50	0.105	0.105	0.101	0.092	0.089	0.088	0.087	0.085	0.084	0.095	0.095	0.095	0.086	0.083	0.082	0.082	0.079	0.078
55	0.100	0.099	0.097	0.088	0.085	0.084	0.083	0.081	0.080	0.091	0.091	0.091	0.082	0.079	0.078	0.078	0.076	0.075
60	0.095	0.096	0.092	0.084	0.081	0.081	0.079	0.077	0.077	0.087	0.087	0.087	0.078	0.075	0.074	0.074	0.072	0.071
65	0.091	0.091	0.089	0.081	0.078	0.077	0.076	0.075	0.074	0.083	0.083	0.083	0.075	0.072	0.071	0.071	0.069	0.068
70	0.088	0.087	0.085	0.078	0.075	0.074	0.073	0.072	0.071	0.080	0.080	0.080	0.072	0.070	0.069	0.069	0.067	0.066
75	0.085	0.085	0.083	0.075	0.072	0.071	0.071	0.069	0.069	0.077	0.077	0.077	0.070	0.067	0.066	0.066	0.065	0.064
80	0.082	0.083	0.080	0.073	0.070	0.069	0.068	0.067	0.066	0.075	0.075	0.075	0.067	0.065	0.064	0.064	0.062	0.062
85	0.080	0.079	0.078	0.070	0.068	0.067	0.067	0.065	0.064	0.073	0.073	0.073	0.065	0.063	0.062	0.062	0.061	0.060
90	0.077	0.077	0.075	0.068	0.066	0.065	0.064	0.063	0.062	0.071	0.071	0.071	0.064	0.061	0.061	0.061	0.059	0.058
95	0.075	0.076	0.073	0.067	0.064	0.063	0.063	0.061	0.061	0.069	0.069	0.069	0.062	0.060	0.059	0.059	0.057	0.057
100	0.073	0.072	0.071	0.065	0.062	0.062	0.061	0.060	0.059	0.067	0.067	0.067	0.060	0.058	0.057	0.057	0.056	0.055

[a]To download this table or any other table in appendix C, follow the link https://agreestat.com/books/cac5/

Table C.6

Maximum standard error of the Percent Agreement p_a and Fleiss' Kappa coefficients by number of raters (r) and number of subjects (n) when the number of categories is $\boldsymbol{q = 4}$[a]

n	Number of raters (r)								
	Percent Agreement p_a								
	2	3	4	5	6	7	8	9	10
10	0.167	0.167	0.167	0.150	0.144	0.143	0.143	0.139	0.137
15	0.133	0.133	0.133	0.120	0.116	0.114	0.114	0.111	0.110
20	0.115	0.115	0.115	0.103	0.099	0.098	0.098	0.096	0.094
25	0.102	0.102	0.102	0.092	0.088	0.087	0.087	0.085	0.084
30	0.093	0.093	0.093	0.084	0.080	0.080	0.080	0.077	0.076
35	0.086	0.086	0.086	0.077	0.074	0.073	0.073	0.071	0.070
40	0.080	0.080	0.080	0.072	0.069	0.069	0.069	0.067	0.066
45	0.075	0.075	0.075	0.068	0.065	0.065	0.065	0.063	0.062
50	0.071	0.071	0.071	0.064	0.062	0.061	0.061	0.060	0.059
55	0.068	0.068	0.068	0.061	0.059	0.058	0.058	0.057	0.056
60	0.065	0.065	0.065	0.059	0.056	0.056	0.056	0.054	0.054
65	0.062	0.062	0.062	0.056	0.054	0.054	0.054	0.052	0.051
70	0.060	0.060	0.060	0.054	0.052	0.052	0.052	0.050	0.049
75	0.058	0.058	0.058	0.052	0.050	0.050	0.050	0.048	0.048
80	0.056	0.056	0.056	0.051	0.049	0.048	0.048	0.047	0.046
85	0.055	0.055	0.055	0.049	0.047	0.047	0.047	0.045	0.045
90	0.053	0.053	0.053	0.048	0.046	0.045	0.045	0.044	0.044
95	0.052	0.052	0.052	0.046	0.045	0.044	0.044	0.043	0.042
100	0.050	0.050	0.050	0.045	0.044	0.043	0.043	0.042	0.041

n	Number of raters (r)								
	Fleiss' Kappa								
	2	3	4	5	6	7	8	9	10
10	0.378	0.342	0.321	0.308	0.296	0.283	0.277	0.275	0.273
15	0.352	0.325	0.308	0.297	0.290	0.285	0.281	0.277	0.272
20	0.342	0.316	0.301	0.292	0.286	0.281	0.278	0.275	0.273
25	0.336	0.312	0.298	0.289	0.283	0.279	0.276	0.273	0.271
30	0.332	0.309	0.295	0.287	0.281	0.277	0.274	0.272	0.270
35	0.330	0.306	0.293	0.285	0.280	0.276	0.273	0.271	0.269
40	0.328	0.305	0.292	0.284	0.279	0.275	0.272	0.270	0.268
45	0.326	0.303	0.291	0.283	0.278	0.274	0.271	0.269	0.268
50	0.325	0.302	0.290	0.282	0.277	0.274	0.271	0.269	0.267
55	0.324	0.302	0.289	0.282	0.277	0.273	0.270	0.268	0.267
60	0.323	0.301	0.289	0.281	0.276	0.273	0.270	0.268	0.266
65	0.322	0.300	0.288	0.281	0.276	0.272	0.270	0.268	0.266
70	0.322	0.300	0.288	0.281	0.276	0.272	0.270	0.267	0.266
75	0.321	0.300	0.288	0.280	0.275	0.272	0.269	0.267	0.266
80	0.321	0.299	0.287	0.280	0.275	0.272	0.269	0.267	0.265
85	0.320	0.299	0.287	0.280	0.275	0.271	0.269	0.267	0.265
90	0.320	0.299	0.287	0.279	0.275	0.271	0.269	0.267	0.265
95	0.320	0.298	0.286	0.279	0.275	0.271	0.269	0.267	0.265
100	0.319	0.298	0.286	0.279	0.274	0.271	0.268	0.266	0.265

[a]To download this table or any other table in appendix C, follow the link https://agreestat.com/books/cac5/

Table C.7

Maximum standard error of Gwet's AC_1 and Brennan-Prediger coefficients by number of raters (r) and number of subjects (n) when the number of categories is $\boldsymbol{q = 5}$[a]

n	Number of raters (r)																	
	Gwet's AC_1									Brennan-Prediger								
	2	3	4	5	6	7	8	9	10	2	3	4	5	6	7	8	9	10
10	0.224	0.224	0.224	0.219	0.206	0.200	0.198	0.197	0.195	0.208	0.208	0.208	0.208	0.194	0.188	0.186	0.185	0.185
15	0.180	0.180	0.179	0.177	0.166	0.161	0.159	0.159	0.157	0.167	0.167	0.167	0.167	0.156	0.151	0.149	0.148	0.148
20	0.154	0.154	0.154	0.151	0.141	0.138	0.136	0.135	0.134	0.143	0.143	0.143	0.143	0.134	0.130	0.128	0.127	0.127
25	0.137	0.137	0.137	0.135	0.126	0.123	0.121	0.121	0.120	0.127	0.127	0.127	0.127	0.119	0.115	0.114	0.113	0.113
30	0.125	0.125	0.124	0.122	0.115	0.111	0.110	0.109	0.109	0.116	0.116	0.116	0.116	0.108	0.105	0.104	0.103	0.103
35	0.115	0.115	0.115	0.113	0.106	0.103	0.102	0.101	0.101	0.107	0.107	0.107	0.107	0.100	0.097	0.096	0.095	0.095
40	0.107	0.107	0.107	0.105	0.099	0.096	0.095	0.094	0.094	0.100	0.100	0.100	0.100	0.093	0.091	0.089	0.089	0.089
45	0.101	0.101	0.101	0.099	0.093	0.090	0.089	0.089	0.088	0.094	0.094	0.094	0.094	0.088	0.085	0.084	0.084	0.084
50	0.096	0.096	0.095	0.094	0.088	0.086	0.084	0.084	0.083	0.089	0.089	0.089	0.089	0.083	0.081	0.080	0.079	0.079
55	0.091	0.091	0.091	0.090	0.084	0.082	0.081	0.080	0.080	0.085	0.085	0.085	0.085	0.079	0.077	0.076	0.076	0.076
60	0.087	0.087	0.087	0.086	0.080	0.078	0.077	0.076	0.076	0.081	0.081	0.081	0.081	0.076	0.074	0.073	0.072	0.072
65	0.084	0.084	0.083	0.082	0.077	0.075	0.074	0.074	0.073	0.078	0.078	0.078	0.078	0.073	0.071	0.070	0.069	0.069
70	0.081	0.080	0.080	0.079	0.074	0.072	0.071	0.071	0.070	0.075	0.075	0.075	0.075	0.070	0.068	0.067	0.067	0.067
75	0.078	0.077	0.078	0.077	0.072	0.070	0.069	0.068	0.068	0.073	0.073	0.073	0.073	0.068	0.066	0.065	0.065	0.065
80	0.075	0.075	0.075	0.074	0.069	0.067	0.067	0.066	0.066	0.070	0.070	0.070	0.070	0.066	0.064	0.063	0.062	0.062
85	0.073	0.073	0.072	0.072	0.067	0.065	0.065	0.064	0.064	0.068	0.068	0.068	0.068	0.064	0.062	0.061	0.061	0.061
90	0.071	0.071	0.070	0.070	0.065	0.063	0.063	0.062	0.062	0.066	0.066	0.066	0.066	0.062	0.060	0.059	0.059	0.059
95	0.069	0.069	0.069	0.068	0.064	0.062	0.061	0.061	0.060	0.064	0.064	0.064	0.064	0.060	0.058	0.058	0.057	0.057
100	0.067	0.067	0.068	0.066	0.062	0.060	0.059	0.059	0.059	0.063	0.063	0.063	0.063	0.059	0.057	0.056	0.056	0.056

[a]To download this table or any other table in appendix C, follow the link https://agreestat.com/books/cac5/

Table C.8

Maximum standard error of the Percent Agreement p_a and Fleiss' Kappa coefficients by number of raters (r) and number of subjects (n) when the number of categories is $\boldsymbol{q = 5}$[a]

n	Number of raters (r)								
	Percent Agreement p_a								
	2	3	4	5	6	7	8	9	10
10	0.167	0.167	0.167	0.167	0.156	0.151	0.149	0.148	0.148
15	0.133	0.133	0.133	0.133	0.124	0.121	0.119	0.119	0.119
20	0.115	0.115	0.115	0.115	0.107	0.104	0.102	0.102	0.102
25	0.102	0.102	0.102	0.102	0.095	0.092	0.091	0.091	0.091
30	0.093	0.093	0.093	0.093	0.087	0.084	0.083	0.083	0.083
35	0.086	0.086	0.086	0.086	0.080	0.078	0.077	0.076	0.076
40	0.080	0.080	0.080	0.080	0.075	0.072	0.071	0.071	0.071
45	0.075	0.075	0.075	0.075	0.070	0.068	0.067	0.067	0.067
50	0.071	0.071	0.071	0.071	0.067	0.065	0.064	0.063	0.063
55	0.068	0.068	0.068	0.068	0.063	0.062	0.061	0.060	0.060
60	0.065	0.065	0.065	0.065	0.061	0.059	0.058	0.058	0.058
65	0.062	0.062	0.062	0.062	0.058	0.057	0.056	0.056	0.056
70	0.060	0.060	0.060	0.060	0.056	0.054	0.054	0.053	0.053
75	0.058	0.058	0.058	0.058	0.054	0.053	0.052	0.052	0.052
80	0.056	0.056	0.056	0.056	0.052	0.051	0.050	0.050	0.050
85	0.055	0.055	0.055	0.055	0.051	0.049	0.049	0.048	0.048
90	0.053	0.053	0.053	0.053	0.049	0.048	0.047	0.047	0.047
95	0.052	0.052	0.052	0.052	0.048	0.047	0.046	0.046	0.046
100	0.050	0.050	0.050	0.050	0.047	0.045	0.045	0.045	0.045

n	Number of raters (r)								
	Fleiss' Kappa								
	2	3	4	5	6	7	8	9	10
10	0.378	0.342	0.321	0.308	0.296	0.283	0.277	0.275	0.273
15	0.352	0.325	0.308	0.297	0.290	0.285	0.281	0.277	0.272
20	0.342	0.316	0.301	0.292	0.286	0.281	0.278	0.275	0.273
25	0.336	0.312	0.298	0.289	0.283	0.279	0.276	0.273	0.271
30	0.332	0.309	0.295	0.287	0.281	0.277	0.274	0.272	0.270
35	0.330	0.306	0.293	0.285	0.280	0.276	0.273	0.271	0.269
40	0.328	0.305	0.292	0.284	0.279	0.275	0.272	0.270	0.268
45	0.326	0.303	0.291	0.283	0.278	0.274	0.271	0.269	0.268
50	0.325	0.302	0.290	0.282	0.277	0.274	0.271	0.269	0.267
55	0.324	0.302	0.289	0.282	0.277	0.273	0.270	0.268	0.267
60	0.323	0.301	0.289	0.281	0.276	0.273	0.270	0.268	0.266
65	0.322	0.300	0.288	0.281	0.276	0.272	0.270	0.268	0.266
70	0.322	0.300	0.288	0.281	0.276	0.272	0.270	0.267	0.266
75	0.321	0.300	0.288	0.280	0.275	0.272	0.269	0.267	0.266
80	0.321	0.299	0.287	0.280	0.275	0.272	0.269	0.267	0.265
85	0.320	0.299	0.287	0.280	0.275	0.271	0.269	0.267	0.265
90	0.320	0.299	0.287	0.279	0.275	0.271	0.269	0.267	0.265
95	0.320	0.298	0.286	0.279	0.275	0.271	0.269	0.267	0.265
100	0.319	0.298	0.286	0.279	0.274	0.271	0.268	0.266	0.265

[a]To download this table or any other table in appendix C, follow the link https://agreestat.com/books/cac5/

Bibliography

Agresti, A. (1988), "A model agreement between ratings on an ordinal scale." *Biometrics*, 44, 539–548.

Agresti, A. (1992), "Modeling patterns of agreement and disagreement." *Statistical Methods in Medical Research*, 1, 201–218.

Aickin, M. (1990), "Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to cohen's kappa." *Biometrics*, 46, 293–302.

Altman, D.G. (1991), *Practical Statistics for Medical Research*. Chapman and Hall.

Axelson, R.D. and C.D. Kreiter (2009), "Reliability." In *Assessment in Health Professions Education* (Steven M. Downing and Rachel Yudkowsky, eds.), 57–74, Taylor and Francis.

Benini, R. (1901), *Principii di Demongraphia*: *Manuali Barbera Di Scienze Giuridiche Sociali e Politiche*. 29, G. Barbera, Firenze, Italy.

Bennett, E. M., R. Alpert, and A.C. Goldstein (1954), "Communications through limited response questioning." *Public Opinion Quarterly*, 18, 303–308.

Berry, K.J. and P.W. Mielke Jr. (1988), "A generalization of cohen's kappa agreement measure to interval measurement and multiple raters." *Educational and Psychological Measurement*, 48, 921–933.

Brennan, R.L. and D.J. Prediger (1981), "Coefficient kappa: some uses, misuses, and alternatives." *Educational and Psychological Measurement*, 41, 687–699.

Byrt, T., J. Bishop, and J.B. Carlin (1993), "Bias, prevalence and kappa." *Journal of Clinical Epidemiology*, 46, 423–429.

Cantor, A.B. (1996), "Sample-size calculations for cohen's kappa." *Psychological Methods*, 1, 150–153.

Carmines, E.G. and R.A. Zeller (1979), *Reliability and Validity Assessment*. Sage Publications.

Cochran, W.G. (1977), *Sampling Techniques*. John Wiley & Sons, Inc., New York.

Cohen, J. (1960), "A coefficient of agreement for nominal scales." *Educational and Psychological Measurement*, 20, 37–46.

Cohen, J. (1968), "Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit." *Psychological Bulletin*, 70, 213–220.

Conger, A.J. (1980), "Integration and generalization of kappas for multiple raters." *Psychological Bulletin*, 88, 322–328.

Cook, R.Dennis and Sanford Weisberg (1982), *Residuals and Influence in Regression*, 1 edition. Monographs on Statistics and Applied Probability, Chapman and Hall, 733 Third Avenue, New York NY 10017.

Cronbach, L.J. (1951), "Coefficient alpha, and the internal structure." *Psychometrika,*, 16, 297–334.

Eckes, T. (2011), *Introduction to Many-Facet Rasch Measurement*. Peter Lang, Internationaler Verlag der Wissenschaften.

Efron, B. (1979), "Bootstrap methods: another look at the jackknife." *Annals of Statistics,*, 7, 1–26.

Everitt, B.S. (1992), *The Analysis of Contingency Tables*, 2 edition. Chapman and Hall, London.

Feinstein, A.R. and D.V. Cicchetti (1990), "High agreement but low kappa: I. the problems of two paradoxes." *Journal of Clinical Epidemiology*, 43, 543–549.

Fenning, S., T.J. Craig, M. Tanenberg-Karant, and E.J. Bromet (1994), "Comparison of facility and research diagnoses in first-admission psychotic patients." *American Journal of Psychiatry*, 151, 1423–1429.

Finn, R.H. (1970), "A note on estimating the reliability of categorical data." *Educational and Psychological Measurement*, 30, 71–76.

Flack, V.F., A.A. Afifi, P.A. Lachenbruch, and H.J.A. Schouten (1988), "Sample size determinations for the two rater kappa statistic." *Psychometrika*, 53, 321–325.

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

Fleiss, J.L. (1981), *Statistical Methods for Rates and Proportions*. John Wiley & Sons.

Fleiss, J.L., J. Cohen, and B.S. Everitt (1969), "Large sample standard errors of kappa and weighted kappa." *Psychological Bulletin*, 72, 323–327.

Fleiss, J.L., J.C.M. Nee, and J.R. Landis (1979), "The large sample variance of kappa in the case of different sets of raters." *Psychological Bulletin*, 86, 974–977.

Goodman, L.A. and W.H. Kruskal (1954), "Measures of association in cross classifications." *Journal of the American Statistical Association*, 49, 1732–1769.

Grove, W.M., N.C. Andreasen, P. McDonald-Scott, M.B. Keller, and R.W. Shapiro (1981), "Reliability studies of psychiatric diagnosis." *Archives of General Psychiatry*, 38, 408–413.

Guttman, L. (1945), "The test-retest reliability of qualitative data." *Psychometrika*, 11, 81–95.

Gwet, Kilem L. (2008a), "Computing inter-rater reliability and its variance in the presence of high agreement." *British Journal of Mathematical and Statistical Psychology*, 61, 29–48.

Gwet, Kilem L. (2008b), "Variance estimation of nominal-scale inter-rater reliability with random selection of raters." *Psychometrika*, 73, 407–430.

Gwet, Kilem L. (2016), "Testing the difference of correlated agreement coefficients for statistical significance." *Educational and Psychological Measurement*, 76, 609–637.

Gwet, Kilem L. (2020), "Large-sample variance of fleiss generalized kappa." *Educational and Psychological Measurement*, URL https://doi.org/10.1177%2F0013164420973080.

Gwet, Kilem L. (2021), *Handbook of Inter-Rater Reliability - Volume 2: Intraclass Correlation Coefficients for Quantitative Ratings*, 5 edition, volume 1. AgreeStat Analytics, Maryland, USA.

Hayes, A.F. and K. Krippendorff (2007), "Answering the call for a standard reliability measure for coding data." *Communication Methods and Measures*, 1, 77–79.

Holley, J.W. and J.P. Guilford (1964), "A note on the g index of agreement." *Educational and Psychological Measurement*, 24, 749–753.

Holsti, O.R. (1969), *Content Analysis for the Social Sciences and Humanities*. Addison-Wesley, Reading, MA.

Hripcsak, G. and A.S. Rothschild (2005), "Agreement, the f-measure, and reliability in information retrieval." *Journal of the American Medical Informatics Association*, 12, 296–298.

Hubert, L. (1977), "Kappa revisited." *Psychological Bulletin*, 84, 289–297.

Janson, H. and U. Olsson (2001), "A measure of agreement for interval or nominal multivariate observations." *Educational and Psychological Measurement*, 61, 277–289.

Janson, H. and U. Olsson (2004), "A measure of agreement for interval or nominal multivariate observations by different sets of judges." *Educational and Psychological Measurement*, 64, 62–70.

Janson, S. and J. Vegelius (1979), "On generalizations of the g index and the phi coefficient to nominal scales." *Multivariate Behavioral Research*, 14, 255–269.

Jung, H.W. (2003), "Evaluating interrater agreement in spice-based assessments." *Computer Standards & Interfaces*, 25, 477–499.

Kendall, M. and A. Stuart (1976), *The Advanced Theory of Statistics*, 3 edition, volume 3. Griffin, London.

Khan, L., G. Mitera, L. Probyn, M. Ford, M. Christakis, J. Finkelstein, A. Donovan, L. Zhang, L. Zeng, J. Rubenstein, A. Yee, L. Holden, and E. Chow (2011), "Interrater reliability between musculoskeletal radiologists and orthopedic surgeons on computed tomography imaging features of spinal metastases." *Current Oncology*, 18, 282–287.

Klein, D. (2018), "Implementing a general framework for assessing interrater agreement in stata." *The Stata Journal*, 18, 871–901.

Kolmogorov, A.N. (1999), "The theory of probability." In *Mathematics - Its Contents, Methods and Meaning* (A.D. Aleksandrov, A.N. Kolmogorov, and M.A. Lavrent'ev, eds.), chapter Chapter XI, 229–264, Dover Publications - Dover Books on Mathematics.

Kottner, Jan and David L. Streiner (2011), "The difference between reliability and agreement." *Jornal of Clinical Epidemiology*, 64, 701–702.

Kraemer, H.C. (1979), "Ramifications of a population model for κ as a coefficient of reliability." *Psychometrika*, 44, 461–472.

Kraemer, H.C., V.S. Peryakoil, and A. Noda (2002), "Kappa coefficients in medical research." *Statistics in Medicine*, 21, 2109–2129.

Krippendorff, K. (1970), "Estimating the reliability, systematic error, and random error of interval data." *Educational and Psychological Measurement*, 30, 61–70.

Krippendorff, K. (1978), "Reliability of binary attribute data." *Biometrics*, 34, 142–144.

Krippendorff, K. (2012), *Content Analysis: An Introduction to Its Methodology*, 3 edition. Thousand Oaks, CA: SAGE Publications, Inc., California, USA.

Krippendorff, Klaus (2004), "Measuring the reliability of qualitative text analysis data." *Quality and Quantity*, 38, 787–800.

Krippendorff, Klaus (2011), "Agreement and information in the reliability of coding." *Communication Methods and Measures*, 5, 93–112.

Krippendorff, Klaus (2018), *Content Analysis: An Introduction to Its Methodology*, 4 edition. SAGE Publications, Inc.

Kuder, G.F. and M.W. Richardson (1937), "The theory of the estimation of test reliability." *Psychometrika*, 2, 151–160.

Landis, J.R. and Koch G. (1977), "The measurement of observer agreement for categorical data." *Biometrics*, 33, 159–174.

Leone, M.A., P. Gaviani, and G. Ciccone (2006), "Inter-coder agreement for icd-9-cm coding of stroke." *Neurological Sciences*, 27, 445–448.

Light, R.J. (1971), "Measures of response agreement for qualitative data: some generalizations and alternatives." *Psychological Bulletin*, 76, 365–377.

Likert, Rensis (1932), "A technique for the measurement of attitudes." *Archives of Psychology*, 22, 3–55.

Lindsay, B.G., M. Markatou, S. Ray, K. Yang, and S. Chen (2008), "Quadratic distances on probabilities: A unified foundation." *The Annals of Statistics*, 36, 983–1006.

Mathet, Yann, Antoine Widlöcher, and Jean-Philippe Métivier (2015), "The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment." *Computational Linguistics*, 41, 437–479.

Maxwell, A.E. (1977), "Coefficient of agreement between observers and their interpretation." *British Journal of Psychiatry*, 130, 79–83.

McCarthy, P.J. (1966), "Replication: An approach to the analysis of data from complex surveys." Technical Report 14, National Center for Health Statistics, Vital and health statistics, Washington, D.C. Series 2: Data evaluation and methods.

McIver, J.P. and E.G. Carmines (1981), *Unidimensional Scaling*. Thousand Oaks, CA: Sage.

Metropolis, N. and S. Ulam (1949), "The monte-carlo method." *Journal of the American Statistical Association*, 44, 335–341.

Neyman, J. (1934), "On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection." *Journal of the Royal Statistical Society*, 97, 558–606.

Nunnally, J.C. (1978), *Psychometric Theory*, 2 edition. McGraw-Hill, New York.

Osgood, C.E. (1959), "The representational model and relevant research methods." In *Trends in Content Analysis* (I. de Sola Pool, ed.), 33–88, University of Illinois Press, Urbana.

Park, H.M. and H.W. Jung (2003), "Evaluating interrater agreement with intraclass correlation coefficient in spice-based software process assessment." In *Proceedings of the Third International Conference On Quality Software*, 308–314, Dallas, TX, USA.

Perreault, W.D. and L.E. Leigh (1989), "Reliability of nominal data based on qualitative judgments." *Journal of Marketing Research*, 26, 135–148.

Quenouille, M.H. (1949), "Approximate tests of correlation in time series." *Journal of The Royal Statistical Society, Series B*, 11, 68–84.

Quenouille, M.H. (1956), "Notes on bias in estimation." *Biometrika*, 61, 353–360.

Rowland, W.J. (1984), "The relationships among nuptial coloration, aggression, and courtship in male threespine sticklebacks." *Canadian Journal of Zoology*, 51, 453–466.

Särndal, C.E., B. Swensson, and J. Wretman (2003), *Model Assisted Survey Sampling*. Springer-Verlag New York, Inc., New York.

Savkov, A., J. Carroll, R. Koeling, and J. Cassell (2016), "Annotating patient clinical records with syntactic chunks and named entities: the harvey corpus." *Lang Resources & Evaluation*, 50, 523–548.

Schuster, C. and A. von Eye (2001), "Models for ordinal agreement data." *Biometrical Journal*, 43, 795–808.

Scott, W.A. (1955), "Reliability of content analysis: the case of nominal scale coding." *Public Opinion Quarterly*, XIX, 321–325.

Shoukri, M.M. (2010), *Measures of Interobserver Agreement and Reliability*, 2 edition. CRC/Biostatistics Series, Chapman and Hall/CRC Press.

Sim, J. and C.C. Wright (2005), "The kappa statistic in reliability studies: use, interpretation, and sample size requirements." *Physical Therapy*, 85, 257–268.

Stein, C.R., R.B. Devore, and B.E. Wojcik (2005), "Calculation of the kappa statistic for inter-rater reliability: The case where raters can select multiple responses from a large number of categories." In *Proceedings of the Thirtieth Annual SAS® Users Group International Conference*, SAS Institute Inc., SAS Institute Inc, Cary, NC.

Tanner, M.A. and M.A. Young (1985), "Modeling agreement among raters." *Journal of American Statistical Association*, 80, 175–180.

Tinsley, Howard E. A. and David J. Weiss (1975), "Interrater reliability and agreement of subjective judgments." *Journal of Counselling Psychology*, 22, 358–376.

Tinsley, Howard E. A. and David J. Weiss (2000), "Journal of counselling psychology." In *Handbook of Applied Multivariate Statistics and Mathematical Modeling* (S. D. Tinsley, H. E. A. & Brown, ed.), 94–124, Academic Press, New York.

Traub, R.E. (1994), *Reliability for the Social Sciences: Theory and Applications*. Sage Publications, Beverly Hills.

Tukey, J.W. (1958), "Bias and confidence in not quite large samples (abstract)." *Annals of Mathematical Statistics*, 29, 614.

von Eye, A. and E.Y. Mun (2006), *Analyzing Rater Agreement: Manifest Variable Methods*, pap/cdr edition. Lawrence Erlbaum Associates.

Wongpakaran, Nahathai, Tinakon Wongpakaran, Danny Wedding, and Kilem L. Gwet (2013), "A comparison of cohen's kappa and gwet's ac1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples." *BMC Medical Research Methodology*, 13, URL https://doi.org/10.1186/1471-2288-13-61.

Zhao, X., J.S. Liu, and K. Deng (2013), "Assumptions behind intercoder reliability indices." In *Communication Yearbook* (C.T. Salmon, ed.), volume 36, 419–480, Routledge.

List of Notations

Author Index

Subject Index

Handbook of Inter-Rater Reliability, 5th Ed.

Vol 1: Analysis of Categorical Ratings

Inter-rater reliability assessment has become an essential component in the process of evaluating the quality of experimental data in almost all fields of research. The 4th edition of the *Handbook of Inter-Rater Reliability* covered the analysis of categorical and quantitative ratings in a single volume. In response to comments on previous editions, the current 5th edition is released in 2 volumes. Volume 2 is devoted to the analysis of quantitative ratings, whereas the current volume 1 focuses on the analysis of categorical ratings.

Here is the link to the webpage of volume 1, where you can find, a link to the errata page, some example workbooks, and other datasets:

www.agreestat.com/books/cac5/

Here are a few topics that are new to the 5th edition:

- Chapter 2 describes various methods for setting up your dataset of ratings before analysis.
- New sample size calculation procedures for chance-corrected agreement coefficients are presented in chapter 6.
- Several new techniques for analyzing categorical ratings are described in chapter 9. Among these are the inter-annotator agreement, useful in Natural Language Processing, the testing of the difference of 2 agreement coefficients for statistical significance, and many others.

About the Author

Kilem L. Gwet, Ph.D.
Statistical consultant, mathematical statistician, researcher, and instructor. Over 20 years of experience in various industries, and several publications in peer-reviewed journals.

$42.99
ISBN 978-1-7923-5463-2
54299>
9 781792 354632

AgreeStat Analytics
PO Box 2696
Gaithersburg, Maryland 20886-2696 – USA