Chapter 819

Confidence Intervals for Kappa

Introduction

The *kappa statistic*, κ, is a measure of the agreement between two raters of N subjects on k categories. This routine calculates the sample size needed to obtain a specified width of a confidence interval for the kappa statistic at a stated confidence level. The kappa statistic was proposed by Cohen (1960). Sample size calculations are given in Cohen (1960), Fleiss et al (1969), and Flack et al (1988).

Technical Details

Suppose that N subjects are each assigned independently to one of k categories by two separate judges or raters. The results are placed in a $k \times k$ contingency table. Each p_{ij} represents the proportion of subjects that Rater A classified in category i, but Rater B classified in category j, with $i, j = 1, 2, \ldots, k$. The proportions $p_{i.}$ and $p_{.j}$ are the marginal frequencies or probabilities of assignment into categories i and j for Rater A and Rater B, respectively. For each rater, the category marginal frequencies sum to one.

	Rater B				
Rater A	1	2	…	k	Total
1	p_{11}	p_{12}	…	p_{1k}	$p_{1.}$
2	p_{21}	p_{22}	…	p_{2k}	$p_{2.}$
⋮	⋮	⋮	⋱	⋮	⋮
k	p_{k1}	p_{k2}	…	p_{kk}	$p_{k.}$
Total	$p_{.1}$	$p_{.2}$	…	$p_{.k}$	1

The proportions on the diagonal, p_{ii}, represent the proportion of subjects in each category for which the two raters agreed on the assignment. The overall proportion of observed agreement is

$$PO = \sum_{i=1}^{k} p_{ii}$$

and the overall proportion of agreement expected by chance is

$$PE = \sum_{i=1}^{k} p_{i\cdot} p_{\cdot i}$$

The value of kappa, which measures the degree of rater agreement, is

$$\kappa = \frac{PO - PE}{1 - PE}$$

A kappa value of 1 represents perfect agreement between the two raters. A kappa value of 0 indicates no more rater agreement than that expected by chance. A kappa value of -1 would indicate perfect disagreement between the raters.

The standard error of the estimated κ is given by Fleiss, Cohen, and Everitt (1969) as

$$SE(\kappa) = \frac{SD(\kappa)}{\sqrt{\mathrm{N}}}$$

where

$$SD(\kappa) = \frac{1}{(1-PE)^2}\left\{PO(1-PE)^2 + (1-PO)^2 \sum_{i=1}^{k} \sum_{j=1,j\neq i}^{k} p_{ij}\left(p_{\cdot i} + p_{j\cdot}\right)^2 - 2(1-PO)(1-PE)\sum_{i=1}^{k} p_{ii}(p_{\cdot i} + p_{i\cdot})^2 - (PO \cdot PE - 2PE + PO)^2\right\}^{\frac{1}{2}}$$

Cohen (1960) gave the following expression for *SD(κ)*

$$SD(\kappa) = \sqrt{\frac{PO(1-PO)}{(1-PE)^2}}$$

Fleiss, Cohen, and Everitt (1969) explain that this expression is an approximation based on faulty assumptions. However, because of its simplicity, it is often used for planning purposes. PASS includes a Kappa calculation tool that lets you compute both values of *SD(κ)* so they can be compared. We have found that Cohen's approximation is often close to Fleiss's more accurate version.

Once *SD(κ)* is calculated, a 100(1 – α)% confidence interval for κ may be computed using the standard normal distribution as follows

$$\kappa \pm z_{\alpha/2} SD(\kappa)$$

The width of the confidence interval is $2z_{\alpha/2}SD(\kappa)$. One-sided limits may be obtained by replacing $\alpha/2$ by α.

Confidence Level

The confidence level, 1 – α, has the following interpretation. If thousands of samples of *N* items are drawn from a population using simple random sampling and a confidence interval is calculated for each sample, the proportion of those intervals that will include the true value of kappa is 1 – α.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Design tab. For more information about the options of other tabs, go to the Procedure Window chapter.

Design Tab

The Design tab contains most of the parameters and options that you will be concerned with.

Solve For

Solve For

This option specifies the parameter to be solved for from the other parameters.

One-Sided or Two-Sided Interval

Interval Type

Specify whether the confidence interval for kappa is two-sided or one-sided. A one-sided interval is often called a **confidence bound** rather than a confidence interval because it only has one limit.

- **Two-Sided**

  The two-sided confidence interval is defined by two limits: an upper confidence limit (UCL) and a lower confidence limit (LCL).

  These limits are constructed so that the designated proportion (confidence level) of such intervals will include the true value of kappa.

- **Upper One-Sided**

  The upper confidence interval (or bound) is defined by a limit above the estimated value of kappa. The limit is constructed so that the designated proportion (confidence level) of such limits has the true population value below them.

- **Lower One-Sided**

  The lower confidence interval (or bound) is defined by a limit below the estimated value of kappa. The limit is constructed so that the designated proportion (confidence level) of such limits has the true population value above them.

Confidence

Confidence Level

The confidence level, $1 - \alpha$, has the following interpretation. If thousands of samples of N items are drawn from a population using simple random sampling and a confidence interval is calculated for each sample, the proportion of those intervals that will include the true population correlation is $1 - \alpha$.

Often, the values 0.95 or 0.99 are used. You can enter single values or a range of values such as *0.90, 0.95* or *0.90 to 0.99 by 0.01*.

Sample Size

N (Number of Subjects)

Enter one or more values for N, the number of subjects rated by the two raters. You can enter a single value or a range of values.

Precision

Width of Confidence Interval

This is the distance between the lower confidence limit (*LCL*) and the upper confidence limit (*UCL*). Its calculation is *UCL* - *LCL*. It is a measure of the precision of the confidence interval. This width usually ranges between 0 and 1.

You can enter a single value or a list of values.

Distance from *κ* to Limit

This is the distance from *κ* to the lower confidence limit (*LCL*) or the upper confidence limit (*UCL*). It is calculated using | *κ* - *LCL*| or |*UCL* - *κ* |. The range is between 0 and 1.

You can enter a single value or a list of values.

Sample *κ* and SD(*κ*)

Specify *κ* and SD(*κ*) using

Select the method used to specify or calculate SD(*κ*), the standard deviation of the estimated κ. Note that the standard error of *κ* is calculated with SD(*κ*)/√N, so SD(*κ*) is that part of the standard error not related to the sample size.

Your choices are

- **κ and SD(κ) of Cohen or Fleiss**

  Specify *κ* and SD(*κ*) directly. You can use the value of SD(*κ*) calculated according to Cohen (1960) or Fleiss (1969).

- **κ and PO (Proportion Agreeing)**

  Specify *κ* and *PO*, the proportion of subjects on which the two raters agreed on the classification. SD(*κ*) is calculated from *κ* and *PO* using the formula of Cohen (1960).

- **κ and Marginal Classification Frequencies**

  Specify *κ* and a set of marginal classification frequencies that will represent both raters. A search is made for the largest SD(*κ*) possible with these frequencies and value of *κ*. SD(*κ*) is calculated using the formula of Fleiss (1969).

- **Rater-by-Rater Frequency Table in Spreadsheet**

  Specify a rater by rater contingency table in the spreadsheet. This table gives the proportion of subjects in each cell (rater-by-rater classification). Only one such table may be specified. *κ* and SD(*κ*) are calculated from this table using the formulas of Fleiss (1969).

κ (Sample Kappa)

Cohen’s κ (kappa) coefficient is a measure of inter-rater agreement between two raters who each classify N subjects into K mutually exclusive classes.

The formula is

$$\kappa = \frac{PO - PE}{1 - PE}$$

where *PO* is the proportion of subjects that are rated identically by both raters and PE is the expect probability of chance agreement.

Kappa ranges between 0 (no more agreement than anticipated by chance) and 1 (perfect agreement).

You can enter a single value and a list of values separated by blanks.

SD(*κ*) of Cohen or Fleiss

SD(κ) is the standard deviation of the estimated kappa. This should not be confused with SE(κ), the standard error of the estimated kappa. The relationship between these two measures is that

$$SE(\kappa) = \frac{SD(\kappa)}{\sqrt{\mathrm{N}}}.$$

That is, SD(κ) is independent of N.

You can use the Kappa Estimator to compute values of SD(κ) from a variety of matrix settings and then enter an appropriate range of SD(κ) values here for use in the confidence interval calculation.

You can enter a single value and a list of values separated by blanks.

Cohen's SD(κ)

Cohen (1960) presented a formula for estimating SD(κ). This was a relatively simple formula involving only PO and PE. Because of its simplicity, it is often used during design planning.

Fleiss's SD(κ)

Fleiss, Cohen, and Everitt (1969) published a new formula for estimating SD(κ) which they claim is more accurate and is based on a better theoretical foundation than Cohen's original formula. Although Fleiss's version is more accurate, it requires the knowledge of the agreement matrix which is usually unavailable during planning. However, once the experiment has been run, there is no reason not to calculate Fleiss's SD(κ).

PO (Proportion Agreeing)

Enter the proportion of subjects on which the raters agree about their classification, PO. SD(κ) is then calculated from κ and PO using the formula

$$SD(\kappa) = \sqrt{\frac{PO(1 - PO)}{(1 - PE)^2}}$$

Since *PO* is a proportion, it must be between 0 and 1. For trained raters, *PO* will often be between 0.70 and 0.90.

You can enter a single value and a list of values separated by blanks.

Specify Frequencies using

Select the method used to specify the marginal classification frequencies.

- **List Input**

  Specify a single set of marginal frequencies as a list. For example, with three categories you might enter "2 3 5."

- **Spreadsheet Column Input**

  Specify more than one set of marginal frequencies (proportions) using the spreadsheet. Each spreadsheet column becomes a set of marginal frequencies. For example, if you have three sets of frequencies in the three columns C1, C2, and C3, you would enter "=C1 C2 C3".

Classification Frequencies

Specify two or more marginal classification frequencies (proportions). These are the relative proportions of subjects in each category as assigned by the two raters or judges. (Note that in practice these frequencies are not exactly equal for the two raters, but the sample size formula makes this simplifying assumption.)

Number of Classes

The number of classes is equal to the number of items you enter in this list.

Frequencies Need Not Sum to 1

Because they are a complete list of marginal proportions, a set of frequencies will sum to 1. However, to make data entry easier, you do not have to enter values that sum to one. The program will rescale the values so they do. For example, an entry of "20 30 50" is rescaled to "0.2 0.3 0.5" and an entry of "1 1 1 1" is rescaled to "0.25 0.25 0.25 0.25". Also, negative values will be made positive. Zero values are ignored.

Syntax

Enter a list of values such as "10 30 60" or "40 60". To rescale the values so they sum to one, each item is divided by the list total.

Spreadsheet Columns

Specify two or more classification frequencies in columns of the spreadsheet. These are the marginal frequencies of subjects in each category as assigned by the two raters. (Note that in practice these frequencies are not exactly equal for the two raters, but the sample size formulas make this simplifying assumption.)

Number of Categories

The number of categories is equal to the number of items in the list.

Frequencies Need Not Sum to 1

In the actual formulas, a set of frequencies must sum to one. However, to make data entry easier, you do not have to enter values that sum to one. The program will rescale the values so they do. For example, an entry of "20 30 50" is rescaled to "0.2 0.3 0.5" and an entry of "1 1 1 1" is rescaled to "0.25 0.25 0.25 0.25". Also, negative values will be made positive. Zero values should not be used.

Spreadsheet Column Input

Several sets of frequencies can be entered in different columns of the PASS spreadsheet. To launch the spreadsheet, click on the Spreadsheet button to the right of the box. To select columns from the spreadsheet, click on the button with the arrow pointing down just to the right of the box. You specify the column (or columns) to be used by beginning your entry with an equals sign, e.g. enter “=C1-C3”.

Specify multiple sets of frequencies using the column input syntax “=[column 1] [column 2] etc.” For example, if you have three frequency sets stored in the spreadsheet in columns C1, C2, and C3, you would enter “=C1 C2 C3” in this box. A separate analysis is conducted for each column.

Each column in the spreadsheet corresponds to a single set of frequencies. The columns may even contain different numbers of frequencies.

Spreadsheet Columns – Rater-by-Rater Frequency Table in Spreadsheet

Enter the column numbers of a matrix in the spreadsheet holding a square contingency table of the ratings of the two raters. The cells in the table will be rescaled to proportions by dividing each cell by the total of all cells, so the actual scale of your table does not matter. The number of rows must match the number of columns selected. The values of κ, PO, PE, and Fleiss's SD(κ) are calculated from this table.

Example Frequency Table

C1	C2	C3
53	11	10
17	42	12
13	15	59

You would enter “=C1 C2 C3” in this box.

Kappa Estimator Tool

A special tool is available to help in preparing this table. Click the Kappa Estimator tool button to the upper right to load the Kappa Estimator tool. Once you have entered a reasonable table, you can copy the table, close the estimator window, open the spreadsheet by clicking the spreadsheet button to the left, and pasting the table into the spreadsheet.

Example 1 – Calculating Sample Size

Suppose a study is planned to estimate kappa with a two-sided 95% confidence interval with a width no wider than 0.1. The researcher would like to examine values of PO from 0.70 to 0.95 in steps of 0.05. From past studies, the researcher wants to use a planning estimate of 0.5 for the sample kappa. The goal is to determine the necessary sample size for each scenario.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Confidence Intervals for Kappa** procedure window by expanding **Correlation**, then clicking on **Kappa Rater Agreement**, and then clicking on **Confidence Intervals for Kappa**. You may then make the appropriate entries as listed below, or open **Example 1** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Solve For	**Sample Size**
Interval Type	**Two-Sided**
Confidence Level	**0.95**
Width of Confidence Interval	**0.1**
Specify κ and SD(κ) using	**κ and PO – Cohen’s SD(κ)**
κ (Sample Kappa)	**0.6**
PO (Proportion Agreeing)	**0.7 0.75 0.8 0.85 0.9 0.95**

Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for a Two-Sided Confidence Interval

                                           Lower     Upper
          Sample                            C.I.      C.I.               Actual    Expect
Conf        Size     Kappa      Cohen      Limit     Limit                Agree     Agree
Level          N         κ      SD(κ)        LCL       UCL      Width        PO        PE
0.950        574     0.600      0.611      0.550     0.650      0.100     0.700     0.250
0.950        738     0.600      0.693      0.550     0.650      0.100     0.750     0.375
0.950        984     0.600      0.800      0.550     0.650      0.100     0.800     0.500
0.950       1394     0.600      0.952      0.550     0.650      0.100     0.850     0.625
0.950       2213     0.600      1.200      0.550     0.650      0.100     0.900     0.750
0.950       4672     0.600      1.744      0.550     0.650      0.100     0.950     0.875
```

References

Cohen, Jacob. 1960. 'A Coefficient of Agreement for Nominal Scales'. Educational and Psychological Measurement. Vol. 20, No. 1, 37-46.

Fleiss, J.L., Cohen, J., and Everitt, B.S. 1969. 'Large Sample Standard Errors of Kappa and Weighted Kappa'. Psychological Bulletin. Vol. 72, No. 5, 323-327.

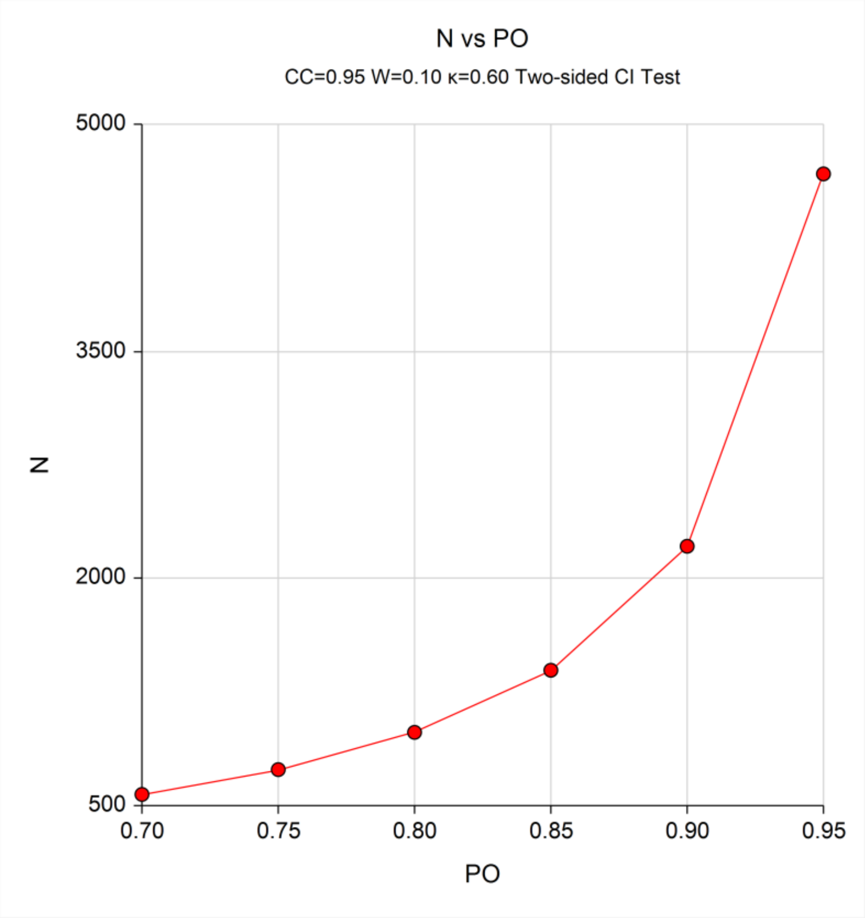
Flack, V.F., Afifi, A.A., Lachenbruch, P.A., and Schouten, H.J.A. 1988. 'Sample Size Determinations for the Two Rater Kappa Statistic'. Psychometrika. Vol. 53, No. 3, 321-325.

Report Definitions
Confidence Level is the proportion of confidence intervals (constructed with this same confidence level) that would contain the true value of κ.
N is the sample size: the number of objects or subjects rated by these raters.
κ is a planning estimate of the value of kappa, the coefficient of agreement.
Cohen SD(κ) is the standard deviation of the estimate of κ calculated using Cohen's 1960 formula. This value is divided by the square root of N to obtain an estimate of the standard error of κ.
Lower and Upper C.I. Limits are the lower and upper limits of the confidence interval.
Width is the distance from the lower limit to the upper limit.
Actual Agree PO is the proportion of all subjects that were classified identically by both raters.
Expected Agree PE is the expected proportion of agreements for both raters. It is the sum of the products of the marginal proportions for each category by each rater.

Summary Statements
A sample size of 574 subjects results in a two-sided 95% confidence interval with a width of 0.100 if the value of κ is 0.600 and the standard deviation SD(κ), is 0.611. This interval uses Cohen's large-sample formula for SD(κ).

This report shows the calculated sample size for each of the scenarios.

Plots Section



This plot shows the sample size versus the value of PO.

Example 2 – Validation using Cohen

Cohen (1960), pages 44 – 45, gives an example in which κ = 0.492, confidence level = 0.95, PO = 0.70, and width = 0.216. The value of N is 200.

We will now validate this routine using this example.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Confidence Intervals for Kappa** procedure window by expanding **Correlation**, then clicking on **Kappa Rater Agreement**, and then clicking on **Confidence Intervals for Kappa**. You may then make the appropriate entries as listed below, or open **Example 2** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Solve For	**Sample Size**
Interval Type	**Two-Sided**
Confidence Level	**0.95**
Width of Confidence Interval	**0.216**
Specify κ and SD(κ) using	**κ and PO – Cohen's SD(κ)**
κ (Sample Kappa)	**0.492**
PO (Proportion Agreeing)	**0.7**

Output

Click the Calculate button to perform the calculations and generate the following output.

```
Numeric Results for a Two-Sided Confidence Interval

                                            Lower     Upper
         Sample                              C.I.      C.I.                Actual    Expect
Conf       Size     Kappa      Cohen        Limit     Limit                 Agree     Agree
Level         N         κ      SD(κ)          LCL       UCL      Width         PO        PE
0.950       199     0.492      0.776        0.384     0.600      0.216      0.700     0.409
```

PASS's N = 199 matches Cohen's result of 200 within rounding.