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* × *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, …, *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

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

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

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

where

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

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 *α*/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.

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

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

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 Reports

Numeric Results

Solve For: Sample Size
Interval Type: Two-Sided

Confidence Level	Sample Size N	Kappa κ	Cohen SD(κ)	Confidence Interval Limits Lower	Upper	Confidence Interval Width	Proportion Agreeing Actual PO	Expected PE
0.95	574	0.6	0.611	0.55	0.65	0.1	0.70	0.250
0.95	738	0.6	0.693	0.55	0.65	0.1	0.75	0.375
0.95	984	0.6	0.800	0.55	0.65	0.1	0.80	0.500
0.95	1394	0.6	0.952	0.55	0.65	0.1	0.85	0.625
0.95	2213	0.6	1.200	0.55	0.65	0.1	0.90	0.750
0.95	4672	0.6	1.744	0.55	0.65	0.1	0.95	0.875

Confidence Level — The proportion of confidence intervals (constructed with this same confidence level) that would contain the true value of κ.
N — The sample size. The number of objects or subjects rated by these raters.
κ — A planning estimate of the value of kappa, the coefficient of agreement.
Cohen SD(κ) — 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.
Confidence Interval Limits — The lower and upper limits of the confidence interval.
Confidence Interval Width — The distance from the lower limit to the upper limit.
PO — Actual Proportion Agreeing. The proportion of all subjects that were classified identically by both raters.
PE — Expected Proportion Agreeing. 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 two-rater design will be used to obtain a two-sided 95% confidence interval for the agreement measure, kappa (κ). The sample kappa is assumed to be 0.6, the proportion agreeing is assumed to be 0.7, and the standard deviation SD(κ) is assumed to be 0.611 (calculated from the stated proportion agreeing). The interval calculation uses Cohen's large-sample formula for SD(κ). To produce a confidence interval with a width of no more than 0.1, 574 subjects will be needed.

Dropout-Inflated Sample Size

Dropout Rate	Sample Size N	Dropout-Inflated Enrollment Sample Size N'	Expected Number of Dropouts D
20%	574	718	144
20%	738	923	185
20%	984	1230	246
20%	1394	1743	349
20%	2213	2767	554
20%	4672	5840	1168

Dropout Rate	The percentage of subjects (or items) that are expected to be lost at random during the course of the study and for whom no response data will be collected (i.e., will be treated as "missing"). Abbreviated as DR.
N	The evaluable sample size at which the confidence interval is computed. If N subjects are evaluated out of the N' subjects that are enrolled in the study, the design will achieve the stated confidence interval.
N'	The total number of subjects that should be enrolled in the study in order to obtain N evaluable subjects, based on the assumed dropout rate. After solving for N, N' is calculated by inflating N using the formula N' = N / (1 - DR), with N' always rounded up. (See Julious, S.A. (2010) pages 52-53, or Chow, S.C., Shao, J., Wang, H., and Lokhnygina, Y. (2018) pages 32-33.)
D	The expected number of dropouts. D = N' - N.

Dropout Summary Statements

Anticipating a 20% dropout rate, 718 subjects should be enrolled to obtain a final sample size of 574 subjects.

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.

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

Plots Section

Plots

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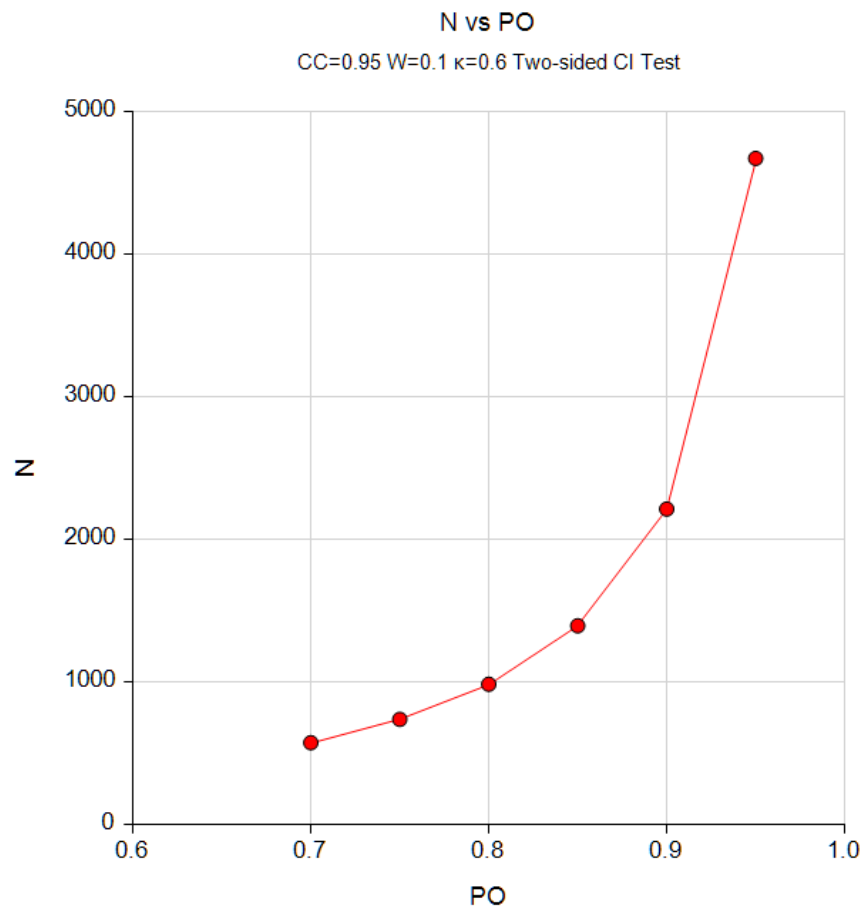
This plot shows the sample size versus the value of PO.

Example 2 – Validation using Cohen (1960)

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

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 2** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

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**

Output

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

Numeric Results
───
Solve For: Sample Size
Interval Type: Two-Sided
───

Confidence Level	Sample Size N	Kappa κ	Cohen SD(κ)	Confidence Interval Limits		Confidence Interval Width	Proportion Agreeing	
				Lower	Upper		Actual PO	Expected PE
0.95	199	0.492	0.776	0.384	0.6	0.216	0.7	0.409

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