How to Calculate Cohen’s Kappa in SPSS: A Step-by-Step Guide

Authored by

stats writer

January 24, 2026

Cohen's Kappa is a sophisticated statistical measure utilized extensively in research to rigorously assess the level of agreement between two independent raters, observers, or measurement methods when classifying items into mutually exclusive categories. Unlike raw percentage agreement, Kappa incorporates an essential correction factor: the probability of agreement occurring purely by chance. To successfully calculate this coefficient in **SPSS**, a specific set of procedures must be executed, starting with data preparation and culminating in the use of the **Crosstabs** dialogue.
It is critically important that the data conforms to the required measurement parameters for this statistic; specifically, the ratings must be measured on a **nominal or ordinal scale**. Furthermore, achieving a reliable and accurate estimation of the Kappa value necessitates the inclusion of a sufficiently large and representative sample size, minimizing the influence of random sampling error.

Calculating Cohen's Kappa Using SPSS Software

Fundamental Definition and Application of Kappa

Cohen's Kappa serves as the standard metric for quantifying inter-rater reliability, particularly when dealing with two raters classifying subjects based on categorical variables. The measure is essential for establishing the credibility of subjective rating systems, ensuring that differences in classification are due to inherent differences in the items being rated, rather than inconsistencies between the raters themselves.

The formula for Cohen's kappa is specifically structured to adjust for the baseline agreement that would naturally be expected if the raters were simply guessing randomly based on the marginal totals of the contingency table. This adjustment makes Kappa a more rigorous indicator of reliability compared to simple percentage agreement.

The widely accepted formula for the coefficient is calculated as:

k = (po – pe) / (1 - pe)

Where the constituent parts define the relationship between observed and expected agreement:

po: This term denotes the Relative **observed agreement** among raters. This is the proportion of observations where the two raters provided identical classifications.
pe: This term denotes the Hypothetical probability of **chance agreement**. This is derived from the marginal totals and represents the expected agreement if the raters' decisions were statistically independent.

By accounting for the probability of chance agreement (pe), Cohen's Kappa provides a

conservative estimate of reliability, offering insight into the true consistency beyond sheer luck.

Interpreting the Coefficient Value Range

The resultant Kappa statistic typically ranges between 0 and 1, providing an easily interpretable measure of reliability. Understanding the interpretation scale is crucial for translating statistical output into meaningful research conclusions.

A value close to **0** indicates minimal or no true agreement. The observed consistency is no better than that predicted by chance factors alone.
A value close to **1** indicates almost **perfect agreement** between the two raters, confirming exceptional consistency and high inter-rater reliability.

The following widely referenced table summarizes how researchers typically categorize and interpret different numerical values of **Cohen's Kappa** in academic and clinical settings:

Cohen's Kappa	Interpretation
0	No agreement
0.10 - 0.20	Slight agreement
0.21 - 0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement
0.81 - 0.99	Near perfect agreement
1	Perfect agreement

Case Study Example: Art Curator Ratings

To demonstrate the practical steps in **SPSS**, consider a hypothetical case where two experienced art museum curators are asked to rate a selection of 30 paintings. Each curator evaluates whether a specific painting is suitable ("Yes") or unsuitable ("No") for display in a new exhibition.

The objective is to calculate Cohen's Kappa to determine the level of consistency in judgment between the two curators, thereby validating the objectivity of their classification system. The dataset must be structured with two categorical variables (Rater 1 and Rater 2), representing the independent ratings for each of the 30 paintings.

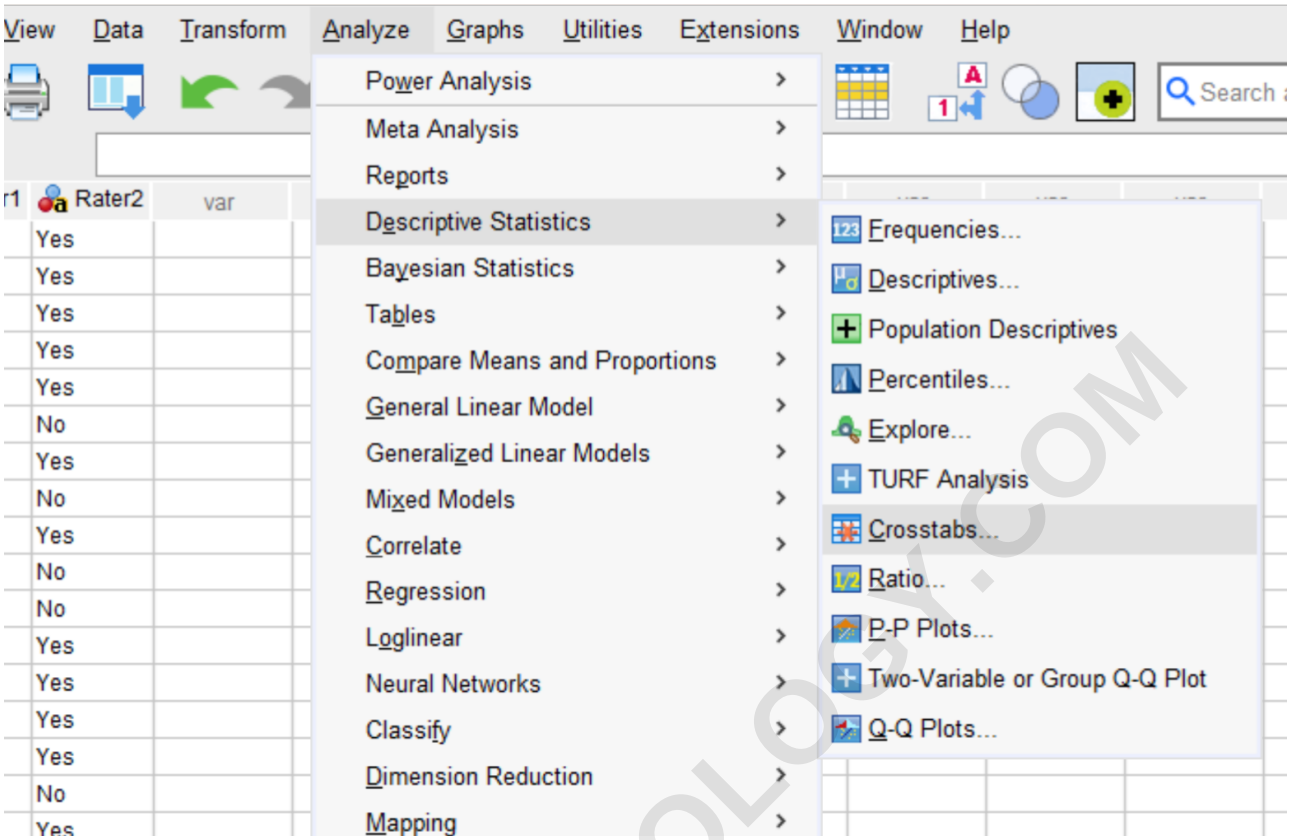
The following data structure represents the classifications provided by each rater:

	Rater1	Rater2	var	var	var	va
1	Yes	Yes				
2	Yes	Yes				
3	Yes	Yes				
4	Yes	Yes				
5	Yes	Yes				
6	Yes	No				
7	Yes	Yes				
8	Yes	No				
9	Yes	Yes				
10	Yes	No				
11	Yes	No				
12	Yes	Yes				
13	Yes	Yes				
14	Yes	Yes				
15	Yes	Yes				
16	Yes	No				
17	Yes	Yes				
18	Yes	Yes				
19	Yes	No				
20	No	Yes				
21	No	Yes				
22	No	No				
23	No	No				
24	No	Yes				
25	No	Yes				
26	No	No				
27	No	No				
28	No	No				
29	No	No				
30	No	No				
31						

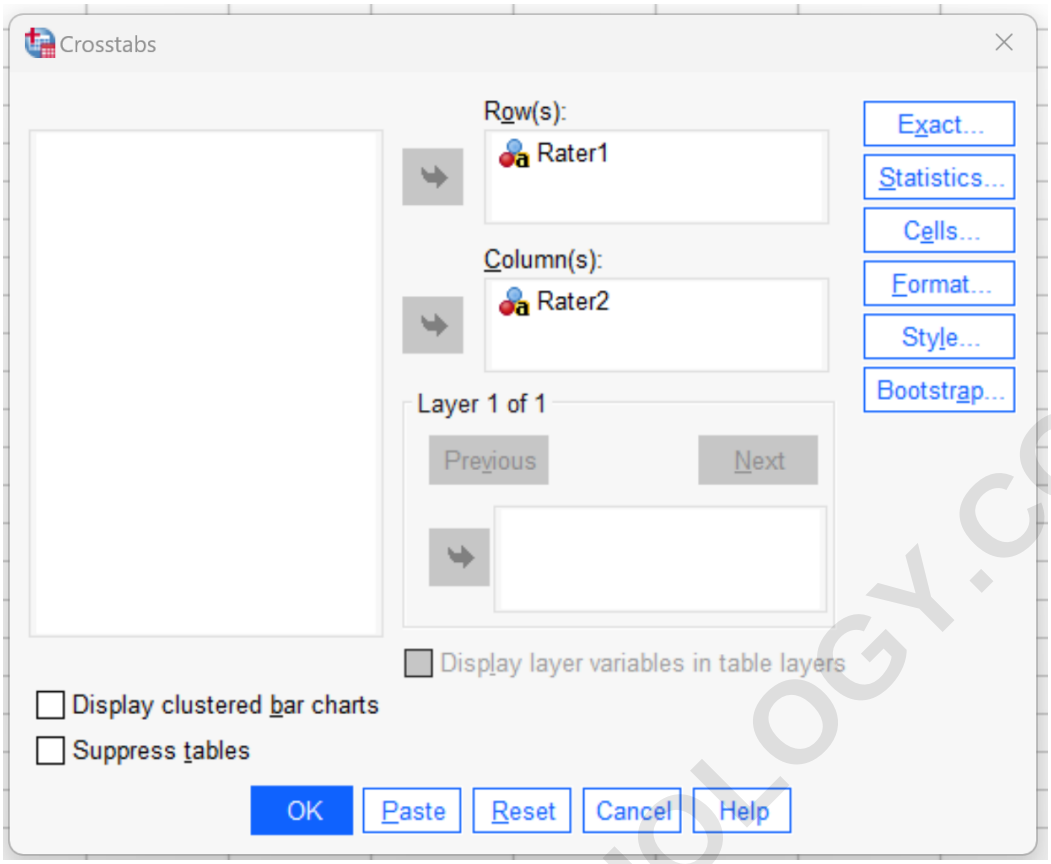
Executing the Crosstabs Command in SPSS

The calculation begins by initiating the appropriate statistical routine within the IBM **SPSS** Statistics software, ensuring that the categorical data is mapped correctly into a bivariate table structure.

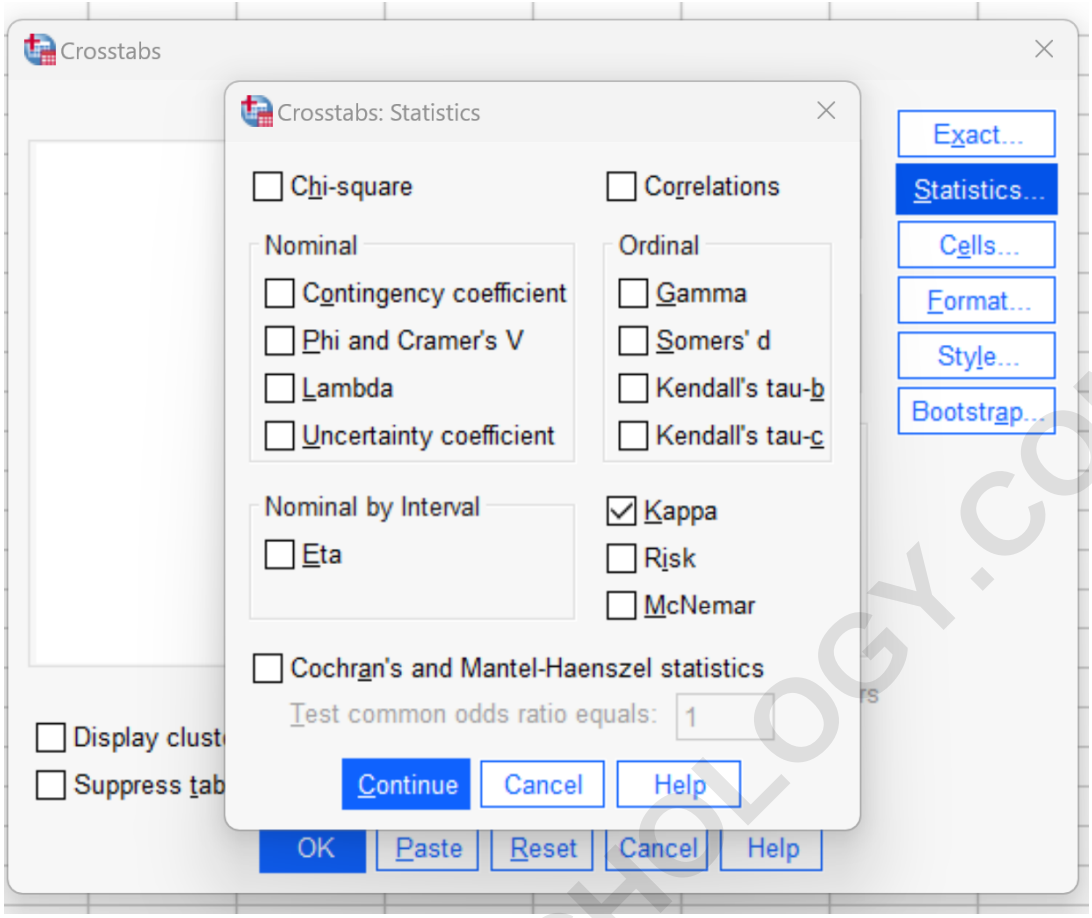
First, access the statistical analysis menu by clicking the **Analyze** tab, then selecting **Descriptive Statistics**, and finally clicking **Crosstabs** to open the primary dialog box:



In the Crosstabs window, drag the **Rater1** variable into the **Rows** field and the **Rater2** variable into the **Columns** field. This action defines the two dimensions of the contingency table that SPSS will construct:



The final step before running the analysis is to explicitly request the Kappa coefficient. Click the **Statistics** button, locate the **Kappa** checkbox, and ensure it is selected to include the inter-rater reliability measure in the output:



After selecting the option, click **Continue** to exit the Statistics sub-menu, and then click **OK** in the main dialog box to generate the results in the SPSS Output Viewer.

Interpreting the SPSS Output Tables

The output provides two main tables of interest: the crosstabulation table (frequency distribution) and the symmetry measures table (which includes the Kappa coefficient). The crosstabulation table summarizes the joint decisions made by the two raters:

Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Rater1 * Rater2	30	100.0%	0	0.0%	30	100.0%

Rater1 * Rater2 Crosstabulation

Count

		Rater2		
		No	Yes	Total
Rater1	No	7	4	11
	Yes	6	13	19
Total		13	17	30

Symmetric Measures

		Value	Asymptotic Standard Error [a]	Approximate T [b]	Approximate Significance
Measure of Agreement	Kappa	.309	.175	1.708	.088
N of Valid Cases		30			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

By inspecting the table diagonals, we can determine the exact points of agreement and disagreement among the 30 paintings rated:

There were **7** instances where both curators rated the painting as **No** (Agreement).
There were **13** instances where both curators rated the painting as **Yes** (Agreement).
Disagreement occurred in $4 + 6 = **10**$ instances.

The overall percentage of observed agreement (po) is $(7+13)/30$, or approximately 66.7%.

Conclusion: Assessing the Level of Agreement

The final table in the output provides the calculated **Cohen's Kappa** value, which accounts for the probability of chance agreement. For this curator example, the calculated Kappa coefficient is determined to be **.309**.

We must reference the standard interpretive guidelines once again to contextualize this numerical result:

Cohen's Kappa	Interpretation
0	No agreement
0.10 - 0.20	Slight agreement
0.21 - 0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement
0.81 - 0.99	Near perfect agreement
1	Perfect agreement

Based on this recognized scale, a Kappa coefficient of 0.309 is categorized as achieving a "Fair" level of agreement. This interpretation signifies that while the curators demonstrated some consistency, the reliability of their joint classification process remains relatively modest, suggesting the need for potentially refining the criteria or providing additional consensus training.