How to Easily Calculate Cohen’s Kappa in Excel

Authored by

stats writer

December 1, 2025

Cohen's Kappa, often denoted by the symbol k, is a robust statistical measure designed to assess the level of agreement between two independent observers or raters when classifying items into mutually exclusive categories. Unlike simple percentage agreement, Kappa is specifically formulated to account for the possibility that agreement may occur purely by chance. This adjustment makes it a much more reliable metric for determining true inter-rater reliability, particularly in fields like psychology, medicine, and data science where consistent human judgment is critical. Understanding and calculating Cohen's Kappa is essential for validating data quality derived from observational or qualitative studies.

While specialized statistical software is often used for this calculation, many researchers prefer to utilize accessible spreadsheet programs. Fortunately, deriving this important statistic can be efficiently performed in Microsoft Excel. Although older versions might require manual calculation steps, modern Excel versions sometimes offer specialized functions. However, the most universally applicable method involves calculating the observed agreement (Po) and the expected agreement by chance (Pe) before combining them into the final formula. This guide provides a comprehensive, step-by-step methodology for deriving Cohen's Kappa directly within Excel, ensuring clarity and accuracy in your statistical findings.

Understanding Cohen's Kappa: Beyond Simple Agreement

Cohen's Kappa is the preferred metric when researchers need to quantify the congruence between two independent judges who are classifying data into predetermined, mutually exclusive categorical items. If two raters agree on 80% of classifications, this raw percentage might seem high. However, if the raters were simply guessing, they might still achieve a high percentage of agreement, especially if the distribution of categories is heavily skewed (e.g., 95% of items belong to Category A). Kappa solves this problem by factoring in the agreement that is attributable solely to random chance based on the marginal frequencies.

The core strength of Cohen's Kappa lies in its rigor and correction for baseline agreement. By providing a measure that is corrected for chance, it offers a more honest assessment of the true consistency inherent in the raters' judgment process. This is vital for maintaining the internal validity of studies that rely on human assessment, such as diagnostic coding, content analysis, or performance evaluation. A high Kappa value confirms that the raters are applying the classification criteria consistently, suggesting the measurement instrument itself is reliable.

The Mathematical Foundation of Kappa (k)

The calculation of Cohen's Kappa relies on comparing the observed proportion of agreement (Po)

against the proportion of agreement expected by chance (Pe). The resulting statistic (k) is essentially the ratio of the observed non-chance agreement to the potential non-chance agreement. The formula provides a standardized way to quantify how much better the observed agreement is compared to what would be expected if the raters were classifying items randomly, based on the marginal probabilities of their ratings.

The fundamental formula for Cohen's Kappa is calculated as:

k = (po - pe) / (1 - pe)

Understanding the components of this equation is critical for accurate manual calculation in Excel, where these elements must be derived separately from the contingency table:

po: This represents the **Relative Observed Agreement** among the raters. It is calculated by summing the counts of items that both raters agreed upon (the diagonal elements of the contingency matrix) and dividing this sum by the total number of observations (N).
pe: This represents the **Hypothetical Probability of Chance Agreement**. It is derived from the marginal totals of the rows and columns. This value estimates how often the raters would align if their judgments were entirely independent and random, essentially modeling what agreement would look like if the raters were simply adhering to the overall category proportions.

By subtracting Pe from Po (the numerator), we isolate the agreement that is truly beyond chance. The denominator (1 - Pe) represents the maximum possible agreement beyond chance. The ratio thus provides a normalized score reflecting the strength of the agreement, ranging from a theoretical minimum (less than 0) to a maximum of 1.0.

Interpreting the Kappa Statistic: A Scale of Agreement

The resulting value for Cohen's Kappa always falls within a standardized range, typically between 0 and 1. A Kappa of 0 signifies that the observed agreement is exactly equal to the agreement expected by random chance--meaning the raters are no better than random guessers. Conversely, a Kappa of 1 indicates **perfect agreement** between the two raters. Negative values, while rare, suggest systematic disagreement, where the raters disagree more often than predicted by chance.

To make the Kappa value useful for actionable insights, it must be interpreted using established guidelines. While different statistical fields may apply slightly varying standards, the following

widely accepted scale, often attributed to Landis and Koch (1977), provides a clear framework for interpreting the strength of inter-rater reliability based on the calculated Kappa value:

The following table summarizes how to interpret different values for Cohen's Kappa:

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

Researchers typically strive for a Kappa value of 0.61 or higher, as this indicates **Substantial** or **Almost Perfect** reliability and suggests that the raters are using the classification criteria effectively and consistently. Values below 0.40 often indicate poor to fair agreement, prompting researchers to critically revisit the training protocols for raters or redefine the categorical definitions, as the level of reliability may be too low to trust the data confidently.

Setting Up Your Data for Kappa Calculation in Excel

To calculate Cohen's Kappa manually in Excel, the essential initial step is to organize the raw rating data into a structured **contingency table**, typically a 2x2 matrix for binary classifications. This contingency table must be constructed because the marginal totals derived from this arrangement are necessary inputs for accurately calculating the expected chance agreement (Pe).

The table layout should clearly delineate the classifications provided by Rater 1 (rows) and Rater 2 (columns). The internal cells of the matrix will contain the count of observations falling into each combination of agreement and disagreement. For a 2x2 table:

Cell (1,1): Count where Rater 1 says Category A and Rater 2 says Category A (Agreement).
Cell (2,2): Count where Rater 1 says Category B and Rater 2 says Category B (Agreement).
Off-Diagonal Cells (1,2) and (2,1): Counts representing disagreement, where classifications do not match.

Once the counts are entered into the matrix, the next necessary step is to calculate the marginal totals: the sum of each row, the sum of each column, and the grand total (N). These summary statistics form the critical basis for deriving the key probabilities Po and Pe within the Excel environment, ensuring the chance correction is properly applied.

Detailed Example: Assessing Curator Reliability

To illustrate the calculation process, let us examine a specific scenario involving professional judgment. Suppose two expert art museum curators are asked to independently evaluate a collection of 70 paintings. Their task is to classify each painting based on whether it is of sufficient quality to be included in a new gallery exhibit (a binary classification of "Yes" or "No"). This assessment is highly subjective, and the museum requires quantifying the level of inter-rater reliability to validate the objectivity of the selection process.

The results of their 70 independent ratings are compiled and summarized in the following 2x2 contingency table, which serves as the starting point for our calculation in Excel:

	A	B	C	D	E	F	G
1			**Rater 2**		Total		
2			**Yes**	**No**			
3	**Rater 1**	**Yes**	25	10	35		
4		**No**	15	20	35		
5	Total		40	30	70		
6							
7							
8							
9							
10							
11							
12							
13							
14							
15							

From this table, we can easily extract the raw counts and marginal totals required for subsequent steps:

Agreed "Yes" (Concordant): 25
Agreed "No" (Concordant): 20
Disagreed (Rater 1 Yes, 2 No): 10
Disagreed (Rater 1 No, 2 Yes): 15
Total Observations (N): 70
Rater 1 Marginals: (Yes=35, No=35)
Rater 2 Marginals: (Yes=40, No=30)

Step-by-Step Calculation of Cohen's Kappa Components in Excel

Achieving the final Kappa statistic requires methodical calculation of its two core components, Po and Pe, directly using the counts from the contingency table set up in Excel.

Calculating Observed Agreement (Po)

The **po** value quantifies the proportion of the total items on which the raters were in agreement. This sum includes all items where both raters concurred, regardless of the category ("Yes" or "No").

Identify the concordant pairs (diagonal cells): 25 + 20 = 45 agreed ratings.
Divide the sum of concordant pairs by the total number of observations (N=70).
Po = 45 / 70 ≈ 0.6429. (The observed agreement is approximately 64.29%).

Calculating Expected Chance Agreement (Pe)

The **pe** value estimates the theoretical probability that the raters would align purely by chance, based solely on their individual marginal distributions. This involves calculating the probability of random agreement for each category and summing them.

Probability of Chance Agreement on "Yes": (Rater 1 Yes Total / N) × (Rater 2 Yes Total / N) = (35/70) × (40/70) = 0.5 × 0.5714 ≈ 0.2857.
Probability of Chance Agreement on "No": (Rater 1 No Total / N) × (Rater 2 No Total / N) = (35/70) × (30/70) = 0.5 × 0.4286 ≈ 0.2143.
Sum these two probabilities to find the total Pe: Pe = 0.2857 + 0.2143.

The resulting probability of expected chance agreement (Pe) turns out to be exactly **0.5**.

Calculating the Kappa Statistic (k)

The final step is to substitute the calculated Po and Pe values into the Kappa formula. The **k** value

represents Cohen's Kappa, quantifying the chance-corrected agreement.

The following screenshot shows how these values are typically organized and calculated within Excel, illustrating the structure needed for accurate spreadsheet analysis:

	A	B	C	D	E	F
1			**Rater 2**		Total	
2			**Yes**	**No**		
3	**Rater 1**	**Yes**	25	10	35	
4		**No**	15	20	35	
5	Total		40	30	70	
6						
7						
8	p_0	0.6429	=(C3+D4)/SUM(C3:D4)			
9	p_e	0.5	=MMULT(C5:D5, E3:E4)/E5^2			
10	**k**	0.2857	=(B8-B9)/(1-B9)			
11						
12						
13						
14						
15						
16						

Using our derived values (Po = 0.6429 and Pe = 0.5):

k = (po - pe) / (1 - pe)
k = (0.6429 - 0.5) / (1 - 0.5)
k = 0.1429 / 0.5
k = **0.2857**

Conclusion and Practical Implications

The calculated Cohen's Kappa value for the two art curators turns out to be **0.2857**. This is the crucial finding: while the raw observed agreement was 64.29%, once the significant degree of agreement expected by chance (Pe = 50%) is factored in, the true, reliable consistency between the curators drops dramatically.

Based on the standard interpretation table provided earlier, a Kappa value of 0.2857 falls squarely

into the "Fair" category of agreement (typically 0.21-0.40). This finding suggests that while the curators agreed slightly more often than they would by random guessing, their consistency is highly insufficient for a critical task like selecting priceless museum pieces.

The practical implication of this result is that the museum should immediately review the criteria used for classifying paintings. The lack of **Substantial** or **Almost Perfect** agreement suggests that the guidelines are either vague, subjective, or the curators received inconsistent training regarding what constitutes "exhibit quality." Further corrective steps might involve intensive retraining, implementing stricter rubrics, or incorporating a third rater to help resolve future major discrepancies.

The following tutorials offer additional resources on Cohen's Kappa: