Understanding Cohen’s Kappa: A Measure of Inter-Rater Agreement

Authored by

Mohammed looti

November 5, 2025

The Cohen's Kappa Statistic ($kappa$) stands as a cornerstone metric in statistical analysis, particularly within fields like psychometrics and data quality assessment. It provides a robust method for quantifying the extent of **agreement between two raters** (or observers) when they classify a set of items into a fixed number of predefined, nominal categories. Unlike basic calculations of percentage agreement, Kappa is specifically engineered to neutralize the influence of agreement that occurs purely by random chance, thereby offering a far more dependable assessment of inter-rater reliability.

This statistic is indispensable whenever subjective judgment is involved, such as in clinical diagnostics, evaluating the quality of manufacturing outputs, or performing intricate content analysis. When researchers rely on human observers, the consistency of those observations must be rigorously verified. A high Kappa score signals that the observed consensus among judges is substantially higher than what would be expected if those judges were merely making classifications randomly, providing confidence in the reliability of the collected data.

The Mathematical Core of Cohen's Kappa

Cohen's Kappa is fundamentally a normalized metric that compares the actual performance of the raters against a theoretical baseline of random guessing. Represented by the symbol $kappa$, the calculation formalizes the difference between the proportion of observed agreement (p_{o}) and the probability of chance agreement (p_{e}). This normalization process is essential for ensuring the resulting coefficient accurately reflects true, non-random concordance.

The core formula that drives Cohen's Kappa is presented simply as:

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

This equation effectively measures how much the raters' agreement improved beyond the baseline level of chance agreement. The denominator, $(1 - p_{e})$, represents the theoretical maximum possible agreement that could be achieved above chance. Correspondingly, the numerator, $(p_{o} - p_{e})$, captures the actual agreement achieved beyond that random baseline. Understanding these components is critical for interpreting the resulting statistic.

The two key variables utilized in the calculation are precisely defined:

po: This is the **observed agreement**, calculated as the proportion of all items for which both raters assigned the exact same category. It is the simple percentage match.
pe: This represents the **hypothetical probability of chance agreement**. It is derived from the marginal totals (row and column sums) of the classification table, modeling the expected frequency of agreement if the classification decisions of the two raters were entirely independent of one another.

Interpreting the Kappa Coefficient ($kappa$)

The calculated value of Cohen's Kappa is constrained to fall between -1 and +1. A value of $kappa = 1$ indicates **perfect agreement**, signifying that the raters classified every single item identically. Conversely, a value of $kappa = 0$ implies that the observed agreement is precisely equivalent to the level of agreement expected purely due to random chance. In essence, the raters were no more reliable than flipping a coin.

While uncommon in practice, a negative Kappa value suggests that the agreement observed is systematically worse than random chance, pointing toward a fundamental disagreement or a profound bias in the application of the rating criteria by one or both judges. To standardize the evaluation of the statistic, researchers typically rely on established benchmarks to categorize the strength of agreement. These thresholds transform the raw number into a meaningful assessment of reliability.

The following widely accepted guidelines, often referenced from the work of Landis and Koch, provide a crucial framework for interpreting various Kappa values in terms of the strength of inter-rater consensus. For most rigorous scientific endeavors, a Kappa value exceeding 0.60 is generally considered indicative of **substantial reliability**, though the required threshold may vary depending on the specific application (e.g., clinical diagnosis often requires higher reliability).

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

Practical Calculation: A Curatorial Example

To demonstrate the mechanics of Cohen's Kappa, let us consider a practical scenario involving the evaluation of visual arts. Suppose two independent expert museum curators are tasked with reviewing 70 paintings. Their classification task is binary: categorize each artwork as "Yes" (suitable for a new exhibit) or "No" (not suitable). The outcomes of their 70 independent assessments are compiled into a 2x2 contingency table, which maps the frequencies of agreement

and disagreement.

The structured data below provides the raw counts based on the combined judgments of the two curators:

		Rater 2	
		Yes	**No**
Rater 1	**Yes**	25	10
	No	15	20

We must now navigate three mandatory computational steps to successfully derive the final Kappa coefficient, adjusting the initial observed agreement for the impact of chance.

Step 1: Calculate Relative Observed Agreement (p_{o})

The initial step requires calculating p_{o}, the simple proportion of cases where the two raters were in perfect concurrence. This value is found by summing the cell counts located on the main diagonal of the contingency table (where both said "Yes" or both said "No") and dividing this sum by the total number of observations (70 paintings).

po = (Agreed Yes + Agreed No) / (Total Ratings)
po = (25 + 20) / (70)
po = 45 / 70 = **0.6429**

Based on this calculation, the two curators exhibited an observed agreement rate of approximately 64.3%. While this seems moderate, we must proceed to the next step to determine how much of this agreement is spurious (due to chance).

Step 2: Calculate the Hypothetical Probability of Chance Agreement (p_{e})

The second step is arguably the most crucial for the integrity of Cohen's Kappa, as it isolates random agreement. We calculate p_{e} by using the marginal totals--the row and column sums--which represent the individual propensity of each rater to assign a specific category. The core assumption here is that the raters' classification habits are statistically independent.

First, we determine the marginal totals (frequency of use for each category by each rater):

Rater 1 "Yes" total: 25 + 10 = 35
Rater 1 "No" total: 15 + 20 = 35

Rater 2 "Yes" total: 25 + 15 = 40
Rater 2 "No" total: 10 + 20 = 30

The total probability of chance agreement (p_{e}) is the sum of the probability that both randomly say "Yes" and the probability that both randomly say "No."

P("Yes" Agreement by Chance) = (Rater 1 Yes Total / Total) × (Rater 2 Yes Total / Total)
P("Yes" Agreement by Chance) = (35/70) * (40/70) $approx$ 0.285714
P("No" Agreement by Chance) = (Rater 1 No Total / Total) × (Rater 2 No Total / Total)
P("No" Agreement by Chance) = (35/70) * (30/70) $approx$ 0.214285
pe = P("Yes" Chance) + P("No" Chance)
pe = 0.285714 + 0.214285 = **0.5000**

This result implies that 50% of the possible agreement could be attributed solely to the curators' underlying tendencies to choose "Yes" or "No," irrespective of the painting's quality.

Step 3: Calculate Cohen's Kappa ($kappa$)

The final step involves substituting the calculated values of p_{o} (0.6429) and p_{e} (0.5000) into the primary Cohen's Kappa formula to yield the chance-corrected agreement score.

k = (po - pe) / (1 - pe)
k = (0.6429 - 0.5000) / (1 - 0.5000)
k = 0.1429 / 0.5000
k = **0.2858**

The resulting Kappa coefficient is approximately 0.2858. When measured against the Landis and Koch interpretation scale, this value indicates a **Fair** level of agreement. Although the curators initially agreed on 64% of the paintings, once chance is factored out, the true reliability is relatively low. This suggests that the criteria used for classification might be ambiguous or that the curators require further training to achieve higher consistency in their subjective evaluations.

Strengths and Contextual Limitations of Kappa

Cohen's Kappa remains a highly respected statistical tool primarily because it provides a conservative and scientifically rigorous measure of reliability, distinguishing true consensus from random coincidence. Its strength lies in preventing researchers from inadvertently inflating the consistency of their categorical ratings by failing to account for chance.

However, users must be aware of two well-documented limitations that can affect the resulting $kappa$ score: the **prevalence problem** and the **bias problem**. The prevalence problem occurs when the categories being rated are highly imbalanced--for instance, if 95% of items fall into

Category A and only 5% into Category B. In these highly skewed scenarios, Kappa tends to artificially yield lower values, even when the observed agreement (p_{o}) is very high.

The bias problem arises when raters exhibit systematic differences in their marginal totals (i.e., one rater uses "Yes" significantly more often than the other). This systematic bias can also suppress the calculated Kappa value. When facing datasets with severe prevalence or bias issues, researchers often consider alternative metrics, such as Gwet's AC1 statistic, which has been demonstrated to be more robust to the prevalence issue, offering a potentially more stable measure of reliability in challenging classification contexts.

Leveraging Software for Reliability Analysis

While performing manual calculation is invaluable for grasping the underlying statistical mechanism of Cohen's Kappa, large-scale studies invariably rely on specialized statistical software packages (such as R, SPSS, or dedicated online tools) for automated analysis. Regardless of the tool utilized, a solid understanding of the mathematical foundation ensures that the outputted $kappa$ value is interpreted accurately, considering the context of the study's design, prevalence rates, and potential rater biases.

Researchers are encouraged to utilize automated tools for efficient computation, allowing them to focus on the proper application and interpretation of this vital measure of inter-rater reliability.

You can use this to automatically calculate Cohen's Kappa for any two raters.