Learn How to Calculate Cohen’s Kappa for Inter-Rater Reliability in Python

Authored by

Mohammed looti

October 29, 2025

In the realm of statistics and data science, accurately quantifying the level of agreement between independent observers or measurement systems is a fundamental analytical challenge. While a simple calculation of percentage agreement is often the intuitive starting point, this metric is inherently flawed because it fails to account for agreements that occur purely by random chance. This limitation is particularly problematic when dealing with heavily skewed or imbalanced data distributions. It is precisely in this context that the Cohen's Kappa statistic rises as an indispensable and widely adopted tool.

Developed specifically for measuring the consistency of categorical judgments, Kappa provides a robust assessment of inter-rater reliability. It evaluates the congruence between two raters or judges who classify items into a defined set of mutually exclusive categories. By factoring in the expected rate of chance agreement, Kappa delivers a more conservative and honest measure than crude agreement percentages.

This chance-corrected approach makes the statistic invaluable across diverse fields, including medical diagnosis, psychological assessment, and, increasingly, in the evaluation of machine learning models for classification tasks. For any data professional working with qualitative annotation or reliability testing, understanding the mathematical basis, interpretation, and practical implementation of Cohen's Kappa in languages like Python is absolutely essential for rigorous analysis.

Understanding the Necessity of Chance Correction

When multiple experts, often referred to as raters or coders, are tasked with classifying a corpus of data--be it patient symptoms, textual sentiment, or image content--the first question is always about the consistency of their judgments. For example, two clinicians diagnosing patients or two researchers coding open-ended survey responses require a consistent baseline. If Rater A classifies 80% of items as 'Positive' and Rater B also classifies 80% as 'Positive', they will naturally agree on a large portion of the data simply due to the high frequency of that single category, irrespective of their true ability to discriminate between categories.

Simple percentage agreement completely masks this underlying issue of base rates and category prevalence. A high percentage agreement might falsely suggest excellent inter-rater reliability when, in reality, the agreement is inflated by the dominance of one category. This misleading metric can lead to overconfidence in the reliability of the data collection process or the calibration of the observers.

Cohen's Kappa, introduced by Jacob Cohen in 1960, specifically addresses this critical limitation. By calculating and subtracting the proportion of agreements expected by chance agreement, the statistic provides a measure of agreement that is significantly more robust. It quantifies how much better the raters agree than if they were assigning categories randomly, based only on the

marginal probabilities of their individual classification tendencies. This makes it a superior metric for evaluating the consistency of human judgments or the performance stability of statistical classification models.

The Mathematical Formulation of Cohen's Kappa

The mathematical foundation of Cohen's Kappa is designed to normalize the observed level of agreement against the maximum possible agreement achievable beyond chance. This normalization process yields a standardized coefficient that is comparable across different datasets, provided they involve two raters and categorical data.

The canonical formula for the Kappa coefficient (k) is expressed as follows:

k = (po – pe) / (1 – pe)

To fully appreciate the utility of this metric, it is necessary to examine the definition and calculation of its primary components:

po (Observed Agreement): This term represents the **relative observed agreement**. It is the simple proportion of items upon which the two raters are in perfect concordance. If there are **N** total items rated, and **A** items where agreement occurs, then **po = A / N**. It is essential to recognize that **po** is synonymous with the simple percentage agreement and serves as the baseline measure before chance correction.
pe (Expected Chance Agreement): This denotes the **hypothetical probability of chance agreement**. It is the proportion of times we would expect the raters to agree purely by random chance, calculated using the marginal probabilities (the totals for each category assigned by each rater individually). Calculating **pe** involves constructing a hypothetical agreement matrix where the probability of agreement on a specific category is the product of Rater 1's marginal probability for that category and Rater 2's marginal probability for that same category. Summing these products across all categories yields the total expected chance agreement.

The crucial element is the numerator, **(po – pe)**, which isolates the portion of agreement that is truly attributable to the reliability of the raters' judgments, subtracting the randomness. The denominator, **(1 – pe)**, represents the maximum agreement possible beyond chance. By dividing the reliable agreement by the maximum potential reliable agreement, Kappa standardizes the result, making the coefficient robust and interpretable.

Interpreting the Cohen's Kappa Coefficient

The resulting Kappa coefficient (k) is a scalar value that typically ranges between 0 and 1. A value of 0 signifies that the observed agreement is no better than what would be expected by chance

alone, indicating a complete lack of reliability. Conversely, a value of 1 represents perfect agreement between the two raters. Although uncommon in practice, a negative Kappa value is theoretically possible if the observed agreement is systematically worse than chance, suggesting a fundamental and consistent disagreement in classification.

While the absolute context of the study (e.g., medical diagnosis versus casual survey coding) must always be considered, researchers often rely on established benchmarks to interpret the strength of the reliability. One of the most frequently cited sets of guidelines for interpreting Cohen's Kappa values was provided by Landis and Koch in 1977.

These guidelines offer a standardized framework for qualitative assessment:

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

Specifically, Kappa values exceeding 0.80 are generally classified as "almost perfect" agreement, demonstrating high confidence in the consistency of the judgments. Values falling between 0.61 and 0.80 suggest "substantial" agreement. Agreement levels below 0.40, such as those between 0.21 and 0.40 ("fair"), indicate that a significant portion of the concordance is still attributable to chance. Critically, any value below 0.20 is typically considered "slight" or "poor," necessitating a re-evaluation of the classification criteria or the training provided to the observers.

Implementing Cohen's Kappa Calculation in Python

Calculating Cohen's Kappa for large or even moderately sized datasets is impractical to perform manually. Fortunately, the Python ecosystem provides highly optimized libraries for this purpose. The industry standard for machine learning and statistical evaluation, scikit-learn (often imported as **sklearn**), offers a specialized, single-function solution that handles the entire computation efficiently.

Let's consider a practical example: Two seasoned art curators are independently assessing a collection of 15 paintings for inclusion in a prestigious exhibition. They must categorize each piece

as '1' (suitable) or '0' (unsuitable). Our objective is to rigorously measure the true consistency of their expert judgments using the chance-corrected Kappa coefficient.

The following Python script imports the necessary function from **sklearn.metrics** and calculates the Kappa score based on the sequential categorical ratings provided by the two curators.

from sklearn.metrics import cohen_kappa_score

#define array of ratings for both raters
rater1 =
rater2 =

#calculate Cohen's Kappa
cohen_kappa_score(rater1, rater2)

0.33628318584070793

The core of this implementation lies in the cohen_kappa_score() function, which accepts the two rating arrays (rater1 and rater2, representing the independent classification outputs) and automatically performs the calculation of po, pe, and the final Kappa value without requiring manual construction of the contingency table.

Analyzing the Results and Alternative Tools

The execution of the Python script yields a calculated Kappa value of approximately **0.33628**. This number is a quantitative representation of the agreement level, corrected for the likelihood that the curators would have agreed simply by chance given their individual rating biases.

Referring back to the Landis and Koch guidelines, a Kappa value of 0.33628 falls squarely into the "fair" range of agreement (0.21 to 0.40). This result suggests that while there is some consistency in the curators' judgments, their criteria are not strongly aligned, and a non-trivial amount of their observed agreement is likely coincidental. Given the high stakes of curating a prestigious exhibit, this finding would typically necessitate an intervention, such as reviewing and standardizing the rating criteria or introducing a third, tie-breaking rater.

While scikit-learn is preferred for its integration into machine learning pipelines, other robust statistical libraries in Python also provide agreement metrics. Notably, the statsmodels library offers comprehensive tools for statistical modeling and hypothesis testing, which can include calculating agreement coefficients, depending on the specific version and functionality required. The choice of library ultimately depends on the complexity of the required analysis and the established workflow of the project.

Limitations and Advanced Considerations

Despite its widespread utility, Cohen's Kappa is not without its critics and limitations. The most frequently discussed drawback is the "Kappa paradox," which demonstrates that high observed agreement (high po) can sometimes result in a surprisingly low Kappa score. This usually occurs when the marginal distributions are extremely unbalanced, a situation known as the **prevalence effect**. If one category (e.g., 'Negative') is overwhelmingly common, the chance agreement (pe) becomes very high, which deflates the numerator **(po – pe)**, leading to a lower Kappa than might intuitively be expected based on the raw agreement percentage.

Another related issue is the **bias effect**. This arises when the two raters exhibit systematic differences in their propensity to assign items to specific categories. For instance, if Rater A consistently classifies items more conservatively than Rater B, this inherent bias reduces the overall agreement that can be achieved, thus lowering the Kappa score, even if their relative judgment within categories is consistent. These effects underscore the necessity of examining the underlying confusion matrix and marginal totals alongside the Kappa value.

Researchers must also consider the context of their analysis. When the study involves more than two observers, **Fleiss' Kappa** is the appropriate generalization of the statistic. Fleiss' Kappa extends the principle of chance-corrected agreement to situations involving three or more raters, assuming that each subject is rated by a potentially different set of raters, thus broadening the applicability of this essential reliability metric.

Summary and Resources for Further Learning

Cohen's Kappa remains a cornerstone in reliability testing for categorical data, offering a crucial correction against agreement by chance. By leveraging powerful Python libraries like scikit-learn, researchers and analysts can seamlessly integrate this robust metric into their data analysis workflows, ensuring confidence in the consistency of their classification data.

For those seeking deeper technical implementation details, the complete documentation for the cohen_kappa_score() function is available on the scikit-learn official website.

To deepen your understanding of Cohen's Kappa, inter-rater reliability, and related statistical topics, the following resources offer further insights: