

Does anyone know how can I calculate Cohen's Kappa with R?

Authored by
stats writer

December 11, 2025

RECOMMENDED CITATION

stats writer (2025). *Does anyone know how can I calculate Cohen's Kappa with R?*.
PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=107150>

The quest for reliable measurement is central to both scientific research and practical data analysis. When evaluating qualitative observations or classifications, it is essential to quantify the consistency achieved by independent observers or measurement tools. This necessity leads us directly to Cohen's Kappa (κ), a powerful statistic designed to measure the degree of agreement between two raters classifying items into categorical data sets.

Unlike simple percentage agreement, which fails to account for the possibility that raters might agree purely by chance, Kappa provides a more robust and conservative estimate of true concordance. Understanding how to calculate and interpret this metric is crucial for any data scientist, especially those utilizing the versatile statistical environment provided by R. This guide will provide a comprehensive, step-by-step methodology for executing this calculation using specialized packages within the R programming language, ensuring both accuracy and efficiency in your statistical workflow.

Understanding Cohen's Kappa: Definition and Importance

Cohen's Kappa is formally defined as a measure of inter-rater reliability for two raters working with mutually exclusive nominal categories. Its significance lies in its ability to adjust for chance agreement, distinguishing genuine consensus from random chance matches. This distinction makes Kappa an indispensable tool in fields such as medical diagnosis, content analysis, psychological assessment, and machine learning model evaluation, where subjective or qualitative judgments are often involved.

When two individuals assess the same set of subjects and assign them to specific categories (e.g., 'Pass' or 'Fail', 'Positive' or 'Negative', 'Mild', 'Moderate', or 'Severe'), the total percentage of identical classifications is the observed agreement. However, if the raters were simply guessing randomly, they would still agree some proportion of the time. Kappa isolates the agreement achieved beyond this baseline expectation. A high Kappa value indicates strong agreement, suggesting that the classification system is reliable and the raters are applying the criteria consistently. Conversely, a low or negative Kappa value signals poor reliability, requiring re-evaluation of either the classification rubric or the raters' training.

It is important to emphasize that while Kappa is generally applied to situations involving two raters, generalized versions exist for more than two observers. However, the foundational Cohen's Kappa statistic remains the benchmark for pairwise reliability assessments. Furthermore, users must ensure their data is truly categorical and nominal before applying this test; ordinal data might benefit more from techniques like Weighted Kappa, which accounts for the magnitude of disagreement.

The Mathematics Behind Kappa: Observed vs. Expected Agreement

The core formulation of Cohen's Kappa is designed to compare the observed agreement (p_o) against the hypothetical agreement expected by chance (p_e). The formula encapsulates the ratio of observed non-chance agreement to maximum possible non-chance agreement.

$$k = \frac{p_o - p_e}{1 - p_e}$$

Here, p_o represents the proportion of observed units where the raters agree. This is calculated simply by summing the diagonal entries of the agreement matrix (the counts where Rater 1 and Rater 2 assigned the same category) and dividing by the total number of items assessed. Calculating p_e , the expected chance agreement, is more intricate. It requires calculating the probability that each rater would assign an item to a specific category based on the marginal totals of the contingency table, and then summing the probability products for each cell where agreement occurs.

The denominator, $(1 - p_e)$, represents the maximum possible agreement beyond chance. If p_o perfectly matches p_e , the numerator becomes zero, yielding a Kappa of zero, meaning agreement is strictly by chance. If p_o reaches perfect agreement (1), the numerator and denominator match, resulting in a Kappa of 1, indicating perfect reliability. Understanding these components is vital not only for manual calculation but also for interpreting the output generated by statistical software like R.

The R Environment: Setting Up for Statistical Analysis

R is the preferred environment for countless statisticians and data analysts due to its open-source nature, vast library of packages, and high capacity for complex statistical modeling. To calculate Cohen's Kappa in R, the foundational environment must first be established. This involves ensuring you have a current installation of R and, ideally, an integrated development environment (IDE) like RStudio, which significantly simplifies package management and script execution.

Before proceeding with specialized reliability calculations, data must be correctly formatted within R. Reliability statistics like Kappa typically require input data to be structured as a contingency table or a matrix where rows represent the items assessed and columns represent the classifications made by the two raters. These classifications must be coded consistently, usually numerically (e.g., 1, 2, 3) or as factors reflecting the nominal categories.

While base R provides extensive statistical functionality, complex metrics often rely on community-contributed packages, which are housed on the Comprehensive R Archive Network (CRAN). For inter-rater reliability measures, the primary resource is the highly respected irr package.

Installing and Utilizing the irr Package

The `irr` package is specifically designed for calculating various forms of inter-rater reliability, including Kappa, Fleiss' Kappa, and Krippendorff's Alpha. Its reliability and ease of use make it the standard choice for researchers working with agreement metrics in R. Installation is straightforward and must be performed once per R environment setup.

To install the package from CRAN, execute the following command in your R console:

```
install.packages("irr")
```

Once installation is complete, the package must be loaded into the current R session using the `library()` command before any of its functions can be called:

```
library(irr)
```

The `irr` package offers several functions, but for standard pairwise Cohen's Kappa, the primary function of interest is `kappa2()`. This function is specifically tailored for assessing agreement between two raters, making it perfectly suited for the task at hand.

Implementing kappa2(): Syntax and Data Preparation

The `kappa2()` function requires the input data to be presented as a matrix or data frame containing the ratings from the two observers. Each row must correspond to a single subject or item being rated, and the two columns must hold the categorical ratings assigned by Rater 1 and Rater 2, respectively. Proper data preparation is paramount for accurate calculation.

Consider a simple example where 10 subjects are rated by two psychologists (P1 and P2) as either 'A' or 'B'. First, we would create the data structure in R:

```
# Create sample data (1 = A, 2 = B)
ratings_P1 <- c(1, 2, 1, 1, 2, 1, 2, 2, 1, 2)
ratings_P2 <- c(1, 2, 2, 1, 2, 1, 1, 2, 1, 2)
data_matrix <- data.frame(Rater1 = ratings_P1, Rater2 = ratings_P2)

# View the data
print(data_matrix)
```

Next, we invoke the `kappa2()` function, passing the data frame as the primary argument. The function returns a comprehensive object containing the Kappa coefficient, the asymptotic standard

error, and the associated p-value for testing the null hypothesis that Kappa is zero:

```
kappa_result <- kappa2(data_matrix)
print(kappa_result)
```

The `kappa2()` function handles the calculation of observed agreement, expected chance agreement, and the final Cohen's Kappa statistic automatically, thereby simplifying the statistical process significantly for the user. Ensure that the input data contains only the ratings themselves and no extraneous identifier columns.

Interpreting the Kappa Statistic: Guidelines and Context

Once the Kappa statistic is calculated via R, its value must be interpreted within the context of established guidelines to determine the strength of the inter-rater reliability. While interpretations can vary slightly across disciplines, the following widely accepted benchmarks (often attributed to Landis and Koch) provide a useful framework for assessment:

Poor Agreement ($\text{Kappa} < 0.00$): Agreement is less than that expected by chance (a rare occurrence, often suggesting systematic disagreement or error).

Slight Agreement ($0.00 \leq \text{Kappa} \leq 0.20$): Minimal reliability exists.

Fair Agreement ($0.21 \leq \text{Kappa} \leq 0.40$): Moderate but often questionable reliability.

Moderate Agreement ($0.41 \leq \text{Kappa} \leq 0.60$): Acceptable reliability for many applied settings.

Substantial Agreement ($0.61 \leq \text{Kappa} \leq 0.80$): High degree of reliability.

Almost Perfect Agreement ($0.81 \leq \text{Kappa} \leq 1.00$): Near-perfect or perfect reliability.

It is crucial to consider the context of the study when interpreting Kappa. In fields where highly subjective judgments are made (e.g., grading essays), a moderate Kappa might be considered excellent. Conversely, in highly objective technical fields (e.g., instrument calibration checks), only substantial or almost perfect agreement may be deemed acceptable. Furthermore, the significance test provided by `kappa2()` tells us if the observed Kappa is statistically greater than zero, but it does not tell us whether the magnitude of agreement is practically useful.

Advanced Considerations: Weighted Kappa and Limitations

While standard Cohen's Kappa is highly effective for nominal categorical data, it has a significant limitation: it treats all disagreements equally. If the categories are ordinal (ordered, such as 'Low', 'Medium', 'High'), disagreeing on 'Low' vs. 'High' is a much more severe discrepancy than

disagreeing on 'Low' vs. 'Medium'. In such cases, the standard Kappa can underestimate true reliability.

For ordinal data, researchers should utilize **Weighted Kappa**. This variant incorporates a weighting scheme that penalizes disagreements based on their distance from the true agreement diagonal. The `kappa2()` function in the `irr` package can also calculate Weighted Kappa by specifying the `weight` argument (e.g., `weight="squared"` for quadratic weights or `weight="linear"` for linear weights). This flexibility allows the R user to apply the most appropriate reliability measure based on the scale of the data.

A final limitation of Kappa is the so-called "Kappa paradox," where high marginal totals (imbalanced category distributions) can lead to a lower Kappa value even when the percentage agreement is high. This highlights the importance of reporting both the percentage agreement and the Kappa statistic, allowing readers to fully appreciate the context of the reliability assessment.

Practical Example: Calculating Kappa Step-by-Step

To solidify the understanding of Cohen's Kappa calculation, let us walk through a typical scenario. Imagine two expert dermatologists reviewing 100 biopsies and classifying them into two categories: 'Malignant' (Yes) or 'Benign' (No). We can summarize their findings in a 2x2 contingency table, which is the structure required for the underlying calculation.

The data below represents the counts for each joint classification. We can use this data to manually understand the components of Kappa, followed by using the interactive tool provided below for quick verification. This particular calculator focuses on the binomial case (Yes/No classifications) and illustrates how the raw counts translate into the final Kappa coefficient.

```
@import url('https://fonts.googleapis.com/css?family=Droid+Serif|Raleway');
```

```
.axis--y .domain {
display: none;
}
```

```
h1 {
color: black;
text-align: center;
margin-top: 15px;
margin-bottom: 0px;
font-family: 'Raleway', sans-serif;
}
```

```
h2 {
```

```
color: black;
font-size: 20px;
text-align: center;
margin-bottom: 15px;
margin-top: 15px;
font-family: 'Raleway', sans-serif;
}
```

```
p {
color: black;
text-align: center;
margin-bottom: 15px;
margin-top: 15px;
font-family: 'Raleway', sans-serif;
}
```

```
#words_intro {
color: black;
font-family: Raleway;
max-width: 550px;
margin: 25px auto;
line-height: 1.75;
}
```

```
#words_intro_center {
text-align: center;
color: black;
font-family: Raleway;
max-width: 550px;
margin: 25px auto;
line-height: 1.75;
}
```

```
#words_outro {
color: black;
font-family: Raleway;
max-width: 550px;
margin: 25px auto;
line-height: 1.75;
}
```

```
#words {  
color: black;  
font-family: Raleway;  
max-width: 550px;  
margin: 25px auto;  
line-height: 1.75;  
padding-left: 100px;  
}  
  
#calcTitle {  
text-align: center;  
font-size: 20px;  
margin-bottom: 0px;  
font-family: 'Raleway', serif;  
}  
  
#hr_top {  
width: 30%;  
margin-bottom: 0px;  
margin-top: 10px;  
border: none;  
height: 2px;  
color: black;  
background-color: black;  
}  
  
#hr_bottom {  
width: 30%;  
margin-top: 15px;  
border: none;  
height: 2px;  
color: black;  
background-color: black;  
}  
  
.input_label_calc {  
display: inline-block;  
vertical-align: baseline;  
width: 350px;  
}
```

```

#button_calc {
border: 1px solid;
border-radius: 10px;
margin-top: 20px;
padding: 10px 10px;
cursor: pointer;
outline: none;
background-color: white;
color: black;
font-family: 'Work Sans', sans-serif;
border: 1px solid grey;
/* Green */
}

#button_calc:hover {
background-color: #f6f6f6;
border: 1px solid black;
}

.label_radio {
text-align: center;
}

```

The **Cohen's Kappa** statistic measures the level of agreement between two raters or judges who each classify items into mutually exclusive categories, adjusted for chance. The definitive formula for calculating Cohen's kappa is given below:

$$k = (p_o - p_e) / (1 - p_e)$$

where the terms are defined as:

p_o : Represents the **Relative Observed Agreement** among the raters (the proportion of items where both raters agreed).

p_e : Represents the **Hypothetical Probability of Chance Agreement** (the proportion of agreement expected if raters were randomly assigning categories based on marginal totals).

To find Cohen's kappa between two raters for a 2x2 matrix, simply fill in the counts for each of the four possible outcomes below and then click the "Calculate" button.

Both raters said 'Yes' (Agreement A)

Both raters said 'No' (Agreement D)

Only the first rater said 'Yes' (Disagreement B)

Only the second rater said 'Yes' (Disagreement C)

Calculate Cohen's Kappa

Cohen's Kappa Result: 0.2857

```
function calc() {
//get input values
var bothyes = document.getElementById('bothyes').value*1;
var bothno = document.getElementById('bothno').value*1;
var yes1 = document.getElementById('yes1').value*1;
var yes2 = document.getElementById('yes2').value*1;

//calculate stuff
var n = bothyes-(-1*bothno)-(-1*yes1)-(-1*yes2)
var po = (bothyes-(-1*bothno))/n;
var pe_1 = ((bothyes-(-1*yes1))/n) * ((bothyes-(-1*yes2))/n);
var pe_2 = ((n-yes1-bothyes)/n)*((n-yes2-bothyes)/n);
var pe = pe_1 - (-1*pe_2);
var k = (po-pe)/(1-pe);

//output
document.getElementById('k').innerHTML = k.toFixed(4);
}
```

The resulting Kappa value shown above, calculated from the example inputs, falls into the "Fair Agreement" category based on the standard interpretation scale. This suggests that while there is agreement beyond chance, the reliability of the raters' classifications is still moderate and potentially needs improvement. By using the [irr package](#) in R, this entire process is streamlined, allowing researchers to quickly quantify the reliability of their [categorical data](#) assessments.