Learn How to Calculate Intraclass Correlation Coefficient (ICC) in Python

Authored by

Mohammed looti

November 5, 2025

The Intraclass Correlation Coefficient (ICC) stands as a paramount statistical tool used extensively in reliability studies. Its fundamental purpose is to quantify the consistency and degree of agreement among two or more quantitative measurements that have been taken on the same subjects or items, often by different observers or raters. Crucially, the ICC moves beyond simple correlation coefficients by explicitly accounting for the inherent structure of the data, recognizing how these measurements are naturally grouped--whether by rater, subject, or item.

The application of the ICC is indispensable across diverse disciplines, including psychology, clinical medicine, and various fields of engineering, where establishing the high reliability of measurement instruments, diagnostic tests, or human judgments is absolutely critical. A high ICC value provides compelling evidence that ratings are largely interchangeable and consistent across different assessors, thereby confirming robust inter-rater reliability. Conversely, a significantly low ICC indicates substantial variability between raters, strongly suggesting that the underlying measurement or assessment process itself may be prone to error or bias.

Numerically, the value of the ICC is constrained to fall within the range of 0 to 1. A score nearing 0 implies negligible agreement among assessors; in this scenario, the majority of the observed variance in ratings is attributed to differences among the raters rather than genuine differences among the items being rated. Conversely, an ICC approaching 1 signifies near-perfect reliability, indicating maximum consistency in the ratings. For practical interpretation, researchers often rely on established benchmarks, such as those suggesting values exceeding 0.75 represent excellent agreement, as popularized by guidelines from Cicchetti or Koo and Li.

To facilitate the efficient and accurate calculation of the ICC within the Python ecosystem, we turn to the specialized statistical package known as Pingouin. This powerful, open-source library provides the concise and highly robust function, `intraclass_corr`, specifically engineered for this task. Utilizing this function vastly simplifies the process, circumventing the complex, manual calculations typically required when relying on Analysis of Variance (ANOVA) components.

The standard syntax for executing the ICC calculation using the Pingouin library is carefully structured to map the components of the measurement study directly to the function's arguments. Understanding this structure is key to successful implementation:

pingouin.intraclass_corr(data, targets, raters, ratings)

This function mandates four essential arguments, which must clearly define the organization of the data housed within the supplied Pandas DataFrame:

data: Specifies the name of the Pandas DataFrame object containing the complete dataset for analysis.
targets: Designates the column name that identifies the "targets"--the primary units being

measured, such as subjects, items, patients, or exams.
raters: Identifies the column name containing the unique identities of the observers, judges, or assessors who performed the measurements.
ratings: Specifies the column name holding the actual quantitative scores or measurements assigned by the raters.

The following comprehensive tutorial provides a practical, step-by-step example demonstrating how to implement this function effectively, from environment setup to final interpretation of the results.

Prerequisites: Installing the Pingouin Library

Before any statistical computation can commence, the specialized `intraclass_corr` function requires that the Pingouin library be correctly installed within your Python environment. Pingouin is recognized as a powerful, dedicated open-source package for statistical analysis, optimized for performance and built upon the foundational numerical capabilities of NumPy and the data handling structures of Pandas.

Installation is highly streamlined, leveraging Python's standard package installer, `pip`. For best results and to ensure all necessary dependencies are properly configured, it is strongly recommended that this command be executed directly in your system terminal or within a relevant cell of a Jupyter Notebook environment.

pip install pingouin

Upon successful completion of the installation process, the essential libraries required for subsequent data manipulation and the core statistical computation can be securely imported, allowing us to proceed with the analysis.

Preparing the Dataset for ICC Analysis

To effectively demonstrate the mechanics of the ICC calculation, we will establish a practical, hypothetical scenario. Consider a research study where six distinct college entrance exams were evaluated and scored by four different judges using a predefined quantitative scale. For ICC computation using Pingouin, this type of measurement structure necessitates that the data be organized in a "long" format. This means that every single row must represent one unique rating instance--for example, Judge A providing a score for Exam 1.

We will harness the capabilities of the Pandas library to construct a DataFrame that precisely captures these measurements. This DataFrame is required to contain the three central columns essential for the `intraclass_corr` function: the items being rated (targets), the identities of the

assessors (raters), and the scores provided (ratings).

import pandas as pd

#create DataFrame structured for ICC analysis
df = pd.DataFrame({'exam': ,
'judge': ,
'rating': })

#view first five rows of DataFrame to confirm structure
df.head()

exam judge rating
0 1 A 1
1 2 A 1
2 3 A 3
3 4 A 6
4 5 A 6

In the resulting DataFrame structure, the `exam` column correctly fulfills the role of the **targets** (the items being assessed), the `judge` column identifies the **raters**, and the `rating` column stores the numerical scores. This organization ensures the data adheres precisely to the input requirements of the Pingouin function.

Calculating the Intraclass Correlation Coefficient

With the required data structure confirmed and the Pingouin library successfully imported, the ICC calculation is ready for execution. This step involves calling the `intraclass_corr` function, passing our DataFrame (`df`) and mapping the corresponding column names to the function's arguments: `targets='exam'`, `raters='judge'`, and `ratings='rating'`.

The execution of this single command yields a comprehensive output table detailing multiple types of ICCs. This multiplicity is essential because the term "ICC" represents a family of related statistics, each derived from differing statistical assumptions regarding the underlying model (e.g., fixed vs. random effects) and the intended generalization of the results (e.g., reliability of a single rater versus the average of all raters).

import pingouin as pg

icc = pg.intraclass_corr(data=df, targets='exam', raters='judge', ratings='rating')

```
icc.set_index('Type')

Description ICC F df1 df2 pval CI95%
Type
ICC1 Single raters absolute 0.505252 5.084916 5 18 0.004430
ICC2 Single random raters 0.503054 4.909385 5 15 0.007352
ICC3 Single fixed raters 0.494272 4.909385 5 15 0.007352
ICC1k Average raters absolute 0.803340 5.084916 5 18 0.004430
ICC2k Average random raters 0.801947 4.909385 5 15 0.007352
ICC3k Average fixed raters 0.796309 4.909385 5 15 0.007352
```

The resulting DataFrame output provides a comprehensive statistical summary. As shown, Pingouin automatically calculates six distinct ICC values, each corresponding to specific assumptions. This detail allows for robust statistical hypothesis testing concerning the calculated reliability of the measurements.

Drawing accurate conclusions about inter-rater agreement necessitates a clear understanding of the key statistical columns in this output table:

Description: Provides a textual label identifying the specific ICC model calculated (e.g., Two-Way Mixed Effects, Consistency).
ICC: The primary measure of reliability, represented by the Intraclass Correlation Coefficient value (ranging from 0 to 1).
F: The F-value derived from the underlying Analysis of Variance (ANOVA) model used in the calculation.
df1, df2: These represent the degrees of freedom associated with the numerator (df1) and denominator (df2) of the F-statistic, respectively.
pval: The p-value associated with the F-value. This value is used to test the null hypothesis that the true ICC in the population is zero. A small p-value (conventionally less than 0.05) suggests the observed reliability is statistically significant.
CI95%: The 95% confidence interval for the ICC. This interval provides an estimate of the range within which the true population ICC is likely to fall, offering crucial context regarding the precision of the calculated point estimate.

Interpreting the ICC Results and Model Selection

The presence of six distinct ICC values in the output table highlights the complexity of reliability analysis. The interpretation of these results critically depends on selecting the appropriate ICC type, which must align precisely with the design of the study and the intended scope of generalization. Selecting an inappropriate model constitutes a common error and can lead to

fundamentally erroneous conclusions regarding measurement reliability.

The six available ICC types are fundamentally differentiated based on three core assumptions that govern the statistical calculation: the underlying statistical model, the definition used for "agreement," and the chosen unit of analysis for the reliability estimate.

1. The Statistical Model Assumption: This determines how the factors (targets and raters) are treated in the calculation:

One-Way Random Effects (ICC1 and ICC1k): This model assumes that the targets (items) were randomly sampled from a larger population, but it only accounts for variance related to the targets, treating differences between raters as measurement error. This model is generally suitable when different raters assess unique subsets of targets.
Two-Way Random Effects (ICC2 and ICC2k): This is the most generalizable model, assuming that both the targets and the raters were randomly selected from larger populations. It allows researchers to generalize the calculated reliability to other raters who were not specifically included in the study.
Two-Way Mixed Effects (ICC3 and ICC3k): This model assumes targets are random, but the specific group of raters used in the study is fixed and exhaustive--meaning the researcher is only interested in the reliability achieved among this particular group of raters.

2. Defining Agreement: This defines the standard by which ratings are considered reliable:

Absolute Agreement: This is the strictest standard, requiring that the absolute numerical values of the scores be identical or nearly identical. This definition is mandatory when the actual magnitude of the measurement is critical (e.g., medical dosage).
Consistency: This standard is less rigid, only requiring that the ratings maintain proportionality or follow the same relative pattern across the targets. This allows for systematic differences or biases between raters (e.g., one rater consistently scoring higher than others).

3. The Unit of Analysis: Specifies the aggregation level for the final reliability estimate:

Single Rater (ICC1, ICC2, ICC3): Estimates the reliability of a typical, individual rater.
Mean of Raters (ICC1k, ICC2k, ICC3k): Estimates the reliability achieved if the final, definitive measurement is derived by averaging all raters' scores. Due to the statistical benefit of aggregating independent observations, these 'k' values are always mathematically higher than their single-rater counterparts.

Returning to our example where four judges evaluated exams, if our research objective is to determine the reliability of the combined, averaged score provided by this specific set of judges (Fixed Raters), the appropriate statistic is the **ICC3k (Average fixed raters)**. With a calculated value of 0.796309, this result suggests an agreement level that is typically classified as good to

excellent when utilizing the mean score across all four judges. This interpretation underscores the importance of correctly mapping the study design to the statistical output.

For researchers seeking a deeper, more rigorous understanding of how these various model assumptions critically influence the calculation, precision, and application of the Intraclass Correlation Coefficient, it is highly recommended to consult authoritative statistical literature. Foundational texts, such as the works of Shrout and Fleiss (1979) or Koo and Li (2016), remain the primary resources for reliability analysis.