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

Authored by

stats writer

December 6, 2025

The Intraclass Correlation Coefficient (ICC) is a powerful statistical measure primarily utilized to quantify the degree of agreement or reliability among multiple raters, judges, or measurement methods when assessing the same set of subjects or items. Unlike simpler agreement metrics, the ICC considers both the consistency of the ratings and the absolute agreement, making it an indispensable tool in fields such such as psychology, clinical research, and quality assurance studies where consistency across observers is paramount.

The numerical result of an ICC calculation spans from 0 to 1, where a value close to 0 signifies minimal or non-existent reliability among the raters, suggesting that the differences between ratings are highly randomized or driven by noise. Conversely, an ICC value approaching 1 denotes near-perfect agreement or highly reliable measurements, indicating that raters are consistently scoring items in a similar manner. Achieving reliable measurements is critical for ensuring the validity of research findings and the trustworthiness of diagnostic procedures.

For data scientists and researchers leveraging Python, the most straightforward and statistically robust method to calculate the ICC is by using the specialized function available within the high-performance statistical package, Pingouin. This library is designed to offer a clean interface for common statistical tests. The primary function for this analysis is structured using the following specific syntax, which maps the data structure directly to the statistical model requirements:

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

This structure requires four key arguments to perform the calculation:

data: This argument specifies the name of the Pandas DataFrame containing the observational data.
targets: This must be the column name that identifies the "targets" or subjects being rated--the entity whose score is being assessed for consistency.
raters: This argument specifies the column name containing the unique identifiers for each rater or judge who provided the scores.
ratings: This is the crucial column containing the actual numerical scores or ratings assigned by the raters to the targets.

The subsequent sections of this comprehensive tutorial will guide you through a practical, step-by-step example, illustrating how to effectively utilize the pingouin.intraclass_corr function to analyze real-world data and interpret the multifaceted results generated by the ICC calculation.

Understanding the Need for Intraclass Correlation

In many research designs, particularly those involving subjective assessments, it is essential to demonstrate that the variability observed in the measurements is due to actual differences

between the items being rated (the targets) rather than inconsistencies stemming from the measurement instrument or the raters themselves. The ICC serves as a powerful metric derived from analysis of variance (ANOVA) principles, allowing researchers to partition the total variance into variance attributable to the subjects, variance attributable to the raters, and residual error variance.

While simpler metrics like Cohen's Kappa or percent agreement can measure agreement between two raters, the ICC is uniquely capable of handling situations involving three or more raters simultaneously. Furthermore, the ICC provides a single, interpretable value that generalizes across different contexts, unlike many other reliability coefficients. This generalization capability makes the ICC a preferred standard in disciplines requiring high stakes reliability, such as psychiatric diagnosis, medical imaging assessment, and complex performance evaluations.

The choice of which specific ICC value to report is fundamentally tied to the experimental design and the nature of the generalization sought by the researcher. For instance, if the goal is to generalize results to a larger population of potential raters (random effects model), a different ICC formulation is required compared to a study where the specific raters used are the only ones of interest (fixed effects model). Understanding these subtle yet critical distinctions is the key to accurate statistical reporting, which the Pingouin package simplifies by providing multiple ICC types in a single output.

Step 1: Setting Up the Python Environment and Installing Pingouin

Before any statistical analysis can commence, the necessary computational environment must be established. For this specific task, we rely on the Pingouin library, a Python package designed for statistical testing that offers an excellent balance of speed, accuracy, and user-friendliness. If you are working in a standard Python environment or a Jupyter Notebook, the installation process is straightforward and utilizes the standard package installer, pip.

To ensure that the library is available within your current Python environment, open your command-line interface or terminal and execute the following installation command. It is highly recommended to perform this installation within a dedicated virtual environment to maintain project dependencies cleanly, although the command itself is universal:

pip install pingouin

Successful execution of this command will download the Pingouin library and its necessary dependencies, preparing your environment for advanced statistical computation. Once installation is complete, you will be able to import the library and proceed with data preparation and analysis steps outlined below, ensuring you have the tools required to calculate the Intraclass Correlation

Coefficient efficiently.

Step 2: Structuring the Data for ICC Analysis

The `intraclass_corr` function requires the data to be structured in a "long" format, which means that each row represents a single rating instance. This structure is essential as it clearly separates the target (item being rated), the rater (who gave the score), and the rating (the score itself). We must first import the Pandas library, which is the cornerstone for data handling and manipulation in the Python ecosystem, allowing us to create and manage the necessary DataFrame.

Consider a scenario where four distinct judges were tasked with evaluating the quality of six different college entrance exams. Each judge provides a single numerical rating for each exam. This setup results in 24 total observations (4 judges * 6 exams). We structure this data into a Pandas DataFrame, explicitly labeling the columns for clarity: 'exam' (the target), 'judge' (the rater), and 'rating' (the score).

The following code block demonstrates how to initialize the DataFrame, which serves as the input for the ICC calculation. Pay close attention to how the data is organized to meet the 'long' format requirement, where rater IDs ('judge') are repeated for each target they rate, and target IDs ('exam') are repeated for each rater who scores them:

import pandas as pd

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

#view first five rows of DataFrame
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

Viewing the initial rows confirms the necessary structure: each exam is linked to a specific judge, and that pairing results in a single rating score. This preparation is a crucial prerequisite, as improper data formatting is the most common pitfall when utilizing specialized statistical functions

like those provided by Pingouin.

Step 3: Executing the Intraclass Correlation Calculation

With the data correctly structured in the `df` DataFrame, the next logical step is to invoke the `intraclass_corr` function. We need to clearly map the columns created in Step 2 to the function's arguments: 'exam' maps to `targets`, 'judge' maps to `raters`, and 'rating' maps to `ratings`. By performing this mapping, we instruct the Pingouin library on how to interpret the variance structure within the dataset.

The following Python code executes the calculation, leveraging the previously imported data. We store the comprehensive results in a variable named `icc`, which itself is returned as a specialized Pandas DataFrame containing all six standard ICC types and their associated statistical parameters:

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

This single function call is highly efficient, generating the results needed for a comprehensive ICC analysis. The output is displayed neatly, with the 'Type' column serving as the index for easier readability and identification of the specific model applied.

Interpreting the Output Variables

The resulting DataFrame produced by `intraclass_corr` is rich in statistical information, far exceeding a single reliability coefficient. Understanding each column is vital for correctly reporting the findings of the reliability study. Six distinct ICC values are presented because the appropriate coefficient depends entirely on the underlying assumptions of the research design.

The output table contains the following key metrics for each of the six computed ICC types:

Description: A detailed label identifying the precise combination of model and unit used for the calculation (e.g., Single raters absolute agreement).
ICC: The calculated Intraclass Correlation Coefficient itself--the primary measure of reliability.
F: The F-value derived from the underlying ANOVA model, used to test the significance of the ICC.
df1, df2: The corresponding degrees of freedom associated with the F-statistic, essential for determining the critical F-value.
pval: The p-value indicating the statistical significance of the ICC. A low p-value (typically < 0.05) suggests that the reliability is significantly greater than zero.
CI95%: The 95% confidence interval for the ICC, providing a range within which the true population reliability coefficient is expected to lie.

In our example, the p-values are all highly significant (ranging from 0.004430 to 0.007352), suggesting strong evidence that the agreement observed among the judges is substantially better than zero. For instance, the ICC1 (Single raters absolute) is 0.505, with a 95% confidence interval spanning . This indicates moderate reliability for any single, randomly selected judge.

A Deep Dive into the Six ICC Types

The presence of six different ICC values necessitates a deeper examination of the theoretical framework. These six values are generated by systematically combining three fundamental design choices: the underlying statistical model, the type of relationship assessed, and the unit of reliability.

These crucial assumptions determine which of the calculated ICC values is the correct one for your specific research question:

Model: There are three primary ANOVA models that can be used:

One-Way Random Effects (ICC1, ICC1k): This model assumes that the raters are different for each target, or that the rater effect is not relevant. It treats only the targets as random variables and is often used when each rater rates only a subset of the targets.
Two-Way Random Effects (ICC2, ICC2k): This is perhaps the most common model. It assumes both the targets and the raters are randomly selected from a larger population. Results are generalizable to other raters not included in the study.
Two-Way Mixed Effects (ICC3, ICC3k): This model treats targets as random but raters as fixed variables. It is used when the specific raters in the study are the only ones of interest, such as when using a specific, trained team of experts.

Type of Relationship:

Absolute Agreement: This type requires that the raters give exactly the same score. Differences in mean scores between raters contribute to the error variance, thus lowering the ICC value. (ICC1, ICC2, ICC3)
Consistency: This type only requires that the ratings are linearly related or consistent in their ordering (i.e., if Rater A scores an item high, Rater B also scores it high, regardless of whether Rater B's absolute scores are generally higher than Rater A's). This ignores systematic differences in rater means. (The `pingouin` output focuses primarily on Absolute Agreement for ICC1, ICC2, ICC3, and then presents the average measures).

Unit:

Single Rater (ICC1, ICC2, ICC3): Estimates the reliability if the decision were based on the measurement of a single, randomly chosen rater.
Average Raters (ICC1k, ICC2k, ICC3k): Estimates the reliability if the decision were based on the mean of all raters involved in the study. Since averaging scores tends to cancel out random error, these values are always higher than their single-rater counterparts.

For a rigorous and detailed explanation of these fundamental assumptions, including guidelines on which ICC type to select based on your specific study design, researchers are strongly encouraged to consult authoritative statistical resources on Intraclass Correlation Coefficient calculation and interpretation. Correct model selection is paramount to deriving meaningful conclusions about measurement reliability.