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

Authored by

stats writer

December 6, 2025

The Intraclass Correlation Coefficient (ICC) serves as a critical statistical measure used to quantify the degree of reliability or consistency among multiple raters, judges, or measurement instruments. It is frequently employed in behavioral science, clinical trials, and psychometrics when assessing how well a set of quantitative ratings on a scale agree or cluster together within predefined groups. Calculating the ICC efficiently can be achieved using the statistical environment R, specifically leveraging the robust functionalities provided by the irr package.

The primary tool within this library is the `icc()` function. This function streamlines the complex calculations required to derive the ICC value, which essentially measures the proportion of variance in the observations that can be attributed to differences between subjects rather than differences between raters. To execute `icc()`, the researcher must provide two core pieces of information: the matrix or data frame containing the ratings provided by each rater, and clear specifications regarding the statistical model, type of agreement sought, and the unit of analysis. The resulting output furnishes the calculated ICC and its associated confidence interval, offering a complete picture of measurement consistency.

The fundamental purpose of calculating the Intraclass Correlation Coefficient (ICC) is to rigorously determine whether items or subjects can be assessed consistently and with appropriate reliability when evaluated by multiple independent raters. This coefficient transitions from a value of 0, indicating a complete absence of measurable consistency or agreement among the judges, up to 1, which signifies perfect agreement and maximal reliability.

For researchers working in R, the most straightforward and accepted method for computing the ICC is through the `icc()` function, which is a core component of the dedicated irr package. Understanding the syntax and the underlying parameters is essential for obtaining a valid and interpretable result that aligns with the specific research design.

The `icc()` function utilizes a generalized syntax that requires the input data and explicit definitions for the intended analysis parameters:

icc(ratings, model, type, unit)

The parameters define the specific statistical assumptions used in the calculation, ensuring the resulting ICC value is appropriate for the study's design. These arguments are further elaborated below:

ratings: This argument requires a **dataframe** or **matrix** containing the raw numerical ratings. Each column typically represents a different rater, and each row represents the item or subject being rated.
model: This specifies the statistical model upon which the ICC calculation is based. Crucial choices include the **"oneway"** model (where raters are nested within subjects) or the **"twoway"**

model (where raters are crossed with subjects).
type: This defines the type of relationship being measured between raters. Options are **"consistency"** (measuring correlation, allowing for systematic differences in mean scores) or **"agreement"** (measuring absolute agreement, where scores must match exactly).
unit: This determines the unit of analysis being used for the reliability estimate. Options include **"single"** (representing the reliability of a single, typical rater) or **"average"** (representing the reliability of the mean of all raters).

The remainder of this tutorial provides a practical, step-by-step walkthrough detailing how to utilize this powerful function effectively within the R environment, starting with data creation and concluding with result interpretation.

Step 1: Preparing the Data Structure in R

Before any calculation can commence, the data must be properly formatted into an appropriate matrix or dataframe that the `icc()` function can process. For illustrative purposes, let us construct a scenario where a psychometric study requires multiple evaluations. Suppose we have four distinct expert judges who have been tasked with rating the overall quality of ten different college entrance exam essays, using a simple numerical scale.

Each rater (Judge A, Judge B, Judge C, and Judge D) provides a score for all 10 essays. This structure requires the data to be organized such that each column represents a judge and each row represents an essay (or subject). This structure is essential because the irr package expects raters to be listed in the columns, allowing the function to calculate variance components across both subjects and raters simultaneously.

We can utilize the following command block in R to instantiate this sample dataset, creating a dataframe that holds the ratings assigned by the four judges (A, B, C, D) across the 10 subjects:

```
#create data
data <- data.frame(A=c(1, 1, 3, 6, 6, 7, 8, 9, 8, 7),
B=c(2, 3, 8, 4, 5, 5, 7, 9, 8, 8),
C=c(0, 4, 1, 5, 5, 6, 6, 9, 8, 8),
D=c(1, 2, 3, 3, 6, 4, 6, 8, 8, 9))
```

Step 2: Selecting the Appropriate ICC Model

Before proceeding to the calculation, researchers must carefully consider which of the many possible ICC types is most appropriate for their specific study design. The choice of the statistical model (One-Way or Two-Way), the desired type of relationship (Consistency or Agreement), and

the unit of interest (Single or Average) fundamentally alters the interpretation of the resulting ICC. Selecting the wrong version can lead to misleading conclusions regarding measurement reliability.

In our scenario, we assume the four judges were randomly selected from a larger population of qualified entrance exam graders. This specific sampling methodology dictates the use of a **Two-Way model**, which assumes that both subjects (essays) and raters (judges) are random effects. Had the raters been fixed (i.e., the only raters of interest), a Two-Way Mixed model might have been used, or if different sets of raters assessed different subjects, a One-Way model would be suitable. The Two-Way model is highly standard when generalizability to other raters is desired.

Furthermore, we must decide whether we seek absolute agreement or consistency. Absolute Agreement requires that the scores match numerically, taking into account systematic biases (like one rater consistently scoring higher than others). Consistency, conversely, only requires that the raters maintain the same rank order for the subjects, ignoring systematic differences in means. Since we are interested in whether the ratings are truly interchangeable and seek the strictest measure, we choose the **Absolute Agreement** type. Finally, if we intend to use the rating of just one randomly selected judge in future studies, we use **"single"** as our unit. If we plan to use the average score of all four judges, we would select "average".

Step 3: Calculating the Intraclass Correlation Coefficient

Based on the research context established in the previous step--where the raters are randomly sampled, we seek absolute agreement, and our interest lies in the reliability of a single rater--we can now execute the `icc()` function. This configuration corresponds to the standard notation ICC(A, 1) if we assume a Two-Way Random Effects Model.

We must first ensure the irr package is loaded into the current R session using the `library()` command. Following the data definition, we specify our parameters: `model = "twoway"`, `type = "agreement"`, and `unit = "single"`.

The execution of the following code generates the comprehensive output detailing the variance components analysis and the final ICC estimate:

```
#load the interrater reliability package
library(irr)

#define data
data <- data.frame(A=c(1, 1, 3, 6, 6, 7, 8, 9, 8, 7),
B=c(2, 3, 8, 4, 5, 5, 7, 9, 8, 8),
C=c(0, 4, 1, 5, 5, 6, 6, 9, 8, 8),
D=c(1, 2, 3, 3, 6, 4, 6, 8, 8, 9))
```

```
#calculate ICC
icc(data, model = "twoway", type = "agreement", unit = "single")

Model: twoway
Type : agreement

Subjects = 10
Raters = 4
ICC(A,1) = 0.782

F-Test, H0: r0 = 0 ; H1: r0 > 0
F(9,30) = 15.3 , p = 5.93e-09

95%-Confidence Interval for ICC Population Values:
0.554 < ICC < 0.931
```

Step 4: Interpreting the Calculated ICC Value

Upon execution, the output clearly states the specifications used--Model: twoway, Type: agreement--confirming that the correct analytical framework was applied. The resulting Intraclass Correlation Coefficient, labeled as ICC(A,1), is calculated to be **0.782**. This single value encapsulates the overall degree of measurement consistency based on the strict criterion of absolute agreement among the randomly sampled judges.

The output also includes a statistical test (F-Test) comparing the observed ICC to a null hypothesis where the true ICC (r0) is zero. In this case, the F-statistic is large (F(9,30) = 15.3) and the p-value is extremely small (p = 5.93e-09). This highly significant result leads us to confidently reject the null hypothesis, concluding that the measured reliability is significantly greater than zero.

Furthermore, the 95% Confidence Interval (CI) provides a range of plausible values for the true population ICC, spanning from 0.554 to 0.931. Because this entire interval is positive and the lower bound is reasonably high, it reinforces the conclusion that the inter-rater reliability is substantial. Generally, the interpretation of the magnitude of the ICC value follows established guidelines:

Less than 0.50: Indicates **Poor reliability**, suggesting that differences between raters account for a large proportion of the total variance.
Between 0.5 and 0.75: Suggests **Moderate reliability**, where the consistency is acceptable for exploratory or screening purposes.
Between 0.75 and 0.9: Represents **Good reliability**, indicating strong consistency suitable for most research applications. Our observed value of 0.782 falls into this desirable category.
Greater than 0.9: Denotes **Excellent reliability**, approaching perfect agreement among raters.

A Detailed Note on Selecting the Correct ICC Version

It is paramount to recognize that the term ICC does not refer to a single statistic but rather a family of related statistics, each based on different statistical assumptions derived from analysis of variance (ANOVA) principles. The appropriate ICC version must be chosen to match the specific sampling scheme and research question of the study. A robust calculation using the `icc()` function requires explicit specification of three essential factors:

Model: Determines whether the raters are fixed or random effects. Options include One-Way Random Effects (each subject rated by a different, random subset of raters), Two-Way Random Effects (subjects and raters are both random factors, allowing generalization to other raters), or Two-Way Mixed Effects (subjects are random, but the raters are fixed and exhaust the population of interest).
Type of Relationship: Dictates the strictness of the agreement required. **Consistency** measures correlation (high ICC if Rater A scores 5 points higher than Rater B across all subjects), whereas **Absolute Agreement** measures the degree to which ratings are exactly the same, penalizing systematic bias.
Unit: Specifies whether the reliability estimate applies to a **Single rater** (ICC(A,1)) or to the mean rating across all available raters (ICC(A,k)). The latter is generally higher because the averaging process cancels out random measurement error.

In the preceding example, the ICC calculation utilized the following combination of assumptions, which is often cited as the strictest measure of absolute agreement for randomly selected raters:

Model: Two-Way Random Effects
Type of Relationship: Absolute Agreement
Unit: Single rater

Understanding these underlying theoretical distinctions is non-negotiable for producing reliable and scientifically defensible measures of inter-rater reliability.