Computing Krippendorff's Alpha for Complex Categorical Data

Elad Zlotnick

2024-12-22

Contents

Introduction to Krippendorff's Alpha **1**

Complex Categorical Data **2**

Kripendorff's α Algorithm **2**

Bibliography **3**

Introduction to Krippendorff's Alpha

Krippendorff's alpha (α) is a versatile and robust statistical measure of inter-rater reliability (IRR), designed to assess the level of agreement among multiple raters or coders. It can be applied to categorical, ordinal, interval, and ratio data, making it a highly flexible tool in research. One of its key advantages is its ability to accommodate missing data, which allows for broader applicability across complex datasets.

Krippendorff's alpha quantifies how much the observed agreement exceeds the agreement expected by chance. Its value ranges from 0 to 1:

- $\alpha = 0$: No agreement beyond chance.
- $\alpha = 1$: Perfect agreement among raters.
- Intermediate values indicate varying degrees of reliability, with values closer to 1 reflecting higher levels of agreement.

Krippendorff's alpha is calculated using the formula:

$$\alpha = 1 - \frac{D_o}{D_e}$$

Where:

- D_o: **Observed Disagreement** – the actual level of disagreement in the data.

- D_e: **Expected Disagreement** – the level of disagreement expected by chance, based on the distribution of categories.

The ratio $\frac{D_o}{D_e}$ represents the proportion of disagreement attributable to chance. Subtracting this value from 1 provides a measure of agreement that exceeds chance expectations.

For further details and a deeper understanding of Krippendorff's alpha, refer to foundational works (Hayes & Krippendorff, 2007; e.g., Krippendorff, 2018; Zapf et al., 2016).

Complex Categorical Data

In the current study we had judges categorize fears. Each fear could be assigned either **one or two** categories. Agreement was defined as having **at least one common category** between judges. For example if if judge a assigns categories 1 and 2, while judge b assigns category 2 and 3, we would consider them as agreeing. This approach captures the multi-faceted nature of fear and allows for stability beyond changes in time and judge interpretation.

While effective, this scoring scheme presented challenges for standard inter-rater reliability (IRR) measures. These measures typically assume **unique category assignments** per unit, but allowing two categories increases the chance of agreement by chance. Additionally, judges assigning one or two categories could have different agreement probabilities, further complicating the analysis.

To address these issues, we designed an algorithm to estimate the likelihood of expected disagreement given the multiple-category nature of our data. This likelihood was then incorporated into Krippendorff's Alpha formula, providing a standard IRR measure for our analysis.

Kripendorff's α Algorithm

This section outlines the algorithm used to estimate Krippendorff's α. While it produces identical results to the original algorithm described by Krippendorff (Krippendorff, 2018, p. 293), it is more computationally expensive. Its primary advantage lies in its flexibility, allowing for the implementation of custom agreement schemes (e.g., "at least one common category selected").

To facilitate understanding of the algorithm, we first define the key concepts and components involved. These definitions provide the foundation for interpreting the steps in the algorithm:

- **Unit**: A unit refers to a single entity or item evaluated by the judges. In our study, each unit corresponds to a single fear. Units serve as the basic elements of analysis and are assigned ratings by judges.
- **Rating**: A rating represents the category or set of categories assigned to a unit by a judge. In this study, a rating is a set of categories that captures the multi-dimensional nature of fears.

The algorithm itself, presented as a step-by-step list, builds on these definitions to compute α by systematically calculating observed and expected disagreement. Each step incorporates these foundational concepts, ensuring

a clear and structured progression.

1. **Clean Data**:
   - Exclude units with fewer than two ratings, as they cannot offer insights into agreement.
2. **Observed Disagreement (D_o):**
   - Agreement Rate per Unit: For each unit, calculate the proportion of pairwise agreements between ratings.
   - $a_u = \frac{\sum_{i,j \in \{1 \dots n_u\}, i \neq j} \delta_{u_i, u_j}}{n_u - 1}$
   - $\delta_{a,b}$ Agreement function (1 for agreement and 0 for disagreement)
   - u_i Values of unit i
   - n_u Number of ratings for unit u
   - Compute observed agreement: $A_o = \sum_u a_u$
   - Compute observed disagreement: $D_o = N - A_o$
     - N total number of ratings
3. **Calculating Expected Disagreement (D_e):**
   - Calculate expected agreement for each *value*:
   - Expected Agreement per rating: For each rating, calculate the proportion of agreement that it has with all other ratings.
     - $e_v = \sum_{i \in \{1 \dots N\}, i \neq v} \delta_{v,i}$
     - v is the value index
     - N total number of ratings
     - $\delta_{a,b}$ Agreement function (1 for agreement and 0 for disagreement)
   - Compute expected disagreement: $D_e = \frac{N^2 - \sum_v e_v}{N-1}$
     - e_v is the individual expected agreement
     - N is the number of ratings overall
4. **Calculating α**
   - $\alpha = 1 - \frac{D_o}{D_e}$
   - Observed Disagreement (D_o): Represents the actual level of disagreement observed between judges in the current data.
   - Expected Disagreement (D_e): Represents the chance of two judges disagreeing by pure chance, based on their past ratings.

Note: For simple nominal comparison, the agreement function δ_{ab} is simply 1 for $a = b$ and 0 for $a \neq b$. For interval metrics $\delta_{ab} = (a - b)^2$. More complex relationships can easily be encoded. For our case we used $\delta_{ab} = b \cap a \neq \emptyset$.

Bibliography

Hayes, A. F., & Krippendorff, K. (2007). Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures*, *1*(1), 77–89. https://doi.org/10.1080/19312450709336664

Krippendorff, K. (2018). *Content Analysis: An Introduction to Its Methodology*. SAGE Publications.

Zapf, A., Castell, S., Morawietz, L., & Karch, A. (2016). Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate? *BMC Medical Research Methodology*, *16*, 93. https://doi.org/10.1186/s12874-016-0200-9