Not for use or reproduction without reference to the author, this source, and its date.

Bootstrapping Distributions for Krippendorff's Alpha

for coding predefined units: single-valued $_c\alpha$ and multi-valued $_{mv}\alpha$

Klaus Krippendorff
kkrippendorff@asc.upenn.edu

Originally posted on 2006.08.16. Last revised 2016.10.10

In the absence of a theoretically motivated distribution of the reliability coefficients α for coding predefined units (Krippendorff, 2013: 277-301; Krippendorff & Craggs, 2017), but more importantly because reliability data tend to be small, have irregular distributions, use diverse metrics (levels of measurement), and may have missing data, the distribution of α is best found by bootstrapping.

Consistent with α's definition as measure of the reliability of coding data (not of observers, coders, or judges), this bootstrapping algorithm resamples hypothetical reliability data from pairs of categories or values found in data generated by any number of independent replications. To get to a hypothetical $\alpha = 1 - D_o^*/D_e$, it computes hypothetical disagreements D_o^* from the resampled data but maintains the generally more stable expected disagreement D_e from the observed data. Numerous repetitions of this resampling process results in a probability distribution within the limits of $-1 \le \alpha \le +1$, which gives rise to α's confidence intervals and the probability q of the Type I error of α failing to reach a minimally acceptable reliability $\alpha_{\min}$.

This significantly simplified algorithm is implemented in the revised software KALPHA (Hayes & Krippendorff, 2007) and is recommended for related applications.

Bootstrapping does not apply when

- $\alpha = 1.000$
- All but one value c in the reliability data are identical and $\alpha = 0.000$ by computation
- Reliability data have no variance whereby $\alpha = 1 - 0/0 = 0$ by definition.

The terms used in the following definitions refer to bootstrapping $_c\alpha$ for coding of single-valued data. When bootstrapping $_{mv}\alpha$ for multi-valued data, c has to be replaced by C, k by K, and $_{\text{metric}}\,\delta^2_{ck}$ by $_{\text{metric}}\Delta_{CK}$ or $_{\text{metric}}\Sigma_{CK}$.

References are made to:

- The original **reliability data**:

Units:	1	2	3	. .	u	. . . .	N_u	
1st Observer	c_{11}	.	.	. .	c_{1u}	. . . .	c_{1N_u}	
:	:				:		:	
ith Observer	c_{i1}	.	.	. .	c_{iu}	. . . .	c_{iN_u}	
jth Observer	c_{j1}	.	.	. .	c_{ju}	. . . .	c_{jN_u}	
:	:				:		:	
mth Observer	c_{m1}	.	.	. .	c_{mu}	. . . .	c_{mN_u}	
Number of pairable values	m_1	.	.	. .	m_u	. . . .	m_{N_u}	$n_{..} = \sum_{u=1}^{n_u} m_u \mid m_u \ge 2$

- The number N_o of **unique pairs** that contribute to α: $N_o = \sum_{u=1}^{N_u} \frac{(m_u - 1)m_u}{2}$
- The **metric difference** ${}_{\text{metric}}\delta_{ck}^2$ used in α_{metric}
- The **expected disagreement** D_e in the denominator of $\alpha_{\text{metric}} = 1 - \frac{D_o}{D_e}$
- The **number *X* of samples** to be assembled, $X = 20{,}000$ suggested
- The **level *p* of statistical significance** (two-tailed test), $p = 0.05$ suggested
- The **minimum reliability** required for data to be acceptable: $\alpha_{\min} = 0.800$ suggested
- The **error function** $E(r)$:

  Given the observed disagreement: $D_o = \frac{1}{n_{..}} \sum_{u=1}^{N_u} \frac{1}{m_u - 1} \sum_{i=1}^{m} \sum_{j \neq i}^{m} {}_{\text{metric}}\delta_{c_{iu} c_{ju}}^2$,

  α can be decomposed into:

  $$\alpha = 1 - \frac{D_o}{D_e} = 1 - \sum_{u=1}^{Nu} \frac{1}{m_u - 1} \sum_{i=1}^{m} \sum_{j>i} 2 \frac{{}_{\text{metric}}\delta_{c_{iu} c_{ju}}^2}{n_{..} \cdot D_e} = 1 - \sum_{u=1}^{N_u} \frac{1}{m_u - 1} \sum_{r=1}^{\frac{(m_u - 1)m_u}{2}} E(r)$$

  Create two lists of N_o items each:

  The r^{th} out of N_o possible **deviations *E*(*r*)** from α = 1 is: $E(r) = 2 \frac{{}_{\text{metric}}\delta_{c_{iu} c_{ju}}^2}{n_{..} \cdot D_e}$

  The r^{th} out of N_o possible **pairs** accounting for $E(r)$ is: $\text{Pair}(r) = \langle c_{iu}, c_{ju} \rangle$

Algorithm for bootstrapping a distribution of hypothetical αs:

Set the integer array $n_\alpha = 0$. $-1 \leq \alpha \leq +1$. (The subscript α needs at least 20001 values)
Do *X* times
  $\alpha = 1$ (As $E(r)$ tends to be very small, α requires high precision)
  Do $u = 1, N_u$
    Do $\frac{(m_u - 1)m_u}{2}$ times (= The number of unique pairs in the u^{th} unit)
      Pick a random integer $1 \leq r \leq N_o$ (uniform distribution)
      If the 2[nd] pick yields $Pair(r_{2^{nd}}) = Pair(r_{1^{st}})$ [return to the random pick]
      $\alpha \Leftarrow \alpha - \frac{E(r)}{m_u - 1}$
  If $\alpha < -1$: $n_{-1} \Leftarrow n_{-1} + 1$
  If $\alpha \geq -1$: $n_\alpha \Leftarrow n_\alpha + 1$ (Here, α is rounded to the number of digits of the subscript)

- The resulting **distribution of α's probabilities** $\frac{n_\alpha}{X}$ is accounted for in terms of:

  The **confidence interval:** $-1 \leq \alpha_{smallest} \leq \alpha \leq \alpha_{largest} \leq 1$
  for a chosen level p of statistical significance (two-tailed):

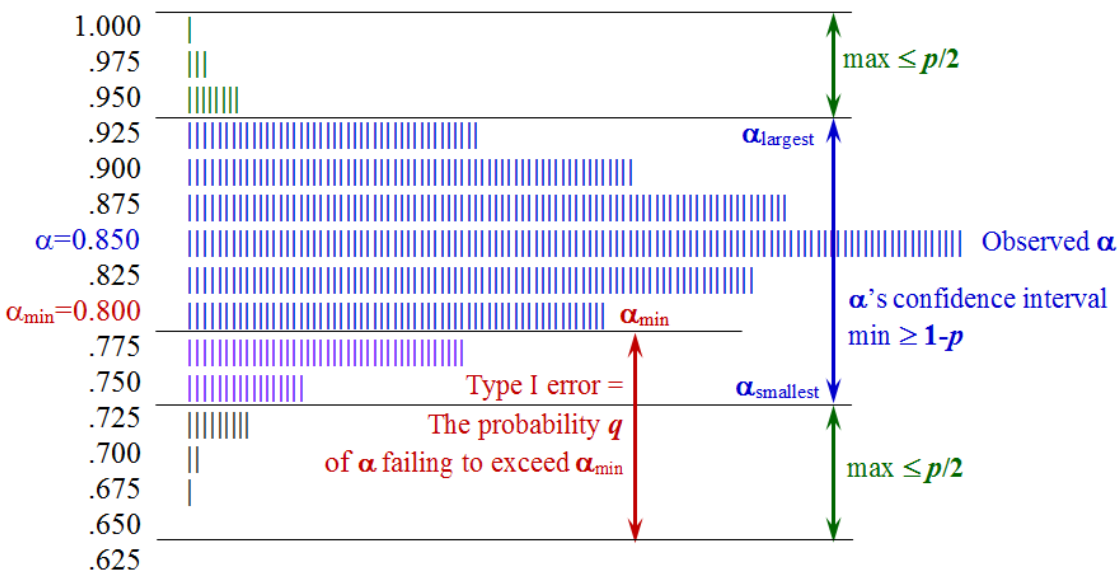
  Where: $\alpha_{smallest}$ = the smallest α for which $\sum_{z=-1}^{z<\alpha_{smallest}} \frac{n_z}{X} \geq \frac{p}{2}$

  $\alpha_{largest}$ = the largest α for which $\sum_{z>\alpha_{largest}}^{z=1} \frac{n_z}{X} \geq \frac{p}{2}$

  The **probability** q of the Type I error, of **α failing** to exceed the required α_{min} (one-tailed):

$$q = \sum_{z=-1}^{z<\alpha_{min}} \frac{n_z}{X}$$

With $\alpha=0.850$ observed, the required minimum $\alpha_{min}=0.800$, and the level of statistical significance $p=0.05$, the following illustrates the two statistical parameters of α's probability distribution:



References:

Hayes, Andrew F. & Krippendorff, Klaus (2007). Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures 1*, 1: 77-89. http://www.afhayes.com/public/cmm2007.pdf (Accessed 2015.9.25).

Krippendorff, Klaus (2013). *Content Analysis; An Introduction to its Methodology*, *3rd Edition*. Thousand Oaks, CA: Sage Publications.

Krippendorff, Klaus & Craggs, Richard (2017 in press). The Reliability of Multi-Valued Coding of Data. *Communication Methods and Measures*. (Includes a link to available software).

Free SAS and SPSS macros called KALPHA for calculating $_c\alpha$ may be downloaded from http://www.afhayes.com/ (Accessed 2015.9.25).