PROPOSAL OF HUBERT_KAPPA.EXE (Version 3)

NOMINAL AGREEMENT MODEL AMONG MANY RATERS

The following is based on the paper Martín Andrés and Álvarez Hernández (2020)

1. Unweighted *kappa*

If $R \geq 2$ raters independently classify n subjects in K categories, a data matrix $\{y_{sr}\}$ is obtained, with $s=1, 2, \ldots, n$, $r=1, 2, \ldots, R$ and $y_{sr}=1, 2, \ldots, K$, in which $y_{sr}=i$ when the rater r classifies subject s into category i. The most common thing to do is to summarize this information in a table of absolute frequencies $x_{i_1 i_2 \ldots i_R}=\#\{s \mid y_{s1}=i_1, \ldots, y_{sR}=i_R\}$ of dimension K^R, where the symbol # refers to "cardinal" and $x_{i_1 i_2 \ldots i_R}$ is the number of subjects classified as type i_1 by rater 1, type i_2 by rater 2, ..., or type i_R by rater R (see Table 1a); so $\sum_{i_1}\sum_{i_2}\ldots\sum_{i_R} x_{i_1 i_2 \ldots i_R} = n$. Let $p_{i_1 i_2 \ldots i_R} = x_{i_1 i_2 \ldots i_R}/n$ be the estimated proportions. The "crude" agreement (without random correction) is $\Sigma_i\, p_i$ where $p_i = p_{ii\ldots i}$ = raw. Let $t_{i(r)}=\sum_{i_1}\ldots\sum_{i_{r-1}}\sum_{i_{r+1}}\ldots\sum_{i_R} p_{i_1 \ldots i_r=i \ldots i_R}$ be the total proportion of responses i of rater r (see Table 1b). Then the Hubert´s *kappa* κ_H (1977) is estimated by:

$$\hat{\kappa}_H = \frac{I_o - I_e}{1 - I_e}, \text{ with } I_o = \sum_{i=1}^{K} p_i,\ I_e = \sum_{i=1}^{K} P_i \text{ and } P_i = \Pi_{r=1}^{R} t_{i(r)},$$

and

$$\hat{V}\left(\hat{\kappa}_H\right) = \frac{D + E - F}{n\left(1 - I_e\right)^2} \text{ con } \begin{cases} D = \sum_{i=1}^{K} p_i \left[1 - \left(1 - \kappa_H\right)\sum_{r=1}^{R} T_{i(r)}\right]^2, \\ E = \left(1 - \kappa_H\right)^2 \sum_{i_1}\sum_{i_2}\ldots\sum_{i_R}\left(1 - \delta_{i_1 i_2 \ldots i_R}\right) p_{i_1 i_2 \ldots i_R}\left(\sum_{r=1}^{R} T_{i_r(r)}\right)^2, \\ F = \left[\kappa_H - \left(R-1\right)\left(1 - \kappa_H\right) I_e\right]^2, \end{cases}$$

where (see Table 1c)

$$T_{i(r)} = \prod_{r' \neq r} t_{i(r')} = P_i / t_{i(r)}, \text{ and so } T_{i_r(r)} = \prod_{r' \neq r} t_{i_r(r')} = P_{i_r} / t_{i_r(r)}, \text{ with } P_{i_r} = \prod_{r'} t_{i_r(r')}$$

Hubert´s *kappa* is based on the following definition"an agreement occurs if and only if all

raters agree on the categorization of an object" or DeMoivre's definition of agreement or definition *R*-wise. When *R*=2 then $\kappa_H=\kappa_C$, $\hat{\kappa}_H=\hat{\kappa}_C$ and $\hat{V}(\hat{\kappa}_H)=\hat{V}(\hat{\kappa}_C)$, where $\hat{\kappa}_C$ is the Cohen's *kappa* coefficient (1960) and $\hat{V}(\hat{\kappa}_C)$ is the variance given by Fleiss *et al.* (1969).

2. Weighted *kappa*

If $v_{i_1 i_2 \ldots i_R}$ (≥ 0) is the weight of the disagreement of the set of responses $\{i_1, i_2, \ldots, i_R\}$, then the Shuster and Smith´s weighted kappa κ_{Hw} (2005) is estimated by

$$\hat{\kappa}_{Hw} = 1-\frac{\overline{I}_o}{\overline{I}_e} \text{ where } \begin{cases} \overline{I}_o = \sum_{i_1}\sum_{i_2}\cdots\sum_{i_R} v_{i_1 i_2 \ldots i_R} p_{i_1 i_2 \ldots i_R}, \\ \overline{I}_e = \sum_{i_1}\sum_{i_2}\cdots\sum_{i_R} v_{i_1 i_2 \ldots i_R} P_{i_1 i_2 \ldots i_R}, \end{cases} \text{ and } P_{i_1 i_2 \ldots i_R} = \Pi_{r=1}^{R} t_{i_r(r)},$$

and

$$\hat{V}(\hat{\kappa}_{Hw}) = \frac{1}{n}\left[\frac{\sum_{i_1}\sum_{i_2}\cdots\sum_{i_R} p_{i_1 i_2 \ldots i_R}\left[v_{i_1 i_2 \ldots i_R}-(1-\kappa_{Hw})\sum_r \overline{v}_{i_r(r)}\right]^2}{\overline{I}_e^2} - \left\{(R-1)(1-\kappa_{Hw})\right\}^2\right],$$

where $\overline{v}_{i_r(r)}=\sum_{i_1}\cdots\sum_{i_{r-1}}\sum_{i_{r+1}}\cdots\sum_{i_R} v_{i_1 i_2 \ldots i_R} T_{i_1 i_2 \ldots i_R(r)}$ con $T_{i_1 i_2 \ldots i_R(r)} = \Pi_{r'\neq r} t_{i_{r'}(r')}$. It is traditional to use the following linear and quadratic weights (Mielke *et al.* 2007 and Schuster and Smith 2005, respectively):

Linear: $v_{i_1 i_2 \ldots i_R} = \sum_r \sum_{r'>r} |i_r - i_{r'}|$,, Quadratic: $v_{i_1 i_2 \ldots i_R} = \sum_r \sum_{r'>r} (i_r - i_{r'})^2$

If $w_{i_1 i_2 \ldots i_R}$ are the traditional weights of agreements ($0\leq w_{i_1 i_2 \ldots i_R}\leq 1$ and $w_{ii\ldots i}=1$) then $v_{i_1 i_2 \ldots i_R} = 1-w_{i_1 i_2 \ldots i_R}$ ($0\leq v_{i_1 i_2 \ldots i_R}\leq 1$ and $v_{ii\ldots i}=0$), but the result of the previous formulas does not change if you put $v_{i_1 i_2 \ldots i_R}$ = Constant×$(1-w_{i_1 i_2 \ldots i_R})$ instead of $v_{i_1 i_2 \ldots i_R}=1-w_{i_1 i_2 \ldots i_R}$. When $v_{i_1 i_2 \ldots i_R}$ = Variance of (i_1, i_2, …, i_R) or $v_{i_1 i_2 \ldots i_R}=\sum_r \sum_{r'>r} (i_r - i_{r'})^2$, then $\kappa_{Hw} = \rho_I$(intraclass correlation coefficient of Pearson, 1901) $= \rho_C$ (concordance correlation coefficient of Lin, 1989). The new expressions, equivalent to the previous ones, are

$$\hat{\kappa}_{Hw}=\frac{I_o-I_e}{1-I_e} \quad \text{with} \quad \begin{cases} I_o=\sum_{i_1}\sum_{i_2}\ldots\sum_{i_R} w_{i_1 i_2 \ldots i_R} p_{i_1 i_2 \ldots i_R}, \\ I_e=\sum_{i_1}\sum_{i_2}\ldots\sum_{i_R} w_{i_1 i_2 \ldots i_R} P_{i_1 i_2 \ldots i_R}, \end{cases}$$

$$\hat{V}\left(\hat{\kappa}_{Hw}\right)=\frac{\sum_{i_1}\sum_{i_2}\ldots\sum_{i_R} p_{i_1 i_2 \ldots i_R}\left[w_{i_1 i_2 \ldots i_R}-\left(1-\kappa_{Hw}\right)\sum_r \overline{w}_{i_r(r)}\right]^2-\left\{\kappa_{Hw}-\left(R-1\right)\left(1-\kappa_{Hw}\right)I_e\right\}^2}{n\left(1-I_e\right)^2}$$

where $\overline{w}_{i_r(r)}=\sum_{i_1}\ldots\sum_{i_{r-1}}\sum_{i_{r+1}}\ldots\sum_{i_R} w_{i_1 i_2 \ldots i_R} T_{i_1 i_2 \ldots i_R(r)}=1-\overline{v}_{i_r(r)}$.

When R=2, then $\kappa_{Hw}=\kappa_{Cw}$, $\hat{\kappa}_{Hw}=\hat{\kappa}_{Cw}$ and $\hat{V}\left(\hat{\kappa}_{Hw}\right)=\hat{V}\left(\hat{\kappa}_{Cw}\right)$, where $\hat{\kappa}_{Cw}$ is the Cohen's weighted *kappa* coefficient (1968) and $\hat{V}\left(\hat{\kappa}_{Cw}\right)$ is the variance given by Fleiss *et al.* (1969). If $w_{i_1 i_2 \ldots i_R}=\delta_{i_1 i_2 \ldots i_R}$ (Kronecker delta), then $\kappa_{Hw}=\kappa_H$.

3. Inferences

3.1. Unrestricted.

- Test statistic for H_0: $\kappa=\kappa_0$: $\left(\hat{\kappa}-\kappa_0\right)\Big/\sqrt{\hat{\mathrm{V}}\left(\hat{\kappa}\right)}$ *vs.* z.

- CI: $\kappa\in\hat{\kappa}\pm z\sqrt{\hat{\mathrm{V}}\left(\hat{\kappa}\right)}=\hat{\kappa}\pm z\times SE$.

- In both cases:

  $\rightarrow \kappa$ is κ_H or κ_{Hw}.

  $\rightarrow z$ is the appropriate standard normal value.

3.2. Restricted (UNweighted).

- Test statistic for H_0: $\kappa=\kappa_0$: $\left(\hat{\kappa}_H-\kappa_0\right)\Big/\sqrt{\hat{\mathrm{V}}_0\left(\hat{\kappa}_H\right)}$, *vs.* z, where $\hat{\mathrm{V}}_0\left(\hat{\kappa}_H\right)=\dfrac{\overline{a}\left(1-\kappa_0\right)^2-2\overline{b}\left(1-\kappa_0\right)}{n\left(1-I_e\right)^2}$

  and $\begin{cases} \overline{a}=\sum_{i_1}\sum_{i_2}\ldots\sum_{i_R} p_{i_1 i_2 \ldots i_R}\left\{\sum_r T_{i_r(r)}\right\}^2-\left\{1+\left(R-1\right)I_e\right\}^2, \\ \overline{b}=\sum_i p_i\left\{\sum_r T_{i(r)}\right\}-\left\{1+\left(2R-1\right)I_e\right\}\Big/2. \end{cases}$

- CI: $\kappa_H\in\dfrac{\hat{\kappa}_H+d\left(\overline{b}-\overline{a}\right)\pm\sqrt{z^2\hat{\mathrm{V}}\left(\hat{\kappa}_H\right)+d^2\overline{b}^2}}{1-d\overline{a}}$, where $d=\dfrac{z^2}{n\left(1-I_e\right)^2}$.

- Test statistic for the independence test= $\hat{\kappa}_H \left(1-I_e\right)\sqrt{n/m_I}$ *vs.* z, where:

$$m_I = \sum_{i_1} \ldots \sum_{i_R} P_{i_1 i_2 \ldots i_R} \left\{\sum_r T_{i_r(r)}\right\}^2 - 2\sum_i P_i \left\{\sum_r T_{i(r)}\right\} + I_e \left\{1-\left(R-1\right)^2 I_e\right\}$$

3.3. Restricted (weighted).

- Test statistic for H_0: $\kappa=\kappa_0$: $\left(\hat{\kappa}_{Hw}-\kappa_0\right)\Big/\sqrt{\hat{\mathrm{V}}_0\left(\hat{\kappa}_{Hw}\right)}$ *vs.* z, where

$\hat{\mathrm{V}}_0\left(\hat{\kappa}_{Hw}\right) = \dfrac{a\left(1-\kappa_0\right)^2 - 2b\left(1-\kappa_0\right)+c}{n\left(1-I_e\right)^2}$ (sometimes this variance can be negative) and:

$\rightarrow$ Test based on weights v: $\begin{cases} a\left(v\right) = \sum_{i_1}\sum_{i_2}\ldots\sum_{i_R} p_{i_1 i_2 \ldots i_R}\left\{\sum_r \overline{v}_{i_r(r)}\right\}^2 - \left\{\left(R-1\right)\left(1-I_e\right)\right\}^2, \\ b\left(v\right) = \sum_{i_1}\sum_{i_2}\ldots\sum_{i_R} p_{i_1 i_2 \ldots i_R} v_{i_1 i_2 \ldots i_R}\left\{\sum_r \overline{v}_{i_r(r)}\right\}, \\ c\left(v\right) = \sum_{i_1}\sum_{i_2}\ldots\sum_{i_R} p_{i_1 i_2 \ldots i_R} v^2_{i_1 i_2 \ldots i_R}. \end{cases}$

$\rightarrow$ Test based on weights w: $\begin{cases} a\left(w\right) = a\left(v\right) - 2R\left(1-I_e\right), \\ b\left(w\right) = b\left(v\right) - \left\{1+R\left(1-\hat{\kappa}_{Hw}\right)\right\}\left(1-I_e\right), \\ c\left(w\right) = c\left(v\right) - 2\left(1-\hat{\kappa}_{Hw}\right)\left(1-I_e\right). \end{cases}$

- <u>CI</u>: $\kappa_{Hw} \in \dfrac{\hat{\kappa}_{Hw} + d\left(b-a\right) \pm \sqrt{z^2\hat{\mathrm{V}}\left(\hat{\kappa}_{Hw}\right) + d^2\left(b^2-ac\right)}}{1-da}$, where $d = \dfrac{z^2}{n\left(1-I_e\right)^2}$ and a, b and c as in the previous paragraph (sometimes the inside of the square root can be negative).

- Test statistic for the independence test= $\hat{\kappa}_{Hw}\left(1-I_e\right)\sqrt{n/m_I}$ *vs.* z, where:

$\rightarrow$ Test based on weights v: $m_{I(v)} = \sum_{i_1}\ldots\sum_{i_R} P_{i_1 i_2 \ldots i_R}\left\{v_{i_1 i_2 \ldots i_R} - \left(\sum_r \overline{v}_{i_r(r)}\right)\right\}^2 - \left\{\left(R-1\right)\left(1-I_e\right)\right\}^2$.

$\rightarrow$ Test based on weights w: $m_{I(w)} = m_{I(v)} = \sum_{i_1}..\sum_{i_R} P_{i_1 i_2 \ldots i_R}\left\{w_{i_1 i_2 \ldots i_R} - \left(\sum_r \overline{w}_{i_r(r)}\right)\right\}^2 - \left\{\left(R-1\right)I_e\right\}^2$.

3.3. Preferred CI and a comment

- The estimators $\hat{\kappa}_H$ and $\hat{V}\left(\hat{\kappa}_H\right)$ are generally unbiased, although with small samples $\hat{\kappa}_H$ underestimates κ_H.

- In the unweighted case ($\kappa_H \geq \kappa_I$): the unrestricted inferences are preferable, but with very small samples ($n \leq 100$) restricted inferences are preferable.
- In weighted case ($\kappa_{Hw} \geq \kappa_I$): the unrestricted inferences are preferable in the small values of κ_{Hw}, but for the large values of κ_{Hw} the restricted inferences based on the weights w are preferable.

REFERENCES

Cohen J (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement 20*, 37-46. DOI: 10.1177/001316446002000104.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreemet or parcial credit. *Psychological Bulletin 70*, 213-220.

Dillon WR and Mulani N (1984). A probabilistic latent class model for assessing inter-judge reliability. *Multivariate Behavioral Research 19*, 438-458. DOI: 10.1207/s15327906mbr1904_5.

Fleiss JL, Cohen J and Everitt BS (1969). Large Sample Standard Errors of Kappa and Weighted Kappa. *Psychological Bulletin 72*, 323-327. DOI: 10.1037/h0028106.

Hubert L (1977). Kappa revisited. *Psychological Bulletin 48(2)*, 289-297. DOI: 10.1037/0033-2909.84.2.289.

Lin, L.I-K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45, 255-268.

Martín Andrés, A. and Álvarez Hernández, M. (2020). Hubert's multi-rater kappa revisited. *British Journal of Mathematical and Statistical Psychology 73 (1)*, 1-22. DOI: 10.1111/bmsp.12167.

Mielke PW Jr., Berry KJ and Johnston, JE (2007). The exact variance of weighted kappa with multiple raters. *Psychological Reports 101*, 655-660. DOI: 10.2466/PRO.l01.2.655-660.

Pearson, K. (1901). Mathematical distributions to the theory of evolution. *Philosophical Transactions of the Royal Society of London, Series A, 197,* 385-497.

Schuster C and Smith DA (2005). Dispersion-weighted kappa: an integrative framework for metric and nominal scale agreement coefficients. *Psychometrika 70(1)*, 135-146.

Table 1

Cognitive response cross-classification of *n*=164 subjects by *R*=3 raters in *K*=3 categories (Dillon and Mulani, 1984, p.449)

(a) Absolute frequencies $x_{i_1 i_2 i_3}$. The relative frequencies are $\hat{p}_{i_1 i_2 i_3} = x_{i_1 i_2 i_3} / n$

Rater 3		*1*			*2*			*3*		
Rater 2		*1*	*2*	*3*	*1*	*2*	*3*	*1*	*2*	*3*
Rater 1	*1*	56	1	0	5	3	0	0	0	1
	2	12	2	1	14	20	4	0	4	2
	3	1	1	0	2	1	7	2	1	24

(b) Data needed to estimate κ_H, which are obtained from the table (a)

Categories (*i*)	*Agreements* $n\hat{p}_i$	*Número de respuestas i del rater r* $n\hat{t}_{i(r)}$		
		Rater=1	*Rater=2*	*Rater=3*
1	56	66	92	74
2	20	59	33	56
3	24	39	39	34
Totals	$n\Sigma\hat{p}_i$=100	*n*=164	*n*=164	*n*=164

(c) Data needed to estimate $V(\hat{\kappa}_H)$, which are obtained from the table (b)

Categories (*i*)	$n^2\hat{T}_{i(r)}$			$n^2\sum_r \hat{T}_{i(r)}$
	Rater=1	*Rater=2*	*Rater=3*	
1	92×74=6,808	66×74=4,884	66×92=6,072	17,764
2	33×56=1,848	59×56=3,304	59×33=1,947	7,099
3	39×34=1,326	39×34=1,326	39×39=1,521	4,173