COMPARING PARTITIONS

(to assess reliability, validity, replicability)

REFERENCES:

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, 66*, 846-850.

Fowlkes, E. B., & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association, 78,* 553-569 .

~~Morey, L., & Agresti, A. (1984). The measurement of classification agreement: An adjustment to the Rand statistic for chance agreement. *Educational and Psychological Measurement, 44*, 33-37.~~

Hubert, L.J., and Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*, 193-218.

Warrens, M. J. (2008). On the equivalence of Cohen's Kappa and the Hubert-Arabie Adjusted Rand Index. *Journal of Classification, 25*, 177-183.

In order to compare two partitions, we first construct the “matching table”:

Notation for Comparing Two Partitions

		Partition V				
	Class	v_1	v_2	...	v_C	Sums
	u_1	n_{11}	n_{12}	...	n_{1C}	$n_{1\cdot}$
	u_2	n_{21}	n_{22}		n_{2C}	$n_{2\cdot}$
	.	.	.		.	.
Partition U	.	.	.		.	.
	.	.	.		.	.
	u_R	n_{R1}	n_{R2}	...	n_{RC}	$n_{R\cdot}$
	Sums	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot C}$	$n_{\cdot\cdot} = n$

The Rand Index

Example: from Rand (1971), also discussed by Morey & Agresti, Hubert & Arabie)

Suppose we have two partitions of the same set of objects:

U = (a b c) (d e f)   V = (a b) (c d e) (f)

"matching table":

		Partition V			
		B_1	B_2	B_3	
Partition U	A_1	2	1	0	3
	A_2	0	2	1	3
		2	3	1	

Let A = the number of "agreements", i.e. the number of object pairs that are either in the same cluster or in different clusters in BOTH partitions. Then,

Rand Index = $A / \binom{n}{2}$ = 5/15 = 1/3 = .333

Values of the Rand index range from 0 (perfect disagreement) to 1 (perfect agreement)

Hubert & Arabie's (1985) adjustment of Rand index for chance agreement

Thus, using the general form of an index corrected for chance:

$$\frac{\text{Index} - \text{Expected Index}}{\text{Maximum Index} - \text{Expected Index}}\,, \tag{4}$$

which is bounded above by 1 and takes on the value 0 when the index equals its expected value, the corrected Rand index would have the form (assuming a maximum Rand index of 1):

$$\frac{\sum_{i,j}\binom{n_{ij}}{2} - \sum_i\binom{n_{i\cdot}}{2}\sum_j\binom{n_{\cdot j}}{2} \Big/ \binom{n}{2}}{\frac{1}{2}\left[\sum_i\binom{n_{i\cdot}}{2} + \sum_j\binom{n_{\cdot j}}{2}\right] - \sum_i\binom{n_{i\cdot}}{2}\sum_j\binom{n_{\cdot j}}{2} \Big/ \binom{n}{2}}\,. \tag{5}$$

1 = perfect agreement, 0 = chance-level agreement, <0 = indicates less than chance agreement

For the example, Hubert & Arabie's measure can be computed as:

$$\frac{\frac{9}{15} - \frac{8\frac{1}{5}}{15}}{1 - \frac{8\frac{1}{5}}{15}} = \frac{2}{17} = .118$$

Hubert & Arabie (1985) – cont.

- H&A also proposed a measure based on the comparison of *object triples*, which has the advantage of a probabilistic interpretation in addition to being corrected for chance (i.e., assuming a constant value under a reasonable null hypothesis)
- "computational formula":

$$\text{Con} - \text{Dis} = 2\left[(n-1)\sum_{i,j} n_{ij}(n_{ij}-1) - \sum_{i,j}(n_{i.}-1)(n_{.j}-1)\,n_{ij}\right] .$$

- Finally, H&A proposed 4 different possible ways of normalizing the above measure so that it is bounded between -1 and +1, and discuss the advantages and disadvantages of each
- This measure has not been widely adopted

Warrens (2008) showed that H&A's adjusted Rand index (equation 5) is related to Cohen's Kappa:

If we represent in a 2x2 table the number of agreements and disagreements between two partitions (i.e. the number of <u>object pairs</u> that are grouped similarly or differently in the two partitions), then the HA adjusted Rand index can be seen to equal Cohen's K:

Table 1. 2×2 Contingency Table Representation of a Matching Table $\mathcal{M}$.

	Second partition		
First partition	Pair in same cluster	Pair in different cluster	Total
Pair in same cluster	a	b	p_1
Pair in different cluster	c	d	q_1
Total	p_2	q_2	N

H&A adjusted Rand index = K $= \dfrac{2(ad - bc)}{p_1 q_2 + p_2 q_1}.$

(Warrens also points out that the Rand index is equivalent to the simple matching coefficient for this table, = (a+d)/N. Note that N = number of <u>object pairs</u>, A (as defined by Rand) = (a+d).

Data Used by Morey and Agresti (1984) from Rand (1971)

		Partition V			
		B_1	B_2	B_3	
Partition U	A_1	2	1	0	3
	A_2	0	2	1	3
		2	3	1	

	V:	same	different	
U:	same	2	4	6
	different	2	7	9
		4	11	15

H&A Adj. Rand Index = K $= \dfrac{2(ad-bc)}{p_1q_2+p_2q_1}$. = $\dfrac{2[(2)(7)-(4)(2)]}{[(6)(11)+(4)(9)]}$ = 12/ 98 = .122

Recall: 1 = perfect agreement, 0 = chance-level agreement, <0 = indicates less than chance agreement