Cohen's Kappa

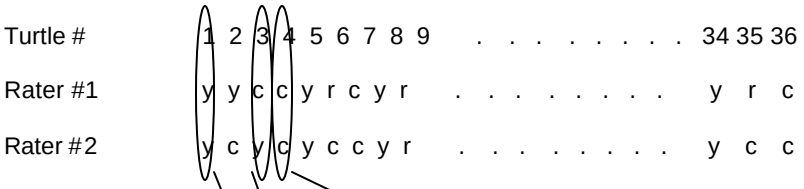
Index of Inter-rater Reliability

Application: This statistic is used to assess inter-rater reliability when observing or otherwise coding qualitative/ categorical variables. Kappa is considered to be an improvement over using % agreement to evaluate this type of reliability.

H0: Kappa is not an inferential statistical test, and so there is no H0:

Interpreting Kappa: Kappa has a range from 0-1.00, with larger values indicating better reliability. Generally, a Kappa > .70 is considered satisfactory.

The data: The research required that the species of each juvenile turtle that was being observed be identified. It can be difficult to correctly discriminate between juvenile Florida Yellow-bellied turtles, Florida Red-bellied turtles, and River Cooters. Working with videotapes of the target behaviors, two raters identified the species of each turtle. Kappa will be used to assess the inter-rater reliability of this identification process. The species will be abbreviated Yellow-bellied = "y", Red-bellied = "r", and Cooters = "c".



Step 1 Organize the scores into a contingency table. Since the variable being rated has three categories, the contingency table will be a 3x3 table.

Rater #2	"y"	"r"	"c"
	Rater #1		
"y"			
"r"			
"c"			

The ratings of each of the 35 turtles will be entered in this contingency table. Agreements between the two raters will be placed in one of the diagonal cells. For example, both raters identified turtle #1 as a yellow-belly, so we would tally one into the upper-left diagonal cell. As another example, turtle #4 was identified as a river cooter by both raters, and so we would tally one into the lower-right diagonal cell.

Disagreements between the raters will be placed in one of the off-diagonal cells. For example, Rater #1 thought turtle #2 was a yellow-belly but rater #2 thought it was a cooter, so we would tally one into the middle cell of the left-hand column. Contrast that with turtle #3, which rater #1 thought was a cooter, but rater #2 thought was a yellow-belly, so it was tallied into the right-hand column of the top row.

Below is the result of tallying the ratings of each turtle by each rater.

Rater #2	"y"	"r"	"c"
	Rater #1		
"y"	9	3	1
"r"	4	8	2
"c"	2	1	6

Step 2 Compute the row totals (sum across the values on the same row) and column totals (sum across the values on the same column) of the observed frequencies.

Rater #2	Rater #1 "y"	"r"	"c"	row totals
"y"	9	3	1	13
"r"	4	8	2	14
"c"	2	1	6	9
				overall total
column totals	15	12	9	36

Step 3 Compute the overall total (shown in the table above). As a computational check, be sure that the row totals and the column totals sum to the same value for the overall total, and that the overall total matches the number of cases in the original data set.

Step 4 Compute the total number of agreements by summing the values in the diagonal cells of the table.

$$\Sigma a = 9 + 8 + 6 = 23$$

Based on this, the % agreement would be 23/36 = 64%. However, this value is an inflated index of agreement, because it does not take into account the agreements that would have agreed by chance.

Step 5 Compute the expected frequency for the number of agreements that would have been expected by chance for each coding category. This is done using the same formula as for computing expected frequencies for Pearson's X^2, but now the formula is applied only to the diagonal cells. Computation of the expected frequency of agreements by chance for the yellow-bellies is shown. Below that is the contingency table with the expected frequencies in each of the diagonal cells shown in parentheses.

$$ef = \frac{\text{row total} * \text{col total}}{\text{overall total}} = \frac{13 * 15}{36} = 5.42$$

Rater #2	Rater #1 "y"	"r"	"c"
"y"	9 (5.42)	3	1
"r"	4	8 (4.67)	2
"c"	2	1	6 (2.25)

Step 6 Compute the sum of the expected frequencies of agreement by chance.

$$\Sigma ef = 5.42 + 4.67 + 2.25 = 12.34$$

Step 7 Compute Kappa

$$K = \frac{\Sigma a - \Sigma ef}{N - \Sigma ef} = \frac{23 - 12.34}{36 - 12.34} = .45$$

Step 8 Evaluate Kappa

-- if the obtained K is less than .70 -- conclude that the inter-rater reliability is not satisfactory.

-- if the obtained K is greater than .70 -- conclude that the inter-rater reliability is satisfactory

For the example data, we would conclude that the inter-rater reliability is not satisfactory, because the obtained Kappa of .45 is less than the commonly applied criteria of .70.

Step 9 Consider the pattern of disagreements for possible ways to focus efforts to improve either the operational definitions upon which the ratings or based, or the training and accuracy of the raters.

For these data, notice that there are more disagreements between red-bellies and yellow-bellies than between either of these species and the river cooters. Thus, we might suggest that re-training would focus on correctly discriminating between these two types of turtles, in order to improve inter-rater reliability.