Interrater Reliability: Cohen's Kappa Statistic (κ)

The Cohen's kappa coefficient, denoted by **κ**, is a statistic that measures the interrater agreement factor of qualitative items in categorical form. It is one of many different approaches to calculating and classifying the "amount of agreement" between two coders.

Calculation

While the calculation can be performed using straightforward linear equation, calculating its differing parts and then combining is often a simpler process.

Setup. The setup of two rater's agreement, one-sided agreement, or disagreement is relatively simple when viewed in tabular form. It is given by an agreement table called a confusion matrix (similar to a Punnett square):

κ		1st Rater/Coder	
		Yes	**No**
2nd Rater/Coder	**Yes**	a	b
	No	c	d

where a count of each agreement, one-sided agreement, and disagreements exist in the corresponding boxes.

Equation. There are four equations to consider when calculating Cohen's Kappa. They are:

Name of Component	Symbol(s)	Equation
Observed agreement	p_0 OR p_{agree}	$\frac{a+d}{a+b+c+d}$
Expected probability of both rater's saying *yes*	p_{yes}	$\frac{a+b}{a+b+c+d} \cdot \frac{a+c}{a+b+c+d}$
Expected probability of both rater's saying *no*	p_{no}	$\frac{c+d}{a+b+c+d} \cdot \frac{b+d}{a+b+c+d}$
Random agreement probability	p_e OR p_{random}	$p_{\text{yes}} + p_{\text{no}}$

Calculation. The calculation of Cohen's Kappa is a combination of the four components and is given by:

$$\kappa = \frac{p_0 - p_e}{1 - p_e}$$

Determining Level. One an output has been determined, the last step is to calculate the *accepted level of agreement*. It is worth noting that to date there **no evidence** to support these standards and they are not universally accepted so threats to statistical validity are always an issue. The list below is simply the one that is commonly used.

Kappa (κ)	**Level of Agreement**
Less than 0.00	Less than chance
0.01 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 0.99	Almost perfect
1.00	Perfect

An example

The Kappa statistic (or value) is a metric that compares an **Observed Accuracy** with an **Expected Accuracy** (random chance). The kappa statistic is used not only to evaluate a single agreement, but also to evaluate agreements amongst themselves. In addition, it takes into account random chance (agreement with a random agreement), which generally means it is less misleading than simply using accuracy as a metric (an **Observed Accuracy** of 80% is a lot less impressive with an **Expected Accuracy** of 75% versus an **Expected Accuracy** of 50%). Computation of **Observed Accuracy** and **Expected Accuracy** is integral to comprehension of the Kappa statistic and is most easily illustrated through use of a confusion matrix.

Let's say **two humans** compiled all of their themes and then begin to discuss them. In this process, they simply took the list from SH and stacked it on top of the list from AH. Then they went through the aggreged list individually and independently and then came together to discuss agreement.

Now the agreement process begins. They start with a confusion matrix from a simple binary classification of agreement between both:

K		Some human (SH)	
		Yes	**No**
Hopefully another human (AH)	**Yes**	10	7
	No	5	8

Step 1: Calculate the occurrences.

From the confusion matrix we can see there are **30** instances of agreement/disagreement in total (10 + 7 + 5 + 8 = 30). According to the first column,

- Column 1: there were **15** occurrences where SH looked at a potential theme and agree that it should be one indicating **Yes** (10 + 5 = 15).
- Column 2: there were **15** occurrences where SH looked at a potential theme and agree that it should NOT be one indicating **No (**7 + 8 = 15).
- Row 1: there were **17** occurrences where AH looked at a potential theme and agree that it should be one indicating **Yes** (10 + 7 = 17).
- Row 2: there were **13** occurrences where AH looked at a potential theme and agree that it should NOT be one indicating **No** (5 + 8 = 13).

Step 2: Calculate the Agreements

The **Observed Agreement** is the number of instances that were classified as **Yes or No by both individuals**, i.e. the number of instances that SH agreed that a potential theme was a theme/not a theme and SH agreed the same way divided by the total number of instances. For this confusion matrix, this would be

$$p_0 = \frac{10 + 8}{30} = 0.6$$

The **Expected Probability of a Yes** is the number of instances that were classified as **Yes by both individuals BUT not necessarily agreed between both people**, i.e. the number of instances that SH agreed that a potential theme was a theme and the number of instances that SH agreed that a potential theme was a theme divided by the total number of instances. For this confusion matrix, this would be

$$p_{yes} = \frac{10 + 7}{30} \cdot \frac{10 + 5}{30} \approx 0.3$$

The **Expected Probability of a No** is the number of instances that were classified as **No by both individuals BUT not necessarily agreed between both people**, i.e. the number of instances that SH agreed that a potential theme was a theme and the number of instances that SH agreed that a potential theme was a theme divided by the total number of instances. For this confusion matrix, this would be

$$p_{no} = \frac{5 + 8}{30} \cdot \frac{7 + 8}{30} \approx 0.2$$

The **Random Probability Agreement** is a sum of the **Expected Probability of a Yes** and the **Expected Probability of a No**.

$$p_e = 0.3 + 0.2 = 0.5$$

Step 3: Calculate the Statistic

Here we simply use the Kappa statistic formula.

$$\boldsymbol{\kappa} = \frac{p_0 - p_e}{1 - p_e} = \frac{0.6 - 0.5}{1 - 0.5} = 0.2$$

Step 4: Interpret the Score

In accordance to the score level explanation, the two individuals have a slight agreement ($\boldsymbol{\kappa = 0.2}$).

What if you have more than two people on a team? Well there are many complex approaches to finding a multi-dimensional agreement score but by far the easiest is to do it twice or thrice. Consider the following:

Groups of three people.

- Step 1: Pick two group members (say P1 and P2) to perform the agreement as outlined above. Remember that each agreement rating must be performed independently or else you introduce bias.
- Step 2: Calculate the individual agreement scores for each theme as well as the aggregated Kappa statistic.
- Step 3: Hide the scores found in step 2.
- Step 4: Send it to the third individual (P3) for their rating.
- Step 5: After finishing, use the hidden scores and find the means. Use the new averages as your final tally.
- Step 6: Adjust the final list of themes based on the final scores in Step 5.

Groups of four people.

- Step 1: Pick two group members (say P1 and P2 that make up a group GA) to perform the agreement as outlined above. Remember that each agreement rating must be performed independently or else you introduce bias.
- Step 2: Have the other two group members (P3 and P4 that make up a group GB) to perform the agreement as outlined above. Again, remember that each agreement rating must be performed independently or else you introduce bias.
- Step 3: Calculate the individual agreement scores for each theme as well as the aggregated Kappa statistic by group. These will be group wise scores (so one for GA and one for GB).
- Step 4: Have each group hide the scores found in step 3.
- Step 5: Exchange your lists so GA sends their list to GB and at the same time, GB sends their list to GA.
- Step 6: After finishing, use the hidden scores and find the means. Use the new averages as your final tally.
- Step 7: Adjust the final list of themes based on the final scores in Step 6.