Cohen's kappa

December 19, 2024

In many instances, calculating the accuracy of the results of some experiment might not be the most ideal choice. After all, knowing whether a set of observations is accurate depends on how well we are able to determine state of the variables of the experiment. In many cases, such as medical/clinical questionnaires, the result might be somewhat subjective. For example, if a psychiatrist were to measure whether a patient is depressed or not might ask the patient to fill out a questionnaire. Now, there must be certain guidelines to determine whether or not the patient is depressed depending on the answers. However, psychiatrists might also use their subjective judgement in order to arrive at a conclusion. Hence, there is a flexibility in the way they arrive at their conclusion.

Suppose 50 people were asked to fill-out a questionnaire that is aimed at determining whether or not they suffer from depression and that two clinicians rate their responses as 'Yes' or 'No'. In effect, based on the features (the questions in the questionnaire) the output is a nominal variable. In a situation like this, the accuracy of the test result might be difficult to gauge and what might be more important is to establish whether or not the two clinicians agree with their independent assessments. This gives rise to the concept of *inter-rater reliability*. Here we talk about one such metric called Cohen's kappa and its two variants – unweighted and quadratically weighted kappas.

1 Cohen's kappa: unweighted

Let us assume that results from the two clinicians, called 'raters', arrive at the results given in Table 1. We observe that the raters agree of 34 out of

		Rater 2	
		No	Yes
Rater 1	No	17	8
	Yes	6	19

Table 1: Diagnostic results from raters 1 and 2.

50 diagnoses. Hence, we can calculate the probability of agreement from the observed data as

$$p_o = \frac{36}{50} = 0.72 \tag{1}$$

However, agreements in the results can also occur due to random chance, which we denote as p_e. With this, Cohen's kappa is defined as

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{2}$$

To calculate p_e, we observe that when raters 1 and 2 conclude 'No' 25 and 23 times respectively. Then

$$p_e^{(\text{No})} = \frac{25}{50} \times \frac{23}{50} = 0.23 \tag{3}$$

Similarly, for 'Yes', we have

$$p_e^{(\text{Yes})} = \frac{25}{50} \times \frac{27}{50} = 0.27 \tag{4}$$

Hence, the total probability of agreement purely through random chance is $p_e = p_e^{\text{No}} + p_e^{\text{Yes}} = 0.5$. This gives us a Cohen's kappa of $\kappa = 0.44$.

The unweighted Cohen's kappa, thus, gives us a measure of inter-rater reliability (for two raters) for nominal variables, *i.e.*, variables that are categorical and have no intrinsic ordering. The Cohen's kappa is a number in the range $-1 \leq \kappa \leq 1$, with the following interpretations of agreement

- $\kappa < 0$ (poor)
- $0 \leq \kappa \leq 0.2$ (slight)
- $0.2 < \kappa \leq 0.4$ (fair)
- $0.4 < \kappa \leq 0.6$ (moderate)
- $0.6 < \kappa \leq 0.8$ (substantial)
- $\kappa > 0.8$ (almost perfect)

2 Weighted Cohen's kappa

The unweighted Cohen's kappa is used for nominal variables. However, categorical variables may have ordering between between. This ordering is not metric by nature, but, rather, arises from an interpretational aspect. For example, when talking about education level, there is an ordering between school, college and grad school. Similarly, there is an order between responses 'bad', 'average' and 'good'. For such cases, we use what is called the weighted Cohen's kappa κ_w.

If we have a collection of n ordinal variables such that $\mathcal{O}_1 < \mathcal{O}_2 < \cdots < \mathcal{O}_n$ such that their ratings are given by the following frequency matrix

$$\begin{pmatrix} f_{11} & f_{12} & \cdots & f_{1n} \\ f_{21} & f_{22} & \cdots & f_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n1} & f_{n2} & \cdots & f_{nn} \end{pmatrix} \tag{5}$$

the weighted kappa is defined as

$$\kappa_w = 1 - \frac{\sum_{i,j} w_{ij} f_{o,ij}}{\sum_{i,j} w_{ij} f_{e,ij}} \tag{6}$$

where $f_{o,ij}$ and $f_{e,ij}$ are the observed and expected frequencies of the outcomes and w_{ij} are elements of the 'weight matrix'. It is easy to check that w_{ij} for the unweighted case is

$$w_{ij} = 1 - \delta_{ij} = \begin{cases} 0, & i = j \\ 1, & i \neq j \end{cases} \tag{7}$$

In matrix form, this is

$$w_{ij} \longrightarrow \begin{pmatrix} 0 & 1 & 1 & \cdots & 1 \\ 1 & 0 & 1 & \cdots & 1 \\ 1 & 1 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & 0 \end{pmatrix} \tag{8}$$

For ordinal variables, we would like to take into account how far apart they are. This is done by consider a nontrivial weighting. For a quadratically

weighted kappa, the weighting matrix is

$$w_{ij} = \left(\frac{i-j}{n-1}\right)^2 \tag{9}$$

where n is the number of ordinal variables.