The University Of Sheffield.

School of Health and Related Research

The Disagreeable Behaviour of the Kappa Statistic

Laura Flight and Steven. A. Julious
E-mail: l.flight@sheffield.ac.uk

Design, Trials & Statistics

Introduction

It is often of interest to measure agreement between a number of raters when an outcome is ordinal. The kappa statistic is a frequently used tool that measures agreement. This statistic however has a number of limitations. The motivating example for this work is used to illustrate some of the disagreeable properties of the kappa statistic.

Cohen's Kappa Statistic

Cohen (1960) proposed the kappa statistic in the context of two raters classifying objects into two categories.

The statistic adjusts the observed agreement (p_o) by agreement expected by chance alone (p_e).

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

A common scale of interpretation for the kappa statistic is given in Table 1 (Altman 1991):

Kappa	Agreement
<0.20	Poor
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Good
0.81-1.00	Very Good

Table 1: Interpretations of kappa

Definitions

The following definitions are used to describe the limitations of the kappa statistic:

- **Symmetrical** – the distribution across g_1 and g_2 is the same as f_1 and f_2 (see Table 2)
- **Asymmetrical** - the distribution across g_1 and g_2 is in the opposite direction to f_1 and f_2
- **Balanced** – the proportion of the total number of objects in g_1 and f_1 is equal to 0.5
- **Imbalance** - the proportion of the total number of objects in g_1 and f_1 is not equal to 0.5
- **Prevalence** – probability with which a rater will classify an object into a category. This is related to the balance of the table
- **Bias** – frequency at which raters choose a particular category. This is related to the symmetry of the table

Paradoxes

Feinstein and Cicchetti (1990) highlight the following issues termed 'paradoxes' of the kappa statistic:

1. For high values of concordance low values of kappa can be recorded.
2. Asymmetric, imperfectly imbalanced tables have a higher kappa than perfectly imbalanced and symmetric tables.

They show that p_e is dependent on the distribution of the marginal totals and consequently kappa is highly sensitive to this.

Byrt (1993) describes kappa as being affected by both prevalence and bias, proposing a **Prevalence and Bias Adjusted Kappa** (PABAK) to over come some of the issues.

$$PABAK = 2p_o - 1$$

Motivational Example

- In the motivational example there are $N = 261$ students categorised by two independent assessors as either 'one' or 'two' in Table 2
- It is important to establish whether there is agreement between the two assessors, as if agreement is poor an additional assessor will be required

		Assessor One		
		1	**2**	**Total**
Assessor Two	**1**	$a = 171$	$b = 72$	g_1 =243
	2	$c = 11$	$d = 7$	g_2 =18
	Total	f_1 =182	f_2= 79	$N = 261$

Table 2: Data from two assessors categorising students into two categories

- For the original data p_o= 0.682 (good agreement) where as $\kappa = 0.038$ (poor agreement) giving conflicting interpretations
- The table is a **symmetrically imbalanced** table with a high **prevalence** for each assessor categorising a student as one. The proportion of students in $g_1 = 0.931$

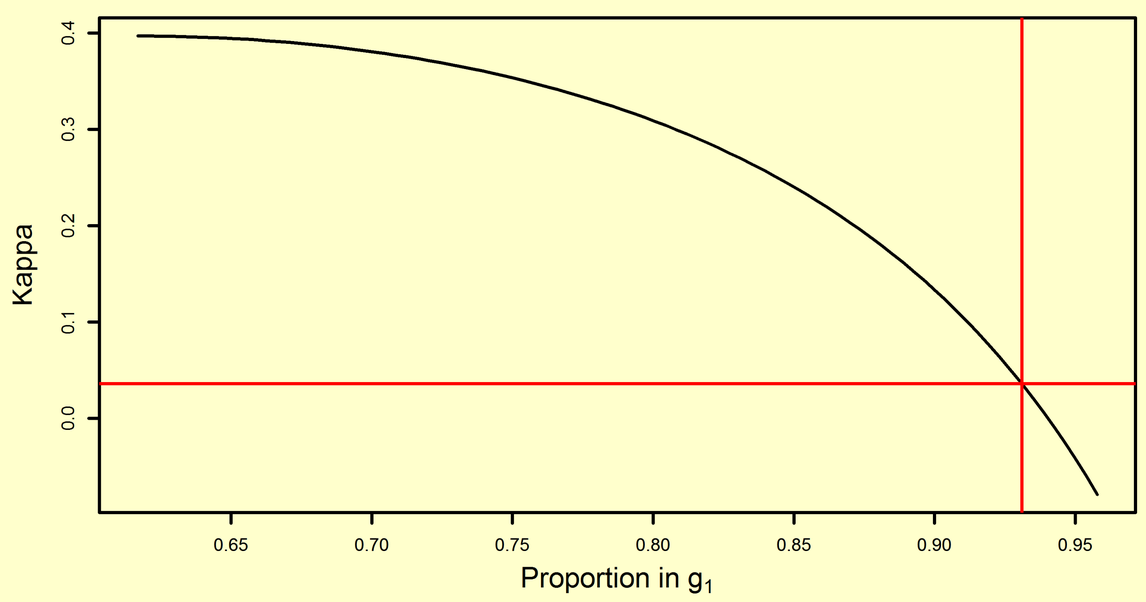

Figure 1: Kappa against proportion in marginal g_1

- Paradox one is occurring with high values of concordance but low kappa
- Figure 1 illustrates as the proportion of students in the marginal g_1 increases kappa decreases rapidly, even falling below zero when the proportion in g_1 reaches over 90%
- $PABAK = 0.364$ gives a more comparable interpretation to the proportion of concordance

Conclusions

- Kappa statistic is sensitive to the distribution of the marginal totals limiting its usefulness
- Kappa from different data are not comparable and a generic scale of interpretation should be used with this in mind
- It is recommended to report other statistics alongside the kappa statistic such as maximum attainable kappa and PABAK
- Kappa should only be considered and interpreted based on the context in hand

References

1. Cohen. J. A Coefficient of Agreement for Nominal Scales *Educational and Psychological Measurement* 1960; 20: 37.
2. Altman. D.G. Practical statistics for medical research. London (1991) : Chapman and Hall.
3. Byrt. T, Bishop. J, Carlin. J.B. Bias, Prevalence and Kappa. *Journal of Clinical Epidemiology.* 1993; 46 (5), 423-429.
4. Feinstein. A.R, Cicchetti. D.V. High agreement but low kappa: I. The Problems of Two Paradoxes. *Journal of Clinical Epidemiology* 1990; 43: 543-548.

This is a summary of independent research funded by the National Institute for Health Research (NIHR)'s Research Methods Fellowship Programme. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

NHS National Institute for Health Research