THE POWER OF KAPPA

A. Blance[1], M. Gilthorpe[1]

[1] *University of Leeds, Leeds, United Kingdom*

Email: *a.blance@leeds.ac.uk*

Cohen's Kappa statistic is commonly used to assess levels of agreement for categorical observations. Kappa has issues pertaining to its calculation that subsequently lead to difficulties in its interpretation. However, many researchers are unaware of this. Often the reason for assessing the level of agreement is so that the observer effect can be ignored in subsequent analyses. In this situation, the standard hypothesis test is erroneous and Kappa should be assessed for non-inferiority to one. This study therefore seeks to provide a summary of these issues and to illustrate that large sample sizes are required to estimate Kappa robustly. A theoretical illustration of the issues that cause difficulty in interpreting Kappa is provided. Simulations are used to create a census mapping of 2x2 Kappa scenarios. For those scenarios where the estimated value of Kappa is deemed to illustrate acceptable levels of agreement, the sample size required to assess non-inferiority to one robustly is calculated. The minimum sample size of the most favourable scenario required for robust calculation of Kappa is 250 pairs of observations, though typically the number of observations required is at least one order of magnitude greater. Kappa should not be blindly assessed against one-size-fits-all criterion. Sample sizes required to estimate Kappa robustly are typically very large. Furthermore, research that ignores the observer effect in analyses may lead to erroneous conclusions being drawn.