A Less Overconservative Method for Reliability Estimation for Cohen's Kappa

Matt He[1], Ryan Baker[2], Stephen Hutt[3], and Jiayi Zhang[2]

[1] Northfield Mount Hermon, Gill, MA 01354, USA
[2] University of Pennsylvania, Philadelphia, PA, 19104, USA
[3] University of Denver, Denver, CO, 80204, USA
email@address.com

Abstract. Cohen's Kappa has been used in interrater reliability calculation for decades, often for small samples. Recently, QE researchers have argued that Kappa cannot validly be used without much larger samples based on very conservative assumptions: treating all degrees of error as equally problematic and conducting an analysis analogous to statistical power analysis using a statistical significance criterion. We present a Monte Carlo analysis assessesing interrater reliability based on distance between the population Kappa and threshold Kappa (i.e., the degree of error), for a range of population Kappa values, threshold Kappa values, and sample sizes. Our findings indicate that Kappa can reasonably be used at the sample sizes often used in practice, either by raising threshold Kappa or by adopting the same stringency as statistical power analysis.

Keywords: Cohen's Kappa, Sample Size Calculation, Monte Carlo Analysis.

1 Introduction

Much quantitative ethonography research relies upon human-coded data, validated by human coders separately coding the same set of examples and checking agreement. The most frequent validation metric is Cohen's Kappa [3], which compares the actual degree of agreement to a base rate that could be expected by chance. However, there is no agreed way to compute standard error [4], making sample size calculation difficult.

Recent work in the quantitative ethnography community argues that Kappa should not be used except with very large sample sizes [1,2]. In this paper, we offer a critique of these recommendations, comparing that approach's stringency and assumptions to power analysis. We propose an alternative analysis, aligning more closely to the assumptions and degree of conservatism of statistical power analysis. We use this analysis to suggest a way for selecting appropriate sample sizes for the use of Kappa.

1.1 An Examination of the Methods in Eagan et al.

Eagan and colleagues [1, 2] presented Monte Carlo analyses testing whether a sample's Kappa is higher than the full population's Kappa. They simulated large sets of codes

and repeatedly sampled from that data set. In each simulation, they specified a simulated coder base rate, and a sample size. True Kappa values (for the full population) varied 0.3-1.0 [1]. A target threshold Kappa was then selected -- 0.65 in [2] and 0.7 in [1]. Among each set of iterations, Eagan and colleagues counted the proportion of cases where population Kappa was below threshold, and sample Kappa was above threshold. They then argued that a sample size must have an error rate under 0.05 for valid use. Within this method, there are thus two steps in evaluating performance across a set of simulations: first, determining how often population Kappa is below threshold and sample Kappa is above threshold. Second, determining if this proportion is above 5%.

Note that in the first step, a coding scheme is treated as invalid if population Kappa is barely above threshold (0.69) and sample Kappa is barely above threshold (.71), treating this case the same as if population Kappa is 0.30 and sample Kappa is 0.71, while treating large differences in Kappa as acceptable if both are below threshold.

We can compare each step to statistical power analysis. For step 1, the statistical test most analogous to the case evaluation procedure in [1, 2] is the Wilcoxon rank-sum test. A statistically significant value can be obtained for Wilcoxon without requiring that fewer than 5% of comparisons be in favor of the lower-valued sample; as such, this first element of [1, 2] is much more stringent than statistical significance testing. For step 2, statistical power analysis typically looks for whether a significant result is seen at least 80% of the time (i.e. a failure mode occurs < 20% of the time). By contrast, Eagan et al. [1, 2] look for whether a failure mode occurs < 5% of the time. As such, the second step of Eagan et al.'s procedure is four times as stringent as power analysis.

Eagan and colleagues argue that Kappa produces erroneous results more than 5% of the time for sample sizes under 400 [2] or 2000 [1], depending on base rate. They therefore argue that the common practice of testing inter-rater reliability using Kappa on samples typically much smaller than these values is flawed and should be abandoned.

In the following sections, we propose a method that more explicitly considers the degree of difference between population Kappa and sample Kappa. We also consider the implications of using a second-step stringency criterion in line with statistical power analysis rather than statistical significance testing.

2 This Paper's Methods

Within this paper, analyze the risks of sample Kappa value over threshold when true population Kappa is under threshold, attempting to achieve a level of conservatism closer to statistical power analysis. Our overall process is similar in structure to [1, 2]. First, we create a simulated population of 1M codes; then we sample from that data set; finally, we test whether that simulated data set represents a false positive.

Each simulation run uses three parameters: a sample size, a threshold Kappa (false positives have Kappa over threshold), and population Kappa, selected in relation to the threshold Kappa. For example, we might select threshold Kappa of 0.65 (as in [2]) and population Kappa 0.2 less than the threshold, making the population Kappa 0.45.

We then repeatedly (100K iterations) sample random data points from the population for the preselected sample size. In each iteration, we test whether the sample Kappa is above or below the threshold Kappa. Choosing both the population Kappa and threshold Kappa (in relation to each other) enables us to avoid treating small levels of variation as a false positive. We then calculate the proportion of time we have a sample Kappa above threshold, despite having population Kappa substantially below threshold.

Several sets of simulations were run. For threshold we used parameters of 0.6, 0.65, 0.7, 0.75, and 0.8; for population Kappa we used threshold (T)-0.05, T-0.1, T-0.2, and T-0.3; for sample size, we used 20, 40, 60, 80, 100, 200, 400, 500, 800, 1000, and 2000.

See [bit.ly/3wRhrt4] for software used in these simulations.

3 Findings

Having created these simulations, we can now check for the proportion of time a specific test produces a Kappa above threshold, despite having a lower population Kappa.

We consider first a sample size of 60 (Table 1) – a small dataset, but of a size seen in QE inter-rater reliability checks. Table 1 reports the proportion of samples with a Kappa value above threshold, when the true value (population Kappa) is some amount (or more) less. We note that for this sample size, across all thresholds, there is a high probability (~30%) that sample Kappa was more than .05 larger than population Kappa. Therefore, for this sample size, there is high risk that a sample Kappa value barely over threshold may represent a population Kappa value barely below threshold, regardless of what that threshold is. As we increase the distance between threshold Kappa and the population Kappa (from .1 to .3), the number of samples that meet the threshold drops, with less than 1% of samples achieving threshold Kappa .3 or more above population Kappa value. These results are fairly consistent between thresholds of .6 and .75 but there is lower error for a .8 threshold. Overall, if a researcher selects a level of conservatism comparable to power analysis (under 20% error), even a small sample of 60 data points is sufficient to be confident that a threshold is unlikely to represent population Kappa over 0.1 below threshold. If a researcher chooses conservatism comparable to statistical significance testing (5%), 60 data points is still sufficient to be confident that a threshold is unlikely to represent a population Kappa more than 0.2 below threshold.

Table 1. The proportion of cases where Population Kappa was more than specific distances (row) below Threshold Kappa (th) (cols), for a sample size of 60.

		Threshold Kappa (th)				
		0.6	**0.65**	**0.7**	**0.75**	**0.8**
Population Kappa	**th - 0.05**	0.316	0.316	0.274	0.300	0.242
	th - 0.1	0.175	0.175	0.139	0.151	0.107
	th - 0.2	0.035	0.035	0.024	0.025	0.014
	th - 0.3	0.005	0.035	0.003	0.002	0.001

We further investigate the impact of different sample sizes in Table 2, considering sample sizes ranging from 20 to 800. We note that very small differences between threshold Kappa and population Kappa can be achieved for large samples.

Table 2. The proportion of cases where Population Kappa was more than a certain distance (cols) below Threshold Kappa, for varying sample sizes (rows)

Sample Size	Population Kappa (Threshold Kappa)				
	0.55 (0.6)	**0.5 (0.6)**	**0.75 (0.8)**	**0.7 (0.8)**	**0.6 (0.8)**
20	0.399	0.307	0.292	0.196	0.081
40	0.350	0.231	0.265	0.144	0.035
60	0.317	0.175	0.243	0.108	0.015
80	0.292	0.145	0.218	0.082	0.007
100	0.269	0.117	0.197	0.061	0.003
200	0.193	0.047	0.126	0.016	<0.001
400	0.112	0.009	0.054	0.001	<0.001
800	0.042	<0.001	0.014	<0.001	<0.001

4 Conclusions

Overall, our Monte Carlo analyses show that the situation for Kappa is not quite so grim as Eagan et al. [1, 2] argue. Whereas they argued that Kappa could only be confidently used for samples of 400 [2] or 2000 [1] data points or higher, we find that only small differences are seen for much smaller samples. Our results indicate is that if we are willing to accept that a sample Kappa may be slightly higher than a population Kappa, fairly small sample sizes are needed to use Kappa with confidence. If we are willing for a threshold Kappa of 0.6 to actually represent a population Kappa of 0.501 5% of the time, then a sample of 200 is sufficient. If we are willing for a threshold Kappa of 0.8 to actually represent a population Kappa of 0.751 20% of the time, or for a threshold Kappa of 0.7 to actually represent a population Kappa of .601 10% of the time, then a sample of 100 is sufficient. Even a very small sample of 40 can be acceptable in some cases -- for instance, if we are willing for a threshold Kappa of 0.8 to actually represent a population Kappa of .701 20% of the time. In other words, our findings provide evidence that Kappa can be acceptable for many uses and assumptions, even with smaller sample sizes than our community typically uses. With a larger sample, our estimate of Kappa can be more precise. But the cost of this is more time spent in coding data for inter-rater checking. [1] argues that so much data must be coded to use Kappa that Kappa is essentially infeasible -- our findings suggest otherwise.

References

1. Eagan, B.R., Brohinsky, J., Wang, J., Shaffer, D.W.: Testing the reliability of inter-rater reliability. In: Proc. of LAK. pp. 454–461 (2020).
2. Eagan, B.R., Rogers, B., Serlin, R., Ruis, A.R., Irgens, G.A., Shaffer, D.W.: Can we rely on IRR? Testing the assumptions of inter-rater reliability. In: Proc CSCL (2017).

3. Cohen, J.: A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement. 20, 37–46 (1960). https://doi.org/10.1177/001316446002000104.
4. Rigby, A.S.: Statistical methods in epidemiology. v. Towards an understanding of the kappa coefficient. Disability and Rehabilitation. 22, 339–344 (2000).