Cohen's kappa

Cohen's kappa coefficient (κ) is a statistic which measures inter-rater agreement for qualitative (categorical) items. It is generally thought to be a more robust measure than simple percent agreement calculation, as κ takes into account the possibility of the agreement occurring by chance. There is controversy surrounding Cohen's kappa due to the difficulty in interpreting indices of agreement. Some researchers have suggested that it is conceptually simpler to evaluate disagreement between items.[1] See the Limitations section for more detail.

Contents

Calculation

Cohen's kappa measures the agreement between two raters who each classify *N* items into *C* mutually exclusive categories. The first mention of a kappa-like statistic is attributed to Galton (1892),[2] see Smeeton (1985).[3]

The definition of κ is:

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

where p_o is the relative observed agreement among raters (identical to accuracy), and p_e is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly seeing each category. If the raters are in complete agreement then $\kappa = 1$. If there is no agreement among the raters other than what would be expected by chance (as given by p_e), $\kappa = 0$. It is possible for the statistic to be negative,[4] which implies that there is no effective agreement between the two raters or the agreement is worse than random.

For categories k, number of items N and n_{ki} the number of times rater i predicted category k:

$$p_e = \frac{1}{N^2} \sum_k n_{k1} n_{k2}$$

The seminal paper introducing kappa as a new technique was published by Jacob Cohen in the journal *Educational and Psychological Measurement* in 1960.[5]

A similar statistic, calledpi, was proposed by Scott (1955).Cohen's kappa andScott's pi differ in terms of how p_e is calculated.

Note that Cohen's kappa measures agreement between **two** raters only. For a similar measure of agreement (Fleiss' kappa) used when there are more than two raters, see Fleiss (1971). The Fleiss kappa, however, is a multi-rater generalization of Scott's pi statistic, not Cohen's kappa. Kappa is also used to compare performance in machine learning, but the directional version known as Informedness or Youden's J statistic is argued to be more appropriate for supervised learning.[6]

Example

Suppose that you were analyzing data related to a group of 50 people applying for a grant. Each grant proposal was read by two readers and each reader either said "Yes" or "No" to the proposal. Suppose the disagreement count data were as follows, where A and B are readers, data on the main diagonal of the matrix (a and d) count the number of agreements and off-diagonal data (b and c) count the number of disagreements:

		B	
		Yes	No
A	Yes	a	b
	No	c	d

e.g.

		B	
		Yes	No
A	Yes	20	5
	No	10	15

The observed proportionate agreement is:

$$p_o = \frac{a+d}{a+b+c+d} = \frac{20+15}{50} = 0.7$$

To calculate p_e (the probability of random agreement) we note that:

- Reader A said "Yes" to 25 applicants and "No" to 25 applicants. Thus reader A said "Yes" 50% of the time.
- Reader B said "Yes" to 30 applicants and "No" to 20 applicants. Thus reader B said "Yes" 60% of the time.

So the expected probability that both would say yes at random is:

$$p_{\text{Yes}} = \frac{a+b}{a+b+c+d} \cdot \frac{a+c}{a+b+c+d} = 0.5 \times 0.6 = 0.3$$

Similarly:

$$p_{\text{No}} = \frac{c+d}{a+b+c+d} \cdot \frac{b+d}{a+b+c+d} = 0.5 \times 0.4 = 0.2$$

Overall random agreement probability is the probability that they agreed on either Yes or No, i.e.:

$$p_e = p_{\text{Yes}} + p_{\text{No}} = 0.3 + 0.2 = 0.5$$

So now applying our formula for Cohen's Kappa we get:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = \frac{0.7 - 0.5}{1 - 0.5} = 0.4$$

Same percentages but different numbers

A case sometimes considered to be a problem with Cohen's Kappa occurs when comparing the Kappa calculated for two pairs of raters with the two raters in each pair having the same percentage agreement but one pair give a similar number of ratings in each class while the other pair give a very different number of ratings in each class.[7] (In the cases below, notice B has 70 yeses and 30 nos, in the first case, but those numbers are reversed in the second.) For instance, in the following two cases there is equal agreement between A and B (60 out of 100 in both cases) in terms of agreement in each class, so we would expect the relative values of Cohen's Kappa to reflect this.However, calculating Cohen's Kappa for each:

		B	
		Yes	No
A	Yes	45	15
	No	25	15

$$\kappa = \frac{0.60 - 0.54}{1 - 0.54} = 0.1304$$

		B	
		Yes	No
A	Yes	25	35
	No	5	35

$$\kappa = \frac{0.60 - 0.46}{1 - 0.46} = 0.2593$$

we find that it shows greater similarity between A and B in the second case, compared to the first. This is because while the percentage agreement is the same, the percentage agreement that would occur 'by chance' is significantly higher in the first case (0.54 compared to 0.46).

Significance and magnitude

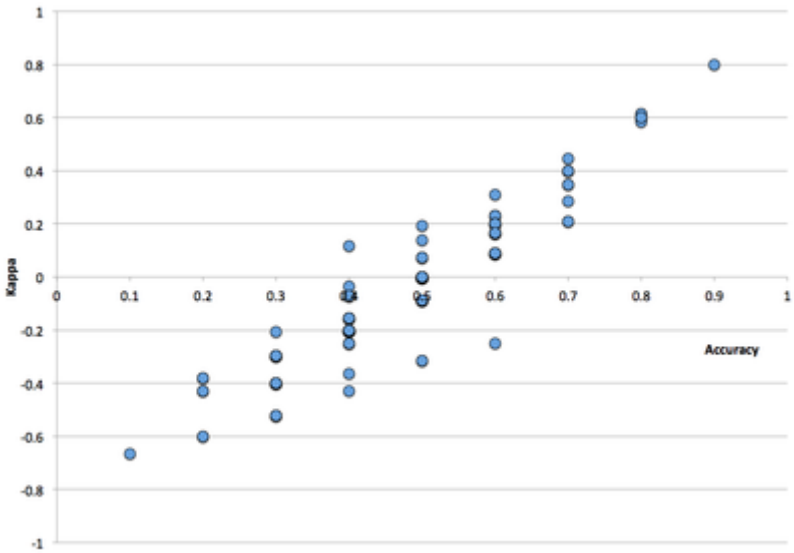
Statistical significance for kappa is rarely reported, probably because even relatively low values of kappa can nonetheless be significantly different from zero but not of sufficient magnitude to satisfy investigators.[8]:66 Still, its standard error has been described[9] and is computed by various computer programs.[10]

If statistical significance is not a useful guide, what magnitude of kappa reflects adequate agreement? Guidelines would be helpful, but factors other than agreement can influence its magnitude, which makes interpretation of a given magnitude problematic. As Sim and Wright noted, two important factors are prevalence (are the codes equiprobable or do their probabilities vary) and bias (are the marginal probabilities for the two observers similar or different). Other things being equal, kappas are higher when codes are equiprobable. On the other hand, Kappas are higher when codes are distributed asymmetrically by the two observers. In contrast to probability variations, the effect of bias is greaterwhen Kappa is small than when it is large.[11]:261–262

Another factor is the number of codes. As number of codes increases, kappas become higher. Based on a simulation study, Bakeman and colleagues concluded that for fallible observers, values for kappa were lower when codes were fewer. And, in agreement with Sim & Wrights's statement concerning prevalence, kappas were higher when codes were roughly equiprobable. Thus Bakeman et al. concluded that "no one value of kappa can be regarded as universally acceptable."[12]:357 They also provide a computer program that lets users compute values for kappa specifying number of codes, their probability, and observer accuracy. For example, given

equiprobable codes and observers who are 85% accurate, value of kappa are 0.49, 0.60, 0.66, and 0.69 when number of codes is 2, 3, 5, and 10, respectively.

Nonetheless, magnitude guidelines have appeared in the literature. Perhaps the first was Landis and Koch,[13] who characterized values < 0 as indicating no agreement and 0–0.20 as slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1 as almost perfect agreement. This set of guidelines is however by no means universally accepted; Landis and Koch supplied no evidence to support it, basing it instead on personal opinion. It has been noted that these guidelines may be more harmful than helpful.[14] Fleiss's[15]:218 equally arbitrary guidelines characterize kappas over 0.75 as excellent, 0.40 to 0.75 as fair to good, and below 0.40 as poor.



Kappa (vertical axis) andAccuracy (horizontal axis) calculated from the same simulated binary data. Each point on the graph is calculated from a pairs of judges randomly rating 10 subjects for having a diagnosis of X or not. Note in this example a Kappa=0 is approximately equivalent to an accuracy=0.5

Weighted kappa

The weighted kappa allows disagreements to be weighted differently[16] and is especially useful when codes are ordered.[8]:66 Three matrices are involved, the matrix of observed scores, the matrix of expected scores based on chance agreement, and the weight matrix. Weight matrix cells located on the diagonal (upper-left to bottom-right) represent agreement and thus contain zeros. Off-diagonal cells contain weights indicating the seriousness of that disagreement. Often, cells one off the diagonal are weighted 1, those two off 2, etc.

The equation for weighted κ is:

$$\kappa = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}x_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}m_{ij}}$$

where *k*=number of codes and w_{ij}, x_{ij}, and m_{ij} are elements in the weight, observed, and expected matrices, respectively. When diagonal cells contain weights of 0 and all off-diagonal cells weights of 1, this formula produces the same value of kappa as the calculation given above.

Kappa maximum

Kappa assumes its theoretical maximum value of 1 only when both observers distribute codes the same, that is, when corresponding row and column sums are identical. Anything less is less than perfect agreement. Still, the maximum value kappa could achieve given unequal distributions helps interpret the value of kappa actually obtained. The equation for κ maximum is:[17]

$$\kappa_{\max} = \frac{P_{\max} - P_{\exp}}{1 - P_{\exp}}$$

where $P_{\exp} = \sum_{i=1}^{k} P_{i+}P_{+i}$, as usual, $P_{\max} = \sum_{i=1}^{k} \min(P_{i+}, P_{+i})$,

k = number of codes, P_{i+} are the row probabilities, and P_{+i} are the column probabilities.

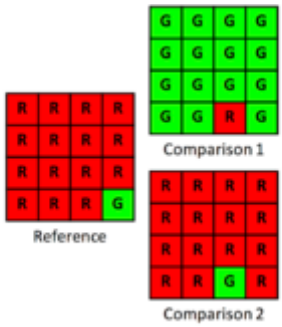
Limitations

Kappa is an index that considers observed agreement with respect to a baseline agreement. However, investigators must consider carefully whether Kappa's baseline agreement is relevant for the particular research question. Kappa's baseline is frequently described as the agreement due to chance, which is only partially correct. Kappa's baseline agreement is the agreement that would be expected due to random allocation, given the quantities specified by the marginal totals of square contingency table. Thus, Kappa = 0 when the observed allocation is apparently random, regardless of the quantity disagreement as constrained by the marginal totals. However, for many applications, investigators should be more interested in the quantity disagreement in the marginal totals than in the allocation disagreement as described by the additional information on the diagonal of the square contingency table. Thus for many applications, Kappa's baseline is more distracting than enlightening. Consider the following example:

Comparison 1

		Reference	
		G	R
Comparison	G	1	14
	R	0	1

The disagreement proportion is 14/16 or .875. The disagreement is due to quantity because allocation is optimal. Kappa is .01.



Kappa example

Comparison 2

		Reference	
		G	R
Comparison	G	0	1
	R	1	14

The disagreement proportion is 2/16 or .125. The disagreement is due to allocation because quantities are identical. Kappa is -0.07.

Here, reporting quantity and allocation disagreement is informative while Kappa obscures information. Furthermore, Kappa introduces some challenges in calculation and interpretation because Kappa is a ratio. It is possible for Kappa's ratio to return an undefined value due to zero in the denominator. Furthermore, a ratio does not reveal its numerator nor its denominator. It is more informative for researchers to report disagreement in two components, quantity and allocation. These two components describe the relationship between the categories more clearly than a single summary statistic. When predictive accuracy is the goal, researchers can more easily begin to think about ways to improve a prediction by using two components of quantity and allocation, rather than one ratio of Kappa.[1]

Some researchers have expressed concern over κ's tendency to take the observed categories' frequencies as givens, which can make it unreliable for measuring agreement in situations such as the diagnosis of rare diseases. In these situations, κ tends to underestimate the agreement on the rare category.[18] For this reason, κ is considered an overly conservative measure of agreement.[19] Others[20] contest the assertion that kappa "takes into account" chance agreement. To do this effectively would require an explicit model of how chance affects rater decisions. The so-called chance adjustment of kappa statistics supposes that, when not completely certain, raters simply guess—a very unrealistic scenario.

See also

- Bangdiwala's B
- Intraclass correlation

References

1. Pontius, Robert; Millones, Marco (2011). "Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment" (http://www.clarku.edu/~rpontius/) *International Journal of Remote Sensing* **32** (15): 4407–4429. doi:10.1080/01431161.2011.552923 (https://doi.org/10.1080%2F01431161.2011.552923)
2. Galton, F. (1892). *Finger Prints* Macmillan, London.
3. Smeeton, N.C. (1985). "Early History of the Kappa Statistic". *Biometrics*. **41** (3): 795. JSTOR 2531300 (https://www.jstor.org/stable/2531300)
4. "The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements". *Physical Therapy*. 2005. doi:10.1093/ptj/85.3.257 (https://doi.org/10.1093%2Fptj%2F85.3.257) ISSN 1538-6724 (https://www.worldcat.org/issn/1538-6724)
5. Cohen, Jacob (1960). "A coefficient of agreement for nominal scales". *Educational and Psychological Measurement*. **20** (1): 37–46. doi:10.1177/001316446002000104 (https://doi.org/10.1177%2F001316446002000104)
6. Powers, David M. W. (2012). "The Problem with Kappa" (https://arquivo.pt/wayback/20160518183306/http://dl.dropbox.com/u/27743223/201209-eacl2012-Kappa.pdf) (PDF). *Conference of the European Chapter of the Association for Computational Linguistics (EACL2012) Joint ROBUS-UNSUP Workshop*. Archived from the original (http://dl.dropbox.com/u/27743223/201209-eacl2012-Kappa.pdf) (PDF) on 2016-05-18.
7. Kilem Gwet (May 2002). "Inter-Rater Reliability: Dependency on Trait Prevalence and Marginal Homogeneity" (http://agreestat.com/research_papers/inter_rater_reliability_dependency.pdf) (PDF). *Statistical Methods for Inter-Rater Reliability Assessment* **2**: 1–10.
8. Bakeman, R.; Gottman, J.M. (1997). *Observing interaction: An introduction to sequential analysis* (2nd ed.). Cambridge, UK: Cambridge University Press. ISBN 978-0-521-27593-4.
9. Fleiss, J.L.; Cohen, J.; Everitt, B.S. (1969). "Large sample standard errors of kappa and weighted kappa". *Psychological Bulletin*. **72** (5): 323–327. doi:10.1037/h0028106 (https://doi.org/10.1037%2Fh0028106)
10. Robinson, B.F; Bakeman, R. (1998). "ComKappa: A Windows 95 program for calculating kappa and related statistics". *Behavior Research Methods, Instruments, and Computers*. **30** (4): 731–732. doi:10.3758/BF03209495 (https://doi.org/10.3758%2FBF03209495)
11. Sim, J; Wright, C. C (2005). "The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements". *Physical Therapy*. **85** (3): 257–268. PMID 15733050 (https://www.ncbi.nlm.nih.gov/pubmed/15733050).
12. Bakeman, R.; Quera, V.; McArthur, D.; Robinson, B. F. (1997). "Detecting sequential patterns and determining their reliability with fallible observers". *Psychological Methods* **2** (4): 357–370. doi:10.1037/1082-989X.2.4.357 (https://doi.org/10.1037%2F1082-989X.2.4.357)
13. Landis, J.R.; Koch, G.G. (1977). "The measurement of observer agreement for categorical data". *Biometrics*. **33** (1): 159–174. doi:10.2307/2529310 (https://doi.org/10.2307%2F2529310) JSTOR 2529310 (https://www.jstor.org/stable/2529310). PMID 843571 (https://www.ncbi.nlm.nih.gov/pubmed/843571).
14. Gwet, K. (2010). "Handbook of Inter-Rater Reliability (Second Edition)" (http://www.agreestat.com/)" ISBN 978-0-9708062-2-2
15. Fleiss, J.L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: John Wiley. ISBN 978-0-471-26370-8.
16. Cohen, J. (1968). "Weighed kappa: Nominal scale agreement with provision for scaled disagreement or partial credit". *Psychological Bulletin* **70** (4): 213–220. doi:10.1037/h0026256 (https://doi.org/10.1037%2Fh0026256) PMID 19673146 (https://www.ncbi.nlm.nih.gov/pubmed/19673146).
17. Umesh, U. N.; Peterson, R.A.; Sauber M. H. (1989). "Interjudge agreement and the maximum value of kappa". *Educational and Psychological Measurement*. **49** (4): 835–850. doi:10.1177/001316448904900407 (https://doi.org/10.1177%2F001316448904900407)
18. Viera, Anthony J.; Garrett, Joanne M. (2005). "Understanding interobserver agreement: the kappa statistic". *Family Medicine*. **37** (5): 360–363.
19. Strijbos, J.; Martens, R.; Prins, F.; Jochems, W. (2006). "Content analysis: What are they talking about?". *Computers & Education*. **46**: 29–48. CiteSeerX 10.1.1.397.5780 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.397.5780). doi:10.1016/j.compedu.2005.04.002 (https://doi.org/10.1016%2Fj.compedu.2005.04.002)
20. Uebersax, JS. (1987). "Diversity of decision-making models and the measurement of interrater agreement" (http://www.na-mic.org/Wiki/images/d/df/Kapp_and_decision_making_models.pdf) (PDF). *Psychological Bulletin*. **101**: 140–146. CiteSeerX 10.1.1.498.4965 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.498.4965) doi:10.1037/0033-2909.101.1.140 (https://doi.org/10.1037%2F0033-2909.101.1.140)

Further reading

- Banerjee, M.; Capozzoli, Michelle; McSweeney, Laura; Sinha, Debajyoti (1999). "Beyond Kappa: A Review of Interrater Agreement Measures". *The Canadian Journal of Statistics* **27** (1): 3–23. doi:10.2307/3315487. JSTOR 3315487.
- Brennan, R. L.; Prediger, D. J. (1981). "Coefficient λ: Some Uses, Misuses, and Alternatives". *Educational and Psychological Measurement* **41** (3): 687–699. doi:10.1177/001316448104100307.
- Cohen, Jacob (1960). "A coefficient of agreement for nominal scales". *Educational and Psychological Measurement* **20** (1): 37–46. doi:10.1177/001316446002000104.
- Cohen, J. (1968). "Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit". *Psychological Bulletin* **70** (4): 213–220. doi:10.1037/h0026256. PMID 19673146.
- Fleiss, J.L. (1971). "Measuring nominal scale agreement among many raters". *Psychological Bulletin* **76** (5): 378–382. doi:10.1037/h0031619.
- Fleiss, J. L. (1981) *Statistical methods for rates and proportions* 2nd ed. (New York: John Wiley) pp. 38–46
- Fleiss, J.L.; Cohen, J. (1973). "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability". *Educational and Psychological Measurement* **33** (3): 613–619. doi:10.1177/001316447303300309.
- Gwet, Kilem L. (2014) *Handbook of Inter-Rater Reliability, Fourth Edition*, (Gaithersburg : Advanced Analytics, LLC) ISBN 978-0970806284
- Gwet, K. (2008). "Computing inter-rater reliability and its variance in the presence of high agreement" (PDF). *British Journal of Mathematical and Statistical Psychology* **61** (Pt 1): 29–48. doi:10.1348/000711006X126600. PMID 18482474.
- Gwet, K. (2008). "Variance Estimation of Nominal-Scale Inter-Rater Reliability with Random Selection of Raters" (PDF). *Psychometrika*. **73** (3): 407–430. doi:10.1007/s11336-007-9054-8.
- Gwet, K. (2008). "Intrarater Reliability." *Wiley Encyclopedia of Clinical Trials, Copyright 2008 John Wiley & Sons, Inc.*
- Scott, W. (1955). "Reliability of content analysis: The case of nominal scale coding". *Public Opinion Quarterly*. **17** (3): 321–325. doi:10.1086/266577.
- Sim, J.; Wright, C. C. (2005). "The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements". *Physical Therapy*. **85** (3): 257–268. PMID 15733050.

External links

- Kappa, its meaning, problems, and several alternatives
- Kappa Statistics: Pros and Cons
- Windows program for kappa, weighted kappa, and kappa maximum
- Java and PHP implementation of weighted Kappa

Online calculators

- Cohen's Kappa for Maps
- Online (Multirater) Kappa Calculator
- Online Kappa Calculator (multiple raters and variables)
- Cohen's Kappa